Genomic Reference Materials for Clinical Application

Justin Zook* and Marc Salit
Biosystems and Biomaterials Division, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899
*jzook@nist.gov; Phone: 301-975-4133

Abstract

Reference Materials are well-characterized, homogeneous, and stable samples that can be used to understand measurement performance. The Genome in a Bottle Consortium is developing whole human genome DNA Reference Materials from large batches of DNA extracted from cell lines to support clinical translation of whole human genome sequencing. This DNA will be characterized using multiple sequencing and bioinformatics methods for SNPs, indels, structural variants, and haplotype phasing across the whole genome. Characterization of the Reference Material will be done with methods being developed by The Consortium to integrate information from multiple datasets to form highly confident genotype calls. These highly confident genotype calls can then be used by clinical and research laboratories to understand and optimize performance of library preparation, sequencing, and bioinformatics methods for genome sequencing, and can be used by accreditation and regulatory bodies to evaluate performance. Because the Reference Materials are homogeneous and stable, they will be able to be used to assess and compare sequencing performance with different methods over time, even as sequencing technologies and bioinformatics methods rapidly improve.

Keywords: Reference Materials, whole genome sequencing, DNA sequencing, bioinformatics, accuracy, homogeneity, stability, Next Generation Sequencing, SRM

Outline:

Introduction:

NIST has recently begun a program to develop whole human genome reference materials. *Reference Materials* (RMs) are frequently used in clinical laboratories to understand accuracy or calibrate instruments. An RM is defined as a "Material, sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process" [1]. Because RMs are homogeneous and stable, they can be used to compare performance in different laboratories at different times. *Standard Reference Materials* ® (SRMs), which are *Certified RMs* produced by the National Institute of Standards and Technology (NIST), are "NIST RMs characterized … for one or more specified properties, accompanied by a certificate that provides the value of the specified property, its associated uncertainty, and a statement of metrological traceability" [1] SRMs, in addition to being homogeneous and stable, have been characterized for certified values, which are values "for which NIST has the highest confidence in its accuracy in that all known or suspected sources of bias have been fully investigated or accounted for by NIST" [1]. SRMs are tools that can be used by any laboratory to benchmark their results against those in which NIST has highest confidence, letting that laboratory establish its performance.

Challenges in Developing a Whole Genome Reference Material

Most SRMs are characterized for a single certified property, usually a quantitative one like the mass fraction of cholesterol (NIST SRM 911c), length of a DNA short tandem repeat (NIST SRM 2399), or DNA concentration (NIST SRM 2366 for cytomegalovirus). SRMs are typically used for calibration of a measurement result to establish metrological traceability to the properties of the SRM or for evaluation of bias by comparison of measured results to certified properties. Metrological traceability lets one compare measurement results across space and time, by referring all results to be compared to a common reference. Evaluating bias with a reference material helps establish validity (providing evidence that "I've measured what I set out to measure."), and understanding bias is critical for meaningful comparison of results – "I am 95% confident that this value is higher/lower/different than/the same as that value.".

Usually only one to a few tens of values are characterized in an RM/SRM, even in complex matrices such as blood serum (e.g., NIST SRM 955c). In contrast, the human genome has billions of properties to be characterized, specifically the genotype at every position in the genome.

Also, these genotype calls are "nominal properties," or values for which no algebraic manipulation is sensible. The development of nominal property RM/SRMs is an immature branch of metrology, and its development is being driven by the application of metrology to biological measurements. Concepts analogous to metrological traceability, measurement uncertainty, or validation are yet to be established in wide practice. In addition, biases in whole genome sequencing measurements are only partially understood. For these reasons, creating a SRM certified for whole genome sequence is a daunting task, unprecedented in the scale and type of measurements.

In consideration of these challenges, NIST will likely release samples of a whole human genome as an RM characterized for homogeneity and stability, but without certified values. A set of "Information values" for SNP and indel genotypes will be released with the RM, but "*all known or suspected sources of bias*" may not be "*fully investigated or accounted for by NIST*". Alternatively, NIST may release the genomic DNA as an SRM with certified values for SNP and/or indel genotypes in well-understood regions of the genome, and information values for the rest of the genome. As additional measurements are made on the RM/SRM, with new technologies and maturing bioinformatics methods, certified values may be added for additional types of variants and more difficult regions of the genome. We expect that the utility of a stable and homogeneous RM/SRM will improve over time as read lengths increase, errors diminish and are better understood, bioinformatics methods improve, and sequencing costs decrease.

