ARTICLE

Precision Metagenomics: Rapid Metagenomic Analyses for Infectious Disease Diagnostics and Public Health Surveillance

Ebrahim Afshinnekoo,[1,2,3,] Chou Chou,[1,2] Noah Alexander,[1,2] Sofia Ahsanuddin,[1,2] Audrey N. Schuetz,[4] and Christopher E. Mason[1,2,5,†]*

[1]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York 10065, USA; [2]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, New York, New York 10021, USA; [3]School of Medicine, New York Medical College, Valhalla, New York 10595, USA; [4]Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine and Science, Rochester, Minnesota 55905, USA; and [5]Feil Family Brain & Mind Research Institute, New York, New York 10065, USA

Next-generation sequencing (NGS) technologies have ushered in the era of precision medicine, transforming the way we treat cancer patients and diagnose disease. Concomitantly, the advent of these technologies has created a surge of microbiome and metagenomic studies over the last decade, many of which are focused on investigating the host-gene-microbial interactions responsible for the development and spread of infectious diseases, as well as delineating their key role in maintaining health. As we continue to discover more information about the etiology of infectious diseases, the translational potential of metagenomic NGS methods for treatment and rapid diagnosis is becoming abundantly clear. Here, we present a robust protocol for the implementation and application of "precision metagenomics" across various sequencing platforms for clinical samples. Such a pipeline integrates DNA/RNA extraction, library preparation, sequencing, and bioinformatics analyses for taxonomic classification, antimicrobial resistance (AMR) marker screening, and functional analysis (biochemical and metabolic pathway abundance). Moreover, the pipeline has 3 tracks: STAT for results within 24 h; Comprehensive that affords a more in-depth analysis and takes between 5 and 7 d, but offers antimicrobial resistance information; and Targeted, which also requires 5–7 d, but with more sensitive analysis for specific pathogens. Finally, we discuss the challenges that need to be addressed before full integration in the clinical setting.

Key Words: pathogen detection, antimicrobial resistance, microbiome, next-generation sequencing

TRANSFORMING THE CURRENT PARADIGM OF INFECTIOUS DISEASE DIAGNOSTICS

Infectious diseases are a leading cause of death among children, adolescents, and adults worldwide, especially in the immunocompromised patient population.[1] Despite significant advancements in infectious disease diagnostics, most current methods still rely on microbial isolation, targeted PCR- or microarray-based assays, and serological methods. In many cases, these techniques suffer from unacceptably slow turnaround (from several days to weeks), as well as have persistent difficulty in detecting infections or coinfections as a result of technical challenges in species abundance or extraction bias because of differential lysis of distinct species. Moreover, a large number of fastidious microorganisms are difficult to grow or study with routine culture-based techniques.[2] Such diagnostic challenges are difficult and ever-more paramount, given the overuse and misuse of antimicrobials in human and animal populations, which have led to a marked increase in organisms with high numbers of AMR determinants that pose an increasing risk of treatment failure in the developed and developing world.[3]

Challenges in nucleic acid extraction, quantification, and characterization from microbial samples are perhaps most critical in clinical settings, wherein delay in diagnosis and treatment of an infection can lead to poor outcomes.[4] Therefore, a molecular diagnostic framework that affords rapid and accurate information regarding pathogen identity and AMRs would greatly reduce the prescription of ineffective antimicrobials, likely decrease AMR rates, and lead to favorable clinical outcomes. Additionally, such an assay could inform the course of infection and control outbreaks by offering real-time, high-resolution strain-typing data. Importantly, the reduction of diagnostic uncertainty and expedition of the time to optimal treatment have been shown to decrease cost of care and improve patient survival.[5]

*Address correspondence to: Ebrahim Afshinnekoo, Dept. of Physiology and Biophysics, Weill Cornell Medicine, 1305 York Ave., New York, NY 10021, USA (Phone: 516-734-1870; E-mail: eba2001@med.cornell.edu).

