



Report of Investigation

Reference Material 8398

Human DNA for Whole-Genome Variant Assessment

(Daughter of Utah/European Ancestry)

This Reference Material (RM) is intended to provide a whole human genome sample and accompanying reference values to assess performance of variant calling from genome sequencing. This RM contains human genomic deoxyribonucleic acid (DNA) extracted from a large growth of the human lymphoblastoid cell line GM12878. A unit of RM 8398 consists of a single vial containing approximately 10 μg of genomic DNA, with the peak of the nominal length distribution longer than 48.5 kb, as referenced by Lambda DNA, and the DNA is in TE buffer (10 mM TRIS, pH 8.0, 1 mM EDTA, pH 8.0).

This material is intended for assessing performance of human genome sequencing, including whole genome sequencing, whole exome sequencing, and more targeted sequencing such as gene panels. Specifically, the material can be used to obtain estimates of true positives, false positives, true negatives, and false negatives for variant calls. This genomic DNA is to be analyzed as any other processed, extracted DNA. Because the RM is extracted DNA, it is not useful for assessing pre-analytical steps such as DNA extraction, but it does assess 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.

Reference Values: Reference values are provided for single nucleotide polymorphisms (SNPs), small indels (insertions and deletions), and homozygous reference genotypes for approximately 77 % of the genome [1]. This report contains variants with respect to the GRCh37 reference assembly. Reference values are noncertified values that are the best estimate of the true value; however, the values do not meet the NIST criteria for certification for which all biases be sufficiently understood. The reference 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. 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 initial high-confidence vcf and high-confidence regions is:

- ftp://ftp-trace.ncbi.nih.gov/giab/ftp/release/NA12878_HG001/NISTv2.18

As sequencing technologies and analysis methods improve, these high-confidence calls and regions will updated with refined versions of the files in a different directory at the Genome in a Bottle ftp site:

- ftp://ftp-trace.ncbi.nih.gov/giab/ftp/release/NA12878_HG001/latest

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 (see “Instructions for Storage and Use”).

Expiration of Value Assignment: RM 8398 is valid, within the measurement uncertainty specified, 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 report is 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.

Anne L. Plant, Chief
Biosystems and Biomaterials 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 L. Vang.

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.

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.

NOTICE AND WARNINGS TO USERS

RM 8398 is from a human lymphoblastoid cell line and is intended for research use. Since there is no consensus on the infections status of extracted DNA, handle RM 8398 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 8398 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 8398 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 only the variants in the test call set inside the high-confidence regions be compared to the high-confidence calls. Due to challenges in comparing variant calls (e.g., around complex variants) 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].

SOURCE PREPARATION⁽¹⁾

Coriell Institute for Medical Research (Camden, NJ) grew a large growth of their cell line GM12878 in multiple stages, produced approximately 83 mg of extracted DNA, and then mixed the DNA and aliquoted it into vials, with the DNA divided approximately equally into vials. Specifically, the pool of cells was split into three separate volumes for DNA extraction, and the extracted DNA was re-pooled and gently mixed at $4\text{ }^{\circ}\text{C}$ for greater than 48 h before the material was aliquoted automatically into vials of 10 μg 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 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. DNA is currently available from this same cell line from Coriell with the number NA12878, but it may contain small differences in the DNA sequence 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 $4\text{ }^{\circ}\text{C}$ for eight weeks, but the size distribution decreased significantly when stored at $37\text{ }^{\circ}\text{C}$ for 8 weeks. In addition, no change was detected after five freeze-thaw cycles, pipetting vigorously, or vortexing. However, because we only measure size distribution, we still recommend storing at $-20\text{ }^{\circ}\text{C}$ 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.

⁽¹⁾ Certain commercial equipment, instrumentation, or materials are identified in this certificate 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.

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 Nov 2015).
- [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 Nov 2015).

Report Revision History: 24 November 2015 (editorial changes); 15 April 2015 (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; fax (301) 948-3730; e-mail srminfo@nist.gov; or via the Internet at <http://www.nist.gov/srm>.