Clinical translation of human genome sequencing calls for well-documented, standard measures of sequencing performance. Homogeneous and stable RMs/SRMs will help make this possible, enabling regulatory oversight by the Food and Drug Administration (FDA) and laboratory accreditation by the College of American Pathologists (CAP) and Clinical Laboratory Improvement Amendments (CLIA) in the United States, or by similar agencies in other countries.

The first RMs NIST plans to develop for these applications will be extracted genomic DNA. The stability of DNA can be assured better than cells (live or fixed), and cells can be measured from a variety of tissues stored in different forms (e.g., frozen vs. formalin-fixed paraffin-embedded (FFPE)). These RMs will be limited in scope to the parts of the generic sequencing process highlighted in Fig. 1. For current sequencing processes, the scope includes library preparation, sequencing, mapping, and variant calling, but does not include pre-analytical steps such as DNA extraction, or clinical interpretation of the variants.

Genome in a Bottle Consortium

The NIST-hosted Genome in a Bottle Consortium was formed to develop the reference materials, reference methods, and reference data needed to enable clinical translation and regulatory oversight of human genome sequencing. NIST organized multiple invitational meetings in 2011 and 2012 to gauge interest in establishing a consortium. The first large public meeting was held at NIST on August 16-17, 2012, with ~100 attendees from government, private companies, academic sequencing centers, and clinical laboratories. Four working groups were formed at this meeting: (1) Reference Material Selection and Design, (2) Reference Material Characterization, (3) Bioinformatics, Data Integration, and Data Representation, and (4) Performance Metrics and Figures of Merit.

Reference Material Selection and Design

The RM Selection and Design Working Group is tasked with selecting genomic DNA for RMs and designing synthetic DNA constructs for RMs. The working group extensively explored a variety of perspectives on the appropriate consent for a genomic RM. The discussion particularly focused on whether older consents, such as the HapMap consent for the highly characterized sample NA12878 [2], are appropriate for a NIST RM. The HapMap consent acknowledged that re-identification may be possible, but the risk was thought to be small at that time, though it also stated that risks may change. The NA12878 sample had been previously characterized extensively by numerous academic studies and is frequently used as a de facto RM by many private companies and clinical laboratories. Therefore, it is ideal for developing bioinformatics methods that can be applied to other RMs, and the consortium currently plans to use it as a pilot RM. NIST received >8300 10ug units of DNA from NA12878 in April 2013, which is candidate NIST RM 8398/SRM 2398. Future RMs will be developed from father-mother-child trios in the Personal Genome Project (PGP) [3]. The PGP genomes have a broad open consent, including consent for re-identification and broad commercial use such as redistribution of derived products from the cell lines.

The working group also discussed potential sources of DNA for RMs, and decided that Epstein Barr Virus-immortalized lymphocyte cell lines were the best option because they can be easily renewed. Mutations can occur in cell lines, so the RMs will be extracted DNA from large homogenized growths of cells. This homogenized DNA may have some de novo or low frequency mutations particular to the batch, but each vial of the RM is expected to be essentially the same. With the consortium, NIST will characterize the homogeneity within and between vials, as well as the stability of the DNA over time. Immortalized cell lines may have some differences from DNA in blood or other tissues, but these differences will be characterized and are expected to be sufficiently small that they should be a reasonable surrogate for assessing performance of sequencing other tissues.

Synthetic DNA constructs are also being discussed as possible RMs to help understand performance. We recently used the NIST SRM 2374 DNA plasmids to analyze and recalibrate base quality scores [4], which was more accurate than recalibrating using the genome. The GIAB Consortium is discussing additional synthetic DNA constructs that could be used to assess DNA sequencing and bioinformatics. Some consortium members have designed DNA plasmids that include known cancer-associated mutations, with a short sequence barcode near the mutation so that the DNA can be spiked-in to any sample in any given ratio. The Consortium has also discussed designed pairs of synthetic DNA sequences that would be modeled after the types of variants and sequence contexts found in the genome, but would have sequence content that is different from any known genome so they could be added to any sequencing experiment. These constructs could allow testing of particular sequencing problems, such as complex variants, structural variants, homopolymers, tandem repeats, and copy number variants.

