National Institute of Standards & Technology

Report of Investigation

Reference Material 8375

Microbial Genomic DNA Standards for Sequencing Performance Assessment (MG-001, MG-002, MG-003, MG-004)

This Reference Material (RM) is intended for validation, optimization, process evaluation, and performance assessment of microbial whole genome sequencing. Particular attention was devoted to assessing the genome sequence. A unit of RM 8375 consists of four vials. Each vial contains a different microbial genomic DNA sample (MG-001 *Salmonella* Typhimurium LT2, MG-002 *Staphylococcus aureus*, MG-003 *Pseudomonas aeruginosa*, and MG-004 *Clostridium sporogenes*).

This material is intended to help assess performance of high-throughput microbial DNA sequencing methods. This genomic DNA has been extracted from cell cultures. Because the RM is extracted DNA, it does not assess pre-analytical steps such as DNA extraction. It does, however, challenge sequencing library preparation, sequencing machines, base calling algorithms, and the subsequent bioinformatics analyses. This RM is not intended to assess other bioinformatics steps such as strain identification, phylogenetic analysis, or genome annotation. The purity of the genomic DNA was assessed to be fit for these purposes.

Information Values: Information values are currently provided for the whole genome sequence to enable performance assessment of variant calling. An information value is considered to be a value that will be of interest and use to the RM user, but insufficient information is available to assess the uncertainty associated with the value. We describe and disseminate our best, most confident, estimate of the genome assembly using the data and methods available at present [1]. Information values cannot be used to establish metrological traceability. The genome sequence files referenced in this Report of Investigation are available at:

https://github.com/usnistgov/NIST_Micro_Genomic_RM_Data/

Expiration of Value Assignment: RM 8375 is valid, until **20 July 2026**, provided the RM is handled and stored in accordance with instructions given in this report (see "Instructions for Storage and Use"). This report is nullified if the RM is damaged, contaminated, or otherwise modified.

Maintenance of RM: This report will be updated periodically to reflect important new releases as the genome assemblies are updated. NIST will monitor this RM over the period of its validity. If substantive technical changes occur that affect the value assignment before the expiration of this report, NIST will notify the purchaser. Registration (see attached sheet or register online) will facilitate notification

Overall direction and coordination of the analyses was performed by N. Olson, J. Zook and M. Salit of the NIST Material Measurement Laboratory.

Technical measurements were conducted by L. Harris, J. McDaniel, E. Romsos, and D. Catoe of the NIST Biosystems and Biomaterials Division. Analyses were conducted by N. Olson, J. Zook, E. Romsos, J. McDaniel, and D. Samarov.

Statistical consultation for this RM was provided by D. Samarov of the NIST Statistical Engineering Division.

Support aspects involved in the issuance of this RM were coordinated through the NIST Office of Reference Materials.

Sheng Lin-Gibson, Acting Chief
Biosystems and Biomaterials Division

Gaithersburg, MD 20899
Report Issue Date: 08 May 2018
Report Revision History on Last Page

Steven J. Choquette, Director
Office of Reference Materials

NOTICE AND WARNINGS TO USERS

RM 8375 is intended for research use. Since there is no consensus on the infectious status of extracted DNA, handle RM 8375 components as Biosafety Level 1 materials capable of transmitting infectious disease, as recommended by the Centers for Disease Control and Prevention (CDC) Office of Safety, Health, and Environment and the National Institutes of Health (NIH) [2].

INSTRUCTIONS FOR STORAGE AND USE

Storage: RM 8375 is stored at –20 °C at NIST but will be shipped with freezer packs and may not arrive frozen. Upon receipt, RM 8375 should be kept in the dark at –20 °C for long-term storage, or in the dark at 4 °C for short-term storage (if use is likely to be within approximately 8 weeks).

Use: It is recommended that the genome sequence be used for assessing the accuracy of variant calls and not genome assembly, because orthogonal measurement methods disagree about the genome sequence structure and the differences could not be resolved. Due to the complexities associated with the method used to assess base call accuracy and the potential for errors in the genome sequence, it is strongly recommended that the user ***manually*** inspect aligned reads around putative regions with high base call error rates. Best practices for evaluating variant calls in microbial genomes were outlined in reference 3.

