AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems™

COMMENTARY
Applied and Environmental Science

Avoiding Pandemic Fears in the Subway and Conquering the Platypus

A. Gonzalez,[a] Y. Vázquez-Baeza,[b] J. B. Pettengill,[c] A. Ottesen,[c] D. McDonald,[d] R. Knight[a,b]

Department of Pediatrics, University of California San Diego, San Diego, California, USA[a]; Department of Computer Science and Engineering, University of California San Diego, San Diego, California, USA[b]; Food and Drug Administration, Center for Food Safety and Applied Nutrition (CFSAN), College Park, Maryland, USA[c]; Institute for Systems Biology, Seattle, Washington, USA[d]

ABSTRACT Metagenomics is increasingly used not just to show patterns of microbial diversity but also as a culture-independent method to detect individual organisms of intense clinical, epidemiological, conservation, forensic, or regulatory interest. A widely reported metagenomic study of the New York subway suggested that the pathogens *Yersinia pestis* and *Bacillus anthracis* were part of the "normal subway microbiome." In their article in mSystems, Hsu and collaborators (mSystems 1(3): e00018-16, 2016, http://dx.doi.org/10.1128/mSystems.00018-16) showed that microbial communities on transit surfaces in the Boston subway system are maintained from a metapopulation of human skin commensals and environmental generalists and that reanalysis of the New York subway data with appropriate methods did not detect the pathogens. We note that commonly used software pipelines can produce results that lack *prima facie* validity (e.g., reporting widespread distribution of notorious endemic species such as the platypus or the presence of pathogens) but that appropriate use of inclusion and exclusion sets can avoid this issue.

Published 28 June 2016

Citation Gonzalez A, Vázquez-Baeza Y, Pettengill JB, Ottesen A, McDonald D, Knight R. 2016. Avoiding pandemic fears in the subway and conquering the platypus. mSystems 1(3): e00050-16. doi:10.1128/mSystems.00050-16.

Address correspondence to A. Gonzalez, antgonza@gmail.com.

A.G. and Y.V.-B. contributed equally to this article.

For the article discussed, see http://doi.org/10.1128/mSystems.00018-16.

The views expressed in this Commentary do not necessarily reflect the views of this journal or of ASM.

The development and validation of novel methods that use next-generation DNA sequence data to detect pathogens from complex ecosystems represent important areas of research. In particular, these methods are important in studies of the built environment and of agricultural systems, where the correct detection of pathogens represents enormous public benefit and where incorrect detection creates fear. For example, in a recent study of the New York subway (1), due to incorrect taxonomic classifications, the authors reported observing *Yersinia pestis* (the causative agent of plague) and *Bacillus anthracis* (the causative agent of anthrax) as part of the "normal subway microbiome." These observations led to high-visibility news reports. But improved reanalysis of the same data by Hsu et al. (2) demonstrated that these results were illusory. Hsu et al. found that these pathogens were not part of the normal subway microbiome, either in New York or in an independent sample set from the Boston subway. They drew the more plausible conclusion that the surfaces were dominated by inputs of normal human skin bacteria, consistent with other studies, and found that the subway was not a reservoir of bacterially encoded toxins or antimicrobial resistance elements. That carefully conducted study added fundamentally to our knowledge of the transmission and expression of microbes in high-traffic built environments.

Another example of the importance of accurate pathogen identification from next-generation sequencing data is the ability to detect *Salmonella* from fresh produce. In a study by Ottesen et al. (3), the authors could not confirm the presence of *Salmonella* on the tomato crops through the use of 16S amplicon sequencing. However, an analysis of shotgun data from samples collected from the roots, leaves, and fruits of the tomato plants performed using the MG-RAST server reported hits corresponding to *Salmonella*. Furthermore, this analysis also showed the surprising presence of *Gallus gallus* (red

TABLE 1 Number of hits to specific taxa, living and extinct, and locations as reported by MG-RAST

Taxonomy	Extinct	Total no. of hits reported by MG-RAST	Main country locations (no. of hits [sorted by abundance])
Ornithorhynchus	No	17,140,078	Brazil (4,338,217), Australia (3,905,173), United States (2,669,553), Italy (2,665,186), Malawi (1,335,746), undefined (585,412), Kyrgyzstan (558,786), Russian Federation (333,978), South Africa (289,642), Belgium (198,052), Finland (168,848), China (50,542), Israel (27,366), Philippines (13,577)
Raphus	Yes	11	Brazil (8), Australia (3)
Salmonella	No	146,842,227	Italy (76,730,072), Brazil (33,956,417), United States (14,178,170), Malawi (3,808,783), China (3,383,261), Australia (3,354,697), undefined (3,257,862), Russian Federation (2,750,106), Finland (1,886,515), Belgium (1,373,668), South Africa (1,105,658), Israel (783,766), Philippines (232,026), Kyrgyzstan (41,226)
Thylacinus	Yes	1,344	Brazil (920), Australia (125), United States (80), Malawi (63), undefined (46), South Africa (32), Finland (23), Belgium (21), Russian Federation (15), Italy (13), Israel (4), China (2)

jungle fowl), *Mus musculus* (house mouse), and even the elusive *Ornithorhynchus anatinus* (duck-billed platypus).