Reference Material Characterization
After selecting which genomes or synthetic DNA constructs to use as RMs, the materials need to be characterized. Testing homogeneity and stability helps ensure that measurements made on different vials at different times are all measuring essentially the same DNA.

For genomic DNA RMs intended to assess sequencing performance, homogeneous means that each vial should have a sufficiently similar mixture of sequences. Since DNA mutations can occur during cell line propagation, the genomes in the cell culture may not be completely homogeneous, and there can be differences between genomes in different expansions of the same cell line [5, 6]. Therefore, the Consortium proposed that NIST purchase a large batch (e.g., ~80 mg of DNA) from a well-mixed combination of expansions of cells for each whole genome RM. An individual unit of the RM may have a mixture of genomes due to mutations occurring during cell line expansions, but each vial should have approximately the same mixture of genomes because the cells and DNA were well-mixed. The homogeneity between vials will be characterized during RM development to determine if there are any detectable differences in allele frequency or copy number variants between vials. The ability to discriminate small differences in allele frequency, and the value in a homogeneous RM, careful attention was paid to mixing the DNA prior to aliquoting, while taking care to avoid shearing the DNA.

Experience suggests that the DNA will be stable when stored frozen, but also that it may become more fragmented when exposed to freeze-thaw cycles or room-temperature storage. Fragmentation may be a secondary consideration for current short read technologies (so long as it is random), but it may have deleterious effect on results from future longer read technologies. Therefore, the stability of the DNA will be tested in a variety of conditions, including after freeze-thaw cycles, stored frozen, and stored at higher temperatures.

In addition to homogeneity and stability, the RMs will be characterized for their sequence so that labs can understand their performance. Since every characterization method has strengths and weaknesses, multiple sequencing technologies, library preparation methods, and other DNA characterization methods will be combined to provide the best, comprehensive, results. Currently planned sequencing methods include Illumina, SOLiD, Ion Torrent Proton, Pacific Biosciences, Complete Genomics, and 454, as well as emerging sequencing technologies such as nanopore sequencing. Library preparation methods will likely include short paired-end, longer mate-pair/paired-end, fosmid sequencing, and limited dilution methods such as those described by Moleculo, Tile-seq [7], Complete Genomics Long Fragment Read [8], and chromosome sorting [9]. Other characterization methods may include genotyping microarrays, array CGH, and optical and nanopore-based mapping techniques. Selected SNP and indel sites may be confirmed by Sanger sequencing, high-depth next generation sequencing, and manual curation of alignments. Structural variants may be confirmed by microarrays, PCR, and mapping technologies.

In addition, the GiaB Consortium decided that sequencing of family members is an important way to understand accuracy and characterize phasing of variants. Mendelian inheritance can be used to identify sequencing errors, particularly when larger pedigrees are used. Haplotype phasing (i.e., identifying whether heterozygous variants fall on the same chromosome or opposite copies of the chromosome) can be achieved through long-read technologies, limited dilution methods, fosmid sequencing, or inheritance patterns, and the consortium plans to use a combination of these methods.

Bioinformatics, Data Integration, and Data Representation

After the experimental characterization of the RMs is performed, the data will be analyzed, integrated, and represented in a useful format. Many bioinformatics methods have been developed to map, re-align, perform local *de novo* assembly, call variants/genotypes, and filter potential false positive variants. For most variant callers, an important first step is mapping reads to the proper location in the reference genome and locally aligning the bases in the read to the bases in the reference genome. Alternatively, some methods have recently been developed to perform global *de novo* assembly of the reads and then call variants with respect to the reference genome. While mapping-only methods are more mature and robust, global and local *de novo* assembly methods can detect larger variants that are difficult to detect with mapping-only techniques, so it will be important to incorporate both types of methods in the characterization of the RMs. In addition, different bioinformatics algorithms are used for small variants (e.g., SNPs and small indels) vs. larger structural variants like copy number variants, inversions, and rearrangements.

