Systematic and Applied Microbiology 38 (2015) 472–482

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb)

Aram Mikaelyan[a,b,*], Tim Köhler[a], Niclas Lampert[a], Jeffrey Rohland[a], Hamadi Boga[a], Katja Meuser[a], Andreas Brune[a,b,*]

[a] Department of Biogeochemistry, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany
[b] LOEWE Center for Synthetic Microbiology (Synmikro), Philipps-Universität Marburg, Marburg, Germany

ARTICLE INFO

Article history:
Received 23 March 2015
Received in revised form 13 July 2015
Accepted 23 July 2015

Keywords:
Insects
Gut microbiota
Deep-sequencing
Termites
Cockroaches
16S rRNA

ABSTRACT

Recent developments in sequencing technology have given rise to a large number of studies that assess bacterial diversity and community structure in termite and cockroach guts based on large amplicon libraries of 16S rRNA genes. Although these studies have revealed important ecological and evolutionary patterns in the gut microbiota, classification of the short sequence reads is limited by the taxonomic depth and resolution of the reference databases used in the respective studies. Here, we present a curated reference database for accurate taxonomic analysis of the bacterial gut microbiota of dictyopteran insects. The Dictyopteran gut microbiota reference Database (DictDb) is based on the Silva database but was significantly expanded by the addition of clones from 11 mostly unexplored termite and cockroach groups, which increased the inventory of bacterial sequences from dictyopteran guts by 26%. The taxonomic depth and resolution of DictDb was significantly improved by a general revision of the taxonomic guide tree for all important lineages, including a detailed phylogenetic analysis of the *Treponema* and *Alistipes* complexes, the *Fibrobacteres*, and the TG3 phylum. The performance of this first documented version of DictDb (v. 3.0) using the revised taxonomic guide tree in the classification of short-read libraries obtained from termites and cockroaches was highly superior to that of the current Silva and RDP databases. DictDb uses an informative nomenclature that is consistent with the literature also for clades of uncultured bacteria and provides an invaluable tool for anyone exploring the gut community structure of termites and cockroaches.

Introduction

Termites and their closest phylogenetic relatives, the cockroaches, represent the majority of species in the insect order *Dictyoptera* [3,16] and are ideal models for studying factors that shape microbial community structure in intestinal ecosystems [7]. During more than 200 million years of evolution, they have diversified to efficiently utilize a wide range of diets and now comprise numerous omnivorous, detritivorous, xylophagous, and humivorous lineages [15]. Previous studies have identified both dietary and phylogenetic patterns in the intestinal community structure of termites and cockroaches [11]. However, understanding the evolution of symbiotic digestion in dictyopteran insects requires a highly resolved analysis of their gut microbiota.

Most studies of bacterial diversity in the guts of termites and cockroaches have employed traditional capillary dideoxy (Sanger) sequencing of cloned 16S rRNA gene amplicons. They provided a wealth of information on the diversity of the gut microbiota and identified numerous novel lineages that are specific for this habitat (e.g., *Elusimicrobia* [25], *Fibrobacteres* subphylum 2 [26], termite gut spirochetes [35,45], and Termite Group 3 [26]). However, cost and effort involved in this approach limit the number of host taxa that can be included in an analysis and the depth to which each community can be sampled.

The development of next-generation sequencing technologies allowed efficient and economical sequencing of multiple 16S rRNA gene libraries with sufficient sampling depth to compare the bacterial communities across a wide host range [10,55]. However, the relatively short length of the sequence reads generated by the most commonly employed Roche 454 and Illumina/Solexa platforms [59] limits the amount of information available for phylogenetic analysis. Therefore, it has become common practice to infer the structure and taxonomic composition of microbial communities by assigning the reads using a pre-defined classification scheme

* Corresponding authors: Department of Biogeochemistry, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany. Tel.: +49 6421178700.
E-mail addresses: arammikaelyan@gmail.com (A. Mikaelyan), brune@mpi-marburg.mpg.de (A. Brune).

http://dx.doi.org/10.1016/j.syapm.2015.07.004
0723-2020/

and the Naïve Bayesian Classifier [60] developed by the Ribosomal Database Project (RDP), which has been implemented in popular workbenches for community analysis [8,52]. Obviously, the quality of such a classification depends strongly on the composition of the reference database and the depth and resolution of its taxonomic framework. The reference taxonomies most commonly used for the classification of short reads are provided by the Silva [65] and RDP [9] databases, which extend the taxonomic outline for cultured organisms [19] by including also phylogenetically coherent groups without cultured representatives.

However, general-purpose reference databases have serious shortcomings when it comes to studying microbial diversity in insect guts [42], particularly in termites and cockroaches [33,62]. One shortcoming is the frequent lack of taxonomic depth in the classification schemes, i.e., the absence of circumscribed taxa particularly at lower taxonomic levels. This is symptomatic for bacterial lineages that are endemic to termites and only rarely encountered in other environments, such as *Fibrobacteres* or the TG3 phylum [11,26]. Another problem is a lack of taxonomic resolution in many genus-level complexes, which comprise highly divergent 16S rRNA gene sequences that are lumped into inflated taxa (e.g., *Treponema* [5,35]) that may even be polyphyletic (e.g., *Ruminococcus* [17]). Finally, a lack of representative bacterial phylotypes from insect guts in general-purpose reference databases seriously affects the taxonomic assignment of short reads using the RDP classifier [42,62].