These assemblies are based on data from multiple measurement methods; we evaluated the material using a number of orthogonal methods to minimize the impact of systematic errors associated with individual platforms. These data are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) as Biosamples SAMN02854572, SAMN02854573, SAMN02854574, and SAMN02854575 for MG-001, MG-002, MG-003, and MG-004, respectively [4]. As sequencing technologies and analysis methods improve, the genome sequence will be updated and available at the Github site (see "Information Values" above). A new version number will indicate updated genome sequences along with a text file describing the changes. Users who identify errors in the reference genome can post their results on the site shown in reference 5, and confirmed errors will be incorporated into the next sequence update.

SOURCE PREPARATION[(1)]

MG-001 is extracted genomic DNA from *Salmonella enterica* subspecies *enterica* serovar Typhimurium strain LT2 isolate CFSAN000639 (NCBI Biosample SAMN02854572). The strain was selected based on public health relevance by the Food and Drug Administration (FDA) and provided by the FDA Center for Food Safety and Nutrition (CFSAN). The genomic DNA for MG-001 was prepared by Lofstrand Labs Limited (Gaithersburg, MD) from a pure culture of *Salmonella enterica* LT2 as follows. The liquid culture received from FDA CFSAN was plated, incubated overnight at 37 °C. Five colonies were suspended in 5 mL Luria Broth, and then inoculated 0.2 mL each into 20 x 150 mm plates (equivalent to 2 L liquid). DNA was isolated by lysing the bacteria in Lysis buffer containing NaCl, Tris, EDTA, Lysozyme and sodium dodecyl (lauryl) sulfate (SDS). Protein was treated with Proteinase K, and RNA was treated by RNase A, followed by ammonium acetate precipitation. DNA was precipitated with isopropanol, washed with 70 % alcohol, drained, and then dissolved in TE (Tris 10 mM, EDTA 0.1 mM, pH 8.0). Proteinase K and RNase A were used to treat protein and RNA. Ammonium acetate was used to remove protein. DNA was recovered by isopropanol precipitation. DNA was washed with 70 % alcohol, drained, and then dissolved in TE (Tris 10 mM, EDTA 0.1 mM, pH 8.0).

MG-002 is extracted genomic DNA from *Staphylococcus aureus* strain NRS100 isolate COL (NCBI Biosample SAMN02854573). The strain was selected based on public health relevance and its low percentage of G and C DNA bases. The strain was selected for use by the FDA and isolated from a clinical sample by Children's National Hospital. The strain is resistant to the antibiotics amoxicillin and tetracycline. The genomic DNA for MG-002 was prepared by Lofstrand Labs Limited (Gaithersburg, MD) from a pure culture of *Staphylococcus aureus* strain NRS100 isolate COL as follows. First, a single colony was obtained from the initial culture stab after an overnight incubation at 37 °C on a plate. A single colony was used to inoculate a new plate. One colony from the new plate was grown in 20 mL LB at 37 °C. The culture was used to inoculate 15 x 150 mm plates which were incubated at 37 °C for 16 h. DNA was isolated by lysing the bacteria in lysis solution containing NaCl, Tris, EDTA and lysostaphin (25 µg/mL) and SDS. Proteinase K and RNase A were used to treat protein and RNA. Ammonium acetate was used to remove protein.

[(1)] Certain commercial equipment, instrumentation, or materials are identified in this report to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

DNA was recovered by isopropanol precipitation. DNA was washed with 70 % alcohol, drained, and then dissolved in TE (Tris 10 mM, EDTA 0.1 mM, pH 8.0).

MG-003 is extracted genomic DNA from a clinical *Pseudomonas aeruginosa* (NCBI Biosample SAMN02854574). The strain was selected based on public health relevance and its high percentage of G and C DNA bases. The strain was selected for use by the FDA and isolated from a clinical sample by Children's National Hospital. The genomic DNA for MG-003 was prepared by Lofstrand Labs Limited (Gaithersburg, MD) from a pure culture of *Pseudomonas aeruginosa* strain as follows. A single colony was used to grow a 20 mL overnight culture at 37 °C. A 1 mL aliquot was cultured for an additional 6 h, to reach log phase. Two 1 L cultures were inoculated with 1 mL of the log culture and incubated for approximately 17 h at 37 °C. Centrifugation was used to pellet the two 1 L cultures. DNA was isolated from the pelleted cultures by lysing the bacteria in lysis buffer containing salt, Tris, EDTA, lysozyme and SDS. Protein was treated with Proteinase K, and RNA was treated by RNase A, followed by ammonium acetate precipitation. DNA was precipitated with isopropanol, washed with 70 % alcohol, drained, and then dissolved in TE (Tris 1 mM, EDTA 0.1 mM, pH 8.0).