†Address correspondence to: Christopher E. Mason, Dept. of Physiology and Biophysics, Weill Cornell Medicine, 1305 York Ave., New York, NY 10021, USA (Phone: 203-668-1448; E-mail: chm2042@med.cornell.edu).

doi: 10.7171/jbt.17-2801-007

Recently, multiple studies have shown that metagenomic (sequencing of all genomes in a sample) technologies are a promising means to identify and track infectious etiologies responsible for outbreaks[6] and can also reveal genetic drivers of AMR or pathogenesis severity.[7] Notably, these studies use massively parallel NGS technologies to generate whole-genome sequence (WGS) data to profile comprehensively hundreds of organisms in a sample, in as little as 12 h,[8] and even have the capability to discover new infectious agents.[9] However, despite their potential, the application of metagenomics methods for characterization of clinical samples in real time for patient care is still greatly underused.[10]

Since the early work of the Human Microbiome Project,[11] the concept that microorganisms live in, on, and around us and ultimately affect our well-being has slowly become integrated into the clinical realm.[12–16] However, many clinical microbiome studies have focused on bacterial 16S rRNA gene sequencing, which although inexpensive and effective for large studies, can miss the putative infectious agent, if fungal, parasitic, or viral in origin. Furthermore, 16S rRNA gene sequencing is limited to most bacteria and does not provide information regarding AMR determinants, virulence factors, or strain type. Although multiplex PCR, microarray, or proteomic assays provide pathogen identification within hours from sample collection, these platforms are constrained by *a priori* assumptions of the expected cause of infection,[17] whereas shotgun-based assays are less biased and open ended, allowing for cross-kingdom analysis.

In a manner similar to how NGS has revolutionized our understanding, assessment, and treatment of cancer—ushering in the era of precision medicine—these same technologies have the potential to launch precision metagenomics and transform our approach to the management of infectious diseases and public health surveillance.

PRECISION METAGENOMICS PIPELINE

NGS technologies have already begun to be introduced into clinical microbiology workflows. Whereas some groups have developed protocols that allow for 48 h turnaround, these methods use 16S rRNA gene sequencing rather than WGS data.[18] **Figure 1** depicts our precision metagenomics pipeline that incorporates 3 major tracks: 1) STAT Track, 2) Comprehensive Track, and 3) Targeted Track. All tracks are initiated at the laboratory, where clinical specimens are collected and processed. Following standardized institutional protocols for sample collection, clinical specimens will be sent to the laboratory, and nucleic acids (DNA and RNA) are extracted. Within 24 h, clinical specimens processed along the STAT Track are sequenced with nanopore and/or single-molecule, real-time sequencing technology and transferred to our bioinformatics pipeline to identify DNA-based microbial pathogens, AMR determinants, and virulence factors within hours. The precision metagenomics pipeline ideally will be integrated into the clinical workflow, with the medical team (clinician, clinical microbiologist, etc.) determining which track is relevant for the patient. Urgent cases are designed to be dealt with by the STAT Track, whereas other less time-sensitive scenarios can be processed through the Comprehensive Track or the Targeted Track if there is a specific organism or amplification needed.

Less clinically urgent cases may undergo further in-depth analysis via the Comprehensive Track and/or Targeted Track to capture all bacteria, mycobacteria, fungi, parasites, and DNA and RNA viruses (Fig. 1). Compared with the STAT Track, the Comprehensive Track will be RNA enriched and offers antimicrobial resistance information, and the Targeted Track will be PCR amplified, allowing for the deeper analysis compared with its rapid counterpart but without antimicrobial resistance information. These samples, in turn, are sequenced using a platform such as MiSeq (Illumina, San Diego, CA, USA) and all sequenced reads transferred for bioinformatics analyses. These 2 tracks would span 5–7 d, still comparable in terms of turnaround time to most routine bacterial culture-based tests.

Upon completion, a report is generated with a list of microbial species, subspecies, and individual strains, identified in the sample (if perceived to be clinically relevant) with their relative abundances. For all analyses, an ensemble approach integrating the best and current bioinformatics tools for accurate identification of microorganisms, AMR determinants, and virulence factors will be used. Moreover, functional analyses can reveal metabolic pathways present and those genes that are transcriptionally active from the extracted RNA, which can further explain the etiology and pathophysiologic mechanisms behind a patient's infection and symptoms. Further research and studies will also integrate this with immunome profiles of patients, allowing one to understand the dynamics of the host response to infection and perhaps assist with our understanding and differentiation of colonization versus active infection.

