REVIEW 10.1111/1469-0691.12181

# Advantages and limitations of genomics in prokaryotic taxonomy

**E. Sentausa and P.-E. Fournier**

*URMITE, UM63, CNRS 7278, IRD 198, Inserm 1095, Aix Marseille Université, Marseille, France*

## Abstract

Taxonomic classification is an important field of microbiology, as it enables scientists to identify prokaryotes worldwide. Although the current classification system is still based on the one designed by Carolus Linnaeus, the currently available genomic content of several thousands of sequenced prokaryotic genomes represents a unique source of taxonomic information that should not be ignored. In addition, the development of faster, cheaper and improved sequencing methods has made genomics a tool that has a place in the workflow of a routine microbiology laboratory. Thus, genomics has reached a stage where it may be used in prokaryotic taxonomic classification, with criteria such as the genome index of average nucleotide identity being an alternative to DNA–DNA hybridization. However, several hurdles remain, including the lack of genomic sequences of many prokaryotic taxonomic representatives, and consensus procedures to describe new prokaryotic taxa that do not, as yet, accommodate genomic data. We herein review the advantages and disadvantages of using genomics in prokaryotic taxonomy.

**Keywords:** Average nucleotide identity, DNA–DNA hybridization, genome-based phylogeny, genomics, prokaryotic taxonomy

*Clin Microbiol Infect*

**Corresponding author:** P.-E. Fournier, URMITE, UM63, CNRS 7278, IRD 198, INSERM U1095, Aix Marseille Université, 27 Blvd Jean Moulin, 13385 Marseille Cedex 5, France
**E-mail: Pierre-Edouard.Fournier@univmed.fr**

## Introduction

Taxonomy, the study of organism classification, is a part of systematics, the study of the diversity and relationships among organisms. Prokaryotic taxonomy is traditionally regarded as consisting of three separate, but interrelated, areas: classification, nomenclature, and characterization. Classification is the arrangement of organisms into taxonomic groups on the basis of similarities; nomenclature is the assignment of names to the taxonomic groups identified in the classification; and characterization is the determination of whether an isolate is a member of a taxon defined in the classification and named in the nomenclature [1]. The influence of prokaryotic taxonomy is tremendous: attaching a name to a microbial strain conveys assumptions and implications associated with that organism, such as routine identification from clinical samples, pathogenicity potential, safety of handling, and cost [2]. However, there is no universal agreement on the rules and criteria used for microorganism classification.

Taxonomic classification has long been based solely on phenotypic characteristics, genetic data having being used only since the 1960s. However, the sequencing of the first bacterial genome in 1995 [3] substantially changed microbiology, by giving access to the whole genetic repertoire of a strain. It is now possible to generate whole prokaryotic genome sequences in a very short period of time, offering the possibility of using the whole genomic sequence of a prokaryote for its taxonomic description. In this review, we explore the benefits and shortcomings of using genomic data in prokaryotic taxonomy.

## Historical Overview and Current Practice in Prokaryotic Taxonomy

Although Carolus Linnaeus set the bases of modern taxonomy in the 18th century by studying plants, it was not before the late 19th century that Ferdinand Cohn classified bacteria into genera and species. Cohn and his contemporaries used

morphology, growth requirements, chemical reactions and pathogenic potential as the basis for bacterial classification [4]. Later, biochemical and physiological properties were also used by the Society of American Bacteriologists (which later became the American Society for Microbiology) in a report on bacterial characterization and classification that became the basis for the first edition of *Bergey's Manual of Determinative Bacteriology* in 1923. In 1947, a Code of Bacteriological Nomenclature was approved at the 4th International Congress for Microbiology [5]. In the 1960s, the technique of DNA–DNA hybridization (DDH) was introduced to measure genetic relatedness [6], but it was only widely accepted for classification purposes more than 20 years later [7]. In the 1980s, the development of PCR and sequencing of the 16S rRNA gene led to major changes in prokaryotic taxonomy [8], and this tool, although already commonly used for the description of new species in the 1990s, was recommended in 2002 as a key parameter in taxonomic classification [9,10].