MG-004 is extracted genomic DNA from *Clostridium sporogenes* strain PA 3679 (NCBI Biosample SAMN02854575). The strain was selected based on public health relevance and its low percentage of G and C DNA bases. The genomic DNA for MG-004 was cultured at FDA CFSAN and prepared by Lofstrand Labs Limited (Gaithersburg, MD) from a pure culture of *C. sporogenes* PA 3679 as follows. *C. sporogenes* was cultured in 2 L of TPGY broth [50 g/L Trypticase (Cat# 211921, BD, Bector Dickinson and Company), 5 g/L Peptone, 20 g/L Yeast extract, and 4 g/L Dextrose, pH adjusted to 7.1 ± 0.2] and Sodium thioglycollate solution (STGS) were freshly prepared and sterilized by autoclaving at 121°C for 15 min. The STGS was added to the tryptone peptone glucose yeast (TPGY) broth after it cooled down to 50 °C at a final concentration of 0.1 %. The 2 L of TPGY broth was inoculated with 1 mL to 2 mL of *C. sporogenes* culture. The media was incubated for 5 d at 35 °C in an anaerobic jar system, containing anaerobic gas generator (AnaeroPack, Thermo Scientific, Inc., Waltham, MA). Cells from the culture media were split into four centrifuge cups, pelleted upon centrifugation at 8000 g for 10 min at 4 °C and transported to Lofstrand Labs for DNA extraction. DNA was isolated from the four pellets by lysing the bacteria in lysis solution containing NaCl, Tris, EDTA and lysostaphin (25 μg/mL) and SDS. Proteinase K and RNase A were used to treat protein and RNA. Ammonium acetate was used to remove protein. DNA was recovered by isopropanol precipitation. DNA was washed with 70 % alcohol, drained, and then dissolved in TE (Tris 10 mM, EDTA 0.1 mM, pH 8.0).

Note: This RM is isolated DNA rather than live cells because cells are less stable and can mutate with each cell division, so that the sequence of live cells may not be stable over time. Extracting DNA from a large batch of cells helps ensure that all vials contain essentially the same sequences of DNA.

Stability: Stability was assessed by measuring the size distribution of DNA with pulsed field gel electrophoresis (PFGE). Using PFGE, no change in the size distribution was detected after storage at 4 °C for eight weeks, but the size distribution decreased significantly when stored at 37 °C for eight weeks. However, because we only measure size distribution, we still recommend storing at –20 °C for long periods of time and limiting freeze-thaw cycles, particularly if the measurement method requires long, undamaged DNA fragments.

Homogeneity: NIST sequenced multiple vials in an experiment designed to assess homogeneity of the samples. No significant differences were detected in terms of proportion of variant or copy number, except for a few in regions known to be susceptible to systematic errors. These results, along with the mixing of DNA before aliquoting, provide confidence that no large differences in small variants or copy number are likely to exist between different vials.

Purity: To assess the genomic purity, each component of RM 8375 was analyzed for the presence of genomic DNA from organisms other than the expected species by performing taxonomic classification of sequencing reads. No sample had more than 2 % of the reads from other species [6].

Size, Concentration, and Volume: Nominal fragment size length and amount of DNA for the RM components are useful for judging whether fragment sizes are appropriate for the application and for sample preparation. The characterization of these properties in this RM was intended to determine if they are fit for this purpose.

The fragment size distribution of DNA is suitable for contemporary short read sequencing methods that use fragments less than 1 kb in length. Long read sequencing may be limited by the size distribution of the molecules, particularly for MG-004, which is substantially smaller than the others. MG-001, MG-002, and MG-003 have a peak in the size distribution higher than a 15 kb fragment in the calibration material supplied by the instrument manufacturer, and MG-004 has a peak in the size distribution near the 15 kb fragment. These size distributions were measured using an automated capillary electrophoresis system, and biases of this method were not characterized.

