Ultra-resolution Metagenomics: When Enough Is Not Enough

Falk Hildebrand

Gut Microbes and Health, Quadram Institute Bioscience, Norwich, United Kingdom
Digital Biology, Earlham Institute, Norwich, United Kingdom

ABSTRACT

Technological advances in community sequencing have steadily increased the taxonomic resolution at which microbes can be delineated. In high-resolution metagenomics, bacterial strains can now be resolved, enhancing medical microbiology and the description of microbial evolution in vivo. In the Hildebrand lab, we are researching novel approaches to further increase the phylogenetic resolution of metagenomics. I propose that ultra-resolution metagenomics will be the next qualitative level of community sequencing, classified by the accurate resolution of ultra-rare genetic events, such as subclonal mutations present in all populations of evolving cells. This will be used to quantify evolutionary processes at ecologically relevant scales, monitor the progress of infections within a patient, and accurately track pathogens in food and infection chains. However, to develop this next metagenomic generation, we first need to understand the currently imposed limits of sequencing technologies, metagenomic strain delineation, and genome reconstructions.

KEYWORDS

metagenomics, sequencing technology, microbiology, microbial evolution

COMMUNITY SEQUENCING IS A RAPIDLY EVOLVING FIELD

Microbial sequencing has drastically increased our understanding of bacterial communities. Amplicon sequencing (metabarcoding) of the 16S rRNA gene can cost-effectively identify and quantify prokaryotes in communities, but it is limited by sequencing errors that will inflate species diversity if not corrected (1). Further, my group previously described phylogenetic restrictions (2, 3) and taxonomic misclassifications in low-biomass scenarios (4). To optimally account for such errors, we maintain and steadily improve the LotuS2 pipeline (5). However, despite active development in read error correction (6), metabarcoding approaches will remain limited to species-level resolution (7).

There is a strong interest to further resolve microbiomes; currently, bacterial strains can be differentiated using high-resolution metagenomics (HRM) (8, 9). But each population of (clonally derived) cells contains de novo mutations segregating in only a fraction of the population—subclonal mutations. These are vital to evolution, as they are either lost or fixed within a population, depending on partly luck and partly relevance to survival. Detecting and linking these to their genome (“phasing”) using reference-free metagenomics will require new approaches, a qualitative leap I refer to as ultra-resolution metagenomics (URM) (Fig. 1). I discuss in this commentary potential use cases of URM and technical challenges to overcome.

APPLICATIONS OF ULTRA-RESOLUTION METAGENOMICS

Contrary to intuitive assumptions, microbial evolution can happen within days: some gut bacteria adapt to antibiotic treatments within days (10), and plant commensals adapt on an “agriculturally relevant evolutionary timescale” (11). Pseudomonas aeruginosa typically loses dozens of genes during cystic fibrosis infections lasting years (12), Bacteroides fragilis adapts to its specific human host (13), and detecting parallel evolution in pathogens can be used to identify pathogenicity genes (14), as demonstrated using variants of...
metagenomic and isolate sequencing. These evolutionary processes can be quantified and predicted using population genetics, dependent on accurate genetic information. In cross-sectional or longitudinal intervention studies, we can develop a new holistic eco-evolutionary understanding of microbes under different external circumstances (reviewed in reference 15).

In medical applications, detecting subclonal mutations is vital to predict and avoid treatment resistances, as used in cancer genetics (16). Awareness of these mutations is also relevant in infections; e.g., HIV treatments are combinatorial to avoid subclonal resistance mutations to become prevalent, and a higher frequency of subclonal mutations was linked to prolonged severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections (17). Subclonal mutations further helped disentangle SARS-CoV-2 transmission chains among multiple hosts (17). Using ancient metagenomes, the divergence times of modern gut bacteria were predicted (18). In URM, much more closely related genomes could be used to time transfer events. Thus, a combination of mutation rates and de novo and subclonal single nucleotide variants (SNVs) obtained from URM could be used to trace “transfer chains” in hospitals, care homes, or production plants, reconstructing the routes and timing of bacterial colonization. However, currently, this resolution is not available in metagenomics.

