International Journal of Infectious Diseases 42 (2016) 24–27

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Feasibility and effectiveness of a brief, intensive phylogenetics workshop in a middle-income country

S. Pollett [a,b,c,*], M. Leguia [a], M.I. Nelson [d], I. Maljkovic Berry [e], G. Rutherford [b], D.G. Bausch [a], M. Kasper [f], R. Jarman [e], M. Melendrez [e]

[a] US Naval Medical Research Unit No. 6, Lima, Peru
[b] Department of Epidemiology and Biostatistics, University of California – San Francisco, California, USA
[c] Marie Bashir Institute for Infectious Diseases and Biosecurity, University of Sydney, New South Wales, Australia
[d] Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA
[e] Walter Reed Army Institute of Research, Silver Spring, Maryland, USA
[f] Global Emerging Infections Surveillance and Response System, Armed Forces Health Surveillance Center, Silver Spring, Maryland, USA

ARTICLE INFO

Article history:
Received 8 September 2015
Received in revised form 29 October 2015
Accepted 4 November 2015

Corresponding Editor: Eskild Petersen, Aarhus, Denmark.

Keywords:
Phylogenetics
Bioinformatics
Training
Low- and middle-income country
Viral pathogens

SUMMARY

There is an increasing role for bioinformatic and phylogenetic analysis in tropical medicine research. However, scientists working in low- and middle-income regions may lack access to training opportunities in these methods. To help address this gap, a 5-day intensive bioinformatics workshop was offered in Lima, Peru. The syllabus is presented here for others who want to develop similar programs. To assess knowledge gained, a 20-point knowledge questionnaire was administered to participants (21 participants) before and after the workshop, covering topics on sequence quality control, alignment/formatting, database retrieval, models of evolution, sequence statistics, tree building, and results interpretation. Evolution/tree-building methods represented the lowest scoring domain at baseline and after the workshop. There was a considerable median gain in total knowledge scores (increase of 30%, $p < 0.001$) with gains as high as 55%. A 5-day workshop model was effective in improving the pathogen-applied bioinformatics knowledge of scientists working in a middle-income country setting.

© 2015 The Authors. Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Dramatic escalations in genome sequencing capacity and the scope of molecular analysis have been seen in recent years. There has been increased use of phylogenetics as a powerful tool in the surveillance and public health control of many pathogens. In the last 5 years alone there were over 750 papers published related to influenza phylogenetics, many of which pertained to low- and middle-income countries (LMIC).[1] Evolutionary analysis of sequence data has proven important in clarifying the epidemiology of emerging and re-emerging pathogens, for instance determining geographic routes of dengue virus diffusion in South East Asia,[2] the origins and transmission dynamics of the latest Ebola epidemic,[3,4] and whether certain tropical locales serve as global 'sources' for influenza A H3NS virus epidemics.[5] The training of scientists in LMIC, which face a significant communicable disease burden, is a global health priority,[6] and phylogenetics is a rapidly advancing field where training needs are quickly evolving. Epidemiologists, molecular biologists, and other scientists are increasingly expected to be able to infer and/or interpret genetic data obtained from sequenced genomes. However, many LMIC scientists may have limited access to the training necessary to gain these skills.

Several websites and publications offer diverse bioinformatic training resources, with SEQWiki alone containing 690 bioinformatics applications.[7–9] However, while many such resources offer detailed 'how-to' instructions, lists of recommended reading, software, or online-classes tend to be more focused on running programs and provide only limited instructions on how to interpret or troubleshoot data output. Some excellent bioinformatics and phylogenetic hands-on workshops exist, including the Multinational Influenza Seasonal Mortality Study (MISMS) workshops organized globally by the Fogarty International Center of the National Institutes of Health.[10] However, these workshops tend to be more advanced and aimed at participants who already have baseline training. Many of the other workshops are located in high-income regions and are prohibitively expensive for most LMIC participants.[11–15]

* Corresponding author.
E-mail address: spollett@med.usyd.edu.au (S. Pollett).