To capture the individual strengths of the different methods (including library preparations, sequencing technologies, mapping/de novo assembly algorithms, and variant calling methods) and provide robustness to their deficits, an integration approach will be established to create NIST's well-characterized RMs. Data can be integrated in multiple ways. Simple voting or "majority rules" methods are easiest to implement and understand, but systematic errors shared across multiple methods can cause the majority of methods to be incorrect. Voting methods can also be biased if one type of sequencing or analysis method is overrepresented. Therefore, we have developed methods that *arbitrate* between genotype calls using information about biases in each dataset. In our arbitration method, if datasets have discordant genotypes at a particular position, we down-weight datasets that have evidence of bias at that position. Evidence of bias includes technical characteristics such as atypical mapping quality, base quality, strand bias, distance from the end of the read (i.e., "soft clipping"), high coverage, variant quality divided by coverage, and other characteristics that are associated with systematic sequencing errors, local alignment errors, and global mapping errors.

As noted in the introduction, an estimate of uncertainty for nominal properties is one for which no widely accepted best practice has been established. For nominal properties, internationally accepted documentary standards allow for an estimate of probability of correctness to be used in place of a quantitative estimate of uncertainty. Several approaches have been proposed to estimate uncertainty for diploid genomes, including expression of uncertainty as the probability of a genotype being incorrect, genotype likelihoods [10, 11, 12], or genotype likelihood ratios [13]. Generally, genotype likelihoods for SNPs and indels are calculated from the pileup of reads at each genomic position, using a Bayesian statistical model that assumes a binomial distribution with a sequencing error rate equal to the quality score of each base. Unfortunately, these models do not currently account well for many systematic sequencing errors, global mapping errors, and local alignment errors. Therefore, genotype likelihoods frequently underestimate uncertainty, particularly with high-depth sequencing.

A better informed estimate will use variety of annotations have been developed to identify potential systematic errors, such as strand bias, base quality score, mapping quality score, and soft-clipping of reads. These annotations can be used in a framework such as GATK's Variant Quality Score Recalibration (VQSR), which identifies variant sites with unusual characteristics. VQSR can potentially be used both to arbitrate between datasets where they have discordant genotypes and to identify sites with lower confidence. In an integrated approach such as the one we propose to use to for our RM characterization, it isn't currently possible to assign accurate quantitative probabilistic uncertainties. Therefore, we currently plan to use qualitative categories of uncertainty for the RM based on genotype likelihoods and characteristics of bias.

A benchmark set of SNP, small indel, and homozygous reference genotypes was recently published [**Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls; Nature Biotechnology 2014;32:246**], and clinical and research laboratories have started to use these high-confidence genotypes to assess genotype accuracy. These genotypes were generated by integrating 14 datasets from five sequencing technologies. Sites with discordant genotypes were arbitrated using VQSR as described above. In addition, regions with known biases in all current sequencing technologies were excluded, including segmental duplications, reported structural variants, and long repeats. The resulting high-confidence genotype calls cover over 77 % of the genome. As sequencing and bioinformatics methods improve, these high-confidence calls will extend to more difficult variants and parts of the genome. In addition, methods have been developed by Genome in a Bottle participants to phase the 11 children of NA12878 and her husband and determine whether the variants are inherited properly as expected from the phased haplotypes. Correctly inherited variants can give more confidence to genotype calls because certain types of systematic errors violate inheritance. Current work includes developing methods to integrate phased inherited variant calls, as well as methods to use long read technologies (e.g., PacBio, BioNano Genomics) and assembled pseudo-long read methods developed by Illumina and Complete Genomics, and other methods to call structural variants. These improved high-confidence genotypes for the pilot Reference Material and for future Reference Materials will be made publicly available through the Genome in a Bottle Consortium.

Data representation

The characterization of the genomic RM could be represented in different formats. Because the characterization will be used to assess false positive variant calls, it is essential that confident homozygous reference locations be distinguished from uncertain locations, which is not typically done in most variant file formats (e.g. VCF). However, recently a new file format called gVCF was proposed to extend VCF to specify regions with homozygous reference calls. Alternatively, standard VCF can be used along with a bed file that specifies genomic regions in which confident genotype calls can be made. Phasing information and structural variants will also need to be represented, which is sometimes difficult in VCF. Some variants can be represented correctly in multiple ways with respect to the reference assembly (see Fig. 2), so standardized ways to represent these variants, or methods to compare different representations of variants is important. To address many of these problems, the RM characterization could be represented as an assembly graph, ideally as paternal and maternal contigs for each pair of chromosomes, since even parental origin of a haplotype can affect function [14]. To assess variant calling in an experiment, these contigs could be mapped to the reference assembly (e.g., GRCh37 or hg19) to determine variants.

