International Interlaboratory Study Comparing Single Organism 16S rRNA gene Sequencing Data: Going Beyond Consensus Sequence Comparisons

Nathan D. Olson[1,*], Justin Zook[2], Steve Lund[3], Fabiola Rojas-Cornejo[3], Brian Beck[4], Jim Huggett[5], Alexandra Whale[5], Zhiwei Shi[6], Anna Baoutina[7], Michael Dobenson[7], Lina Partis[7], Jayne B. Morrow[1]

1 Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA

2 Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, USA

3 Instituto de Salud Pblica de Chile,

4 American Type Culture Collection, Manassas, VA, USA

5 Science and Technology Division, LGC, Teddington, Middlesex, UK

6 National Institute of Metrology, Beijing, China

7 National Measurement Institute, West Lindfield, NSW Australia

∗ **E-mail: nolson@nist.gov**

Abstract

The Consultative Committee for Amount of Substance established a working group to advance the metrology of microbial identification and quantification. This study represents the initial step towards international comparability and measurement assurance for microbial identity. Six laboratories sequenced the 16S rRNA gene from Escherichia coli O157:H7 strain EDL933 and *Listeria monocytogenes* serovar 4b strain NCTC 11994. The 16S rRNA is an orthologous gene, frequently also present as paralogues, containing both conserved and variable bases. Participants used either Sanger sequencing, Roche 454 pyrosequencing®("454"), or Ion Torrent PGM®. To rigorously challenge comparability, the sequencing data was evaluated on three levels: 1) identity of conserved bases, 2) ratio of bases at the variable positions, and 3) the set of likely variant combinations for an individual gene copy. Regardless of sequencing method biologically conserved positions were correctly identified. To determine the abundant base copy ratio and the likely variant combinations for a gene copy analytical methods were developed using Bayesian and maximum likelihood statistics. This study presents novel methods for comparing sequencing data from different sequencing platforms for a multi-copy gene by evaluating both conserved and variable bases, paving the way towards measurement assurance for DNA sequence based microbial identification. Additionally, this study demonstrates the power in evaluating both biologically conserved and variable positions when comparing multicopy gene sequence data.

Introduction

The Consultative Committee for Amount of Substance (CCQM) was established in 1993 to address metrological challenges in chemistry including biochemistry and clinical chemistry [1]. At a workshop "Metrology and the need for traceable microbiological measurements to ensure food quality and safety" hosted by CCQM in 2011, a need to advance metrology of microbiological systems was identified. These advances would have broad impact beyond food safety, including human health, water quality, biodefense, environmental monitoring, and basic microbiological research. This study presents the results from the first international interlaboratory study to establish foundational measurement capabilities in microbial genomics. The 16S ribosomal RNA (rRNA) gene is the most commonly used marker in bacterial genotypic identification, and there are a number of benefits and challenges associated with its use [2,3]. Two primary benefits to using the 16S rRNA gene for identification are: (1) the 16S gene is present in all prokaryotic organisms allowing for universal microbial identification, and (2) the16S gene has a series of

variable and conserved regions [4]. The variable regions allow for genus and sometimes species level identification [5]. Universal 16S rRNA primers targeting the conserved regions can be used in combination with PCR to amplify the 16S gene from a diverse selection of prokaryotes [6]. Three challenges are associated with 16S rRNA microbial identification. (1) The diversity of 16S genes within a taxonomic group varies greatly from highly conserved to highly variable. (2) A number of organisms have more than one copy of the 16S gene within their genome and the diversity of these paralogues can be greater than the diversity between the orthologous sequences from different species [7,8]. (3) 16S rRNA genes from microbial DNA contamination found in the laboratory or reagents are also commonly sequenced, leading to erroneous results [9-11]. A number of sequencing technologies are currently used for microbial identification, each with different advantages making them suitable for different applications [12]. Three main parameters determine suitability: read length, number of reads, and error rates. Of the established methods Sanger sequencing has the longest read length (800 bp) and sequences are derived from a mixed population of PCR amplicons in combination with cloning. Sanger sequencing is limited to low throughput applications (typically 100s of reads) sequencing small regions of individual genomes [12]. On the other hand, next generation sequencing (NGS) platforms, including the Roche 454 pyrosequencing®("454"), Illumina MiSeq®and HiSeq®, and the Ion Torrent PGM®have relatively shorter reads (75 bp to 250 bp) but a much higher throughput per run (1 x 104 to over 1 x 108 reads). Sanger sequencing has a lower, better-characterized error rate compared to NGS [13-15]. For NGS platforms the disadvantages of higher error rates are outweighed by the advantages of higher coverage (i.e., higher number of sequences covering each position). All sequencing platforms have systematic errors; errors that occur in a predictable (non-random) manner for a given sequence context.[13,15]. A number of previous interlaboratory sequencing studies used for proficiency testing [16] focused on comparability of bacterial identification methods [17]. However, to our knowledge no interlaboratory studies have focused on next generation and Sanger sequencing of a multi-copy gene in single organisms. Therefore, interlaboratory comparisons and proficiency testing are needed to establish confidence in measurement capabilities. Proficiency testing is critical for policy and clinical decision-making, as it instills confidence in the results generated by the participating laboratories [18]. The results from the CCQM microbial identity working group interlaboratory study are presented here and are a first step toward comparability and measurement assurance. The objective of the study was to evaluate the comparability of 16S rRNA sequencing data among six laboratories across the globe using both Sanger and NGS platforms. Sequencing data were evaluated at three levels: (1) correct identification of each base at biologically conserved positions (identical between paralogues) in the 16S rRNA gene; (2) correct identification of the proportion of bases at biologically variant positions (different between paralogues) in the different gene copies; (3) identification of likely variant combinations in an individual gene copy sequence. This study shows that for multicopy genes the correct consensus sequence is obtainable with any of the sequencing methods used in this study. However, higher coverage and longer reads are required to obtain sequences for individual gene copies.

