NRC Research Press

ARTICLE

bcgTree: automatized phylogenetic tree building from bacterial core genomes

Markus J. Ankenbrand and Alexander Keller

Abstract: The need for multi-gene analyses in scientific fields such as phylogenetics and DNA barcoding has increased in recent years. In particular, these approaches are increasingly important for differentiating bacterial species, where reliance on the standard 16S rDNA marker can result in poor resolution. Additionally, the assembly of bacterial genomes has become a standard task due to advances in next-generation sequencing technologies. We created a bioinformatic pipeline, bcgTree, which uses assembled bacterial genomes either from databases or own sequencing results from the user to reconstruct their phylogenetic history. The pipeline automatically extracts 107 essential single-copy core genes, found in a majority of bacteria, using hidden Markov models and performs a partitioned maximum-likelihood analysis. Here, we describe the workflow of bcgTree and, as a proof-of-concept, its usefulness in resolving the phylogeny of 293 publically available bacterial strains of the genus *Lactobacillus*. We also evaluate its performance in both low- and high-level taxonomy test sets. The tool is freely available at github (https://github.com/iimog/bcgTree) and our institutional homepage (http://www.dna-analytics.biozentrum.uni-wuerzburg.de).

Key words: bacteria, phylogeny, genome, phylogenomics, multi-gene.

Résumé : Le recours à des analyses multigéniques dans divers champs scientifiques comme la phylogénétique et le codage à barres de l'ADN s'est accru récemment. En particulier, ces approches sont de plus en plus importante pour distinguer les espèces bactériennes du fait que le recours au seul marqueur de l'ADNr 16S peut occasionner une résolution limitée. De plus, l'assemblage de génomes bactériens est devenue une opération courante en raison des avancées en matière de séquençage à haut débit. Les auteurs ont créé un pipeline bioinformatique, bcgTree, lequel utilise des génomes bactériens assemblés provenant soit de banques de données ou nouvellement séquencés par les chercheurs pour reconstruire leur phylogénie. Le pipeline extrait automatiquement 107 gènes essentiels présents en simple copie, lesquels sont retrouvés chez la majorité des bactéries, à l'aide de modèles de Markov cachés et réalise une analyse de vraisemblance maximale partitionnée. Dans ce travail, les auteurs décrivent le processus de travail de bcgTree et, à titre de preuve de concept, son utilité en vue de résoudre la phylogénie de 293 souches disponibles du genre *Lactobacillus*. Les auteurs ont évalué sa performance tant au sein de jeux de données ciblant des niveaux taxonomiques fins ou grossiers. Cet outil est disponible librement sur github (https://github.com/iimog/bcgTree) ainsi que sur le site web de l'institut des auteurs (http://www.dna-analytics.biozentrum.uni-wuerzburg.de). [Traduit par la Rédaction]

Mots-clés : bactéries, phylogénie, génome, phylogénomique, multigénique.

Introduction

Resolving the evolutionary and taxonomic relationships of organisms by DNA sequence data has a long history in bacteria (Woese and Fox 1977; Woese 1987; Cavalier-Smith 1993). Morphologically, bacteria are hard to distinguish and classify, making DNA barcoding and molecular phylogenetics the methods of choice for researchers attempting to determine the relationships of bacterial strains. However, resolving phylogenetic relationships through the use of DNA sequences can be a challenging task. Selecting an appropriate genetic marker, one with both sufficient information for distinguishing taxa and with sufficient homology to make comparisons valid and conclusive (Wu et al. 2013; Capella-Gutierrez et al. 2014), is essential for a correct reconstruction. Often, different markers are used for low- (strain/species/genus) and high-level (family/class/order/phylum) phylogenetic analyses to compensate for this trade-off between information and conservation (Wu et al. 2013; Capella-Gutierrez et al. 2014).

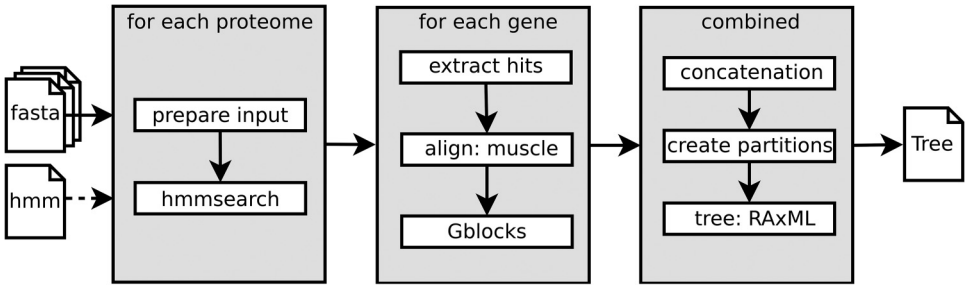
In bacteria, the 16S rDNA is currently the unrivaled and universally applied marker of choice for most phylogenetic and ecological studies. In this marker, several

Received 26 November 2015. Accepted 21 April 2016.