Performance Metrics and Figures of Merit

Perhaps the key application of genomic RMs is to understand performance of the sequencing process, including library preparation, sequencing, and bioinformatics (mapping and variant calling), as depicted in Fig. 1. Because the RM is characterized for homogeneity and stability, the RM provides a constant benchmark that can be used to compare performance of different methods, including new methods developed in the future.

The Genome in a Bottle Performance Metrics and Figures of Merit Working Group is tasked with developing a framework for assessing performance of a sequencing process. This framework would allow any laboratory that has sequenced the RM to compare their variants, genotype calls, and/or assembly to the consensus characterization of the RM. Regulatory and accreditation bodies can use standard methods of performance assessment and reporting to establish a consistent enterprise-scale way to compare performance and make confident decisions. Laboratories could refine and optimize their protocols and procedures and learn about the different types of biases and errors affecting their results.

Assessing performance of genome sequencing poses a variety of challenges:

1. Sensitivity, specificity, false positive, and false negative rates for variant calls are oft-used measures to specify performance, where “positives” typically refer to any type of variant and “negatives” refer to homozygous reference. These two categories over-simplify performance assessment, since at least three possible genotypes exist at any genome position (homozygous reference, heterozygous, and homozygous variant). At sites with more than one possible alternate allele, even more than three possible genotypes exist. Therefore, genotype comparison tables in which genotype calls from two methods are compared give a more comprehensive description of different types of genotyping error rates.

2. In most current clinical genetic tests, samples with the mutation(s) of interest are used as “positives” and samples without the mutations are “negatives”. In this way, laboratories can measure accuracy for each mutation. This paradigm becomes untenable for whole genome, whole exome, and even multi-gene panels because it is not possible to have RMs with every possible variant that might be seen in clinical samples. Fortunately, since a single sample or a few samples will generally have many examples of most variant types, it becomes possible to test sequencing performance for different classes of variants with a limited number of samples. However, dividing variants into different classes is not trivial, since sequencing accuracy can be affected by variant type, sequence context, and genome region in complex ways.

3. Current sequencing technologies and bioinformatics methods have different accuracies for different variant types (e.g., SNPs, indels, CNVs, rearrangements). For example, SNPs tend to be easier to detect than indels and CNVs, and bioinformatics methods are more mature for SNPs. Complex variants (clustered SNPs and indels) and moderately long indels (~10-100 nucleotides) can be particularly difficult to detect with current mappers and variant callers, though new local and global *de novo* assembly methods often help.

4. Some sequence contexts are particularly difficult for current sequencing technologies, particularly repeat sequences (e.g., homopolymers and tandem repeats) that are longer than the sequencing read length. Certain sequence contexts can also cause systematic sequencing errors for different platforms (e.g., homopolymers for 454, Ion Torrent, and Sanger capillary sequencing, or GGT motifs in Illumina).

5. Some regions of the genome are more difficult. Regions with high or low GC content often have low or no coverage due to PCR bias when sequencing with NGS. In addition, a small fraction of the human reference assembly is not finished, so that reads cannot be mapped to it. The functionally important HLA region requires specialized bioinformatics methods due to its high sequence diversity. Centromeres and telomeres have low sequence diversity,

> which makes them difficult to sequence and map. Large tandem duplications, mobile element insertions, pseudogenes, and other regions of the genome with high homology also cause significant problems for most current sequencing technologies. For these regions, it is often impossible to determine from which copy a particular sequencing read originates. It is important to identify these duplicated regions in the reference assembly as well as duplications in the reference material sample that differ from the reference assembly. Duplicated regions in the reference assembly can be identified from low mapping quality of reads, but duplicated regions in the sample of interest require specialized methods for copy number variant analysis.

