National Institute of Standards & Technology

Report of Investigation

Reference Material 8393

Human DNA for Whole-Genome Variant Assessment

(Son of Chinese Ancestry)
(HG-005)

This Reference Material (RM) is intended for validation, optimization, and process evaluation purposes. It consists of a male whole human genome sample of East Asian (Chinese) ancestry from the Personal Genome Project (ID hu91BD69), and it can be used to assess performance of variant calling from genome sequencing. A unit of RM 8393 consists of a vial containing human genomic DNA extracted from a single large growth of human lymphoblastoid cell line GM24631 (labeled as HG-005) from the Coriell Institute for Medical Research (Camden, NJ). The vial contains approximately 10 µg of genomic DNA, and the DNA is in TE buffer (10 mM TRIS, 1 mM EDTA, pH 8.0).

This material is intended for assessing performance of human genome sequencing variant calling by obtaining estimates of true positives, false positives, and false negatives. Sequencing applications could include whole genome sequencing, whole exome sequencing, and more targeted sequencing such as gene panels. This genomic DNA is intended to be analyzed in the same way as any other sample a lab would process and analyze extracted DNA. Because the RM is extracted DNA, it is not useful for assessing pre-analytical steps such as DNA extraction, but it does challenge sequencing library preparation, sequencing machines, and the bioinformatics steps of mapping, alignment, and variant calling. This RM is not intended to assess subsequent bioinformatics steps such as functional or clinical interpretation.

Information Values: Information values are provided for single nucleotide variations (SNVs), small insertions and deletions (indels), and homozygous reference genotypes. The v3.3.2 benchmark set covers approximately 89 % of the GRCh37 assembly and 83 % of the GRCh38 assembly (excluding gaps in the assemblies), using methods described in reference 1. 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 genotypes using the data and methods currently available. These data and genomic characterizations will be maintained over time as new data accrue and measurement and informatics methods become available. Data for HG-005 can be found under BioSample SAMN03283350 in the National Center for Biotechnology Information (NCBI) Sequence Read Archive. The information values are given as a variant call file (vcf) that contains the benchmark SNVs and small indels, as well as a tab-delimited "bed" file that describes the benchmark regions in which any additional variants not in the benchmark vcf should be errors. Information values cannot be used to establish metrological traceability. The files referenced in this report are available at the Genome in a Bottle ftp site hosted by the National Center for Biotechnology Information (NCBI). The Genome in a Bottle ftp site for the benchmark vcf and benchmark regions is:

https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/ChineseTrio/HG005_NA24631_son/latest

Expiration of Value Assignment: RM 8393 is valid, until **23 December 2029**, provided the RM is handled and stored in accordance with instructions given in this report (see "Instructions for Storage and Use"). This material and associated information values are nullified if the RM is damaged, contaminated, or otherwise modified.

Overall direction and coordination of the analyses was performed by J. Zook and M. Salit of the NIST Biosystems and Biomaterials Division.

Sheng Lin Gibson, Chief
Biosystems and Biomaterials Division

Gaithersburg, MD 20899
Report Issue Date: 22 August 2024
Report Revision History on Last Page

Steven J. Choquette, Director
Office of Reference Materials

Maintenance of RM: This report will be updated periodically to reflect important new releases as the benchmark set is 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.

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

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

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

NOTICE AND WARNINGS TO USERS

RM 8393 IS A HUMAN DNA SOURCE MATERIAL. RM 8393 is a Biosafety Level 1 material and should be handled according to applicable federal, state, and/or local regulations and according to policies and procedures of recipient's organization [2].

INSTRUCTIONS FOR STORAGE AND USE

Storage: RM 8393 is stored at –20 °C at NIST but will be shipped with freezer packs and may not arrive frozen. Upon receipt, RM 8393 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 imminent).

Use: It is recommended that after comparing a vcf to the benchmark vcf, only the variants inside the benchmark regions be considered as true positives, false positives, and false negatives. In addition, to understand the causes of false positives and false negatives, including the potential for errors in the benchmark set, it is strongly recommended that the user manually inspect aligned reads around a subset of putative false positive and false negatives using a genome browser. To develop standardized definitions for performance metrics and tools to compare variant calls with different representations, the Global Alliance for Genomics and Health Benchmarking Team published best practices for benchmarking germline small-variant calls in human genomes, which we strongly recommend following [3].

As sequencing technologies and analysis methods improve, these benchmark calls and regions will be updated with refined versions of the files in a different directory, and this Report of Investigation will be updated periodically to reflect important new releases. The current release contains small variant benchmark sets with respect to the GRCh37 and GRCh38 reference assemblies from the Genome Reference Consortium. Datasets from a variety of technologies for this genome are described in reference 4.