Corresponding Editor: Frédéric Chain.

M.J. Ankenbrand and A. Keller. Department of Animal Ecology and Tropical Biology, University of Würzburg, Germany.
Corresponding author: Alexander Keller (email: a.keller@biozentrum.uni-wuerzburg.de).

Fig. 1. Schematic workflow of the bcgTree pipeline.



variable and conserved regions are present, allowing good amplification of sufficiently informative regions and differentiation of closely related taxa. The 16S rDNA is well conserved across all prokaryotic species, allowing comparisons between phyla. But still this approach has its limitations, both regarding high- and low-level analyses (Wu et al. 2013; Capella-Gutierrez et al. 2014). The bacterial phylogeny is still unresolved at its basal branches and it is unlikely that these will be resolved using a single marker. Furthermore, 16S sequences can be identical between strains despite massive genomic reorganisation, precluding the ability of this marker to differentiate certain genetically different strains with varied ecological functions (Jaspers and Overmann 2004). In addition, ribosomal rDNA is present in multiple copies in each genome and intra-genomic variability is possible (Větrovský and Baldrian 2013). Both effects may confuse the interpretation of taxonomic assignments, especially in functional or ecological analyses, as well as mutualistic or pathogenic host – bacteria associations.

While 16S rDNA-based phylogenetic analysis has been of great importance in understanding the identity and evolutionary associations of bacteria, there are several drawbacks to calculating a tree on a single marker sequence. Current advances in high throughput sequencing technologies allow broader analysis beyond this single marker convention. Bacterial genomes are usually of limited size relative to the majority of eukaryotes, ranging from 130 kbp to 14 Mbp. The drops in price per basepair and the small genome sizes make it feasible to sequence a complete bacterial genome even for working groups with limited funding (Metzker 2009; Keller et al. 2014). This also leads to increasing numbers of complete bacterial genomes being sequenced and deposited in public databases (Pruitt et al. 2007; Uchiyama et al. 2013).

However, deriving a phylogeny from whole genomes bears its own challenges. For example, there are usually large genomic regions with no apparent similarities. Also, those regions with homologies need to be extracted for downstream phylogenetic analysis, a process requiring extensive bioinformatic expertise. One way to address this challenge is to build a database of pre-calculated alignments of tight genomic clusters (ATGC, Novichkov et al. 2009), enabling high resolution micro-evolutionary analyses. Yet, this approach is limited as only a fraction of the available genomes are included and it is not possible to supplement the analysis with user-provided data. One solution to this problem is to concentrate on the conserved regions present in a majority of organisms of interest (Ciccarelli et al. 2006).

Here we present bcgTree, a tool that identifies and extracts a set of 107 essential single-copy genes from amino-acid sequences of whole-genome data. The definition of "essential core genes" used here is based on the work of Dupont et al. (2012) and follows a statistical, not biological, argument. Our software automatically compiles the core gene sequence data and uses it to reconstruct a phylogenetic tree using a partitioned maximum likelihood analysis. For validation purposes, we applied bcgTree to resolve the phylogeny of the genus *Lactobacillus*, including most genomes currently available for the Lactobacillales. Additionally, test sets of high- and low-level phylogenetic analysis were directly compared to corresponding 16S rDNA trees for evaluation purposes.

Materials and methods

The bcgTree pipeline

The principal workflow of bcgTree is as follows (Fig. 1): As input files, protein fasta (often defined as *.faa) sequences can be used directly, for example, those deposited in the Genome database of NCBI (Pruitt et al. 2007) or obtained by protein reading frame prediction tools. Each of those proteome sets are then searched for 107 essential bacterial single-copy genes (Dupont et al. 2012) using hmmsearch (version 3.1b1) (Eddy 2010). After completing this search for each organism, the tool generates an overall presence/absence table, which can be used for validation purposes, such as whether the majority of genes have been found (compare to supplementary data, File S1[1]).

For each gene, the sequences of the best hit above a gene-specific cut-off are obtained from each proteome and stored in a gene-specific fasta file using the SeqFilter obtained from proovread (Hackl et al. 2014). Those gene-wise sequence sets are then aligned using muscle

[1]Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/gen-2015-0175.

(v3.8.31) (Edgar 2004). Alignments are refined using Gblocks (version 0.91b) (Castresana 2000; Talavera and Castresana 2007) to avoid over-extensive gapped areas and highly misaligned regions obtained via the automation procedure. In case a gene was not found within a specific proteome, alignments of these genes are supplemented with completely gapped sequences for this organism.

