Biomolecular Detection and Quantification 3 (2015) 17–24

Contents lists available at ScienceDirect

Biomolecular Detection and Quantification

journal homepage: www.elsevier.com/locate/bdq

Original Article

International interlaboratory study comparing single organism 16S rRNA gene sequencing data: Beyond consensus sequence comparisons

Nathan D. Olson[a,*], Steven P. Lund[b], Justin M. Zook[a], Fabiola Rojas-Cornejo[c], Brian Beck[d,1], Carole Foy[e], Jim Huggett[e], Alexandra S. Whale[e], Zhiwei Sui[f], Anna Baoutina[g], Michael Dobeson[g], Lina Partis[g], Jayne B. Morrow[a]

[a] Biosystems and Biomaterials Division, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD 20899, USA
[b] Statistical Engineering Division, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD 20899, USA
[c] Instituto de Salud Pública de Chile, Chile
[d] American Type Culture Collection, 10801 University Boulevard, Manassas, VA 20110, USA
[e] Science and Innovation Division, LGC, Queens Rd, Teddington, Middlesex TW11 0LY, UK
[f] National Institute of Metrology, Beijing 100013, China
[g] National Measurement Institute, West Lindfield, NSW 2070, Australia

ARTICLE INFO

Article history:
Received 30 October 2014
Received in revised form 27 January 2015
Accepted 27 January 2015
Available online 5 March 2015

Keywords:
DNA sequencing
16S rRNA
Interlaboratory study

ABSTRACT

This study presents the results from an interlaboratory sequencing study for which we developed a novel high-resolution method for comparing data from different sequencing platforms for a multi-copy, paralogous gene. The combination of PCR amplification and 16S ribosomal RNA gene (16S rRNA) sequencing has revolutionized bacteriology by enabling rapid identification, frequently without the need for culture. To assess variability between laboratories in sequencing 16S rRNA, six laboratories sequenced the gene encoding the 16S rRNA from *Escherichia coli* O157:H7 strain EDL933 and *Listeria monocytogenes* serovar 4b strain NCTC11994. Participants performed sequencing methods and protocols available in their laboratories: Sanger sequencing, Roche 454 pyrosequencing®, or Ion Torrent PGM®. The sequencing data were evaluated on three levels: (1) identity of biologically conserved position, (2) ratio of 16S rRNA gene copies featuring identified variants, and (3) the collection of variant combinations in a set of 16S rRNA gene copies. The same set of biologically conserved positions was identified for each sequencing method. Analytical methods using Bayesian and maximum likelihood statistics were developed to estimate variant copy ratios, which describe the ratio of nucleotides at each identified biologically variable position, as well as the likely set of variant combinations present in 16S rRNA gene copies. Our results indicate that estimated variant copy ratios at biologically variable positions were only reproducible for high throughput sequencing methods. Furthermore, the likely variant combination set was only reproducible with increased sequencing depth and longer read lengths. We also demonstrate novel methods for evaluating variable positions when comparing multi-copy gene sequence data from multiple laboratories generated using multiple sequencing technologies.

1. Introduction

The 16S ribosomal RNA gene (16S rRNA) is the most commonly used marker in bacterial genotypic identification, and there are a number of benefits and challenges associated with its use [1,2]. The 16S rRNA gene is an ideal target due to its ubiquitous presence in prokaryotic organisms and is characterized by a series of variable and conserved regions [3]. Universal PCR primers targeting different conserved regions can amplify the intermittent variable regions from a diverse selection of prokaryotes [4]. The amplified regions are subsequently sequenced allowing for genus and sometimes species level identification [5].