SOURCE PREPARATION[(1)]

This individual is the son in a family trio available from Coriell, where the son, father and mother are cell lines GM24631, GM24694, and GM24695, respectively. Although only the son is available as a NIST RM, the parents have been characterized as well [1]. Coriell Institute for Medical Research grew a large batch of their cell line GM24631 to produce approximately 103 mg of total extracted DNA, divided equally into 10 329 vials. To produce this large quantity of DNA, Coriell started with five aliquots of cells from their stock. These aliquots were pooled, cultured, and split into 50 aliquots. One of these aliquots was taken for quality control, and ten of the aliquots were pooled, split into 21 flasks, grown, and combined. A small amount of these cells was saved for potential future sequencing. The combined 21 growths were mixed and the pool was split into 5 roller bottles, which were again grown and combined. A small amount of these cells was also saved for potential future sequencing. Finally, this large pool was mixed and split into 25 roller bottles, which were grown and combined. A small amount of these cells was also saved for potential future sequencing. This final pool of cells was split into 3 pools for DNA extraction, and the extracted DNA was re-pooled and gently mixed at 4° C for >48 hours before automated aliquoting into vials of 10 ug of DNA.

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 may not be stable over time for live cells. Extracting DNA from a large batch of cells

[(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.

helps ensure that all vials contain essentially the same sequences of DNA. DNA is currently available from this same cell line from Coriell, but it may contain small differences in the DNA due to different mutations occurring in different batches of the cells.

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 2 weeks or longer. In addition, for similar human DNA materials, we have found that no change is detected after five freeze-thaw cycles, pipetting vigorously, or vortexing. Because we only measure size distribution, we still recommend storing at –20 °C for long periods of time and limiting freeze-thaw cycles, pipetting, and vortexing, 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.

Size, Concentration, and Volume: Nominal fragment size length and amount of DNA 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 and linked read sequencing may be limited by the size distribution of the molecules. HG-005 has a peak in the size distribution longer than 48.5 kb, as referenced by Lambda DNA. These size distributions were measured using PFGE, and biases of this method were not characterized.

The nominal concentration of DNA was measured by fluorescence. The mean measured concentration for HG-005 was approximately 280 ng/µl. The mean volume, as measured by pipette, was 38 µl for HG-005. Biases for these measurements were not characterized. It is expected that the user will characterize these properties using measurement methods appropriate for use in their application.

REFERENCES

[1] Zook, J.M.; McDaniel, J; Olsen, N.D.; Wagner, J.; Parikh, H.; Heaton, H.; Irvine, S.A.; Trigg, L.; Truty, R.; McLean, C.Y.; De La Vega, F.M.; Xiao, C.; Sherry, S.; Salit, M.; *An Open Resource for Accurately Benchmarking Small Variant and Reference Calls;* Nat. Biotech., Vol. 37, pp. 561-566 (2019) available at https://www.nature.com/articles/s41587-019-0074-6 (accessed Aug 2024).

[2] CDC/NIH: *Biosafety in Microbiological and Biomedical Laboratories,* 6th ed.; HHS publication No. (CDC) 21-1112; Meechan, P.; Potts, J.; Eds.; US Government Printing Office: Washington, D.C. (2020); available at https://www.cdc.gov/labs/bmbl/index.html (accessed Aug 2024).

[3] Krusche, P.; Trigg, L.; Boutros, P.C.; Mason, C.E.; De La Vega, F.M.; Moore, B.L.; Gonzalez-Porta, M.; Eberle, M.A.; Tezak, Z.; L;babidi, S.; Truty, R.; Asimenos, G.; Funke, B.; Fleharty, M.; Chapman, B.A.; Salit, M.; Zook, J.M.; *Best Practices for Benchmarking Germline Small-variant Calls in Human Genomes*; Nat. Biotechnol. Vol. 37, pp. 555-560 (2019).

[4] Zook, J.M; Catoe, D., McDaniel, J; Vang, L; Spies, N.; Sidow, A.; Weng, Z.; Liu, Y.; Mason, C.E.; Alexander, N.; Henaff, E.; McIntyre, A.B.R.; Chandramohan, D.; Chen, F.; Jaeger, E.; Moshrefi, A.; Pham, K.; Stedman, W.; Liang, T.; Saghbini, M.; Dzakula, Z.; Hastie, A.; Cao, H.; Deikus, G.; Schadt, E.; Sebra, R.; Bashir, A.; Truty, R.M.; Chang, C.C.; Gulbahce, N.; Zhao, K.; Ghosh, S.; Hyland, F.; Fu, Y.; Chaisson, M.; Xiao, C.; Trow, J.; Sherry, S.T.; Zaranek, A.W.; Ball, M.; Bobe, J.; Estep, P.; Church, G.M.; Marks, P.; Kyriazopoulou-Panagiotopoulou, S.; Zheng, G.X.Y.; Schnall-Levin, M.; Ordonez, H.S.; Mudivarti, P.A.; Giorda, K.; Sheng, Y.; Rypdal, K.B.; Salit, M.; *Extensive Sequencing of Seven Human Genomes to Characterize Benchmark Reference Materials;* Sci. Data 3, 160025 (2016); available at https://www.nature.com/articles/sdata201625 (accessed Aug 2024).

Report Revision History: 22 August 2024 (Change of expiration date; editorial changes); 01 October 2019 (Updated title; updated released benchmark set information; editorial changes); 26 October 2016 (Corrections to cell line number and ftp site address; editorial changes); 08 September 2016 (Original certificate 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; e-mail srminfo@nist.gov; or via the Internet at https://www.nist.gov/srm.