PERFORMANCE OF HIGH-RESOLUTION METAGENOMICS

Short read (e.g., Illumina) sequencing enables HRM through a combination of large read numbers and a relatively good quality. Illumina reads typically have $10^{-3}$ to $10^{-4}$ errors/nucleotide, meaning that 1 in 1,000 to 1 in 10,000 nucleotides (nt) are wrong. In large sequencing runs with typically billions of bases, this accumulates to millions of reported errors. Dedicated SNV calling tools can correct some of these but are limited by available computational resources, nonrandom errors in Illumina reads, and mapping biases (19). Thus, sequencing error rate does not directly translate to metagenomic resolution.

The earliest HRM approaches were reference based; i.e., sequencing reads were mapped to reference genomes to identify SNVs per species and metagenome (20). However, reference-based metagenomics has several disadvantages, primarily that the majority of unknown microorganisms is not represented in reference databases (21, 22). Further, mapping to a distantly related reference genome will bias and increase the SNV calling error (23). Last, the bacterial genome is very dynamic: our experiments showed that on average, only 72% of genes are shared among intraspecific strains from 155 bacterial species (24). This might also apply to gut bacteria: when we de novo reconstructed the genome of *Parabacteroides distasonis*, this genome was better...
covered in time series metagenomes from the same host (median, 98%) than those from other hosts (<80% coverage).

For the same species, two extremely similar strains were detected, occurring in cohabiting family members. These strains were only 185 nt distant (99.996% average nucleotide identity [ANI]) but could be reliably assigned to dozens of time series samples from their respective human hosts. Thus, using de novo-reconstructed genomes, we could effectively resolve strains at <4 × 10^-4 errors/nucleotide (25). This workflow is implemented in our MATAFILER pipeline and was recently used to track 440 species in 5,278 metagenomes (9). To estimate the metagenomic resolution on real data, we compared the nucleotide sequence of 26,000 strains persisting between longitudinal samples. The average ANI was 99.98% (estimated error rate, 2 × 10^-3), although 55% of these strains were reconstructed at 100% ANI (9). This resolution compares favorably to other strain-level pipelines when evaluated on real metagenomes (see Fig. 2d in reference 26), exemplifying resolution of current HRM.

HRM typically use a “consensus” SNV calling approach, considering only the major (>0.5 frequency) alleles in a population. Typically, samples are excluded if ≥2 conspecific strains of a species are present. In contrast, HRM approaches such as STRONG phase SNVs to reconstruct strain-specific genomes, even from mixtures of conspecific strains co-occurring in the same samples, delineating strains at 99.95% ANI (5 × 10^-3 errors/nucleotide) (27). In a different strategy, inStrain (26) reports minor and major alleles indiscriminately, reporting high strain similarities in simulations (99.9998% ANI, 2 × 10^-5 errors/nucleotide) but also reporting sequencing errors.

To distinguish true biological variation from sequencing errors is one of the biggest challenges to enabling URM.

ENABLING ULTRA-RESOLUTION METAGENOMICS

The bacterial mutation rate is estimated at between 10^-8 and 10^-5 mutations/nucleotide/year (28), and population genetics theory predicts that neutral mutations get fixed at a similar rate. To identify bacterial de novo mutations fixed after 1 year would thus require a metagenomic resolution of <1 e-5 errors/nucleotide. Detecting subclonal mutations at a minor allele frequency (MAF) of =0.5 would require a similar error rate but a much higher read coverage, approximated recently at >5× coverage for Illumina reads (26). Therefore, URM requires a combination of (i) increased sequencing coverage, (ii) read accuracy, (iii) phasing of SNVs, and (iv) using orthogonal genetic markers, as I briefly discuss below.

Increased read coverage is enabled through inexpensive, high-throughput sequencing. However, to cover more species from an ecosystem, a linear increase in sequencing depth is unlikely to cover its rare members, as species abundance curves in ecosystems follow a power law. We are therefore exploring alternatives to systematically increase the coverage (and thereby resolution) of underrepresented microbes. One of our bioinformatic solutions is to fine-tune strain delineation at ≥2× coverage (9). Experimentally, we demonstrated the potential of microfluidics to enrich particular microbes (29), while single-cell sequencing can increase read coverage of a single cell, allowing phasing of subclonal mutations.

