Type Strain Genome Server Tutorial

Correspondence: Jan Meier-Kolthoff (E-mail: jan.meier-kolthoff@dsmz.de)

last change: May 3, 2019

List of Figures

1  The submission form ................................................................. 3
2  The confirmation dialog ............................................................ 4
3  First part of the results page including TYGS Table 1 ....................... 5
4  The viewer for the genome-based phylogenetic tree ......................... 6
5  The viewer for the 16S rRNA gene sequence-based phylogenetic tree .... 7
6  Table 2 of a selected TYGS result page ....................................... 8
7  Pairwise comparisons ............................................................... 9
8  Strain overview ........................................................................ 10
9  Publication-ready text for an exemplary TYGS run ......................... 11
Summary

To demonstrate the various capabilities and characteristics of TYGS, we prepared a dataset of the medically and economically relevant Bacillus cereus group (BCG), also known as B. cereus sensu lato. The BCG harbours species such as B. anthracis, the aetiologic agent of anthrax, a common disease of livestock and, in few cases, of humans [Liu et al., 2015]. This dataset comprised 20 user-defined strains and was a smaller—though representative—subset of one of our previous phylogenomic analyses of the BCG [Liu et al., 2015].

TYGS yielded a final dataset of 35 strains (20 user-defined strains plus 15 automatically determined closest neighbours). Species and subspecies clustering was conducted using the dedicated clustering algorithm and established thresholds [Meier-Kolthoff et al., 2014, 2013] analogous to 70 % and ca. 79% DDH, respectively, thus producing a conservative subdivision into 23 species and 29 subspecies clusters. The 20 user-defined strains were found to represent 17 (sub-)species clusters.

Among the user-defined strains, 14 did not cluster together with a known type strain. The lack of a match between a user-defined strain and a type strain can have several causes: (i) the respective user-defined strain represents a new species, (ii) the user-defined strain would match an existing type strain but its genome has not been sequenced or is not publicly available (iii) the user-defined strain would match an existing genome-sequenced type strain but due to deficiencies in the metadata of the genome sequence this sequence could not be identified in GenBank. As in all instances of genome-based taxonomy, users are advised to check the 16S rRNA gene databases to rule out cases (ii) and (iii) prior to proposing taxonomic consequences. Publicly available type-strain genome sequences are rarely missing in the TYGS database but if so could be added by hand.

In addition to previous results [Liu et al., 2015], multiple species clusters were found to harbour multiple subspecies, demonstrating how TYGS can elucidate intra-species diversity. The genome-based GBDP tree was well-resolved at the species level and below, thus supporting the obtained (sub-)species clusters.

Citation: Meier-Kolthoff JP, Göker M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun. 2019 (in press). DOI: 10.1038/s41467-019-10210-3
Figure 1: The submission form

(A) Upload of up to 20 nucleotide FASTA or GenBank files. (B) GenBank accession IDs can be specified as well. (C) By restricting job to user genomes only, the automatic determination of closest type strain genomes is disabled. (D) Further type-strain genomes can be selected. The list can also be used to check whether or not TYGS contains all type strains relevant to the analysis as genome sequences might not be publicly available yet for all type strains. (E) The right-pointing arrowhead can be used to select strains. (F) The left-pointing arrowhead can be used to remove strains from the selection. (H) All selected strains are shown in this area. (G) The whole selection can be removed via the left-pointing double arrowhead. (I) Specification of contact data. (K) On job submission the input data are checked and a job summary is shown afterwards, as explained below.
After a positive validation outcome, TYGS will show a confirmation dialog, displaying detailed information about the submitted data as well about the next steps. If GenBank accession IDs were provided by the user, this view will update after TYGS has finished downloading sequence data from the NCBI servers (this might take up to a minute per accession depending on the load on the NCBI servers) – The extracted organisms names will be shown in column “Genome”, whereas a checkmark in the column “Source” will indicate that the download has succeeded.
The unique result ID is displayed at the top the TYGS job. The overall job runtime is given in minutes and can vary depending on factors such as job type, dataset size and the overall load on the servers and computing clusters. The green-colored quick navigation bar allows the user to directly jump to certain parts of the analysis. The first result table (Table 1) contains an overview of the inferred phylogenies along with download options and additional information such as average branch support and $\delta$ statistics.
Figure 4: The viewer for the genome-based phylogenetic tree

