ISSN: 0974-5823 Vol. 7 (Special Issue 5, April-May 2022)

International Journal of Mechanical Engineering

Bioinformatics and Data Science in Industrial Microbiome Applications: A Review

Sreenivasa Rao Veeranki and Manish Varshney

Department of Computer Science and Engineering, School of Engg. & Tech., Maharishi university of Information Technology, Lucknow, INDIA

In recent years, progress in sequencing and computational biology has greatly expanded our capacity for studying microbial communities' taxonomic and functional components, which are critical in industrial processes of all kinds and sizes. As a result, commercial interest has increased in applications where microbial populations play a significant role. Probiotics, cosmetics, and enzyme research are a few examples. Gut microbiome data may also be used in commercial applications, such as software that provides evidence-based, automated and individualised food recommendations for maintaining healthy blood sugar levels. Data integration for predictive machine learning requires strain-level precision in community profiles and many bioinformatic and data science difficulties. In this viewpoint, we touch on numerous industrial fields and briefly address the developments and future prospects of bioinformatics and data science in microbiome research in order to share our perspectives on such difficulties.

Keywords: *DNA sequencing, microbiome, industrial biotechnology, probiotics, 16S rRNA gene profiling, metagenomics, bioinformatics, data science*

Introduction

Industrial activities such as the manufacture of food, drinks, probiotics, paper, and cleaning goods rely heavily on microbial populations (for a review, see Singh et al., 2016). Marker gene (16S rRNA) and shotgun metagenome sequencing for product development, optimization, and quality control have become industry standards for studying the taxonomic makeup and functional capabilities of these microorganisms (Costessi et al., 2018). Metatranscriptomics and metabolomics data, for example, can be employed in integrative investigations to produce leads in enzyme discovery. In several of these microbiome research, strain-level analysis of community composition is required to determine the effectiveness of probiotics (McFarland et al., 2018). Studies examining microbiome capacity to synthesise specific substances and requiring bacterial genome recovery from complicated microbiomes (e.g. dirt) are also being conducted (Howe et al., 2014). Bioinformatics, data mining, and machine learning approaches must be used to extend microbiome applications to the general population for practical findings, such as controlling blood sugar levels (Zeevi et al., 2015).

Bioinformatic and data science problems are highlighted in this review of industrial microbiome applications. In addition, we discuss some of the new developments that might shed light on the issues that these applications face. We'll wrap things off by discussing where we see industrial microbiome applications going in the future and what computational components they'll need.

Current Applications and Products

Dairy Starter Cultures

Cheese, yoghurt, pork, and wine all benefit from the utilisation of microbial populations (e.g. lactic acid bacteria) in various food and beverage manufacturing processes, including the creation of these products. Particularly during cheese ripening, when flavour and structure are formed, their contribution is critical. Enzymes unique to a particular strain control these processes (Escobar-Zepeda et al., 2016). Such enzymes may be challenging to isolate and study since cultivating strain representatives can be time-consuming or tedious (Lagier et al., 2016). An alternative method of examining these enzymes is by metagenome sequencing, assembly, and annotation in the context of product improvement, for example (De Filippis et al., 2017). As a result of the importance of metagenome assembly in understanding bacteriophage populations and the abundance, diversity, and development of these microorganisms, not only can it help prevent fermentation failures due to viral infections but it can also help unlock the potential of these microorganisms to fight food-borne pathogens (Fernández et al., 2017).

Probiotics

When taken in sufficient quantities, probiotics are beneficial bacteria meant to promote the health of the host. To find new probiotics, researchers must first build a strain library using a method known as culturomics (Lagier et al., 2016). Following in vitro and computational study on the acquired strains, such as for their bile resistance and ability to survive the transit of the stomach, is the next step. After completing each of these phases, a smaller pool of candidates remains to be evaluated by regulatory agencies like the European Food Safety Authority (EFSA, FEEDAP et al., 2018). When combined with other datasets such as metabolomic, demographic, dietary, and lifestyle data, we believe that findings from comparative studies of the gut microbiome highlight associations between phenotypic traits such as inflammation (Andoh et al., 2012).