16S rRNA microbial identification has a number of well-documented challenges including orthologue (between organisms) and paralogue (within an organism's genome) sequence diversity [6,7]. Another major challenge occurs due to differential microbial DNA contamination found in the laboratory or reagents, leading to erroneous results [8–10]. In addition, disparities between different laboratories lead to poor reproducibility [3,7]. 16S rRNA gene sequencing is currently performed using both traditional Sanger sequencing and Next Generation Sequencing (NGS). Sequence read

* Corresponding author.
E-mail address: nathanael.olson@nist.gov (N.D. Olson).
[1] Present address: Microbiologics, 200 Cooper Ave N, St. Cloud, MN 56303, USA.

http://dx.doi.org/10.1016/j.bdq.2015.01.004

lengths, throughput, and base call accuracy vary by sequencing platform.

NGS platforms, which are increasingly being used, have relatively short reads (75 base pairs (bp) to 500 bp), but much deeper coverage (i.e., higher number of sequence reads covering each position) per run (from approximately 1×10^4 to over 1×10^8 reads). Sanger sequencing offers long read lengths (~800 bp) and lower, better-characterized error rate compared to NGS [11–13]. The disadvantage of higher error rates in NGS is often mitigated by deeper coverage. Regardless of error rate all sequencing platforms have systematic errors [11,13]. To date, there have been no comparisons between 16S rRNA sequences from single organisms obtained using multiple platforms from different laboratories that consider the diversity of 16S rRNA gene copies (paralogues).

The objective of this study was to compare 16S rRNA sequencing data among six international laboratories using both Sanger and NGS platforms. The newly formed Microbiology Steering Group (MBSG) of the Consultative Committee for Amount of Substance (CCQM) conducted this study (http://www.bipm.org/en/committees/cc/wg/mbsg.html). 16S rRNA sequencing data were evaluated at three levels (Fig. 1, Definitions):