http://dx.doi.org/10.1016/j.ijid.2015.11.001
1201-9712/© 2015 The Authors. Published by Elsevier Ltd on behalf of International Society for Infectious Diseases. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

To help address this training gap, an intensive 5-day pathogen-focused phylogenetics workshop was offered free of charge in Lima, Peru. The syllabus and learning objectives are presented here to assist others who may want to develop similar programs in LMIC. In order to assess effectiveness and improve potential future workshops for this and other LMIC, the participants' knowledge in applied phylogenetics and bioinformatics was measured objectively before and after the workshop. Specifically, it was sought to describe the participants' baseline knowledge in core domains of pathogen-applied bioinformatics/phylogenetics and to demonstrate any gains in these domains (and overall) after completing the workshop.

2. Methods

2.1. Setting, population, and participant recruitment

The 5-day workshop took place in Lima, Peru, in January 2015 and was advertised via direct e-mails and social media to local Peruvian universities and research institutions.[16] As workshop capacity was limited, participants were selected on a competitive basis, with emphasis on the relevance of the workshop to their current scientific area of study or work (i.e., pathogen research with public health impact requiring phylogenetic analyses). Participants were required to bring their own laptop and have at least intermediate-level English.

2.2. Workshop format

The workshop was offered free of charge to participants. It was designed to train scientists in basic evolutionary analysis of pathogen nucleotide sequence data in a public health context, with dengue and influenza viruses being the exemplar pathogens. The specific objectives are listed in Table S1 (**Supplementary Material**) and the course syllabus is given in Table S2 (**Supplementary Material**). The workshop comprised morning lectures followed by hands-on afternoon analysis tutorial sessions with three instructors (MM, MN, and IMB, all PhD scientists with a background in pathogen molecular epidemiology, bioinformatics, and genomics). All lecture and tutorial materials were made available to participants via Google Docs before the workshop commenced (https://www.google.com/docs/about/), thus enabling a paperless workshop. Datasets and links for the download of free and demonstration evolutionary analysis software were also provided, and participants were encouraged to bring their own data.[17–21] The workshop and materials were provided in English, although two of the instructors were fluent in Spanish.

2.3. Measurement of participant characteristics and knowledge

A written 20-item knowledge assessment questionnaire was administered to all closed-session participants (n = 21) on day 1 and the same questionnaire was administered immediately after the workshop concluded on day 5. The questionnaire contained a range of questions in five core domains: sequence quality/cleaning, sequence alignment/formatting, database retrieval, evolution models, tree building, and other similar analytical methods, and results interpretation (see **Supplementary Material**). Each correctly answered question received a 1-point score, with a maximum possible score of 20. All information collected was non-identifiable with a code used to link pre- and post-workshop questionnaire results. The questionnaire was available in English and Spanish and was administered under examination conditions. This study was deemed neither research nor human subjects research by the Walter Reed Army Institute of Research (WRAIR) Institutional Review Board.

2.4. Data analysis

Frequencies and means (with the standard deviation (SD)) of participant characteristics and medians of total pre-workshop and post-workshop questionnaire scores (total and by each core domain) were calculated, and changes in questionnaire scores before and after the workshop were assessed. Statistical significance was calculated with non-parametric tests (Wilcoxon signed-rank test) due to the skewed data distributions and small number of observations. Associations of baseline scores (and changes in scores) with previous phylogenetic experience and years of working/training in science were also examined by Wilcoxon rank-sum test and Spearman rank correlation. The analysis was performed using Stata version 13 software (Stata Corp., College Station, TX, USA).

3. Results

Table 1 presents the characteristics of the 21 workshop participants, selected from 120 total applications. All were Peruvian and the majority (86%) were working/training in Peru, with the remainder working/training in Brazil. Students made up the greatest proportion of the participants (33%), followed by employees of the US Government (29%), Peruvian Government employees (24%), and those in academia (10%). The majority (52%) stated that they had some previous experience in phylogenetic analysis. The mean time spent working and training in science was 4.9 years (SD 3.4 years).