CHALLENGES

Whereas the protocol outlined in Fig. 1 integrates all of the components of metagenomics and offers a roadmap to its application, there are still many challenges that must be addressed. These challenges are associated with key steps in the pipeline: specimen collection, specimen processing, and bioinformatic analyses.

Specimen Collection

Similar to complex environmental samples, which have a plethora of variables and uncertainties in sample

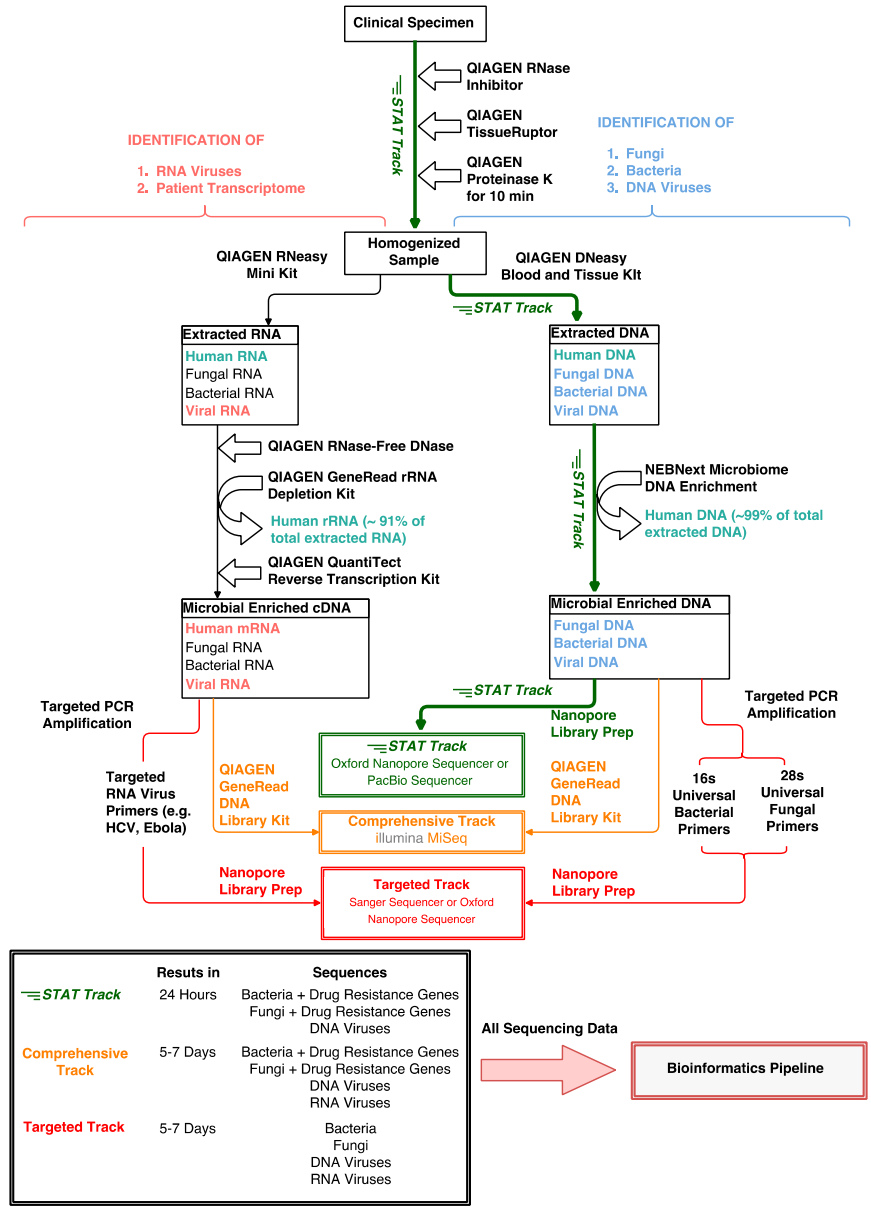

FIGURE 1

Precision metagenomics pipeline. Workflow for application of precision metagenomics for infectious disease diagnostics. There are 3 tracks through which a specimen can be processed: STAT (24 h), Comprehensive, and Targeted. All tracks feed into a bioinformatics pipeline for taxa classification, screening of antimicrobial resistance and virulence markers, and functional analysis.

collection, composition of clinical samples can greatly vary based on the type of specimen collected. The development of a standardized operating procedure for specimen collection is crucial for reducing risk of contamination and avoiding potential biases. It will also be essential to have various sample collection devices and protocols for the different types of specimens that can be submitted into the various tracks. As the field grows, and our understanding of the human microbiome expands, methods will likely need to be developed.