Because accuracy can vary by variant type, sequence context, and genomic region, overall performance assessment may change as characterization of the RM improves. As more difficult variants, sequence contexts, and genomic regions can be characterized and included in performance assessment, the overall accuracy of a particular method will likely decrease when assessed against the RM characterization. To avoid accuracy changing as the RM characterization improves, the genome and variants could be divided into different regions and types of variants. However, the genome and variants could be divided in numerous ways, and some divisions could depend on sequencing platform and library preparation. For example, longer reads can resolve longer repeat regions, long mate-pair can help resolve duplications, and some platforms have higher error rates for homopolymers or other specific sequence motifs.

Reference Data

In addition to distributing the physical genomic DNA as a NIST Reference Material, data collected for the RM will be made available as Reference Data. These data will likely include mapped and unmapped sequence reads (e.g., bam files), and genotype and variant calls across the whole genome. In addition, these data may be visualized through a genome browser to view alignments and variants in a particular region (e.g., the browser being developed by NCBI for the Genetic Testing Reference Materials project GeT-RM). A lab sequencing and analyzing the RM could look in this browser at any locations at which they differ from the integrated consensus genotype calls to help determine why their answer differs.

The Reference Data can also be used to help understand performance of bioinformatics pipelines. Typically, bioinformatics pipelines are assessed using synthetic "*in silico*" generated reads. Synthetic reads are used so that the truth about the location of the reads and variants in the genome are known. Unfortunately, synthetic read generators do not model all systematic error processes that occur during sequencing, so they generally overestimate performance of the bioinformatics programs. Nevertheless, synthetic reads can be useful for understanding errors, particularly in mapping and alignment. Alternatively, the genome reference assembly to which the reads are mapped can be altered in strategic ways. For example, variants can be introduced that are not in the genome being sequenced, and the ability to detect these variants can be assessed. For microbial genomes, the reads from one strain can be mapped to the genome reference assembly from a related strain, and the variants found can be compared to the variants between the two strains [15]. While changing the genome reference assembly can be useful for assessing variant detection by microbial bioinformatics software, it can only assess detection of homozygous variants, so it is less useful for diploid organisms like humans.

Reference Data from the RMs can provide a useful way to assess bioinformatics pipelines with real human sequence data from a well-characterized genome. These Reference Data could include datasets from multiple sequencing platforms, so that bioinformatics pipelines could be tested with multiple datasets from different platforms or versions of library preparation and sequencing chemistry. Because the RMs will be well-characterized, the results of the bioinformatics analyses of the Reference Data could be compared to the integrated consensus variant calls to assess accuracy. Using the Reference Data as a benchmark, the effect of changing parameters in the bioinformatics software could be analyzed. A challenge of assessing performance of bioinformatics pipelines with Reference Data is that the sequencing platforms and library preparation methods are changing rapidly, which can strongly interact with the bioinformatics results. An advantage of having a homogeneous and stable RM is that Reference Data for this RM will continue to be accumulated for new versions of sequencing methods, as users of the RM choose to deposit their data in public databases.

Other Reference Materials for Genome-scale Measurements

Microbial genome RMs

NIST is also working with the FDA to develop whole genome microbial DNA RMs (and/or SRMs) similar to the human DNA RMs the Genome in a Bottle Consortium is developing. These whole genome RMs will be DNA extracted from large-scale cultures of several bacterial organisms, across a range of GC content (to enable evaluation of sequencing platform performance for low- and high-GC content genomes, a challenge to some current platforms). Similar to human RMs, a significant value of these RMs will be homogeneity and stability; each RM vial will contain a sample of the same DNA, so it will not be subject to changes over time due to mutations that occur during growth of the organisms. The DNA will also be characterized on a whole genome scale with multiple methods, with the expectation of a highly confident *de novo* assembly for the particular genome(or genomes – despite the care taken in preparing the samples from a clonal population, these single strain samples might in fact contain mixtures of genomes arising from mutation on culture) contained in the vials. These RMs could then be used to understand performance of sequencing instruments and bioinformatics pipelines used for microbial sequencing.