All gene alignments are then concatenated and a partitioning file is generated to mark the boundaries of each gene. A tree is calculated on this concatenated alignment using RAxML (version 8.2.4) (Stamatakis 2014). Models are estimated individually for each original gene region by using the partition file. The final output is a maximum-likelihood tree with bootstrap support values. Several parameters for the internal programs (e.g., number of bootstraps, number of threads) can be adjusted by the user.

The tool is executed as a Perl script from the command line or with a graphical user interface written in Java. It is available as source code and executable via https://github.com/iimog/bcgTree. This page also includes detailed installation instructions and lists the dependencies on other software tools.

Selection of hidden Markov models (HMMs) used for searches

The HMMs of the 107 essential single copy genes were taken from TIGRFAM (Haft et al. 2003) and Pfam (Finn et al. 2010) as described by Dupont et al. (2012). In Dupont et al. (2012), these HMMs were found to be present in more than 95% of all bacteria. Further, all but four of the genes (*glyS*, *proS*, *rpoC*, and *pheT*) were represented by only one HMM, with the remaining four being represented by two HMMs. As such, the latter HMMs are treated separately in the workflow due to the high sequence dissimilarity. Approximately half of these genes encode ribosomal proteins, and given that we found all of them on the chromosome of *Lactobacillus acidophilus* strain 30SC, they are unlikely to be found on plasmids.

Case study: *Lactobacillus* phylogeny

As a case study, the phylogeny of the Lactobacillales has been reconstructed using bcgTree. Genomes for this study have been taken from the EzGenome database (http://www.ezbiocloud.net/ezgenome, accessed January 2016). The 2225 genomes found by searching for Lactobacillales included 293 genomes of the genus *Lactobacillus*. The most dominant groups were *Streptococcus* (1188 genomes) and *Enterococcus* (622 genomes). As the focus of this case study was the analysis of the *Lactobacillus* phylogeny, only 50 random genomes of *Streptococcus* and *Enterococcus* each were used, but all of the remaining groups. The resulting dataset contained the protein sequences from 515 Lactobacillales genomes. Then, bcgTree was used with default parameters. All computation was performed on an Ubuntu 12.04 LTS 64 bit machine with an 80 core Intel® Xeon® CPU E7-4850 processing system and 512 GB of RAM. The resulting tree is discussed here to address several questions regarding the *Lactobacillus* phylogeny and the current All-Species Living Tree (Yarza et al. 2008).

Comparison with 16S phylogeny and multi-marker benefit

Two evaluation sets of smaller sample size were used for the evaluation of our tool beyond the case study. This was done in direct comparison with a corresponding 16S tree that was reconstructed, as described below, using sequence data from the same bacterial strains.

Evaluation set 1

To demonstrate the utility of bcgTree on low-level taxonomy, the software was applied to a subset of the case study sequences, i.e., only the genus *Lactobacillus*, and only those genomes represented at the NCBI genome database (accession date: October 2015) (Pruitt et al. 2007), which resulted in 68 *Lactobacillus* genomes. In this database, the 16S data are readily accessible alongside the proteomes (Pruitt et al. 2007). As an outgroup, 11 genomes from the genus *Paenibacillus* were added. The amino acid sequences were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.faa.tar.gz). The different files for plasmids and chromosomes were combined into a single fasta file for each genome. Then bcgTree analysis was performed on those 68 genomes with default parameters.

Evaluation set 2

The high-level taxonomy evaluation set contains two arbitrarily chosen genomes each from most of the distinguished bacterial high-level groups. These include the two gram-positive clades Firmicutes and Actinobacteria, the PVC group and the FCB group, five subgroups of Proteobacteria (alpha, beta, gamma, delta, and epsilon), and the Thermotogae. Two archaeal genomes were used as an outgroup.

For both evaluation sets, the corresponding 16S rDNA gene tree was calculated for exactly the same genomes using the same steps of alignment with muscle, refinement with Gblocks, and tree building with RAxML to maximize comparability between the approaches. 16S rRNA genes were extracted from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/all.frn.tar.gz, accession date: October 2015). In cases where multiple reference 16S rRNA sequences were available, only the longest was used. For one organism (*Lactobacillus brevis* KB290) the *.frn file was missing. Thus, we used RNAmmer (version 1.2) (Lagesen et al. 2007) to extract 16S sequence of this organism from the whole genome sequence.