Although prokaryotic nomenclature is regulated in the International Code of Nomenclature of Prokaryotes or the 'Bacteriological Code' [11], which is the latest edition of the Code of Bacteriological Nomenclature and is overseen by the International Committee on Systematics of Prokaryotes (ICSP), there has been no officially recognized system for the characterization and classification of prokaryotes until now. However, the most widely used system of characterization relies on a polyphasic approach, which is also used in the most widely accepted classification presented in *Bergey's Manual of Systematic Bacteriology* [4,12].

The term 'polyphasic taxonomy' was introduced in 1970 to refer to a taxonomy that brings together and incorporates many levels of information, from ecological to molecular, and includes several distinct types of information to yield a multidimensional classification. Currently, polyphasic taxonomy refers to a taxonomy that aims to utilize all available data [13]. These data include both phenotypic information, such as chemotaxonomic features (cell wall compounds, quinones, polar lipids, etc.), morphology, staining behaviour, and culture characteristics (medium, temperature, incubation time, etc.), and genetic properties, such as G+C content, DDH value, and 16S rRNA gene sequence identity with other closely related species with validated names [14].

Currently, the most commonly used tool for evaluating the phylogenetic position of a prokaryote is 16S rRNA gene sequence comparison. Likewise, the latest whole taxonomic schema for prokaryotic diversity presented in *Bergey's Manual* uses 16S rRNA phylogeny as its main basis [15]. However, there is growing interest in the use of other genes (protein-encoding genes) to resolve issues that are not solved by 16S rRNA gene sequencing. For example, some housekeeping genes (e.g. *dnaJ*, *dnaK*, *gyrB*, *recA*, and *rpoB*) have been used instead in multilocus sequence typing/multilocus sequence analysis (MLSA) [16]. One limitation of 16S rRNA is that it is rather conserved, and hence is not universally reliable for determination of taxonomic relationships at the species level. Furthermore, both nucleotide variations within multiple rRNA operons in a single genome and the possibility of 16S rRNA genes being derived from horizontal gene transfer may distort relationships between taxa in phylogenetic trees [17]. Nevertheless, 16S rRNA is currently the first-line tool for evaluating the taxonomic status of a prokaryotic strain at the same genus or species levels. It is currently assumed that two strains are members of the same species if their 16S rRNA gene sequence identity is >99%, and it may provide the first indication that a novel species has been isolated if an identity of <98.7% is found [18]. Similarly, a 16S rRNA identity of <95% with the phylogenetically closest species with a validated name may suggest that the isolate is a representative of a new genus.

Another widely used taxonomic criterion is DDH. A DDH value of $\geq$70% has been recommended as a threshold for the definition of members of a species, and DDH is deemed necessary when strains share >98.7% 16S rRNA gene sequence identity [12,14]. However, the DDH cut-off used is not applicable to all prokaryotic genera. For example, when applied to *Rickettsia* species, a DDH of 70% would not discriminate *Rickettsia rickettsii*, *Rickettsia conorii*, *Rickettsia sibirica*, and *Rickettsia montanensis* [19]. In addition, DDH protocols are considered to be tedious and complicated, with inherently large degrees of error, and only a few laboratories are equipped for this method, which remains expensive and is clearly not adapted to routine microbiology [2,20]. Furthermore, DDH studies can provide only a rough measurement of average genetic relationship, only closely related species or subspecies can be distinguished, and incremental databases cannot be developed for this method [4].

## The Prokaryotic Genomic Era

The sequencing of the *Haemophilus influenzae* genome in 1995 by conventional Sanger sequencing was a landmark in modern biology, as it marked the beginning of the genomic era [3]. However, in the next decade, bacterial genome sequencing remained time-consuming and expensive, and was reserved to a few sequencing centres worldwide. Thanks to the next-generation sequencing (NGS) technologies introduced from 2005, the number of sequenced prokaryotic genomes has rapidly increased, as new platforms are much faster and cheaper [21]. As of 18 September 2012, the Genome online Database listed 3381 prokaryotic genomes available as either