Quality Control

Live organisms are found in items like probiotics and dairy starter cultures, which are either marketed to customers or utilised in the production of consumer goods. Quality control of the finished product is just as important as checking the raw materials for proper strains and the absence of harmful microorganisms (Fenster et al., 2019). Due to the wide variety of phenotypes that may exist across various strains of the same species of microorganism, strain-level identification throughout the quality control process is essential for identifying any potential contaminants (Huys et al., 2013).

Cosmetics

Skin microbiome research is becoming more popular in the cosmetics sector as a possible treatment target for conditions such as acne, eczema, and Malassezia folliculitis (Wallen-Russell, 2019). Despite this, these investigations are sometimes impeded by the low biomass of skin samples, which may lead to inaccurate results due to contaminations (e.g., from nearby skin or reagents) (Kong et al., 2017). Human skin microbiota (Zeeuwen et al., 2012) is subject-specific and difficult to generalise about the effects of skin products on a large group. As a result of this, there is a need for personal longitudinal studies, where statistical methods such as redundancy analysis and principle response curve (Van den Brink and Braak, 1999) can be used to assess correlations between taxonomic or functional composition and the characteristics of the samples being studied (environmental variables). A further benefit is that data may be adjusted for one of the covariates before the actual analysis is conducted, making it easier to determine the treatment's impact.

Enzyme Discovery

Cleaning agents, laundromat chemicals, paper and textile enzymes, and many more are among the many types of industrial chemicals whose costs, environmental impact, and efficacy are under constant scrutiny. Enzymes with desirable qualities may be found in a variety of microbiomes, including those found in seawater and soil, as well as those found in lakes and wetlands. In addition, new enzymes that can perform complex reactions may also be found in these microbiomes (Popovic et al., 2015). Two enzymes recently discovered that allows a sustainable alternative to toluene, a petrochemical with an annual market of 29 million tonnes, to be produced by complex microbial communities living in sewage and lakes are significant examples of the latter (Beller et al., 2018).

Microbiome-Based Health and Personalized Nutrition

Services such as MyMicroZoo1, Biovis2, and American Gut3 make microbiome analysis accessible to the general public at low prices. They must pay significantly greater attention to the clarity of their findings even if the results are declared not to be construed as a diagnosis, even though they are operationally identical to those employed in research. As a practical matter, this implies that skilled healthcare experts [such as dieticians and general practitioners (GPs)) should assist the end-user in making sense of the (actionable) data to avoid misinterpretation.

Based on published research findings is an excellent practice, but the fact that most studies concentrate on a set cohort and present "averaged" population trends makes it doubtful whether results can be applied to individuals. Function-based techniques using metagenomics may make such individual translations less difficult since the 'personalised' impacts in these datasets are less evident (Lloyd-Price et al., 2017). Zeevi and colleagues (2015) demonstrated that a person's gut microbiome can be used to predict post-meal glycemic responses by integrating data from their blood parameters, anthropometrics and physical activity as well as the gut microbiome into a machine learning algorithm that predicted the post-meal glycemic responses of their subjects. The final prediction model incorporated 72 taxonomic or functional microbiome characteristics. An example of how huge datasets from scientific research and data science may be merged in commercial settings for providing consumers with evidence-based health-related advice is DayTwo4, which is now available to the public.

Current Advances

Metagenome Assembly, Binning, and Annotation

Because it allows for gene prediction, annotation, and abundance profiling, metagenome assembly is an essential computational step in studies of microbiome function. There is a variety of (de Bruijn graph-based) metagenome assembly techniques available, and it is critical to choose the one that best suits the research topic at hand in terms of simplicity of use, scalability, run time, and memory requirements. (Van der Walt et al., 2017). Measuring the influence of probiotic supplementation on the abundance of gene groups and pathways in large cohorts requires computationally less demanding techniques like MEGAHIT (Li et al., 2015). In contrast, investigations with a modest number of samples, such as enzyme discovery applications, may employ assembly tools like metaSPAdes (Nurk et al., 2017) that incorporate optimizations such as error correction but with a consequent runtime trade-off.