The upper part of this view displays information on the underlying genome data, the GBDP distance formula and algorithm. Additional annotation is displayed to the right-hand side of the interactive tree viewer including species and subspecies clusters, genomic G+C content (in %), δ values, total sequence length (in bp), number of proteins and an indication whether or not the strain was provided by the user. (C) The tree can be downloaded in either a vector graphics format or as a PNG file. Branch lengths are scaled in terms of the used GBDP distance formula and can be displayed via the button ‘show branch length values’ on the left-hand side.
Figure 5: The viewer for the 16S rRNA gene sequence-based phylogenetic tree

The upper part of this view displays information on the GBDP distance formula and algorithm. Additional annotation is displayed to the right-hand side of the interactive tree viewer including genome-based species and subspecies clusters, genomic G+C content (in %), total sequence length (in bp), number of proteins and an indication whether or not the strain was provided by the user. (C) The tree can be downloaded in either a vector graphics format or as a PNG file. Branch lengths are scaled in terms of the used GBDP distance formula and can be displayed via the button ‘show branch length values’ on the left-hand side. An extended 16S rRNA gene sequence analysis via Maximum Likelihood (ML) and Maximum Parsimony (MP) can be requested via the “Request now!” button (red circle).
The table ‘Identification’ shows the results from matching the user-defined strains against the type-strain database. Since not all type strains have been genome-sequenced yet, the lack of a match does not necessarily indicate a novel species and is thus labelled as ‘potential new species’. Both the number of species clusters and the number of subspecies clusters are indicated at the bottom of this table.
Figure 7: Pairwise comparisons

The table shows all pairwise comparisons between user-defined strains (printed in bold face) as well as between type strains and user-defined strains. The table can be exported in various formats (CSV, PDF, Excel, etc.) and the data can be filtered using keywords. The table is originally sorted by the recommended distance formula $d_4$ but can be rearranged by the user. Results include digital DDH estimates and confidence intervals for the three distinct formulae but also differences in genomic G+C content. Whereas formulae $d_0$ and $d_6$ measure the similarity in gene content, the formula $d_4$ reports a similarity based on sequence identity. This information can be used for genome-based delineation of prokaryotic species. (E) Names of deposits in suitable culture collections are directly linked to the web page of the respective culture collection for this deposit. Supplementary Table S2 contains a complete export of the depicted table.
Figure 8: Strain overview

Detailed list of strains used in this TYGS run. The entire table can be exported to various formats. Extended type strain info is provided via a link to the BacDive database (A). Names of deposits in suitable culture collections are directly linked to the web page of the respective culture collection for this deposit (B). Where possible taxon names are directly linked to literature source (C). User-defined strains are highlighted. Strains sharing a (sub-)species cluster ID belong to the same (sub-)species. Supplementary Table S2 contains a complete export of the depicted table.
Materials and Methods

Determination of closely related type strains

The 16S rRNA gene sequences were extracted from the user genomes using RNAmmer [1] and each sequence was subsequently BLASTed [2] against the 16S rRNA gene sequence of each of the approx. 7000 type strains available in the TYGS database. This was used as a proxy to find the best 20 matching type strains (according to the bit score) for each user genome and to subsequently calculate pairwise distances using the Genome BLAST Distance Phylogeny approach (GBDP) under the algorithm 'coverage' and distance formula $d_2$ [3]. The distances were finally used to determine the 10 closest type strain genomes for each of the user genomes.

Pairwise comparison of genome sequences

All pairwise comparisons among the set of genomes were conducted using GBDP and intergeneric distances inferred under the algorithm 'trimming' and distance formula $d_2$ [3] and 100 distance replicates were calculated each. Digital DEH values and confidence intervals were calculated using the recommended settings of the GGDC 2.1 [5].

Phylogenetic inference

The resulting intergeneric distances were used to infer a balanced minimum evolution tree with branch support via FASTME 2.1.4 including SPR postprocessing [4]. Branch support was inferred from 100 pseudo-bootstrap replicates each. The trees were rooted at the midpoint [5] and visualized with PhyDI [6].

Type-based species and subspecies clustering

The type-based species clustering using a 70% DEH radius around each of the 14 type strains was done as previously applied [7]. The resulting groups are shown in Table 1 and 4. Subspecies clustering was done using a 70% DEH threshold as previously introduced [8].

Results

Type-based species and subspecies clustering

The resulting species and subspecies clusters are listed in Table 4, whereas the taxonomic identification of the query strains is found in Table 1. Briefly, the clustering yielded 23 species clusters and the provided query strains were assigned to 17 of these. Moreover, user strains were located in 17 of 29 subspecies clusters.

References


Figure 9: Publication-ready text for an exemplary TYGS run

The text is auto-generated, based on the various results from the underlying TYGS job, and covers the typical manuscript sections: Materials and Methods, Results and References.
References