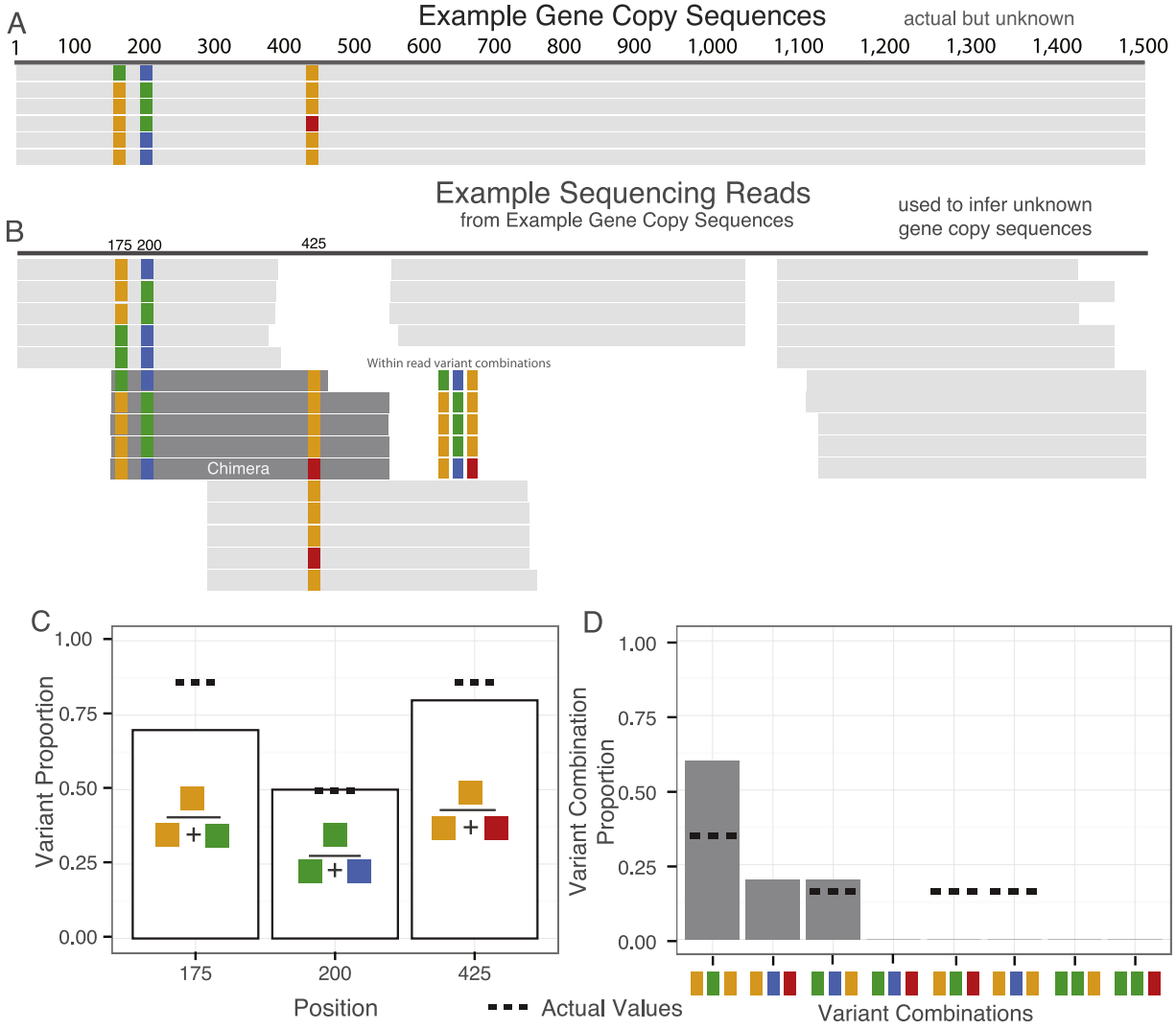

Fig. 1. Example of the three levels of sequence analysis for multi-copy genes. An example set of six 16S rRNA gene copies with three biologically variable (colored) positions (175, 200, and 425) and example-sequencing reads are used to depict the three levels of analysis. The collection of variants, which in this example is comprised of six triplets; define the identity and provide a complete picture of the six 16S rRNA gene copies. (A) Six 16S rRNA gene copies represent the actual, but unknown, sequences within the example genome. Gray horizontal boxes stretching between 1 and 1500 bp represent individual genes; colored boxes, widened to aid visualization, indicate three variable positions. (B) Example set of "454" sequencing reads generated from the actual 16S rRNA gene copies (A) aligned to the reference. Sequencing data generated from the 16S rRNA copies can be used to make inference about the unknown gene copies from which they originated. For the first level of analysis, the identity of the biologically conserved positions, indicated in gray for both (A) and (B), is assessed using single nucleotide polymorphism calling pipelines. The second level of analysis is estimating the variant copy ratios (unknown) for the variable positions from the observed variant proportions (proportion of variants at biologically variable positions for a sequencing dataset). A statistical model was developed to estimate the variant copy ratio from the observed variant proportions. (C) The observed variant proportions, white bars, is calculated for the example read set (B), the variant proportion equation for each position is shown inside the bar. The true but unknown variant proportion (proportion of variants at biologically variable positions for a set of gene copies), calculated from the gene copies (A) is indicated with a dashed line. For the third level of analysis, the gene copy variant combination set is estimated from the sequencing data. Using a statistical model, the likely variant combination set is estimated from the observed variant combination proportions. (D) Observed variant combination proportions for reads covering the three variant positions (dark gray) in the example read set is indicated by gray bars. Dashed lines are used to indicate the true but unknown gene copy variant combination proportions. A chimeric read is included in the example read set. Chimeras are the hybrid product of two parent sequences, in this study two 16S rRNA gene copies. The combination of variants found in the chimeric read (depicted as 'gold, blue, and red') is not present in any of the actual 16S rRNA gene copies. The read is, therefore, the product of a chimera event between a PCR product from a gene copy with 'gold, blue, and gold' variant combination and a PCR product from a gene copy with a 'gold, green, and red' variant combination. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)