Trends in
Microbiology

CellPress
OPEN ACCESS

Forum

Bioinformatics challenges for profiling the microbiome in cancer: pitfalls and opportunities

Nicholas A. Bokulich [1,*] and Michael S. Robeson II [2]

Increasing evidence suggests that the human microbiome plays an important role in cancer risk and treatment. Untargeted 'omics' techniques have accelerated research into microbiome–cancer interactions, supporting the discovery of novel associations and mechanisms. However, these techniques require careful selection and use to avoid biases and other pitfalls. In this essay, we discuss selected challenges involved in the analysis of microbiome data in the context of cancer, including the application of machine learning (ML). We focus on DNA sequencing-based (e.g., metagenomics) methods, but many of the pitfalls and opportunities generalize to other omics technologies as well. We advocate for extended training opportunities, community standards, and best practices for sharing data and code to advance transparency and reproducibility in cancer microbiome research.

Background

The human body is inhabited by trillions of microorganisms, collectively known as the human microbiome, with the majority inhabiting the intestinal tract. These communities exhibit a high degree of variation between individuals and can exert a profound impact on host health and disease. Increasing evidence suggests that the gut microbiome also influences an individual's risk of cancer, as well as treatment response, for example, via modulating immune responses and xenobiotic metabolism [1]. In addition to the gut microbiome, intracellular bacteria resident within the tumor microenvironment (TME) appear to play a role in tumorigenesis and antitumor response [1]. Although the association of microbes and TME has been known for over a century, it has only recently become tenable to investigate microbe–tumor interactions. Recent advancements in 'omics' technologies, including use of high-throughput DNA sequencing for marker-gene and whole metagenome sequencing, allow for the investigation of microbiome–cancer interactions and clinical responses to therapeutics through untargeted characterization of host and microbial activities *in situ*. However, these powerful techniques require careful selection, evaluation, and control to avoid spurious observations and common pitfalls in cancer microbiome research. There are myriad challenges associated with different bioinformatics approaches for microbiome multi-omics analysis; in this short essay, we highlight three selected challenges that are persistent across basic research in general, but which we see as particularly prominent for cancer microbiome research: covariates, contaminants, and (lack of) standards for reproducibility, consistency, and interoperability.

Controlling covariates

Driven by the power of 'omics' techniques, a leading goal in current cancer–microbiome research is the discovery of putative microbial biomarkers, for example, for predicting cancer risk or treatment outcomes. A primary challenge regarding the assessment of predictive biomarkers is cohort heterogeneity (e.g., age, diet, lifestyle, geography, ethnicity, comorbidities), which, if not properly accounted for, can mask the detection of microorganisms linked to the primary outcome and introduce biases that can detrimentally impact statistical testing and ML model performance. As the microbiome is correlated with certain host characteristics, including diet, lifestyle factors, and to some extent genetics, identifying robust putative microbiome-based biomarkers can be challenging, particularly within individual cohorts, and hence multicenter studies and meta-analyses are needed. Prior studies [2,3] have demonstrated improved model performance and generalizability for predicting colorectal cancer (CRC)-associated biomarkers when increasing numbers of distinct metagenomic data sets are combined. To improve generalizability of ML models, cross-validation can be performed across study sites or datasets in meta-analyses such that each round of model testing is always performed using an independent dataset [4]. This approach allows better generalization by ensuring that biological variation inherent in a given dataset does not leak and lead to overfitting. Importantly, associations detected using ML or statistical methods are not necessarily causative, and rigorous validation is required due to significant inter-individual variation and host heterogeneity.

Another issue related to covariate control with ML is class imbalance, that is, when samples are not evenly distributed across target classes. This is a common challenge for many ML tasks but can be particularly common in cancer microbiome research, as patient groups (e.g., cancer patients versus healthy controls; or responders versus non-responders in a clinical trial) frequently present highly skewed distributions. In these cases, predictive models often favor the overrepresented class(es), leading to high false-negative rates for the underrepresented class(es). Techniques for addressing class imbalance have been discussed previously [4] and we recommend assessing class distributions and applying these techniques when appropriate. This is not always possible, for example, when

sample sizes are too low to enable robust over- or undersampling, or when prior probabilities cannot be inferred; in these cases (and in general), we advise investigators to consult multiple evaluation metrics, including those that are less sensitive to class imbalances, for example, area under the precision–recall curve or Matthew's correlation coefficient.

Low biomass, contamination, and host reads

Issues with accurate microbiome profiling and putative biomarker detection are exacerbated when investigating the TME, which generally contains very low microbial biomass, making it difficult to distinguish genuine observations from contaminants. This makes data interpretation more prone to methodological biases that complicate discovery of microbe–tumor interactions [5]. Inadequate detection and filtering of contaminants, including microbial species that are implausible in human samples (and hence most likely represent misclassification of contaminant sequences), as well as host reads, can yield misleading results, and require careful control [6].