To overcome the challenges, we constructed a customized rRNA reference database for an accurate taxonomic analysis of the gut microbiota of termites and cockroaches. The Dictytoperan gut microbiota reference Database (DictDb) is based on the skeleton structure of the Silva database [65] and on the collation of published rRNA sequences obtained from termites and cockroaches and rigorous phylogenetic curation of the existing taxonomic framework. Initial, so far undocumented versions of DictDb were successfully used to improve the analysis of bacterial communities in termite guts (version 1.0; [33,48]) and subsequently both in termites and cockroaches (versions 2.3 and 2.4; [11,29,40,46,50,58]).

Here, we document for the first time the general architecture of DictDb and present the latest version (DictDb v. 3.0). This substantially expanded version includes more than 1000 novel phylotypes that were obtained from 11 host species in the context of this study. They represent severely under-sampled host groups among cockroaches (*Blaberidae*, *Polyphagidae*), lower termites (*Mastotermitidae*, *Kalotermitidae*), and higher termites (*Termitidae*), including representatives with fundamentally different diets. An improved taxonomic framework based on thorough phylogenetic analyses provided an unprecedented depth and resolution in termite-specific taxa, particularly among *Fibrobacteres* and candidate phylum TG3, and hitherto unresolved taxonomic complexes, such as the genera *Treponema* and *Alistipes*. The performance of the taxonomic framework of DictDb in the genus-level classification of deep-sequenced rRNA gene libraries of bacterial communities in termites and cockroaches is compared to that of the SILVA and RDP reference databases.

Materials and methods

Sample preparation

Termites used in this study were taken from laboratory colonies or were collected in the field. Only worker termites or pseudergates were used for this study. Cockroaches were purchased from a commercial breeder and maintained on leaf litter for several months. Only female cockroaches were used. The origin and other details of the samples are summarized in Table 1.

The guts of termites (10–20 individuals) and cockroaches (3 individuals) were dissected with sterilized fine-tipped forceps. Pools of guts, hindguts, or hindgut compartments (see Table 1 for sample details) were suspended in 750 μl sodium phosphate buffer (120 mM; pH 8.0) in 2-ml tubes and homogenized. DNA was extracted and purified using a bead-beating protocol as previously described [47].

Clone libraries

16S rRNA genes were amplified using the universal bacterial primers 27f and 1492r [34]. PCR products were purified and cloned as described by Thompson et al. [58]. Clones were tested for correct insert size, and inserts were sequenced bidirectionally with M13 vector primers using automated Sanger sequencing (GATC Biotech, Konstanz, Germany). In the case of *Cubitermes ugandensis* and *Ophiotermes* sp., the clone libraries were pre-screened by partial sequencing, and only novel phylotypes (<98% sequence identity to previously published full-length sequences) were sequenced in both directions. Partial sequences from the same clones were assembled using Seqman (version 4.05; DNA Star, Madison, WI, USA). Chimeric sequences were identified using the mothur [52] implementation of UCHIME [14] and confirmed by fractional treeing [37].

Construction of the reference database

Quality-checked sequences from the new clone libraries were aligned using the mothur aligner against the Silva reference alignment available on the mothur website (http://www.mothur.org/wiki/Silva_reference_alignment/) and imported into the Silva database (release 119; http://www.arb-silva.de/documentation/release-119/) using the ARB software package [38]. When necessary, alignments were manually refined using the ARB alignment editor.

Because of the inconsistent and varied usage of the fields "isolation_source" and "host" in the sequence-associated information in the INSDC databases (EMBL, DDBJ, and GenBank), we introduced the fields "DictDb_source" and "DictDb_specific_host" in DictDb v. 3.0. "DictDb_specific_host" indicates the insect host from which a given 16S rRNA sequence was derived. "DictDb_source" clarifies the preparation from which it was derived (e.g., a pool of flagellates, a particular gut compartment, or a capillary-picked bacterial filament). Additionally, we introduced a field "DictDb_diet" to describe the diet of the insect host from which the rRNA sequence was obtained.

The taxonomic framework of DictDb is based on the phylogenetic taxonomy described by the guide tree in the Silva database. All bacterial clades in the Silva database that contained a substantial fraction of sequences derived from the guts of termites and cockroaches were phylogenetically analyzed to redefine or further resolve the node-based taxonomy. Conservative column filters were applied to the alignments to eliminate highly variable positions in the alignment. Filtered alignments comprising approximately 1200 valid columns were exported for tree calculations using the maximum-likelihood (ML) method as implemented in PhyML (version 3.0.1; [23]) and a general time-reversible (GTR) model of sequence evolution. ML trees were inferred by subtree pruning and regrafting (SPR) of five random starting trees, and node support was estimated using the Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT) [1].

The topologies, branch lengths, and node supports from the calculated maximum-likelihood trees were grafted onto the main guide tree. The hierarchy of well-supported nested clusters obtained in the analyses was then used to enhance the taxonomic skeleton of the Silva database. This phylogenetic framework