The robustness of the trees generated by bcgTree and 16S was evaluated by bootstrap values obtained with both approaches. Bootstrap support values for all nodes together were statistically compared by using a Student's

Fig. 2. Case study tree calculated with bcgTree containing 515 Lactobacillales genomes. Numbers at nodes designate bootstrap support values resulting from 100 bootstrap replicates. Outgroup is *Aerococcus*. Monophyletic genera and species have been collapsed as <G> and <S>, respectively, (represented as a triangle considering intra-group variation as distance, with number of included genomes in curly brackets).

t test in R (R Development Core Team 2010). Tree topologies for both evaluation sets were compared between bcgTree and 16S using the R package dendextend (Galili 2015) for tanglegrams. Differences in topologies were highlighted using the same package with dashed lines.

The influence of the number of genes used for the tree-building process on the accuracy of the resulting tree was also assessed. For this, we used the final alignment files obtained through bcgTree for both evaluation sets with all 109 partitions (two of the 107 genes have two partitions each). These were randomly subsampled using RAxML and their corresponding partitions in the alignment excluded. For both evaluation sets, we used this approach to create concatenated alignment files with 1 to 108 random HMMs with 10 replicates per number. RAxML-derived trees were constructed using these alignments without bootstrapping and with the same parameters used in the default bcgTree analysis. The quartet distances between the resulting trees and the full gene set tree was calculated using qdist (version 2.0) (Mailund and Pedersen 2004). The quartet distance is a measure of the topological distance between two phylogenetic trees (Mailund and Pedersen 2004; Keller et al. 2010). For visualization purposes, the number of genes was rounded to increments of five. The quartet distance of the 16S rDNA tree was also calculated for comparison.

Computational performance

The computational performance of bcgTree on sets with different numbers of complete genomes was assessed on a standard dual-core desktop computer (Intel® Core™2 Duo CPU E8500, 4 GB RAM, Ubuntu 14.04.3 LTS 64 bit). Since bcgTree is designed to work on complete genomes and has a fixed set of 107 essential genes, the only variable is the number of genomes. Genome size variation is not expected to change the overall runtime substantially and was not evaluated.

The pipeline was executed on a variable number of genomes: for 5–15 genomes, each step-size of one was repeated with five replicates, for 5–50 genomes only one replicate was done with a step-size of five. For all steps, random proteomes were selected from all data downloaded from the genome database of NCBI. The total runtime of bcgTree was further separated into time before and after start of RAxML, to estimate the proportion of preparation and tree calculation of the total runtime. Linearity of the runtime increase with number of genomes was tested using a linear model for complete and pre-RAxML runtime.

Results and discussion

Case study: *Lactobacillus* phylogeny

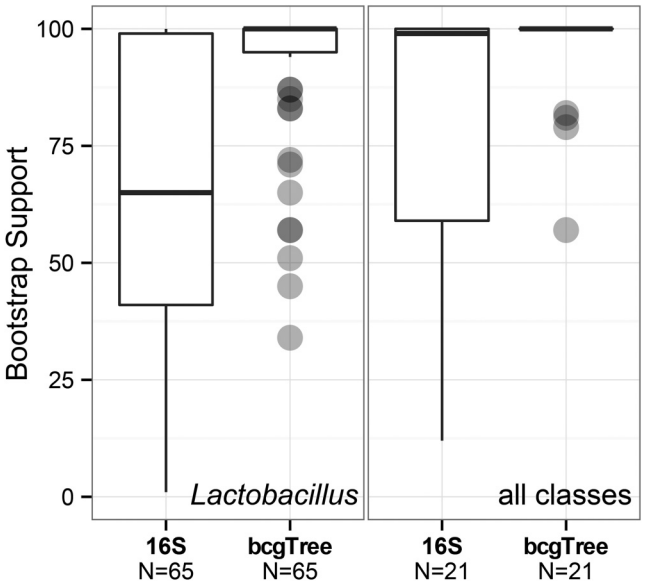
The case-study tree automatically generated by bcgTree largely supports the monophyly of most genera within the Lactobacillales (Fig. 2). However, *Pediococcus* (9 genomes) and the family Leuconostocaceae (51 genomes that form a monophylum and include the genera *Weissella*, *Oenococcus*, *Fructobacillus*, *Leuconostoc*) get inserted into the *Lactobacillus* genus, thus violating the monophyly of the latter. This is a known phenomenon that can also be observed in the All-Species Living Tree (Yarza et al. 2008). Within the genus *Lactobacillus*, the tree is well resolved and consistent with previously published results on the genus (Kant et al. 2011). Most species are well resolved into monophyletic clusters with high support values, thus providing better assignments than 16S only based analyses (Yarza et al. 2008), yet the All-Species Living Tree contains only a single representative of most *Lactobacillus* species.