Citation of Einstein's paper [**?**].

Materials and Methods

Study overview

Participants included five national metrology institutes (NMIs): National Institute for Standards and Technology (NIST, USA), LGC (LGC, UK), National Measurement Institute Australia (NMIA, AUS), National Institute of Metrology China (NIMC, CHN), Chilean Public Health Institute (ISP, Instituto de Salud Pblica ,CHL) and a stakeholder laboratory, American Type Culture Collection (ATCC, USA.) Study participants sequenced the 16S rRNA gene from two freeze-dried genomic DNA certified reference materials from IRMM (Institute for Reference Materials and Measurements, Belgium), IRMM 449 Escherichia coli O157:H7 EDL933 and IRMM 447 Listeria monocytogenes strain 4B NCTC 11994.

Escherichia coli O157:H7 EDL933 has seven copies of the 16S rRNA gene in its genome and Listeria monocytogenes strain 4B NCTC 11994 has six.

Sequencing methods

Study participants sequenced the reference materials using different sequencing platforms and strategies (Fig 1, detailed protocols in Supplemental Sequencing Methods). Sequencing platforms included Sanger Sequencing and two NGS platforms (Table 1, Supplemental Results Table 1): 454 Pyrosequencing®("454", 454 Life Sciences, Branford, CT, USA) and Ion Torrent PGM®(Life Technologies, San Francisco, CA, USA). All PCR primer sequences and thermocycler protocols used in this study were previously used by the Human Microbiome Project Jumpstart Consortium Group [4].

Sequence Data Analysis

Raw sequence data were submitted to NIST for analysis. Sequence data were compared at three levels: base calls for the biologically conserved positions, base ratios for positions that are variant between gene copies, and the likely variant combinations in an individual gene copy. All scripts and reference sequences used for data analysis are available at https://github.com/nate-d-olson/ccqm$_m bwg_1 6S. The 16S rRNA gene reference sequen$
$//www.ncbi.nlm.nih.gov/genome). The individual 16S rRNA gene copies were extracted from whole genome sequences for E.$
$H7 EDL933 (accession number NC_0 02655) and L.\ monocytogenes strain 4B LL195 (accession number NC_0 19556), since the wh$

Biologically Conserved Positions

A single nucleotide polymorphism (SNP) calling pipeline was used to evaluate biologically conserved positions. To validate the bioinformatics pipeline, eight pipelines were evaluated using a full factorial experimental design (Supplemental Results Table 2). Datasets were compared using the TMAP mapping algorithm, duplicate removal, realignment around indels, and the UnifiedGenotyper variant caller bioinformatics pipeline.

Biologically Variant Positions

Next, the datasets were compared based on the estimated ratio of bases at the biologically variant positions. As the true base ratio is unknown, the estimated ratios were compared to the consensus base ratio estimates. We defined the consensus base ratio as the estimated base ratios predicted by a majority of the datasets generated in the study. At each biologically variable position, the number of copies for the two bases or base ratio, were estimated from the observed proportions using Bayesian statistics and the binomial sampling theory. Additionally, power analysis was performed to determine the desired coverage for abundant base ratios predictions. See the supplemental methods for a detailed description of the statistical methods (Supplemental Statistical Methods).

