Systematic and Applied Microbiology 38 (2015) 79–83

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Large-scale evaluation of experimentally determined DNA G+C contents with whole genome sequences of prokaryotes

Mincheol Kim[a,1], Sang-Cheol Park[c,1], Inwoo Baek[b], Jongsik Chun[b,c,*]

[a] Division of Polar Life Sciences, Korea Polar Research Institute, Incheon 406-840, Republic of Korea
[b] School of Biological Sciences, Seoul National University, Seoul 151-742, Republic of Korea
[c] Interdisciplinary Program in Bioinformatics and Bioinformatics Institute, Seoul National University, Seoul 151-742, Republic of Korea

ARTICLE INFO

Article history:
Received 30 July 2014
Received in revised form
18 November 2014
Accepted 24 November 2014

Keywords:
DNA G+C content
Whole genome sequences
Prokaryote taxonomy
Average nucleotide identity

ABSTRACT

Historically, DNA G+C content has played a critical role in the description of bacterial and archaeal species. Despite its importance in prokaryote taxonomy, its accuracy has been questioned due to methodological heterogeneity and measurement errors of conventional methods. Here we investigated the extent of accuracy of experimentally determined DNA G+C contents by comparing the reference values calculated from whole genome sequences. The large-scale comparison revealed that G+C contents determined by high-performance liquid chromatography and buoyant density centrifugation methods were more similar to the genome-derived reference values than those generated by thermal denaturation method. However, there was a substantial degree of discrepancy in DNA G+C contents between values obtained by conventional methods and genome-derived reference values. The majority of the differences between them fell out of the acceptable range (i.e. 1 mol% G+C content difference) for species delimitation of prokaryotes. In contrast, when average nucleotide identity (ANI) was correlated to G+C difference among genomes, most G+C difference was confined to less than 1% within species. Therefore, erroneous conventional methods are not meaningful in the description of bacterial and archaeal species. For taxonomic purposes, DNA G+C content should be determined by calculating directly from high-quality genome sequences with at least 16× or higher sequencing depth of coverage.

Introduction

In *Bacteria* and *Archaea*, DNA G+C contents of genomes vary dramatically across phylogenetic lineages, ranging widely from 13.5% in Candidatus Zinderia insecticola [17] to 74.9% in *Anaeromyxobacter dehalogenans* [29]. G+C contents have been widely used as one of the key taxonomic parameters in a higher-level taxonomic classification, especially for characterizing bacterial phyla such as *Actinobacteria* (high G+C) and *Firmicutes* (low G+C) [31]. G+C contents have also been recognized as an important taxonomic information when circumscribing prokaryote species. For example, determination of G+C contents has been recommended as a minimum standard for describing new species of the family *Halomonadaceae* [1]. Generally, it is known that the G+C content is quite constant within a taxon, and its range does not exceed 10% within genus and members of a species differ by no more than 3% [30] or 5% [6]. However, the G+C content itself cannot be a unique taxonomic marker as many phylogenetic lineages share similar range of G+C contents, although they are distantly related to one another [27]. Furthermore, the suggested variation range in G+C contents at both genus and species levels was established on the basis of data mostly determined by experimental measurements which inherently have a certain level of errors [6,30].

Traditionally, DNA G+C content was determined experimentally by using buoyant density centrifugation (BD) [26], thermal denaturation (Tm) [16], and high-performance liquid chromatography (HPLC) [19]. In general, HPLC-based measurements are known to be among the most precise methods (±0.1 in standard deviation) and generate less experimental errors than other two methods (±0.4 in Tm and ±1.0 in BD) [19]. However, a certain level of errors always remains during the course of measurements. Discrepancies between experimentally determined- and genome sequence-derived values were recently reported. The difference between two approaches ranges from 1.2% (Tm) to 2.0% (HPLC, BD) on the basis of 80 case comparisons when measurement conditions were standardized [20]. However, it is almost impossible to

* Corresponding author at: School of Biological Sciences, Seoul National University, Seoul, Republic of Korea. Tel.: +82 2 876 8153; fax: +82 2 875 7250.
E-mail address: jchun@snu.ac.kr (J. Chun).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.syapm.2014.11.008
0723-2020/

check if all G+C contents available are measured under standard conditions. Moreover, a large proportion of G+C content values has been frequently re-used without further validation. Previous G+C content data determined by experimental measurements should be re-evaluated by comparing with reference values derived from their corresponding genomes.