The bcgTree phylogeny also supports the three major groups of *Lactobacillus* species as listed in *Bergey's Manual of Systematic Bacteriology* (Vol. 3) (De Vos et al. 2011) and Kant et al. (2011) (*L. delbrueckii* group or NCFM, *L. reuteri* group or WCFS1, *L. salivarius* group or WCFS2, and *L. rhamnosus*/*casei* group or GG) as monophyla with some new species added. The WCFS group reported by Kant et al. (2011) is split in the bcgTree tree into two groups separated by the Leuconostocaceae, which were not considered in that study. Furthermore, there are *Lactobacillus* strains that do not belong to any of these groups. Within the *L. rhamnosus*/*casei* group, *L. casei* (22 genomes) and *L. paracasei* (37 genomes), the leaves of the tree were highly intermixed, suggesting that it might not be appropriate to assign these to different species.

Comparison to 16S topology

For the low-taxonomy evaluation set of *Lactobacillus*, bootstrap support values were consistently higher with bcgTree than 16S data only (File S2[1], $t = 2.25$, df = 27, $p < 0.05$*; Fig. 3). The presence/absence results of the bcgTree procedure with *Lactobacillus* show two genomes that lack a high proportion of the included genes (File S1[1]). These were *Lactobacillus fermentum* CECT 5716 and *Lactobacillus salviarius* CECT 5713. Both genomes are of typical size, suggesting that they have been well assembled. They showed, however, a reduction by approximately one half of predicted open reading frames, indicating incomplete annotation. Thus the proteome files were smaller than those of closely related taxa. Still the tree reconstruction with bcgTree assigned them adequately at their expected positions in the trees, and it

Fig. 3. Boostrap support values for bcgTree and 16S rDNA only for direct comparison. Both evaluation sets, i.e., high- and low-level taxonomy are displayed.



was robust with high bootstrap values. This indicates that the tool is resilient to missing gene predictions in the proteome files. Also, some genes are represented by two HMMs and in almost all cases a hit was found for only one HMM per gene.

For the second evaluation set on high-level taxonomy, the bcgTree tree outperformed the 16S rDNA tree in terms of robustness with bootstrap support values (File S3[1], t = 6.31, df = 93, $p < 0.001^{***}$; Fig. 3). The general topology of the bcgTree tree was in concordance with the currently prevailing opinion of bacterial phylogeny. The bcgTree tree results provided support for monophyly of the two gram-positive groups Firmicutes and Actinobacteria. These groups were not resolved using 16S sequences alone. Also the Spirochaetes, PVC and FCB cluster together only with bcgTree, which is the current consensus opinion of the relatedness of these clades according to Bergey's Taxonomic Outlines (Ludwig et al. 2010). The remaining clades were resolved consistently with both methods, although with slightly different arrangements between groups. Please consider, however, that this study does not intend to resolve the molecular phylogeny of bacterial high-level groups and that taxa were arbitrarily chosen to validate the utility of the bcgTree approach for higher-level taxonomy.

Multi-marker benefits

For both evaluation sets, we subsampled the numbers of genes randomly and compared the resulting trees with the tree calculated on the complete gene set to infer influences of gene number on the accuracy of the method. In both evaluation sets, the quartet distance of trees calculated on subsets of the genes strongly decreases with an increase in the number of genes included (see Fig. 4). This shows that including more genes has great impact and benefits the accuracy of obtaining a tree similar to the full gene set. In both sets it can be seen that including five genes already leads to a great improvement in comparison to single gene analyses and this benefit appears to continually increase with more genes. The quartet distance of the 16S rDNA *Lactobacillus* tree is higher than most of the bcgTree-derived trees calculated with even small subsets of the essential genes. In contrast for the high-level taxonomy example, the 16S rDNA tree has a lower quartet distance than most of the small subset trees but a higher quartet distance than the bcgTree trees constructed using the full 107 genes and large subsets of the genes. This observation highlights the suitability of 16S rDNA as a marker for high-level taxonomy while demonstrating that the single-marker 16S rDNA analyses can be improved upon through multi-marker approaches, such as bcgTree.

For our case study the mean number and standard deviation of genes identified and used per genome was 104.0 ± 6.4. For the low-level as well as the high-level evaluation sets this was 104.9 ± 9.0 and 99.4 ± 21.5, respectively. The low values in the low-level evaluation set is explained by the inclusion of a parasitic organism into the test set, which has a reduced genome.

