



National Institute of Standards & Technology

Report of Investigation

Reference Material 8392

Human DNA for Whole-Genome Variant Assessment
(Family Trio of Eastern European Ashkenazim Jewish Ancestry)
(HG-002, HG-003, HG-004)

This Reference Material (RM) is intended for validation, optimization, and process evaluation purposes. It consists three whole human genome samples from a son-father-mother family trio of Eastern European Ashkenazim Jewish ancestry, and it can be used to assess performance of variant calling from genome sequencing. A unit of RM 8392 consists of three vials containing human genomic DNA from a specific family member (son, father, mother); extracted from three large growths of human lymphoblastoid cell lines (GM24385, GM24149, and GM24143, respectively) from the Coriell Institute for Medical Research (Camden, NJ). Each vial contains approximately 10 µg of genomic DNA; with the peak of the nominal length distribution longer than 48.5 kb, as referenced by Lambda DNA; 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, true negatives, 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 polymorphisms (SNPs), small insertions and deletions (indels), and homozygous reference genotypes for approximately 88 % of the genome, using methods similar to 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. The information values are given as a variant call file (vcf) that contains the high-confidence SNPs and small indels, as well as a tab-delimited “bed” file that describes the regions that are called high-confidence. Information values cannot be used to establish metrological traceability. The files referenced in this Report of Investigation 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 high-confidence vcf and high confidence regions is:

<ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio>

Expiration of Value Assignment: RM 8392 is valid, until **23 December 2024**, 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.

Maintenance of RM: This report will be updated periodically to reflect important new releases as the high-confidence calls and regions 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 J. Zook and M. Salit of the NIST Biosystems and Biomaterials Division.

Anne L. Plant, Chief
Biosystems and Biomaterials Division

Gaithersburg, MD 20899
Report Issue Date: 08 September 2016

Steven J. Choquette, Director
Office of Reference Materials

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

Technical measurements were conducted by L. Vang, J. McDaniel, and D. Catoe of the NIST Biosystems and Biomaterials Division. Analyses were conducted by J. Zook, 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 8392 is from a human lymphoblastoid cell line and is intended for research use. Since there is no consensus on the infectious status of extracted DNA, handle RM 8392 components as Biosafety Level 1 material potentially 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 8392 is stored at $-20\text{ }^{\circ}\text{C}$ at NIST but will be shipped with freezer packs and may not arrive frozen. Upon receipt, RM 8392 should be kept in the dark at $-20\text{ }^{\circ}\text{C}$ for long-term storage, or in the dark at $4\text{ }^{\circ}\text{C}$ for short-term storage (if use is imminent).

Use: It is recommended that after comparing a vcf to the high-confidence vcf, only the variants inside the high-confidence regions be considered as true positives, false positives, and false negatives. In addition, due to challenges in comparing variant calls (e.g., around complex variants with different representations) and the potential for errors in the high-confidence calls, 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 [3]. To address challenges in defining performance metrics and comparing variant calls with different representations, the Global Alliance for Genomic Health formed a Benchmarking Team. This working group has been developing standardized definitions of performance metrics and methods for assessing accuracy of variant calls for this RM and other well-characterized genomes. These definitions and tools are described in reference 4.

As sequencing technologies and analysis methods improve, these high-confidence 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 variants with respect to the GRCh37 reference assembly. These calls and future callsets use a variety of datasets described in reference 5. It is important to recognize that there is currently no standardization of definitions for true positive, false positive, true negative, and false negative. For example, genotyping errors can be counted as true positives, false positives, or false negatives, and no-calls can be treated as uncertain or homozygous reference regions.

SOURCE PREPARATION⁽¹⁾

The DNA for the son in NIST RM 8392 is from the same growth as the DNA in RM 8391 Human DNA for Whole-Genome Variant Assessment (Son of Eastern European Ashkenazim Jewish Ancestry), which contains only the son. Coriell Institute For Medical Research grew a large batch of their cell lines GM24385, GM24149, and GM24143 to produce approximately 107 mg, 35 mg, and 30 mg of total extracted DNA, divided equally into 10 728, 3475, and 3037 vials for the son, father, and mother, respectively. 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 $4\text{ }^{\circ}\text{C}$ for >48 hours before automated aliquoting into vials of 10 μg of DNA.

⁽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.

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 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.

REFERENCES

- [1] Zook, J.M.; Chapman, B.; Wang, J.; Mittelman, D.; Hofmann, O.; Hide, W.; Salit, M.L.; *Integrating Human Sequence Data Sets Provides a Resource of Benchmark SNP and Indel Genotype Calls*; Nature Biotechnology Vol. 32, pp. 246–251 (2014).
- [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 <http://www.cdc.gov/biosafety/publications/bmb15/> (accessed Sep 2016).
- [3] To facilitate viewing aligned reads for this and other RMs, the *GeT-RM* genome browser is an online tool developed by the NCBI for the CDC and NIST whole genome RM projects. To address challenges in defining standard performance metrics and variant comparison methods, the Global Alliance for Genomic Health formed a Benchmarking Team in July 2014. This Team is developing standardized performance metrics and methods for assessing accuracy of variant calls for this RM and other well-characterized genomes. The *GeT-RM* genome browser is available at <http://www.ncbi.nlm.nih.gov/variation/tools/get-rm/> (accessed Sep 2016).
- [4] Global Alliance for Genomic Health Benchmarking Team; *Benchmarking Tools and Standards*; available at <https://github.com/ga4gh/benchmarking-tools> (accessed Sep 2016).
- [5] 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 <http://www.nature.com/articles/sdata201625> (accessed Sep 2016).

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 <http://www.nist.gov/srm>.