Hypothesis-Driven Functional Analyses

A microbiome dataset's functional components and prospective longitudinal and cross-sectional aspects are often thought hopeless to analyse exhaustively and query. Even if it is technically possible, several testing difficulties reduce the analytical power significantly. Determining the important functional elements is a significant step toward addressing these restrictions, even while measures like the elimination of collinear variables and confirmation of putative correlations in separate datasets may in part solve these concerns (Falony et al., 2016). Such a strategy involves using a specific database to answer a specific hypothesis, such as in the case of select enzyme classes or an entire set of enzymatic pathways. To provide three examples, Resfams (Gibson et al., 2015) and dbCAN (Yin and al., 2012) both concentrate on antibiotic resistance while antiSMASH (Blin et al., 2017) examines secondary metabolite syntheses. A method like STRING (Szklarczyk et al., 2014) that uses guilt by association approaches to identify genes that are not directly flagged by comparison to specific functional datasets, such as those described above, can be used to identify genes whose distribution patterns are similar enough to genes that are represented in the reference set. Functional studies that need protein sequences have the problem of requiring assembly and gene prediction, which may be computationally costly, as explained above. This is a limitation. For profiling protein family abundance, tools such as HUMAnN2 (Franzosa et al., 2018) operate directly with short-read data.

Assembly-Independent Strain-Level Characterization

Both Bifidobacterium long subsp. longum and Bifidobacterium longum subsp. infantis, which have two unique phenotypes with major functional implications in infant nutrition, vary just slightly in their 16S rRNA gene sequences (Lawley et al., 2017). In OTU clustering-based taxonomic analysis, these distinctions are not preserved. Unorthodox techniques such as UNOISE2 and DADA2 (Callahan et al., 2016) avoid the need for clustering and sequence filtering processes, allowing single-nucleotide-level differentiation between sequences (ASVs). The phylogenetic depth at which microbiome investigations may be understood may be improved greatly as a result of this. Among the notable uses of these new algorithms were fresh insights into oral (Mukherjee et al., 2018) and vaginal microbiomes at the sub-species level (Callahan et al., 2017).

These methods may be used to perform strain-level studies using shotgun metagenomic datasets without the necessity for metagenome assembly in circumstances when many strains of the same species have identical 16S rRNA sequences (Truong et al., 2017; PanPhlAn (Scholz et al., 2016). (Figure 1). Routine compositional analysis may now be performed to confirm the presence of desired strains or to detect possible infections in finished goods using these approaches.

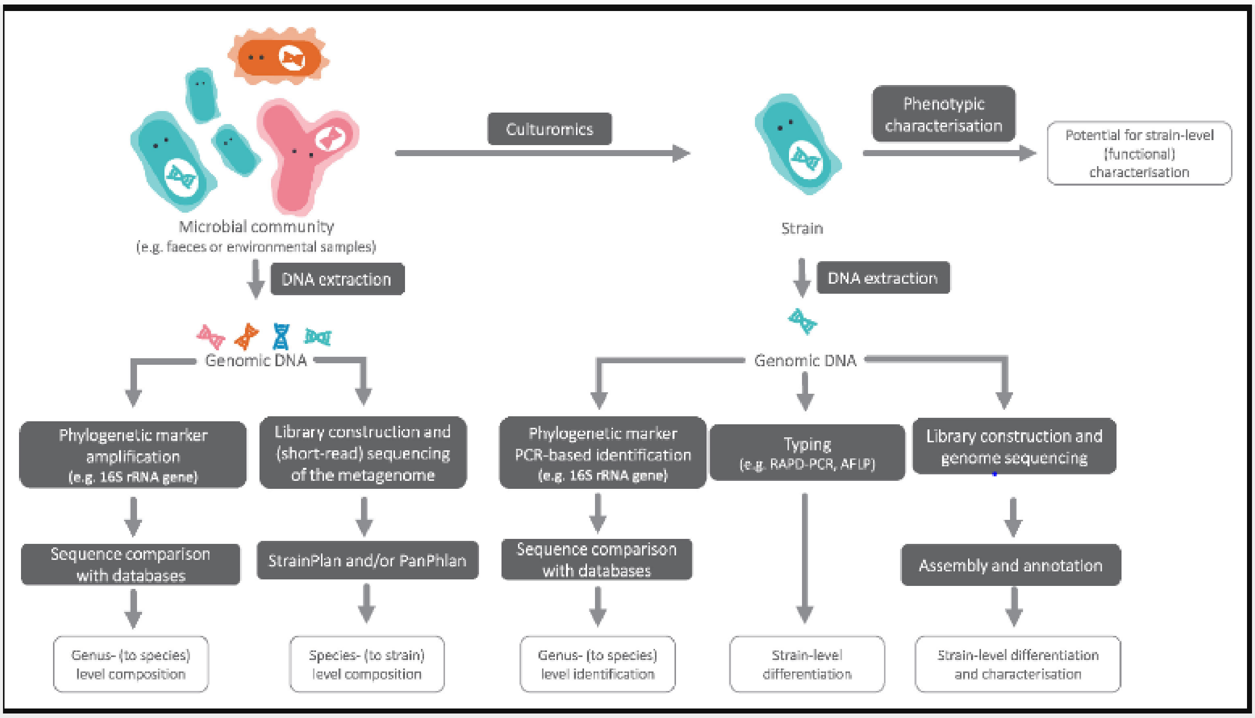
Culturomics
Microbial community
(e.g. faeces or environmental samples)
Strain
Phenotypic characterisation
Potential for strain-level (functional) characterisation
DNA extraction
DNA extraction
Genomic DNA
Genomic DNA
Phylogenetic marker amplification (e.g. 16S rRNA gene)
Library construction and (short-read) sequencing of the metagenome
Phylogenetic marker PCR-based identification (e.g. 16S rRNA gene)
Typing (e.g. RAPD-PCR, AFLP)
Library construction and genome sequencing
Sequence comparison with databases
StrainPlan and/or PanPhlan
Sequence comparison with databases
Assembly and annotation
Genus- (to species) level composition
Species- (to strain) level composition
Genus- (to species) level identification
Strain-level differentiation
Strain-level differentiation and characterisation