Computational performance

The computational time for the preliminary preparation steps, including HMM searches and alignments, increased linearly with each additional genome (t = 55.597, df = 63, $p < 0.001^{***}$, R^2 = 0.98) and the best fit line for this relationship exhibited a slope of 16.9 s/genome. The total runtime was also found to be significantly correlated with a linear model (t = 45.9, df = 63, $p < 0.001^{***}$, R^2 = 0.97) and this relationship exhibited a slope of 709.9 s/genome, although non-linear increases were observable for low-genome numbers (Fig. 5). This may be due to general RAxML initialization steps that are independent of data amount and are thus proportionally overrepresented with the low genome number analyses. In general, it can be assumed that the tree-building step, not the bcgTree specific tasks, consumes the largest fraction of the runtime.

In the current setup, a phylogenetic tree from core genomes of 50 organisms can be calculated on a standard desktop computer in less than 24 h. For larger analyses, the runtime can be decreased by using alternative tree calculation software or the high-performance computing variant ExaML that parallelizes tasks on different computer nodes.

Comparison with existing tools

The challenge of comparing whole genomes of bacteria has been undertaken through different approaches. One approach is to limit the scope to very closely related species to have large sets of orthologous genes. This approach is used by ATGC (Novichkov et al. 2009), which provides pre-calculated alignments for whole genomes

Fig. 4. Accuracy errors according to the quartet distance between trees of variable numbers of genes (1–108) and the trees based on the complete set. For each number of genes, 10 replicates were performed. For visualization purposes, the number of genes was rounded to increments of five. In addition, the quartet distance of the 16S rDNA tree is shown.

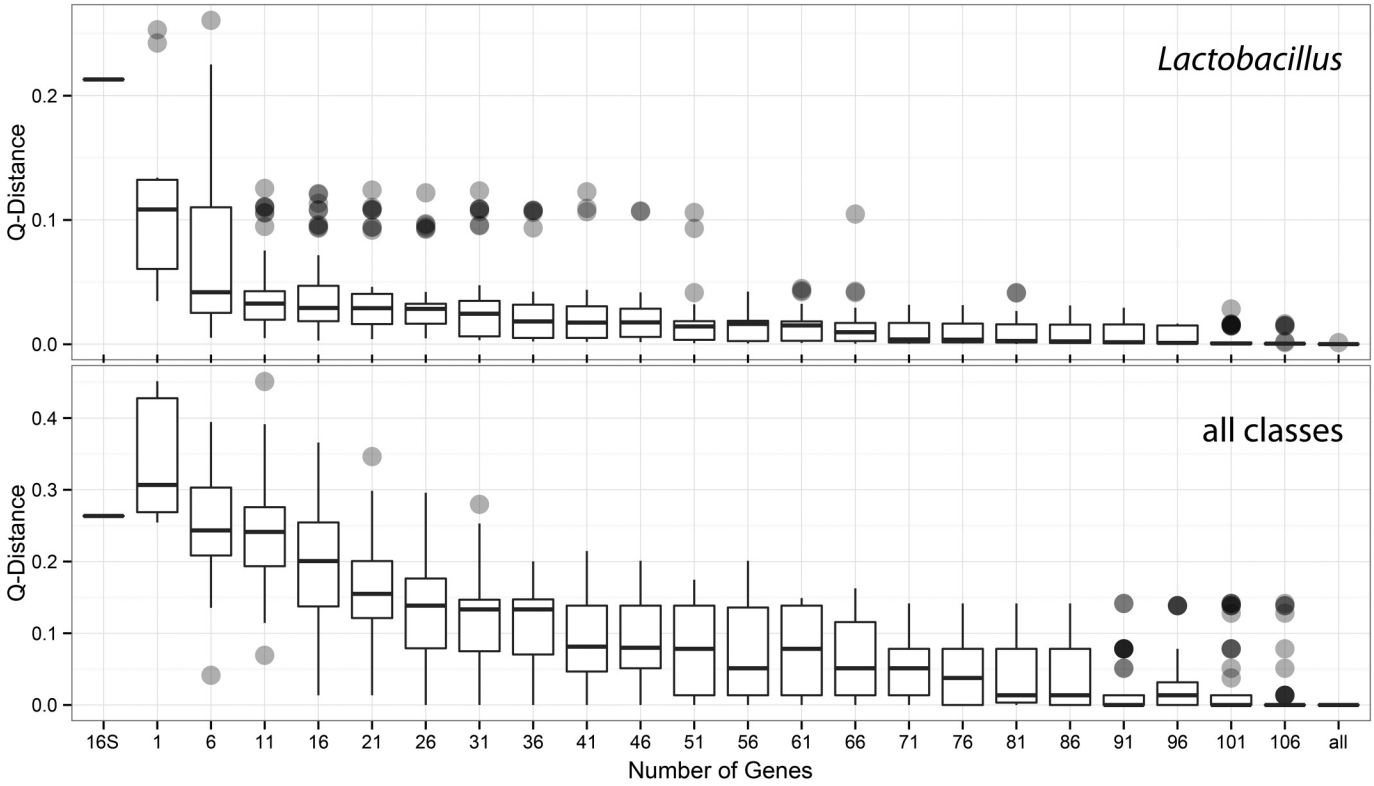
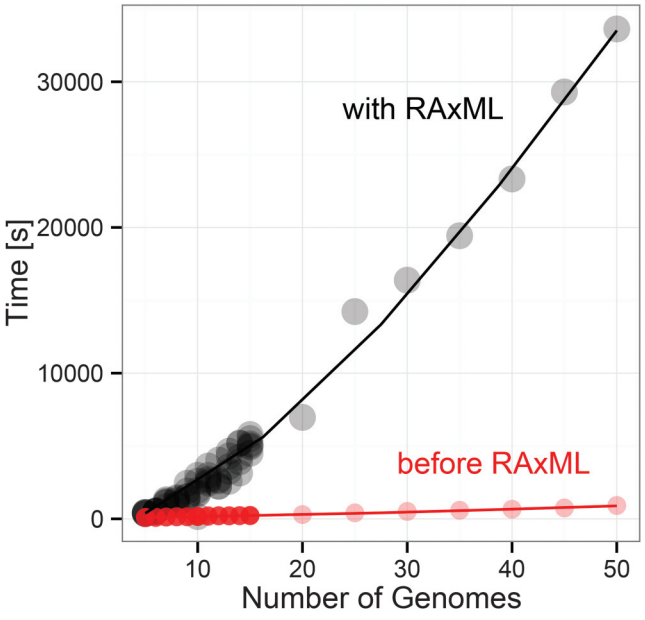