Likely Gene Copy Variant Combinations

Finally, the individual datasets were evaluated based on the estimated set of likely variant combinations in an individual 16S gene copy sequence. The phasing of the biologically variant positions, namely the variants on the same gene copy, similar to haplotype phasing for diploid genomes, was used to determine the individual gene copy sequences. To determine phasing, biologically variant positions in individual "454" sequencing reads and Sanger clones were concatenated to form variant strings for individual 16S rRNA gene copies, and the variant string proportions were determined for each dataset. Reads in the Ion Torrent datasets were too short and Sanger amplicon sequencing datasets were too small for variant

combination analysis. The statistical analysis using maximum likelihood was used to evaluate the probability of all possible combinations of variant strings for the seven *E. coli* and six *L. monocytogenes* gene copies and is described in the supplemental statistical methods. The analysis considered the probability that the variant strings from any individual read were the product of a chimera event. In the absence of the truth or the gene copy variant combinations results from the individual datasets were compared to the values obtained using all the datasets combined.

Results and Discussion

All datasets generated by the study concurred at biologically conserved positions, however, the precision of the results for the biologically variable positions and likely variant combinations in a gene copy were dependent on the length and number of reads in the datasets.

Summary of Datasets

Sequencing datasets generated by the participating laboratories varied in both read length and number, depending on the sequencing platform and method (Table 1, Supplemental Results Table 1). As expected the differences in read length were dependent on the sequencing chemistry, and the number of reads was platform dependent [12,19].

Conserved Positions Method Validation

Biologically conserved positions were evaluated using a whole genome variant calling pipeline. Only false positive and no false negative SNPs were identified using the eight pipelines, indicating robust base calls for the conserved positions (Table 2, Supplemental Results).

Conserved Positions Dataset Comparison

The TMAP UnifiedGenotyper pipeline with realignment around indels and duplicate read removal was used to compare the biologically conserved positions. The TMAP algorithm was chosen because the shotgun method used to generate the Ion Torrent datasets presented the greatest challenge to the mapping algorithm. Additionally, "454" sequencing reads have similar error profiles to Ion Torrent reads, e.g. higher rate of insertions and deletions compared to substitutions [13]. The UnifiedGenotyper variant caller was chosen for dataset comparison because it provides a number of statistics for use in evaluating variant calls, e.g. Fisher Exact strand bias test. Regardless of the read length or number of reads in the dataset, the base calls assigned by the variant calling pipeline for the conserved positions was identical to the reference gene sequence, excluding false positive variant calls (Table 2). There were two primary causes of the false positive variant calls: (1) strand bias, due to the amplicon sequencing strategy and bioinformatics indicated by high Fisher Strand bias statistics (FS), (2) mapping errors around biologically variable positions.

Biological Variant Ratios Method Development

The second level of sequence analysis was the abundant base copy ratio for the biologically variant positions. Traditionally 16S rRNA sequence comparisons are based on individual sequences ignoring differences between gene copies. Ignoring gene copies can result in the overestimation of microbial diversity within a sample or sequence misclassification [8]. To determine the ratio of bases at the biologically variant positions, a novel Bayesian analysis based on binomial sampling theory was developed. According to the binomial distribution, the observed base ratios, while precise (due to high coverage), differed significantly from all potential copy ratios (Supplemental Results Figure 1 and 2). Subsequently given

the observed base ratios a Bayesian approach was used to identify the most probable copy ratio out of the possible abundant base ratios assuming *E. coli* and *L. monocytogenes* have seven and six 16S gene copies respectively (Table 2). The *L. monocytogenes* strain 4b NCTC 11994 had 3 variant positions in its six 16S rRNA gene copies. The *E. coli* O157:H7 EDL933 strain had 11 variant positions in its seven16S rRNA gene copies. The model also assumes a single organism homogenous sample and no bias in sequencing the individual gene copies.