Figure 1 An overview of approaches to achieve taxonomic resolution at different levels.

Long-Read Sequencing and Other Advances

Even though long-read sequencing systems PacBio and Oxford Nanopore Technologies (ONT) are yet to be widely used in microbiome investigations, they provide interesting prospects for a variety of commercial applications. As an example, PacBio's circular consensus sequencing application offers the essential phylogenetic precision for applications like fermentation investigations, which is impossible with short-read amplicon sequencing. Quality control applications for pathogen identification may benefit from ONT's on-demand sequencing, but the high rate of error prevents reliable strain-level detection.

It is usual for short-read datasets to be quite fragmented even with high dataset coverage and sophisticated algorithms, particularly in samples from diverse ecosystems like soil. Long-read sequencing will likely become more widely used in research aimed at assembling whole microbial genomes at the chromosomal level shortly. The possibility to apply hybrid assembly techniques such as hySPADES for long- and short-read metagenome datasets is something we're excited about. 10x Genomics (http://10xgenomics.com) is a company that provides barcoded short reads with long-range information for microbiome research. Athena assembler

(Bishara et al., 2018) is an example of a customised bioinformatics tool that exploits barcode information in short reads and enhances the congruency of metagenome assemblies.

Machine Learning and Data Science

In microbiome investigations, the number of datasets and the depth of sequencing per sample have grown as the cost of sequencing has fallen. Studying OTU tables and functional profiles as starting material for further analyses such as machine learning (ML) applications became possible because of the increased statistical power of the investigations (Pasolli et al., 2016). Random forests (RF) methods have been employed effectively by many in the context of illness, for example, properly predicting IBS (Saulnier et al., 2011) and bacterial vaginosis (Beck and Foster, 2014) based on taxonomic profiles (for a review, see LaPierre et al., 2019 and Qu et al., 2019). A median accuracy of only 56.68 percent was found when Sze and Schloss (2016) used 10 previously published obesity datasets to train RF ML models on one dataset and test them on the other nine, suggesting that the method may not be applicable for some diseases or ii) the disease signal may be more apparent at the level of differentially expressed functions (gene transcripts) of the microbiome.