Specimen Processing

This phase can be divided into two key components: nucleic acid extraction, library preparation, and sequencing. Sample processing is often the bottleneck in microbiome and metagenomic pipelines, and depending on the number of samples to be processed, the kits and sequencing platform used can take anywhere from hours to days.

Nucleic acid extraction

Nucleic acid extraction poses a particular challenge, as ideally both DNA and RNA would be extracted from a specimen to capture all of the microorganisms and provide an opportunity for more meaningful analyses. This is the strength of WGS data compared with amplicon sequencing for infectious disease diagnostics,

but it is also a double-edged sword, as some clinical specimens (*e.g.,* vaginal, nasal, and throat) may contain extremely high levels of human DNA/RNA[19] that dominate the total nucleic acid composition of a sample. Therefore, these samples may require microbial nucleic acid enrichment through depletion of host DNA/RNA or use of methylation or selective lysis as a method to remove human DNA.[20] As previous researchers have revealed, certain extraction methods may alter the natural composition of the microbiome and introduce bias[21]—the so-called "kit-ome"; thus, providing a list of potential organisms that could be recovered by the various tracks that will assist with these potential kit-based biases. Recently, methods and kits are being developed that allow simultaneous DNA and RNA extraction from a single sample[22–24] or use of a poly-enzyme cocktail that can more thoroughly digest different microbes.[25] Further developments on physical, enzymatic, or chemical extraction and DNA/RNA yield optimizations will help to improve metagenomics and metatranscriptomics studies and aid the adoption of these techniques to clinical care.

Library preparation and sequencing

Additional consideration of the chemistries and steps in library preparation will need to be made, and they will depend on the sequencing platform to be used (*e.g.,* Illumina; Pacific Biosciences, Menlo Park, CA, USA; and Oxford Nanopore Technologies, Oxford, United Kingdom). Oxford Nanopore Technologies sequencing technology has already been established as a tool for rapid sequencing,[26] making it a viable option for the STAT Track. One key challenge inherent to DNA-based NGS methods is organism viability, as even detection of viable RNA is not necessarily a measure of the biology (the cell may have died during processing). To resolve this question, some platforms have integrated live/dead assays[27] into their molecular workflows to address this question; a similar approach could be used with our proposed precision metagenomics pipeline. Even if viability is addressed, several other distinctions need to be made, such as among colonization versus infection, estimated pathogenicity, host immune system state, and co-occurring species.

Bioinformatics Analyses

With the advent of NGS, clinical specimens can be characterized rapidly and richly, but a gap remains in terms of the computational methods for accurate classification of metagenomic samples and dissemination of clinically useful information. Some of our ongoing work has shown the best way to address this challenge is to use an ensemble approach of analysis that uses 2–3 tools with different bioinformatics approaches (k-mer, marker and alignment based) to ensure the highest sensitivity and specificity for taxonomic classification.[28] Moreover, positive-control mock communities can help ensure that proper parameters and filters are applied to rule out false positives.[28] Likewise, for detection and accurate annotation of AMR and virulence factors, a comprehensive, well-curated database is essential. Functional analysis involves the identification of biochemical and metabolic pathways and their relative abundance, which can help to explain the molecular mechanisms that the organisms use and on which they thrive.

The power of metagenomics and WGS is that it allows us to not only study what organisms are in a sample but also what they are doing and how they are doing it. As previously suggested, human reads can be a challenge both in terms of analysis and privacy concerns,[18, 29] but they also can confirm the patient's identify to rule out sample mislabeling or contamination.[29] The optimization of protocols in the sample-processing phase, coupled with a computational filtering process in the bioinformatics pipeline, will ensure that these reads do not impact subsequent analyses or interpretations, as well as guarantee that the patient's privacy and genomic rights are respected. One of the ultimate challenges facing the fields of metagenomics and microbiome is performing effective subspecies and strain-level identification, which could play a role in the patient's infection or be appropriated for epidemiologic purposes. Indeed, just as the human genome's refinement over the years has led to improved use in precision medicine,[30, 31] the same is likely to occur with expanding and improved reference genomes for the microbiome.

