Robustness of phylogenetic analysis for detecting clusters of new HIV infections

Authors:

August Guang[1], Mark Howison[2], Mia Coetzer[3], Lauren Ledingham[3], Matt D'Antuono[3], Philip A. Chan[3], Charles Lawrence[4], Casey W. Dunn[5], Rami Kantor[3]

[1] Computing and Information Services, Brown University, Providence, RI, USA

[2] Research Improving People's Lives, Providence, RI, USA

[3] Division of Infectious Diseases, The Alpert Medical School, Brown University, Providence, RI, USA

[4] Division of Applied Mathematics, Brown University, Providence, RI, USA

[5] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

1 Abstract

2 **Background:** Phylogenetic analysis of HIV sequences obtained as part of clinical care is increasingly applied to

3 detect clustering of new HIV infections and inform public health interventions to disrupt transmission. Conventional

4 approaches summarize the within-host HIV diversity using only a single consensus sequence of only the HIV pol gene

5 per individual, typically from Sanger or next-generation sequencing (NGS).

6 **Methods:** In all newly HIV diagnosed individuals in the first half of 2013 from the largest HIV center in Rhode

7 Island, USA, we evaluate the robustness of the consensus approach and the potential benefit of considering full within-

8 host HIV diversity, available via NGS, and of near whole HIV genome for phylogenetic inference. We compare Sanger

9 and NGS-derived pol consensus sequences to an alternate approach that samples many sequences per individual from

10 a profile hidden Markov model of their NGS data for pol and near-whole HIV genomes.

11 **Results:** The space of phylogenies inferred from sampling is multi-modal, suggesting that a consensus-inferred

12 phylogeny is not an appropriate summary of within-host variation. Cluster inference differs in phylogenies from

13 consensus sequences in three clinically-relevant regions (prrt, int, env) versus the whole genome, and using Sanger

14 versus NGS data.

15 **Discussion:** The choice of gene region and sequencing and summary methods affects the detection of clusters, and

should be considered carefully in public health applications of phylogenetic analysis to disrupt HIV transmission.

Background

Methods

Results

Discussion