Biological Variants Dataset Comparison

The agreement between base ratio estimations for the individual datasets and the overall consensus was correlated with coverage, which is dependent on the number of reads and read length (Table 3). The estimated base ratios for all "454" datasets agreed with the consensus for the biologically variant positions. Out of the methods with base ratio estimations in disagreement with the consensus, Sanger clone libraries had the fewest disagreements, followed by Ion Torrent datasets and Sanger amplicon datasets. The most probable ratios for the Ion Torrent sequencing datasets were not in agreement with the consensus base ratios for 10 out of the 28 total variant positions and had lower overall coverage then the "454" datasets (Table 3). All nine of the *E. coli* 6:1 variants in the NIST Ion Torrent dataset were in disagreement with the consensus ratio, with 5:2 as the most probable ratio. For the NIMC Ion Torrent dataset, the estimated abundant base ratios were in agreement with the consensus for the 9 variants with the 6:1 ratio. As only the NIST and not the NIMC Ion Torrent abundant base ratios estimations for the positions with consensus base ratios of 6:1 were not in agreement with the consensus the bias is unlikely due the to sequencing platform or library preparation method indicating an unknown run specific bias. Despite having lower coverage compared to the Ion Torrent datasets, the most probable base ratios for the Sanger clone libraries were in agreement with the consensus base ratios for all but one of the positions. The Sanger amplicon-sequencing dataset contained only four reads, which was inadequate to precisely determine the base ratio. To estimate the required coverage to determine the abundant base ratio a power analysis was performed. Based on power analysis for the 6:1 and 5:1 abundance base ratios 96 and 80 x coverage is desired and 196 and 144 x coverage is required for 4:3 and 3:3 abundant base ratios respectively. Likely variants for individual gene copies The third level of sequence analysis is the comparison of the full-length sequences for the individual gene copies. Previously only Sanger clone libraries were used to evaluate likely variants for individual gene copies [8]. A novel method for evaluating likely variants from NGS datasets was developed using maximum likelihood statistics and taking into consideration the rate of a chimera event occurring between two gene copies. The likely variants for individual gene copies for the two strains were determined using "454" and Sanger clone library datasets. The sample size of the Sanger clone library datasets was not large enough to confidently identify the variant strings for all 11 positions for the individual 16S rRNA gene copies in *E. coli*, so the individual gene copy sequences were only determined for 10 of the 11 positions. Since the true combination of likely variants for the individual gene copies is not known consensus values were used for dataset comparison. The consensus values were defined as the variant combination obtained using the combined Sanger clone library and "454" datasets. The dataset sizes were not normalized, biasing the "454" data, but reducing the impact of chimeras present in the Sanger clone libraries.

Comparison of Likely Variant Combinations

For *L. monocytogenes* the most likely combination of variant strings when all datasets were combined was the same as the most likely combination for the datasets individually excluding the LGC Sanger clone library (Supplemental Results Table 5). For the LGC Sanger clone library none of the clones had the GTA variant string. The most likely variant string combination for the LGC Sanger clone library dataset had a copy of the GCA variant string instead of GTA. The GCA variant string was most likely the product of a chimera PCR amplicon. For the individual *E. coli* datasets, the consensus combination of likely

variants was not estimated for either Sanger clone library. The estimated likely variant combinations for the clone libraries had an additional copy of the gene with the variant string "ACCGGATTGTG'. The difference between the likely variant combinations of a gene copy for the Sanger clone libraries and the consensus was due to the small sample size compared to the "454" datasets (Figure 3).

Conclusions

As a first step in establishing comparability in microbial identity sequencing data, we developed methods to compare results from multiple sequencing platforms, for a gene with multiple copies, on three levels. The methods presented here could be applied to future multicopy gene sequence comparison studies. The sequence analysis methods presented here provide additional levels of comparability between 16S rRNA gene sequences beyond comparison of the gene consensus sequence alone. This could potentially allow for higher resolution 16S rRNA gene taxonomic identification and reduced overestimation of microbial diversity. As a follow-up study, the microbial identity group plans to perform whole genome sequencing for a genomic DNA reference material and establish methods for comparing whole genome sequencing datasets among laboratories.

Acknowledgments

The authors would like to thank the members of the CCQM Metrology of Microbial Systems Steering Group and the Microbial Identity working group. We also thank Mirjana Trifunovic from ATCC as well as Dr. Jorge Fernndez and Dr. Abel Vsquez from Instituto de Salud Pblica de Chile for assistance with sequencing. Jenny McDaniel and Lindsay Vang at NIST helped with the sequencing at NIST. Additionally, we would like to acknowledge Drs. Nancy Lin and Erica Seifert for help during the writing process. The Department of Homeland Security (DHS) Science and Technology Directorate under the Interagency Agreement HSHQPM-12-X-00078 with NIST supported this work.

Disclaimers

Opinions expressed in this paper are the authors and do not necessarily reflect the policies and views of DHS, NIST, or affiliated venues. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendations or endorsement by NIST or NMIA, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Official contribution of NIST; not subject to copyrights in USA.

References

Figure Legends

Figure 1. Bold the first sentence. Caption should be left justified, as specified by the options to the caption package.

Tables