Simply increasing sequencing coverage does not scale linearly to SNV calling accuracy, since PCR and Illumina sequencing errors are not random (30). Single-molecule sequencing like PacBio has the advantage of having random read errors, allowing for better error correction. This is implemented in PacBio HiFi sequencing, reaching under optimal conditions an error rate below 10^-5. Illumina read accuracy can be further increased to potential error rates of ≤10^-7 using molecular barcoding techniques (31). Bioinformatically, read error profiles can be generated to further increase accuracy, as used in DADA2 (6). In metagenomes this is not applicable, but instead, spiking artificial DNA into a sample can be used to calculate sample-specific error profiles.

Phasing SNVs is possible with short read sequencing but inefficient when very similar genomes are co-occurring. In the URM context, other techniques need to be
employed, such as single-cell sequencing or barcoding DNA molecules (32). However, using long read sequencing (Oxford Nanopore or PacBio) at high read accuracy might be the most straightforward approach.

Instead of SNVs, structural variants (SVs) of microbial genomes can be used to track bacterial evolution. Even short read data can reliably track small SVs such as indels and microsatellites, and the confidence in observing a specific, e.g., 5-bp, deletion across samples is significantly higher than repeatedly observing a single SNV. SVs can inform on the evolution of bacteria; for example, we used microsatellites to reconstruct the demography and spread of *Mycobacterium tuberculosis* (33). With long read technologies, the detection of larger inversions, insertions, and deletions will become more commonplace, allowing increased resolution of evolutionary processes.

**CONCLUSION**

Ultra-resolution metagenomics (URM), here defined by the ability to detect ultra-rare mutations within a microbial population, is the next logical step in metagenomics. For evolutionary and epidemiological science, this resolution will revolutionize our understanding of microbial ecosystems. Given the importance of this technology, the work of my group is focused on providing both experimental and bioinformatic solutions to get the most out of current HRM and future URM data. Given the rapidly increasing accuracy and throughput of long read sequencing, I predict that standard metagenomic sequencing data will be URM compatible within 5 years.

**ACKNOWLEDGMENTS**

I am grateful to Marnix H. Medema for encouraging me to write this article and to C. Quince, W. Haerty, and the Hildebrand lab for discussions and help in conceiving this article.

The lab is supported by European Research Council H2020 StG (erc-stg-948219, EPYC) and the Biotechnology and Biological Sciences research Council (BBSrC) Institute Strategic Program Gut Microbes and Health BB/r012490/1 and its constituent project BBS/e/F/000Pr10355.