Finally, recent work in the field epigenetics and RNA modifications (epitranscriptome) has shown an expansive catalog of modified DNA and RNA in the microbiome and metagenome. This includes modified DNA or RNA bases, such as methyl-6-adenosine, which has been discovered on many bacteria,[7, 32] and also on all examined RNA viruses to date, including HIV, dengue, Zika, yellow fever, West Nile, and influenza.[33–35] The use of these nascent discoveries of microbial regulation in a clinical context requires further development of current tools and methods of enriching, detecting, and computationally specifying the precise sites of modified bases (epigenetic and epitranscriptomic) and then eventual integration with other tools that are currently improving the detection of microbes.[36–39] Just as the genome is examined in the context of the epigenome for its regulation, viral RNAs and modified bases in bacteria will eventually be examined and also understood

through these additional lenses of host and microbial regulation.

CONCLUSION

Despite the challenges and complexity of metagenomics, it has the potential to offer a more comprehensive molecular profile of a patient's metagenome and microbiome. These techniques would allow clinicians and researchers to identify the etiological agent(s) of infection, AMR determinants, the presence of virulence factors, disease-specific host biomarkers, and microbial metabolic activity. DNA sequencers have essentially become "molecular microscopes" that empower scientists to examine and explore clinical samples, and nature in general, in a novel way. It is time we better translate this to the clinical practice of infectious diseases and public health.

AUTHORSHIP

E.A. and C.E.M. led the writing of the manuscript and developed the idea of precision metagenomics and its experimental design. C.C. and N.A. developed the precision metagenomics pipeline. E.A, C.E.M., C.C., A.N.S., and S.A. edited the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

The authors thank Rita Colwell, Nur Hasan, and Manoj Dadlani for their help in discussions of the ideas for this manuscript and Sofia Ahsanuddin for help in formatting the manuscript for submission. The authors also thank the following for funding: Starr Cancer Consortium (Grants I7-A765, I9-A9-071), Irma T. Hirschl and Monique Weill-Caulier Charitable Trusts, Bert L. and N. Kuggie Vallee Foundation, WorldQuant Foundation, Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH50G, NNX17AB26G), U.S. National Institutes of Health (R25EB020393, R01NS076465, R01AI125416, R01ES021006), Bill and Melinda Gates Foundation (OPP1151054), and Alfred P. Sloan Foundation (G-2015-13964).

DISCLOSURES

The authors herein declare that this research was conducted in the absence of any financial or commercial interests that could be potentially regarded as a conflict of interest.