Exogenous contaminants can be introduced during sample collection and laboratory procedures, for example, from environmental sources, reagents, or cross-contamination. Laboratory best-practices are paramount, but cannot prevent all sources, for example, reagent contaminants. Rigorous use of controls is necessary to identify contaminants, and bioinformatics approaches can detect and remove contaminants in some experimental settings (e.g., [5]), though these methods alone are not a suitable replacement for laboratory best practices.

Host-derived sequence reads also can be highly prevalent, particularly in very-low-biomass samples such as tissue biopsies. Methods for host nucleotide depletion exist, but some methods can introduce biases and must be used cautiously. Host reads can introduce analytical biases as well as ethical issues for data sharing and reuse, and must be removed using appropriate computational methods. Sensitive personal information (e.g., sex, ancestry, re-identification) can be derived from human sequence reads found in shotgun metagenome sequence surveys from human body sites when read filtering is not adequately performed, and a recent evaluation of read filtering methods found that most fail to remove 100% of host sequences [7]. The risk of inadequate filtering is particularly high when only a single human reference genome is used for filtering, underrepresenting the degree of genetic diversity inherent in the wider human population. Failure to remove host reads can introduce false-positives and other errors [6]. Hence, stringent filtering should be performed using (i) multiple human reference genomes and (ii) multiple filtering algorithms to minimize the probability that human reads pass filtering and are mistakenly handled as non-human reads [6].

One solution to avoid retention of non-target contaminants and host reads in DNA sequence data is to map sequences to specific microorganisms and/or curated microbial reference genomes [8]. This approach will increase the robustness of establishing the residency of given microbes within the TME, though this comes at the cost of losing reads that may represent genuine (and potentially important) biological information but fail to map, for example, because they are from novel strains or species outside of the reference.

Contaminants and host reads can introduce bias in ML or statistical methods when they covary with specific groups, introduce batch effects, or are mistaken as meaningful biological signals (e.g., host reads misannotated as microbial reads [6]). When properly annotated and used to identify, for example, specific somatic mutations, host DNA can contain valuable diagnostic markers. However, detection of (misannotated) contaminant DNA on its own is of dubious value as a biomarker or therapeutic target, and could artificially inflate predictive accuracy if these contaminants covary with patient classes [6]. Hence, careful contaminant removal should be performed when applying ML or statistical methods to cancer microbiome datasets.

Bioinformatics methods, reproducibility, and community standards

Currently there is significant heterogeneity in terms of the methodology (both laboratory and bioinformatics approaches) used in analysis of cancer microbiome datasets. Meta-analysis and comparison of datasets from published studies is significantly hampered by inaccessibility, inconsistency, and/or inadequacy of (meta)data and use of heterologous methods. In addition to wet lab methods that influence microbiome profiles (starting with sample collection and DNA extraction), bioinformatics methods selection as well as parameterization significantly impact performance and results from bioinformatics [9], statistical, and ML methods [10], limiting comparability and requiring rigorous reporting standards.

Appropriate laboratory and bioinformatics techniques should be selected, ideally adhering to community standards, best practices, and benchmarks. Consistent methodology, including use of standards for data and metadata reporting (e.g., www.gensc.org/pages/standards-intro.html), is needed to facilitate data re-use and comparison across studies. Community benchmarks [11] can be useful guides to select methods with validated performance. Commonly used bioinformatics workflows and software platforms that integrate tools with validated benchmarks, including platforms with provenance tracking systems [12], are recommended to facilitate comparability and reproducibility of results. Researchers should make their code and

Box 1. Recommended solutions and opportunities for profiling the microbiome in cancer

1. Use multi-center studies and meta-analyses to identify robust biomarkers. In these cases, when ML methods are applied perform leave-one-study-out or cross-study cross-validation strategies [4]. These may not be suitable study designs in all cases, however, for example, when studying rare cancers or specific patient subpopulations, for which low sample size may constrain experimental design.
2. Use longitudinal study designs to evaluate temporal relationships between microbiota and patient outcomes. This can identify temporal dynamics associated with treatment response [8] and identify potentially causal temporal relationships. However, longitudinal designs may not be appropriate for all experimental questions, and may reduce study power if the number of subjects must be reduced to accommodate additional timepoints.
3. Integration of multiple omics techniques can yield insight, for example, into potential mechanisms of host–microbiome interactions and the relative predictive value of different markers [13]. However, different omics datasets often require specific pre-processing steps prior to analysis and integration [14] that must be checked to avoid common pitfalls.
4. Use dedicated facilities and/or a cleanroom environment to minimize exogenous contamination and cross-contamination during sample preparation and analysis, as well as the RIDE principles for handling of low-biomass samples [15].
5. Sharing data and analytical workflows is important for transparency and integrity in cancer microbiome research, as well as facilitating data re-use. Follow community-accepted standards and best practices for data and metadata reporting to the extent ethically permissible, including use of controlled-access data repositories when appropriate. Data and metadata publication and management should be considered before the start of the project and included in ethics approvals and informed consent.
6. Ensure transparency and reproducibility in bioinformatics analyses by sharing code in executable formats, for example, as Jupyter notebooks and/or version-controlled code repositories. Consider sharing processed data (e.g., as supplemental files in publications or code repositories) with integrated provenance information [12] and/or together with detailed metadata for full transparency in line with the FAIR principles (www.go-fair.org/fair-principles/).
7. Promote team science to enhance integrative omics and translational research. Close integration of clinical and technical expertise is often needed to interpret predictions from high-dimensional microbiome datasets, including from ML approaches. Moreover, the multi-omics data explosion has led to a commensurate rise in myriad analytical approaches to parse and analyze data, often requiring synergistic expertise to fully leverage results. This also has the effect of making -omics fields less intimidating to new researchers as it becomes a 'team sport'.