**REFERENCES**

1. Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LFF, Sloan WT. 2009. Accurate quantification of microbial diversity from 454 pyrosequencing data. Nat Methods 6:639–641. https://doi.org/10.1038/nmeth.1361.
2. Tedesoro L, Anslan S, Bahram M, Pólme S, Riit T, Liiv I, Kõljalg U, Kisand V, Nilsson H, Hildebrand F, Bork P, Abarenkov K. 2015. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metatranscriptomic analyses of fungi. MycoKeys 10:1–43. https://doi.org/10.3897/mycokeys.10.4852.
3. Bahram M, Anslan S, Hildebrand F, Bork P, Tedesoro L. 2018. Newly designed 16S rRNA barcoding primers amplify diverse and novel archaeal taxa from the environment. Environ Microbiol Rep 11:487–494. https://doi.org/10.1111/1758-2229.12684.
4. Bedarf JR, Beraza N, Khazneh H, Öz Kurt E, Baker D, Borger V, Wüllner U, Hildebrand F. 2021. Much ado about nothing? Off-target amplification can lead to false-positive bacterial brain microbiome detection in healthy and Parkinson’s disease individuals. Microbiome 9:75. https://doi.org/10.1186/s40668-021-01012-1.
5. Hildebrand F, Tadeo R, Voigt A, Bork P, Raes J. 2014. LotusSu: an efficient and user-friendly OTU processing pipeline. Microbiome 2:30. https://doi.org/10.1186/2049-2618-2-30.
6. Calahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869.
7. de Oliveira Martins L, Page AJ, Mather AE, Charles IG. 2019. Taxonomic resolution of the ribosomal RNA operon in bacteria: implications for its use with long-read sequencing. NAR Genom Bioinform 2:qz0016. https://doi.org/10.1093/nargab/qz0016.
8. Frioux C, Singh D, Korcsmanos T, Hildebrand F. 2020. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Comput Struct Biotechnol J 18:1722–1734. https://doi.org/10.1016/j.csbj.2020.06.028.
9. Hildebrand F, Gossmann Tl, Frioux C, Öz Kurt E, Myers PN, Ferretti P, Kuhn M, Bahram M, Nielsen HB, Bork P. 2021. Dispersal strategies shape persistence and evolution of human gut bacteria. Cell Host Microbe 29:1167–1176.e9. https://doi.org/10.1016/j.chom.2021.05.008.
10. Roodgar M, Good B, Garud N, Martis S, Avula M, Zhou W, Lancaster S, Lee H, Babveyh A, Nesamoney S, Pollard K, Snyder M. 2019. Longitudinal linked read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. bioRxiv. https://doi.org/10.1101/2019.12.21.886093.
11. Li É, de Jonge R, Liu C, Jiang H, Friman V-P, Pieterse CMJ, Bakker PAHM, Jousset A. 2021. Rapid evolution of bacterial mutualism in the plant rhizosphere. Nat Commun 12:3829. https://doi.org/10.1038/s41467-021-24005-y.
12. Dingemans J, Ye L, Hildebrand F, Tontodounati F, Crags M, Biloqc F, De Vos D, Crabbee A, Van Houdt R, Malfroot A, Cornelis P. 2014. The deletion of TonB-dependent receptor genes is part of the genome reduction process that occurs during adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis lung. Pathog Dis 71:26–38. https://doi.org/10.1111/1409-632X.12170.
13. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ. 2019. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe 25:656–667.e8. https://doi.org/10.1016/j.chom.2019.03.007.
14. Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of TonB-dependent receptor genes. Nat Genet 44:1275–1280. https://doi.org/10.1038/ng.2632X.
15. Garud NR, Pollard KS. 2020. Population genetics in the human microbiome. Trends Genet 36:53–67. https://doi.org/10.1016/j.tig.2019.10.010.
16. Schmitt MW, Kennedy SR, Poyet M, Alarcon M, Groussin M, Davis MR, Schmitt MW, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 109:14508–14513. https://doi.org/10.1073/pnas.1208715109.
23. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, Stoeßer N, PETA, Crook DW, Walker AS. 2020. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. Giga-science 9:653774. https://doi.org/10.1093/gigascience/giaa007.

24. Maisterenko OM, Mende DR, Luefte M, Hildebrand F, Schmidt TS, Li SS, Rodrigues JFM, von Mering C, Pedro Coelho L, Huerta-Cepas J, Sunagawa S, Bork P. 2020. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J 14:1247–1259. https://doi.org/10.1038/s41396-020-0600-z.

25. Hildebrand F, Moitinho-Silva L, Blasche S, Jahn MT, Gossmann T, Huerta-Cepas J, Hercog R, Luefte M, Bahram M, Przy狭ak A, Alves RJ, Waszak SM, Zhu A, Ye L, Costea P, Aalvink S, Belzer C, Forslund SK, Sunagawa S, Hentschel U, Merten C, Patil KP, Benes V, Bork P. 2019. Antibiotics-induced monodominance of a novel gut bacterial order. Gut 68:1781–1790. https://doi.org/10.1136/gutjnl-2018-317715.

26. Olm MR, Crits-Christoph A, Bouma-Gregson K, Fike BA, Morowitz MJ, Banfield JF. 2021. inStrain profiles population microbiome diversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol 39:727–736. https://doi.org/10.1038/s41587-020-00797-0.

27. Tracanna V, Ossowicki A, Petrus MLC, Overduin S, Terlouw BR, Lund G, Robinson SL, Weill F-X, Le Hérel E, Hawkey J, Edwards DJ, Fourment M, Holmes EC. 2016. Genomic-scale rates of evolutionary change in bacteria. Microb Genom 2:e000094. https://doi.org/10.1099/mgen.0.000094.

28. Duchêne S, Holt KE, Weill F-X, Le Hérel E, Hildebrand F, Merten C. 2021. Enrichment of gut microbiome strains for cultivation-free genome sequencing using droplet microfluidics. arXiv 2106.01455 [q-bio.GN].

29. Pryšlak A, Wenzel T, Seitz KW, Hildebrand F, Kartal E, Cosenza MR, Benes V, Bork P, Merten C. 2019. Enrichment of gut microbiome strains for cultivation-free genome sequencing using droplet microfluidics. Nat Biotechnol 372:eabg0821. https://doi.org/10.1038/s41579-020-0368-1.