REFERENCES

1. Fishman JA. Infections in immunocompromised hosts and organ transplant recipients: essentials. *Liver Transpl* 2011;17 (Suppl 3):S34–S37.
2. Vartoukian SR, Palmer RM, Wade WG. Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol Lett* 2010;309:1–7.
3. Liu YY, Wang Y, Walsh TR, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect Dis* 2016;16:161–168.
4. Kumar A, Roberts D, Wood KE, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med* 2006;34:1589–1596.
5. Kaplan RS, Porter ME. How to solve the cost crisis in health care. *Harv Bus Rev* 2011;89:46–52, 54, 56–61 passim.
6. Rasko DA, Webster DR, Sahl JW, et al. Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* 2011;365:709–717.
7. Fang G, Munera D, Friedman DI, et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* 2012;30:1232–1239.
8. Mason CE, Porter SG, Smith TM. Characterizing multi-omic data in systems biology. *Adv Exp Med Biol* 2014;799: 15–38.
9. Stenglein MD, Sanders C, Kistler AL, et al. Identification, characterization, and in vitro culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: candidate etiological agents for snake inclusion body disease. *MBio* 2012;3:e00180–e12.
10. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis* 2012;6:e1485.
11. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; 486:207–214.
12. Kuntz TM, Gilbert JA. Introducing the microbiome into precision medicine. *Trends Pharmacol Sci* 2017;38:81–91.
13. Zaborin A, Smith D, Garfield K, et al. Membership and behavior of ultra-low-diversity pathogen communities present in the gut of humans during prolonged critical illness. *MBio* 2014;5:e01361–e14.
14. Shogan BD, Smith DP, Christley S, Gilbert JA, Zaborina O, Alverdy JC. Intestinal anastomotic injury alters spatially defined microbiome composition and function. *Microbiome* 2014;2:35.
15. Gilbert JA, Quinn RA, Debelius J, et al. Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* 2016;535:94–103.
16. Bluemel S, Williams B, Knight R, Schnabl B. Precision medicine in alcoholic and nonalcoholic fatty liver disease via modulating the gut microbiota. *Am J Physiol Gastrointest Liver Physiol* 2016;311:G1018–G1036.
17. Schubert RD, Wilson MR. A tale of two approaches: how metagenomics and proteomics are shaping the future of encephalitis diagnostics. *Curr Opin Neurol* 2015;28: 283–287.
18. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. *Nat Rev Genet* 2014;15:409–421.
19. Quinn RA, Navas-Molina JA, Hyde ER, et al. From sample to multi-omics conclusions in under 48 hours. *mSystems* 2016;1: pii: e00038–16.
20. New England Biolabs. NEBNext Microbiome DNA Enrichment Kit. Available at: https://www.neb.com/products/e2612-nebnext-microbiome-dna-enrichment-kit.
21. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014;12:87.
22. Qiagen. AllPrep DNA/RNA Mini Kit. Available at: https://www.qiagen.com/us/shop/sample-technologies/combined-sample-technologies/preparation/allprep-dnarna-mini-kit/.
23. PrimerDesign.genesig Easy. DNA/RNA Extraction Kit. Available at: http://www.primerdesign.co.uk/assets/files/extraction_kit_handbook.pdf?timestamp=148 5225049.

24. Thermo Fisher Scientific. Available at: https://www.thermofisher.com/us/en/home/life-science/dna-rna-purification-analysis.html.
25. The MetaSUB Consortium. MetaSUB—The Global Metagenomics and Metadesign of the Subways and Urban Biomes Consortium. *Microbiome* 2016;4:24.
26. Greninger AL, Naccache SN, Federman S, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* 2015;7:99.
27. Qiagen. BLU-V Viability PMA Kit. Available at: https://www.qiagen.com/us/shop/sample-technologies/dna/dna-preparation/blu-v-viability-pma-kit/#orderinginformation
28. Afshinnekoo E, Meydan C, Chowdhury S, et al. Geospatial resolution of human and bacterial diversity from city-scale metagenomics. *Cell Syst* 2015;1:72–87.
29. Callaway E. Microbiome privacy risk. *Nature* 2015;521:136.
30. Rosenfeld J, Mason CE, Smith T. Limitations of the human genome reference. *PLoS One* 2012;7:e40294.
31. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data.* 2016;3:160025.
32. Kurylo CM, Alexander N, Dass RA, et al. Genome sequence and analysis of *Escherichia coli* MRE600, a colicinogenic, nonmotile strain that lacks RNase I and the type I methyltransferase, EcoKI. *Genome Biol Evol* 2016;8:742–752.
33. Lichinchi G, Gao S, Saletore Y, et al. Dynamics of the human and viral m(6)A RNA methylomes during HIV-1 infection of T cells. *Nat Microbiol* 2016;1:16011.
34. Gokhale NS, McIntyre AB, McFadden MJ, et al. N6-m=Methyladenosine in Flaviviridae viral RNA genomes regulates infection. *Cell Host Microbe* 2016;20:654–665.
35. Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE. The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol* 2012;13:175.
36. Li S, Garrett-Bakelman FE, Akalin A, et al. An optimized algorithm for detecting and annotating regional differential methylation. *BMC Bioinformatics* 2013. 14 (Suppl 5), S10.
37. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;12:902–903.
38. Eppley JM, Tyson GW, Getz WM, Banfield JF. Strainer: software for analysis of population variation in community genomic datasets. *BMC Bioinformatics* 2007;8:398.
39. Choi SY, Rashed SM, Hasan NA, et al. Phylogenetic diversity of *Vibrio cholerae* associated with endemic cholera in Mexico from 1991 to 2008. *MBio* 2016;7:e02160.