Table 2 presents the median baseline questionnaire scores by individual core domains and overall. The lowest scoring domain at the start of the workshop was in evolutionary models, tree-building, and other analytical methods (20%, median score 1 of a maximum possible 5 points), which was also the lowest scoring domain after the workshop (60%, 3/5 points). There were statistically significant gains in scores for each of the domains tested, with the exception of database retrieval for which the median score was 100% (2/2 points) before and after the workshop and the number of items tested was small. Of the four other domains with statistically significant changes in scores, the greatest gains were seen in results interpretation and evolution models/tree-building/other analytical methods. There was considerable gain in total overall knowledge scores after the workshop (30%, 6/20 points; $p < 0.001$) with gains as high as 55% (11/20 points).

Higher baseline median scores were seen in those with previous phylogenetic experience compared to those without ($p = 0.04$), but there was no statistically significant difference in magnitude of

Table 1
Characteristics of participants

	n	(%)
Total	21	100
Nationality		
Peruvian	21	100
Current location of work or training		
Peru	18	86
Brazil	3	14
Current work or training position		
Student	7	33
Academia (pre-faculty)	2	10
Academia (faculty)	0	0
Government (Peru)	5	24
Government (USA)	6	29
Industry	0	0
Other	1	5
Previous phylogenetic experience	11	52
Years training and working in science, mean (SD)	4.9 (3.4)	

SD, standard deviation.

Table 2
Participant baseline and post-workshop questionnaire scores in various phylogenetic core domains[a]

Core domain	Maximum possible score	Pre-workshop score[b]	Post-workshop score[c]	Change in paired scores[c]	*p*-Value[d]
Sequence quality and cleaning	5	3 (1; 1 to 5)	4 (1; 3 to 5)	1 (2; −1 to 3)	0.002
Sequence alignment and formatting	3	2 (0; 1 to 3)	3 (1; 2 to 3)	1 (1; −1 to 2)	0.002
Database retrieval	2	2 (1; 1 to 2)	2 (0; 1 to 2)	0 (0; 0 to 1)	0.083
Evolution models, tree building, and other methods	5	1 (2; 0 to 4)	3 (1; 0 to 5)	2 (2; −2 to 4)	<0.001
Results interpretation	5	2 (3; 0 to 5)	5 (1; 2 to 5)	2 (3; 0 to 5)	<0.001
Total (all domains)	20	10 (6; 5 to 17)	17 (4; 9 to 19)	6 (6; 0 to 11)	<0.001

IQR, interquartile range.
[a] Results are given as the median (IQR; range).
[b] $n = 21$.
[c] $n = 19$.
[d] Determined by Wilcoxon signed-rank test using 19 paired observations.

overall knowledge gains between these two groups. The participants' duration of time working and training in science had no correlation with baseline scores or magnitude of knowledge gains.

4. Conclusions

A feasible model for a brief, intensive, pathogen-focused phylogenetic workshop held in a middle-income country is described, and a workshop syllabus for others to consider using in the development of similar training programs in other LMIC is provided. To the authors' knowledge, this is the first study to provide some indication of the bioinformatics training needs of middle-income country scientists working in Latin America with pathogens of global health significance. This is also the first to assess the effectiveness of a pathogen-focused phylogenetics workshop in a LMIC, although there have been notable publications regarding the progress and challenges of pathogen bioinformatics training in Nigeria.[22] Despite the small sample size, the knowledge gained from the workshop was sufficiently large to be detected in this study. An intensive 5-day workshop model thus appears to be effective in improving pathogen-applied bioinformatics knowledge of scientists working in a LMIC setting.

Beyond the small sample size, several other caveats must be considered when interpreting these data. First, the number of questionnaire items was small (limited by time and resources). Second, while the questionnaires were administered under examination conditions (and were available in English and Spanish), administering the questionnaires at the end of the workshop may not reliably infer longer-term changes in participant knowledge and skills. A follow-up questionnaire at 6 and 12 weeks would be preferable; however, administering further questionnaires was not feasible in this study. In addition, questionnaires per se are a limited construct to measure scientific knowledge and skill. Due to logistics and limited workshop spaces to optimize the ratio of instructors to participants, it was only possible to select under 20% of applicants, and the workshop was limited to those with intermediate-level English. For these reasons, the present findings may not be representative of the phylogenetic training needs of other scientists in Peru.