Microbiome uses of ML in the industrial sector include constructing classification models for oil sites based on soil microbiome data, and the previously mentioned customised health-related lifestyle advice services that are in part based on gut microbiome data. When it comes to screening new probiotics, we anticipate data integration and machine learning to have an influence. Microbiome analysis tools like MicrobiomeAnalyst (Dhariwal et al., 2017), QIIME 2 (Bokulich et al., 2018), and USEARCH (Edgar, 2010) have begun incorporating ML methods that can be used by researchers who aren't necessarily bioinformaticians to meet the general demand for user-friendly ML in microbiome research.

Conclusions and Outlook

Microbiome research has a wide range of options due to the wide range of experimental and computational approaches accessible. While standards and standardisation are critical for improving comparability and reproducibility, reaching worldwide agreement on the methodologies utilised is still a difficulty. To a large extent, researchers are constrained by the time and effort required to implement new procedures, which might in turn compromise the comparability of results between investigations, or even within studies that last for a long time. We agree with Knight et al. (2018) that standardising the documentation of procedures, tools, data formats, and data processing settings should be a fundamental goal of microbiome investigations, and these "logs" should be published alongside the final findings and interpretations. BaseClear5, NIZO food research6, Clinical Microbiomics7, Vedanta Biosciences8, and COSMOSID9 are among the microbiome analysis providers that are concerned about revealing a major portion of their intellectual property if they were to be be completely disclose their findings.

For microbiome investigations, we predict long-read sequencing to become more widespread as prices continue to fall. This will allow for better taxonomic resolution, as well as greater functional analysis, due to more continuous metagenome assemblies. Bioinformatic procedures currently defined for short readings, such as denoising and read categorization, will likely be translated to long-read versions as the primary emphasis of future advances.

The use of costly computations in de novo assembly and annotation for shotgun metagenome analysis in big studies may also cause capacity concerns. If your company can't afford a significant on-premise computer infrastructure, the cloud offers a flexible solution where cloud computing knowledge is vital.

It will also continue to be stimulated by the fast translation of microbiome research into key commercial uses in the healthcare, energy, and food industries. In this interaction, we anticipate the role of bioinformatics and data science to grow in importance.

References

1. (FEEDAP), E., Rychen, G., Aquilina, G., Azimonti, G., Bampidis, V., Bastos, M., et al. (2018). Guidance on the characterisation of microorganisms used as feed additives or as production organisms. *EFSA J.* 16 (3), e05206. DOI: 10.2903/j.efsa.2018.5206
2. CrossRef Full Text | Google Scholar
3. Andoh, H., Kizuoka, H., Tsujikawa, T, Nakamura, S, Hirai, F., Suzuki, Y., et al. (2012). Multicenter analysis of fecal microbiota profiles in Japanese patients with Crohn's disease. *J. Gastroenterol.* 47 (12), 1298–1307. DOI: 10.1007/s00535-012-0605-0
4. PubMed Abstract | CrossRef Full Text | Google Scholar
5. Antipov, D., Korobeynikov, A., McLean, J., Pevzner, P. (2015). hybrid spaces: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32 (7), 1009–1015. DOI: 10.1093/bioinformatics/btv688
6. PubMed Abstract | CrossRef Full Text | Google Scholar
7. Arboleya, S., Bottacini, F., O'Connell-Motherway, M., Ryan, C., Ross, R., Van Sinderen, D., et al. (2018). Gene-trait matching across the Bifidobacterium longum pan-genome reveals considerable diversity in carbohydrate catabolism among human infant strains. *BMC Genomics* 19 (1), 33. DOI: 10.1186/s12864-017-4388-9
8. PubMed Abstract | CrossRef Full Text | Google Scholar
9. Ayling, M., Clark, M., Leggett, R. (2019). New approaches for metagenome assembly with short reads. *Brief Bioinform.* 1–11. DOI: 10.1093/bib/bbz020
10. CrossRef Full Text | Google Scholar
11. Beck, D., Foster, J. (2014). Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. *PloS One* 9 (2), e87830. DOI: 10.1371/journal.pone.0087830
12. PubMed Abstract | CrossRef Full Text | Google Scholar
13. Beller, H., Rodrigues, A., Zargar, K., Wu, Y.-W., Saini, A., Saville, R., et al. (2018). Discovery of enzymes for toluene synthesis from anoxic microbial communities. *Nat. Chem. Biol.* 14 (5), 451. DOI: 10.1038/s41589-018-0017-4
14. PubMed Abstract | CrossRef Full Text | Google Scholar
15. Berendsen, E., Boekhorst, J., Kuipers, O., Wells-Bennik, M. (2016). A mobile genetic element profoundly increases heat resistance of bacterial spores. *ISME J.* 10 (11), 2633. doi: 10.1038/ismej.2016.59
16. PubMed Abstract | CrossRef Full Text | Google Scholar
17. Bishara, A., Moss, E., Kolmogorov, M., Parada, A., Weng, Z., Sidow, A., et al. (2018). High-quality genome sequences of uncultured microbes by the assembly of reading clouds. *Nat. Biotechnol.* 36, 1067–1075. doi: 10.1038/nbt.4266
18. CrossRef Full Text | Google Scholar
19. Blin, K., Wolf, T., Chevrette, M., Lu, X., Schwalen, C., Kautsar, S., et al. (2017). antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* 45 (W1), W36–W41. doi: 10.1093/nar/gkx319
20. PubMed Abstract | CrossRef Full Text | Google Scholar
21. Bokulich, N., Dillon, M., Bolyen, E., Kaehler, B., Huttley, G., Caporaso, J. (2018). q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J. Open Source Softw.* 3 (30), 934. doi: 10.21105/joss.00934
22. CrossRef Full Text | Google Scholar
23. Callahan, B., DiGiulio, D., Goltsman, D., Sun, C., Costello, E., Jeganathan, P., et al. (2017). Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of US women. *Proc. Natl. Acad. Sci* 114 (37), 9966–9971. doi: 10.1073/pnas.1705899114
24. CrossRef Full Text | Google Scholar
25. Callahan, B., McMurdie, P., Rosen, M., Han, A., Johnson, A., Holmes, S. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13 (7), 581. doi: 10.1038/nmeth.3869
26. PubMed Abstract | CrossRef Full Text | Google Scholar