Genome sequence-based G+C content determination undoubtedly has an advantage over those experimental methods as it allows for the direct counting of nucleotides throughput the whole genome sequences, thus the precise ratio of G+C content over total bases can be obtained. In earlier studies, G+C contents have been estimated using a single gene [4] or several genes [15] due to the high cost of genome sequencing. It is now possible to calculate more accurate G+C contents from whole genome sequences thanks to the rapidly increasing number of genomes in public databases. For example, the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project has greatly increased the number of genomes of prokaryote type strains [32]. Of course, even genome-predicted G+C contents may not be 100% accurate. Although the effect is not always large, both GC-rich and AT-rich regions are generally underrepresented in sequencing reads, so G+C contents of poorly assembled genomes do not always corroborate that of the complete genome [3,25]. Overall G+C content of a genome is known to be not considerably affected by those factors [3]. However, it is unclear the depth of genome sequencing to provide an accurate estimate of the G+C content.

Here we tested how reliable the experimentally determined G+C contents are when compared to those directly calculated from genome sequences. Almost all G+C contents data determined by traditional methods were surveyed by a comprehensive literature search, and overall error rate of each method and distributional differences were determined by comparing with the genome-derived reference values. To test the possible effect of sequencing efforts on the G+C content values, the changing pattern of G+C contents was examined on six bacterial genomes at differing sequencing depths of coverage. Lastly, given the taxonomic importance of G+C content in the description of genus and species of prokaryotes, we tested if there is any recognizable relationship between pairwise ANI and G+C difference values by performing a large-scale genome comparison.

Materials and methods

Data collection

25,944 genomes of *Bacteria* and *Archaea* were downloaded from the GenBank database (as of July 2014). To filter out low-quality genomes, draft genomes generated by single-cell genomics and metagenomic assembly were excluded. The genomes with possible contamination were also checked by the presence of abnormal G+C content and inconsistent classification result when taxonomically identified using genome-extracted 16S rRNA gene sequences. In total, 25,428 high-quality genomes, including 2814 type strains and 3679 strains identified with valid names, were used for the final analysis. The experimentally determined G+C contents data (1120 by HPLC; 1070 by Tm; and 174 by BD method, respectively) was obtained from the literature, mostly the International Journal of Systematic and Evolutionary Microbiology (IJSEM) and Bergey's Manual of Systematic Bacteriology (Table S1).

Supplementary Table S1 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.syapm.2014.11.008.

Average nucleotide identity and G+C content calculation

G+C contents and pairwise average nucleotide identity (ANI) were obtained from high-quality genomes with validly published species names. ANI values were calculated following the algorithm described by Goris et al. [7] and 95% ANI cut-off, which corresponds to 70% DNA–DNA hybridization (DDH), was used as a boundary for species circumscription [11,23]. Genomic G+C content was calculated by simply counting the proportion of guanine and cytosine among the total nucleotide sequences per genome.

Genome assembly at multiple sequencing depths of coverage

To test what extent overall G+C contents of genomes change with differing sequencing depths of coverage, we downloaded Illumina-generated raw reads of six bacterial genomes from NCBI SRA database. They include two low G+C genomes, *Clostridium difficile* 630 (PRJNA57679) and *Campylobacter jejuni* subsp. *jejuni* NCTC 11168 (PRJNA57587), two medium G+C genomes, *Escherichia coli* K12 strain MG1655 (PRJNA57779) and *Salmonella enterica* subsp. *enterica* serovar Heidelberg strain CFSAN002069 (PRJNA212974), and two high G+C genomes, *Burkholderia pseudomallei* 668 (PRJNA58389) and *Mycobacterium tuberculosis* CDC1551 (PRJNA57775). Raw Illumina reads were randomly subsampled to reach multiple coverages ranging from 1× to 64× with five replicates at each depth. Each subsampled subset was assembled using Velvet 1.20.1 [33] with default parameters and G+C contents for the newly assembled genomes were calculated using the method described above.

Statistical analyses

In order to evaluate how well experimentally determined G+C contents agree to the genome-derived equivalents, several statistical methods were performed. Values exhibiting over 10% difference to the reference values were first removed as this difference level is too high to be considered as measurement error given that 10% difference has generally been thought of a threshold for differentiating genus. After removing outliers, violin plot and mountain plot were generated for visualizing the data structure. The mountain plot (folded empirical cumulative distribution plot) is prepared by computing a percentile for each ranked difference between the new and reference methods [14]. It does not need any distributional assumption and is particularly useful when comparing several distributions simultaneously and estimating percentiles for large differences. Overall mean and standard deviation of the difference values was calculated and Bland–Altman plot was additionally generated for aiding visual interpretation of the result. Bland–Altman method is a graphical tool to evaluate the agreement between two different methods on a single subject [2]. It plots the difference against the mean of two measurements with 95% limits of agreement (LOA) and the smaller range of LOAs represents the better agreement between two methods. Prior to applying Bland–Altman method, normality of residuals and constant variance were checked using 'car' package [5] in R (www.r-project.org). A modified Bland–Altman method, which differences between two methods are plotted against the reference values [13], was performed as genome sequence-derived G+C contents are almost the 'true' values although there may be a minute level of technical errors. Linear regression was additionally performed to support the result and 95% prediction limits were estimated. All statistical analyses and plotting were performed using R.