Fig. 5. Runtime of bcgTree for varying numbers of genomes in the tree calculation process with or without RAxML. [Colour online.]



from NCBI RefSeq (Pruitt et al. 2007). This way, a large amount of information is available for each cluster (e.g., 1500 orthologous genes for *Lactobacillus*), providing a solid basis for micro-evolutionary analyses. Yet, only a very limited number of taxa can be included in ATGC analyses, for example only 11 *Lactobacillus* genomes. Also, ATGC is not designed to do analyses across clusters. The 11 *Lactobacillus* genomes are split over four clusters, so a comparison across the whole genus or with closely related genera is not practical. Further, including user-provided datasets in ATGC is not possible. In summary, ATGC can help answering micro-evolutionary questions, but it is limited for broader phylogenetic research questions.

Another approach is to use alignment-free methods using composition vectors, as implemented in CVTree (Zuo and Hao 2015). By dropping the alignment step many potential bioinformatic challenges (like length-hypervariable genes) can be avoided and the distances between genomes can be rapidly calculated. However, whilst more sequences are included (typically all proteins of a genome), the position information of each amino acid is dropped and thus a great amount of information ignored. As a consequence, the overall information content is quite different to the approach described in this article and might yield different results. We suggest using both the alignment-free and our alignment-based approach together for phylogenetic studies on whole genomes. Both methods are valid and may provide complementary and supportive viewpoints on bacterial phylogenies.

Conclusions

As demonstrated by the case study and evaluation, bacterial phylogenies can be accurately and robustly

reconstructed using our automated pipeline implemented in bcgTree. By using 107 single-copy essential genes, the resolution is not limited to either lower or higher taxonomic ranks. The good results on both a fine scale (*Lactobacillus*) and a coarse scale (major bacterial groups) demonstrate its potential and versatility. It circumvents the restrictions that apply to single-marker phylogenies and also eases and standardizes the processes to perform whole-genome phylogenies with bacteria. The tool is freely available for download and use at the github repository https://github.com/iimog/bcgTree and our institutional homepage http://www.dna-analytics.biozentrum.uni-wuerzburg.de.

Acknowledgements

M.J.A. was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. We are grateful to Wiebke Sickel and Rodney T. Richardson for proofreading the manuscript. We greatly appreciate also the comments of the two anonymous reviewers and the editor.

References

Capella-Gutierrez, S., Kauff, F., and Gabaldón, T. 2014. A phylogenomics approach for selecting robust sets of phylogenetic markers. Nucleic Acids Res. **42**: e54. doi:10.1093/nar/gku071. PMID:24476915.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**: 540–552. doi:10.1093/oxfordjournals.molbev.a026334. PMID:10742046.

Cavalier-Smith, T. 1993. Kingdom Protozoa and its 18 phyla. Microbiol. Rev. **57**: 953–994. PMID:8302218.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science, **311**: 1283–1287. doi:10.1126/science.1123061. PMID:16513982.