27. Callahan, B., Wong, J., Heiner, C., Oh, S., Theriot, C., Gulati, A., et al. (2018). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* gkz569. https://doi.org/10.1093/nar/gkz569
28. Google Scholar
29. Costessi, A., van den Bogert, B., May, A., Ver Loren van Themaat, E., Roubos, J., Kolkman, M., et al. (2018). Novel sequencing technologies to support industrial biotechnology. *FEMS Microbiol. Letters* 365 (16), fny103. doi: 10.1093/femsle/fny103
30. CrossRef Full Text | Google Scholar
31. De Filippis, F., Parente, E., Ercolini, D. (2017). Metagenomics insights into food fermentations. *Microb. Biotechnol.* 10 (1), 91–102. doi: 10.1111/1751-7915.12421
32. PubMed Abstract | CrossRef Full Text | Google Scholar
33. Dhariwal, A., Chong, J., Habib, S., King, I., Agellon, L., Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 45 (W1), W180–W188. doi: 10.1093/nar/gkx295
34. PubMed Abstract | CrossRef Full Text | Google Scholar
35. Dutilh, B., Schmieder, R., Nulton, J., Felts, B., Salamon, P., Edwards, R., et al. (2012). Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* 28 (24), 3225–3231. doi: 10.1093/bioinformatics/bts613
36. PubMed Abstract | CrossRef Full Text | Google Scholar
37. Edgar, R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26 (19), 2460–2461. doi: 10.1093/bioinformatics/btq461
38. PubMed Abstract | CrossRef Full Text | Google Scholar
39. Edgar, R. (2016). UNOISE2: improved errorcorrection for Illumina 16S and ITS amplicon sequencing. *BioRxiv* 081257. doi: 10.1101/081257
40. CrossRef Full Text | Google Scholar
41. Escobar-Zepeda, A., Sanchez-Flores, A., Baruch, M. (2016). Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiol.* 57, 116–127. doi: 10.1016/j.fm.2016.02.004
42. PubMed Abstract | CrossRef Full Text | Google Scholar
43. Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., et al. (2016). Population-level analysis of gut microbiome variation. *Science* 352 (6285), 560–564. doi: 10.1126/science.aad3503
44. PubMed Abstract | CrossRef Full Text | Google Scholar