data available (to the extent ethically possible) to maximize transparency and reproducibility. Reporting checklists and guidelines for laboratory techniques, bioinformatics (e.g., www.stormsmicrobiome.org/), machine learning [10], and other analytical methods can guide researchers in best practices, but many of these standards have yet to be widely adopted in the field. In the short term, researchers should at least adhere to community standards for methods, code, and (meta)data reporting and sharing to maximize data re-use potential and research transparency; in the long term, further standardization will help advance the field by facilitating large-scale meta-analysis of increasingly large and complex microbiome datasets.

Improved training and career development opportunities are also needed to foster widespread adoption of best practices for research data management and community standards. Training programs for research data management should be integrated in relevant biomedical education programs alongside related training in ethics and good clinical practice. Many such programs are hosted by various universities and institutions. Continuing education programs, including workshops from the National Microbiome Data Collaborative (NMDC) (https://microbiomedata.org/), Massive Open Online Courses (MOOCs) (www.mooc.org/), and Software Carpentry (https://software-carpentry.org/) will also support practicing bench researchers, clinicians, and others involved in cancer-associated microbiome research (and more generally) to develop these skill sets, keeping pace with rapid advancements in the field. Wider recognition and career pathways for research data managers will facilitate adoption of best practices for management of sensitive cancer microbiome datasets.

Conclusion

Cancer microbiome research presents a combination of challenges that complicate accurate use of omics techniques. We recommend some solutions and opportunities to be considered in study design and analysis (Box 1). Above all, we recommend use of community-accepted standards and tools for bioinformatics analysis and data management to enhance the transparency, reproducibility, integrity, generalizability, and re-usability of cancer microbiome datasets in pursuit of the common goal of discovering novel mechanisms of host-microbiome interaction that impact the risk, prevention, and treatment of cancer.

Acknowledgments

N.A.B. gratefully acknowledges support from the grant [#2021-362] of the Strategic Focus Area 'Personalized Health and Related Technologies (PHRT)' of the ETH Domain (Swiss Federal Institutes of Technology) and by Dr Walter and Edith Fischli via the ETH Zurich Foundation. M.S.R. gratefully acknowledges support from the following NIH grants: R01CA143130, R01CA245083, and R01CA282198.

Declaration of interests

No interests are declared.

[1]Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland
[2]University of Arkansas for Medical Sciences, Department of Biomedical Informatics, Little Rock, AR, USA

*Correspondence:
nicholas.bokulich@hest.ethz.ch (N.A. Bokulich).

https://doi.org/10.1016/j.tim.2024.08.011

References

1. Park, E.M. *et al.* (2022) Targeting the gut and tumor microbiota in cancer. *Nat. Med.* 28, 690–703
2. Wirbel, J. *et al.* (2019) Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25, 679–689
3. Thomas, A.M. *et al.* (2019) Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25, 667–678
4. Asnicar, F. *et al.* (2024) Machine learning for microbiologists. *Nat. Rev. Microbiol.* 22, 191–205

5. Ghaddar, B. *et al.* (2022) Tumor microbiome links cellular programs and immunity in pancreatic cancer. *Cancer Cell* 40, 1240–1253.e5
6. Gihawi, A. *et al.* (2023) Major data analysis errors invalidate cancer microbiome findings. *MBio* 14, e0160723
7. Tomofuji, Y. *et al.* (2023) Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. *Nat. Microbiol.* 8, 1079–1094
8. Battaglia, T.W. *et al.* (2024) A pan-cancer analysis of the microbiome in metastatic cancer. *Cell* 187, 2324–2335.e19
9. Bokulich, N.A. *et al.* (2020) Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. *Comput. Struct. Biotechnol. J.* 18, 4048–4062
10. Heil, B.J. *et al.* (2021) Reproducibility standards for machine learning in the life sciences. *Nat. Methods* 18, 1132–1135
11. Meyer, F. *et al.* (2022) Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* 19, 429–440
12. Bolyen, E. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857
13. Bokulich, N.A. *et al.* (2022) Multi-omics data integration reveals metabolome as the top predictor of the cervicovaginal microenvironment. *PLoS Comput. Biol.* 18, e1009876
14. Graw, S. *et al.* (2021) Multi-omics data integration considerations and study design for biological systems and disease. *Mol. Omics* 17, 170–185
15. Eisenhofer, R. *et al.* (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol.* 27, 105–117