The nominal concentration of DNA was measured by fluorescence with confirmatory measurements by droplet digital PCR. The median measured concentrations by fluorescence are 45 ng/µL for MG-001, 36 ng/µL for MG-002, 43 ng/µL for MG-003, and 58 ng/µL for MG-004. The median volumes, as measured by pipette, are 61 µL for MG-001, 61 µL for MG-002, 59 µL for MG-003, and 49 µL for MG-004. Biases for these measurements were not extensively characterized, and some measurement methods appeared to have substantial GC-related bias, though methods that appeared to have this bias were not used to assign nominal concentrations. The approximate GC content (fraction of cytosine and guanine bases in the genome) of MG-001 is 52 %, MG-002 is 32 %, MG-003 is 66 %, and MG-004 is 28 %. It is expected that the user will characterize these properties using measurement methods appropriate for use in their application. A small proportion of vials were outside the expected ranges of volume and concentration. If a customer receives vials with volume or a concentration much lower than expected and unsuitable for the customer's application, please email srminfo@nist.gov to report the issue and receive a replacement.

Properties of interest for using this material: Table 1 includes nominal fragment size length, amount of DNA, and GC content for each genome, provided solely for information to determine they are fit for purpose, but uncertainty and bias of these measurements could not be established.

Table 1. Properties of Interest for Using This Material

Genome	Nominal Size Distribution Peak (kb)	Nominal Concentration[(a)] (ng/µL)	Nominal Volume[(b)] (µL)	Nominal GC Content
MG-001	> 15	45	61	52
MG-002	> 15	36	61	32
MG-003	> 15	43	59	66
MG-004	~ 15	58	49	28

[(a)] Nominal concentration is the median value measured by fluorescence, and a small proportion of vials with concentration less than 20 ng/µL were found. Note that some measurement methods have substantial GC-related bias. If a customer receives vials with a concentration less than 20 ng/µL, they will be replaced upon request.

[(b)] Nominal volume is the median value measured by pipette, and a small proportion of vials with volume less than 40 µL were found. If a customer receives vials with volume lower than 40 µL, they will be replaced upon request.

REFERENCES

[1] Olson, N.D.; Zook, J.M.; Samarov, D.V.; Jackson, S.A.; Salit, M.L.; *PEPR: Pipelines for Evaluating Prokaryotic References*; Anal. Bioanal. Chem., Vol. 408, pp. 2975−2083 (2016).

[2] CDC/NIH: *Biosafety in Microbiological and Biomedical Laboratories*, 5th ed.; HHS publication No. (CDC) 21-1112; Chosewood, L.C.; Wilson, D.E.; Eds.; US Government Printing Office: Washington, D.C. (2009); available at https://www.cdc.gov/biosafety/publications/bmbl5/ (accessed May 2018).

[3] Olson, N.D.; Lund, S.P.; Colman, R.E.; Foster, J.T.; Sahl, J.W.; Schupp, J.M.; Keim, P.; Morrow, J.B.; Salit, M.L.; Zook, J.M.; *Best Practices for Evaluating Single Nucleotide Variant Calling Methods for Microbial Genomics*; Front. Gen., Vol. 6, p. 235 (2015).

[4] NCBI; SRA; available at https://www.ncbi.nlm.nih.gov/sra (accessed May 2018).

[5] NIST Micro Genomic RM Data wiki available at https://github.com/usnistgov/NIST_Micro_Genomic_RM_Data/wiki (accessed May 2018).

[6] Olson, N.E.; Zook, J.M.; Morrow, J.B.; Lin, N.J.; *Challenging a Bioinformatic Tool's Ability To Detect Microbial Contaminants Using in silico Whole Genome Sequencing Data*; PeerJ, 5:3729 (2017).

Report Revision History: 08 May 2018 (Revised staff name; editorial changes); 12 March 2018 (Revised estimates for DNA fragment size and amount provided; editorial changes); 08 September 2016 (Original report date).

Users of this RM should ensure that the Report of Investigation in their possession is current. This can be accomplished by contacting the SRM Program: telephone (301) 975-2200; fax (301) 948-3730; e-mail srminfo@nist.gov; or via the Internet at https://www.nist.gov/srm.