Detecting the presence of specific taxa from MG-RAST public datasets. To exemplify the pervasiveness of false positives in MG-RAST, we downloaded all public samples (25,943 samples; accessed 22 April 2015), searched each report for *Salmonella*, *Raphus* (dodo bird), *Thylacinus* (Tasmanian tiger), and *Ornithorhynchus* (duck-billed platypus), and summarized the findings by the countries in which these organisms were observed on the basis of the latitude and longitude fields in the associated metadata (Table 1). A Jupyter (8) Notebook reproducing this report can be found in http://goo.gl/UlhBjf.

Conquering the platypus. To demonstrate how the problem of confirming the presence of specific taxa in metagenomic samples can be addressed, we created Platypus Conquistador (https://github.com/biocore/Platypus-Conquistador), a BSD-licensed Python package based on BLAST (4) and SortMeRNA (5). Platypus Conquistador confirms the presence or absence of a taxon of interest within shotgun metagenomic datasets by relying on two reference sequence databases: an inclusion database, which includes the sequences of interest (e.g., *Salmonella*), and an exclusion database, which includes any known sequence background (e.g., platypus). The reference sequence databases are expected to be mutually exclusive. In general, these two databases can be created by partitioning an existing database, such as the gene data provided by the Integrated Microbial Genomes (IMG) (6) system. These partitions can be customized to include taxa of specific interest. This method has been used by Ottesen et al. (7) to describe the efficacy of enrichment steps in the effort to culture *Salmonella* from tomatoes. For that analysis, the authors ran Platypus Conquistador on shotgun metagenomic data using the IMG database split into a reference database, including only those sequences assigned to *Salmonella*, and an exclusion database containing all remaining sequences, demonstrating the absence of this pathogen.

Conclusions. Simple bioinformatics solutions exist to detect taxa of interest and to resolve incorrect taxonomic classifications for shotgun sequencing data. Incorrect but pervasive taxonomic classifications can lead to conclusions that lack *prima facie* validity (for example, environments in which the platypus was reportedly found include environments from the built environment to the human gut). Worse, these incorrect assignments have great potential to spark unwarranted public concern, as was seen in the case of the NYC subway microbiome paper noted above.

These examples should also serve as a reminder that, although analytical software pipelines and computational methods can be thoroughly tested and validated, their results are based on user-specified parameters that change the results and, as a consequence, their validity. Researchers must always question the rationality of the parameters and meaning of the results to reduce the possibility of incorrect conclusions. Moving toward standardized and reproducible pipelines of analysis that can be scrutinized by our peers will greatly help avoid similar problems in the future. For

pathogen detection, it is critical to additionally define taxon inclusion and exclusion criteria based on the studied environment in order to discard misleading results. This is especially important in cases of intense public interest, such as exposure in systems used by millions of people every day to apparent pathogens that are as illusory as the benthic Platypus.

ACKNOWLEDGMENTS

We thank James Robert White for his suggestions and Elaine Wolfe for the design of the Platypus Conquistador logo.

This work was supported in part by the Sloan Foundation and the Crohn's and Colitis Foundation of America.

REFERENCES

1. **Afshinnekoo E, Meydan C, Chowdhury S, Jaroudi D, Boyer C, Bernstein N, Maritz JM, Reeves D, Gandara J, Chhangawala S, Ahsanuddin S, Simmons A, Nessel T, Sundaresh B, Pereira E, Jorgensen E, Kolokotronis SO, Kirchberger N, Garcia I, Gandara D, Dhanraj S, Nawrin T, Saletore Y, Alexander N, Vijay P, Henaff EM, Zumbo P, Walsh M, O'Mullan GD, Tighe S, Dudley JT, Dunaif A, Ennis S, O'Halloran E, Magalhaes TR, Boone B, Jones AL, Muth TR, Paolantonio KS, Alter E, Schadt EE, Garbarino J, Prill RJ, Carlton JM, Levy S, Mason CE**. 2015. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. Cell Syst **1:**72–87. http://dx.doi.org/10.1016/j.cels.2015.01.001.
2. **Hsu T, Joice R, Vallarino J, Abu-Ali G, Hartmann EM, Shafquat A, DuLong C, Baranowski C, Gevers D, Green JL, Morgan XC, Spengler JD, Huttenhower C**. 2016. Urban transit system microbial communities differ by surface type and interaction with humans and the environment. mSystems **1**(3)**:**e00018-16. http://dx.doi.org/10.1128/mSystems.00018-16.
3. **Ottesen AR, González Peña A, White JR, Pettengill JB, Li C, Allard S, Rideout S, Allard M, Hill T, Evans P, Strain E, Musser S, Knight R, Brown E**. 2013. Baseline survey of the anatomical microbial ecology of an important food plant: Solanum lycopersicum (tomato). BMC Microbiol **13:**114. http://dx.doi.org/10.1186/1471-2180-13-114.
4. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. 1990. Basic local alignment search tool. J Mol Biol **215:**403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.
5. **Kopylova E, Noé L, Touzet H**. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics **28:**3211–3217. http://dx.doi.org/10.1093/bioinformatics/bts611.
6. **Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC**. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res **40:**D115–D122. http://dx.doi.org/10.1093/nar/gkr1044.
7. **Ottesen AR, Gonzalez A, Bell R, Arce C, Rideout S, Allard M, Evans P, Strain E, Musser S, Knight R, Brown E, Pettengill JB**. 2013. Co-enriching microflora associated with culture based methods to detect salmonella from tomato phyllosphere. PLoS One **8:**e73079. http://dx.doi.org/10.1371/journal.pone.0073079.
8. **Pérez F, Granger BE**. 2007. IPython: a system for interactive scientific computing. Comput Sci Eng **9:**21–29. http://dx.doi.org/10.1109/MCSE.2007.53.