Reply to "Expanding the Conversation on High-Throughput Virome Sequencing Standards To Include Consideration of Microbial Contamination Sources"

Jason T. Ladner, Michael R. Wiley, Gustavo Palacios

Center for Genome Sciences, United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, Maryland, USA

We thank Strong et al. (1) for this opportunity to continue the discussion of high-throughput (HT) sequencing standards for viruses. Along with colleagues from several other genome sequencing centers, we recently published an editorial proposing a common nomenclature for describing levels of finishing for viral genomes assembled with HT sequencing data (2). In addition, we discuss two ways of characterizing viral samples that go beyond consensus genome assembly, namely, the description of population-level diversity and the identification of contaminants or adventitious agents.

Due to limitations of space, we focused primarily on the importance of these types of characterizations and the strengths of HT sequencing for conducting such analyses. Strong et al. (1) expand upon this discussion by raising an issue associated with the interpretation of contaminating reads, specifically in relation to their true source. They rightly point out that the high sensitivity of HT sequencing, which is one of its primary strengths, can also be a liability since low levels of microbial nucleic acids introduced from reagents during library preparation will often be detected, and without a proper understanding of the types and sources of such contamination, the presence of these reads can lead to incorrect interpretations.

We completely agree. While HT sequencing is an incredibly powerful and transformative technology, caution needs to be taken in the interpretation of all HT sequencing results, especially when basing conclusions on reads or variants present at low frequencies. A similar but distinct source of potential contamination, briefly mentioned in our article, is from multiplexed samples run together on an instrument. In order to take full advantage of the current throughput offered by HT technologies, indexing (i.e., bar coding) and pooling samples have become commonplace. While in theory these distinct samples can be sorted bioinformatically after sequencing, several different types of errors have been shown to result in incorrect assignment of reads to samples, thus resulting in "bleed-through" between samples (3). It is important to appropriately quantify this source of error and to implement control measures to minimize the impact (e.g., dual indexes and index quality filtering) (3). While it may be impossible to completely eliminate such sources of error, precautions taken during experimental design and analysis can help mitigate the potential for incorrect interpretations.

Furthermore, the reverse of the warning from Strong et al. (1) is also true: the lack of detected contaminant reads does not guarantee the absence of contaminating microbes. In nearly all sequencing runs, there is a set of sequences that cannot be attributed to any particular source. Many of these are likely artifacts of library preparation, but given the incompleteness of the current reference databases, contamination from an unsequenced microbe cannot be ruled out. There is great interest in automating the analysis of HT sequencing data for the detection and characterization of viruses and microbes in clinical samples and other biological specimens. However, given the current state of the technology, it remains important for an expert to examine and interpret the results to avoid false-positive and -negative calls.

REFERENCES

1. **Strong MJ, Lin Z, Flemington EK.** 2014. Expanding the conversation on high-throughput virome sequencing standards to include consideration of microbial contamination sources. mBio 5(6):e01989-14. http://dx.doi.org/10.1128/mBio.01989-14.
2. **Ladner JT, Beitzel B, Chain PSG, Davenport MG, Donaldson EF, Frieman M, Kugelman J, Kuhn JH, O'Rear J, Sabeti PC, Wentworth DE, Wiley MR, Yu G-Y, The Threat Characterization Consortium, Sozhamannan S, Bradburne C, Palacios G.** 2014. Standards for sequencing viral genomes in the era of high-throughput sequencing. mBio 5(3):e01360-14. http://dx.doi.org/10.1128/mBio.01360-14.
3. **Kircher M, Sawyer S, Meyer M.** 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 40:e3. http://dx.doi.org/10.1093/nar/gkr771.

Address correspondence to Jason T. Ladner, jtladner@gmail.com.