Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods - DTU Orbit (30/01/2019)

Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods

Background
Whole genome sequencing (WGS) is increasingly used in diagnostics and surveillance of infectious diseases. A major application for WGS is to use the data for identifying outbreak clusters, and there is therefore a need for methods that can accurately and efficiently infer phylogenies from sequencing reads. In the present study we describe a new dataset that we have created for the purpose of benchmarking such WGS-based methods for epidemiological data, and also present an analysis where we use the data to compare the performance of some current methods.

Results
Our aim was to create a benchmark data set that mimics sequencing data of the sort that might be collected during an outbreak of an infectious disease. This was achieved by letting an E. coli hypermutator strain grow in the lab for 8 consecutive days, each day splitting the culture in two while also collecting samples for sequencing. The result is a data set consisting of 101 whole genome sequences with known phylogenetic relationship. Among the sequenced samples 51 correspond to internal nodes in the phylogeny because they are ancestral, while the remaining 50 correspond to leaves. We also used the newly created data set to compare three different online available methods that infer phylogenies from whole-genome sequencing reads: NDtree, CSI Phylogeny and REALPHY. One complication when comparing the output of these methods with the known phylogeny is that phylogenetic methods typically build trees where all observed sequences are placed as leafs, even though some of them are in fact ancestral. We therefore devised a method for post processing the inferred trees by collapsing short branches (thus relocating some leafs to internal nodes), and also present two new measures of tree similarity that takes into account the identity of both internal and leaf nodes.

Conclusions
Based on this analysis we find that, among the investigated methods, CSI Phylogeny had the best performance, correctly identifying 73% of all branches in the tree and 71% of all clades. We have made all data from this experiment (raw sequencing reads, consensus whole-genome sequences, as well as descriptions of the known phylogeny in a variety of formats) publicly available, with the hope that other groups may find this data useful for benchmarking and exploring the performance of epidemiological methods. All data is freely available at: https://cge.cbs.dtu.dk/services/evolution_data.php.

General information
State: Published
Organisations: Department of Bio and Health Informatics, Genomic Epidemiology, Disease Intelligence and Molecular Evolution, National Food Institute, Research group for Genomic Epidemiology, University of Copenhagen
Number of pages: 13
Publication date: 2017
Peer-reviewed: Yes

Publication information
Journal: BMC Genomics
Volume: 18
Article number: 19
ISSN (Print): 1471-2164
Ratings:
BFI (2019): BFI-level 1
Web of Science (2019): Indexed yes
BFI (2018): BFI-level 1
Web of Science (2018): Indexed yes
BFI (2017): BFI-level 1
Scopus rating (2017): CiteScore 4.08 SJR 2.11 SNIP 1.151
Web of Science (2017): Impact factor 3.73
Web of Science (2017): Indexed yes
BFI (2016): BFI-level 1
Scopus rating (2016): CiteScore 4.05 SJR 2.163 SNIP 1.096
Web of Science (2016): Impact factor 3.729
Web of Science (2016): Indexed yes
BFI (2015): BFI-level 1
Scopus rating (2015): CiteScore 4.3 SJR 2.348 SNIP 1.159
Web of Science (2015): Impact factor 3.867
Web of Science (2015): Indexed yes
BFI (2014): BFI-level 1
Electronic versions:
DOIs:
10.1186/s12864-016-3407-6

Bibliographical note
This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Source: PublicationPreSubmission
Source-ID: 128124692