Laptop ownership and English proficiency may be challenges to running this workshop in certain geographic areas. The authors think a key aspect to the success of the workshop model was partnering with local institutions and including local scientists fluent in Spanish in the curriculum. Future workshops could include the entire syllabus in both English and the local language if required. If sufficiently funded, this workshop model could also provide laptops as needed, although this may be a logistical and financial constraint.

Almost half of those participants selected had no phylogenetic experience before, emphasizing a significant training gap for local scientists studying pathogens in this region. The specific baseline training needs in several domains of phylogenetics was highlighted, particularly in models of evolution/tree-building methods, which was the lowest scoring domain at baseline, followed by result interpretation and sequence cleaning. Post-workshop training deficits were also demonstrated, particularly in models of evolution/tree-building methods. These findings could be used to guide future LMIC phylogenetic and bioinformatic workshops adopting the model described here.

Acknowledgements

The authors thank Ana Sofia Rengifo and the Universidad Peruana Cayetano Heredia for their assistance. This work was supported by the Global Emerging Infections Surveillance and Response System (GEIS), a division of the Armed Forces Health Surveillance Center (AFHSC).

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the US Government. Some of the authors are US Government employees. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17 U.S.C.§101 defines a US Government work as a work prepared by a military service member or employee of the US Government as part of that person's official duties.

Conflict of interest: The authors declare no conflicts of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijid.2015.11.001.

References

1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/?term=influenza+phylogenetic (Accessed March, 2015).
2. Rabaa MA, Ty Hang VT, Wills B, Farrar J, Simmons CP, Holmes EC. Phylogeography of recently emerged DENV-2 in southern Viet Nam. *PLoS Negl Trop Dis* 2010;**4**:e766.
3. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SL, Andersen KG, et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* 2015;**161**:1516–26.
4. Dudas G, Rambaut A. Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak. *PLoS Curr* 2014;6.
5. Pollett S, Nelson MI, Kasper M, Tinoco Y, Simons M, Romero C, et al. Phylogeography of influenza A(H3N2) virus in Peru, 2010-2012. *Emerg Infect Dis* 2015;**21**:1330–8.
6. Available at: http://blogs.plos.org/speakingofmedicine/2015/04/09/training-next-generation-scientists-disease-endemic-countries-high-priority-disease-elimination-efforts/.
7. Available at: http://seqanswers.com/wiki/Software.
8. Available at: http://mygoblet.org/.

9. Brazas MD, Lewitter F, Schneider MV, van Gelder CW, Palagi PM. A quick guide to genomics and bioinformatics training for clinical and public audiences. *PLoS Comput Biol* 2014;**10**:e1003510.
10. Available at: http://www.origem/info/misms.
11. Available at: http://treethinkers.org/2014-workshop/.
12. Available at: http://evomics.org/workshops/workshop-on-molecular-evolution/2015-workshop-on-molecular-evolution-cesky-krumlov/.
13. Available at: http://www.mbl.edu/education/special-topics-courses/workshop-on-molecular-evolution/.
14. Available at: http://abacus.gene.ucl.ac.uk/CoME/.
15. Available at: http://www.eid.ed.ac.uk/event/20th-international-bioinformatics-workshop-virus-evolution-and-molecular-epidemiology.
16. Available at: https://www.facebook.com/NAMRU6.
17. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011;**28**:2731–9.
18. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010;**27**:221–4.
19. Available at: http://tree.bio.ed.ac.uk/software/figtree/.
20. Stucky BJ. SeqTrace: a graphical tool for rapidly processing DNA sequencing chromatograms. *J Biomol Tech* 2012;**23**:90–3.
21. http://www.genecodes.com/.
22. Fatumo SA, Adoga MP, Ojo OO, Oluwagbemi O, Adeoye T, Ewejobi I, et al. Computational biology and bioinformatics in Nigeria. *PLoS Comput Biol* 2014;**10**:e1003516.