De Vos, P., Garrity, G., Jones, D., Krieg, N.R., Ludwig, W., Rainey, F.A., et al. (*Editors*). 2011. Bergey's Manual of Systematic Bacteriology. Vol. 3: The *Firmicutes*. Springer, New York. doi:10.1007/978-0-387-68489-5.

Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.-J., Richter, R.A., Valas, R., et al. 2012. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J. **6**: 1186–1199. doi:10.1038/ismej.2011.189. PMID:22170421.

Eddy, S. 2010. HMMER3: a new generation of sequence homology search software. Available from http://hmmer.janelia.org [accessed July 2010].

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**: 1792–1797. doi:10.1093/nar/gkh340. PMID:15034147.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., et al. 2010. The Pfam protein families database. Nucleic Acids Res. **38**: D211–D222. doi:10.1093/nar/gkp985. PMID:19920124.

Galili, T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics btv428.

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. 2014. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics, **30**(21): 3004–3011. doi:10.1093/bioinformatics/btu392. PMID:25015988.

Haft, D.H., Selengut, J.D., and White, O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. **31**: 371–373. doi:10.1093/nar/gkg128. PMID:12520025.

Jaspers, E., and Overmann, J. 2004. Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl. Environ. Microbiol. **70**: 4831–4839. doi:10.1128/AEM.70.8.4831-4839.2004. PMID:15294821.

Kant, R., Blom, J., Palva, A., Siezen, R.J., and de Vos, W.M. 2011. Comparative genomics of *Lactobacillus*. Microb. Biotechnol. **4**, 323–332. doi:10.1111/j.1751-7915.2010.00215.x.

Keller, A., Förster, F., Müller, T., Dandekar, T., Schultz, J., and Wolf, M. 2010. Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. Biol. Direct, **5**: 4. doi:10.1186/1745-6150-5-4. PMID:20078867.

Keller, A., Horn, H., Förster, F., and Schultz, J. 2014. Computational integration of genomic traits into 16S rDNA microbiota sequencing studies. Gene, **549**: 186–191. doi:10.1016/j.gene.2014.07.066. PMID:25084126.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35**: 3100–3108. doi:10.1093/nar/gkm160. PMID:17452365.

Ludwig, W., Euzéby, J., and Whitman, W.B. 2010. Road map of the phyla *Bacteroidetes*, *Spirochaetes*, *Tenericutes* (*Mollicutes*), *Acidobacteria*, *Fibrobactere4*, *Fusobacteria*, *Dictyoglomi*, *Gemmatimonadetes*, *Lentisphaerae*, *Verrucomicrobia*, *Chlamydiae*, and *Planctomycetes*. *In* Bergey's manual of systematic bacteriology. Springer. pp. 1–19.

Mailund, T., and Pedersen, C.N.S. 2004. Qdist—quartet distance between evolutionary trees. Bioinformatics, **20**: 1636–1637. doi:10.1093/bioinformatics/bth097. PMID:14962942.

Metzker, M.L. 2010. Sequencing technologies—the next generation. Nat. Rev. Genet. **11**: 31–46. doi:10.1038/nrg2626. PMID:19997069.

Novichkov, P.S., Ratnere, I., Wolf, Y.I., Koonin, E.V., and Dubchak, I. 2009. ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. Nucleic Acids Res. **37**: D448–D454. doi:10.1093/nar/gkn684. PMID:18845571.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35**: D61–D65. doi:10.1093/nar/gkl842. PMID:17130148.

R Development Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, **30**: 1312–1313. doi:10.1093/bioinformatics/btu033. PMID:24451623.

Talavera, G., and Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. **56**: 564–577. doi:10.1080/10635150701472164. PMID:17654362.

Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. 2013. MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. Nucleic Acids Res. **41**: D631–D635. doi:10.1093/nar/gks1006. PMID:23118485.

Větrovský, T., and Baldrian, P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS ONE, **8**: e57923. doi:10.1371/journal.pone.0057923. PMID:23460914.

Woese, C.R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221–271. PMID:2439888.

Woese, C.R., and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U.S.A. **74**: 5088–5090. doi:10.1073/pnas.74.11.5088. PMID:270744.

Wu, D., Jospin, G., and Eisen, J.A. 2013. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. PLoS ONE, **8**: e77033. doi:10.1371/journal.pone.0077033. PMID:24146954.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., et al. 2008. The All-Species Living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. **31**: 241–250. doi:10.1016/j.syapm.2008.07.001. PMID:18692976.

Zuo, G., and Hao, B. 2015. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. Genomics Proteom. Bioinform. **13**: 321–331. doi:10.1016/j.gpb.2015.08.004.