Because bacterial genomes are haploid and do not have heterozygous variants, the genome reference assembly to which reads are mapped can indeed be changed to understand performance of bioinformatics pipelines. If the reads from the RM are mapped to the genome reference assembly generated from the RM, no variants should be detected. While this could help understand certain types of errors, mapping to a genome reference assembly different from the RM assembly is a more realistic test of how bioinformatics software is typically used. Therefore, the genome reference assembly generated from the RM could either be modified with variants, or the reads from the RM could be mapped to the genome reference assembly from a related strain or species that has known differences from the RM [16]. By sequencing the RM and mapping to different genome reference assemblies, multiple steps in the sequencing process could be systematically tested, including library preparation, sequencing, mapping/alignment, and variant calling.

As with the human genomic DNA samples, these microbial DNA RMs will not test pre-analytical steps such as DNA extraction, so they would not be useful for understanding the effect of (differential) DNA extraction on quantitation (such as in metagenomic studies). In addition, these RMs will be from known strains of only a few species, so they will not comprehensively assess the ability of laboratories to assign identity to an unknown microbial sample. However, they will provide a way to understand performance of sequencing and bioinformatics, including random and systematic errors introduced by these methods.

Gene expression RMs

NIST has recently released SRM 2374 – “DNA Sequence Library for External RNA Controls” as an RM to support confidence in genome-scale gene expression measurements. This reference material is intended to be used as a library of templates from which RNA controls can be *in vitro* transcribed (IVT) and added (“spiked-in”) to samples of interest.

Genome-scale gene expression measurements are impractical to calibrate; there are too many mRNAs to prepare exogenous calibration materials, and there are no reliable methods to establish the purity of calibration transcripts. While unable to provide a calibration, the addition of exogenous control RNA molecules is a reasonable approach to building evidence to assess confidence in a gene expression experiment. This approach was described and initiated at a NIST hosted industry workshop, out of which grew the External RNA Control Consortium (ERCC). This consortium was established to develop a common set of RNA controls for use in gene expression assays, standard methods for their use, and standard, objective, quantitative analysis approaches that would allow the technical performance of an experiment to be reported in a comparable fashion.

The controls are designed to mimic natural mammalian mRNA, and to be useful in a variety of assay formats. There are 96 different controls represented in the reference set, averaging ~1000 nucleotides in length, and ranging in GC content from ~33% to ~54%. Each control is a DNA sequence inserted in a common plasmid vector, engineered for simple IVT of either “sense” or “anti-sense” RNA, flanked with restriction and sequencing promoter sites. The 96 controls contain 86319 bases of certified sequence, with a confidence estimate for each base.

This SRM was a pilot for further developments in sequence RMs. It was the first material with a scale of many thousands of certified properties, and was the material for which an "ordinal scale" was developed to describe confidence in the certified properties (this scale is "Most Confident," "Very Confident," "Confident," and "Ambiguous"). It was also developed in partnership with the ERCC, which was composed of the end-user community, reagent manufacturers, technology developers, other federal agencies (including regulators), academic labs, and professional societies. The consortium model ensured that the RM would be relevant and useful, and the partnership with the technology and reagent developers assured that assay content would be available for the standard.

We are hopeful that this model proves useful in the context of the Genome in a Bottle Consortium, as that effort gets fully underway.

Another type of reference material appropriate for genome-scale gene expression measurements is a mixed-tissue reference material, first described by Thompson et al. [17]. Such a material relies on the fractions of materials from different tissues mixed into a sample pair in different known proportions. While the absolute abundances of the mRNA molecules are unknown in the sample pair, their relationship can be established through the mixing proportions and characterization of the signals from assay of the pure components of the mixture. NIST is actively evaluating this approach as a way to establish reference materials for validation of genome-scale measurements.

Conclusions

Reference Materials can play an important role in enabling clinical translation of new sequencing technologies. The Genome in a Bottle Consortium and NIST are developing well-characterized whole human and microbial genomes as NIST Reference Materials, which will be used by clinical and research laboratories to understand performance of sequencing and bioinformatics pipelines. In the future, these pure DNA Reference Materials could be supplemented by additional types of Reference Materials for genome-scale measurements, such as whole transcriptome, and proteome materials, which might be developed from induced pluripotent stem cell lines from the same individuals from which DNA Reference Materials are being developed. Reference Materials for genome-scale measurements, including the genomic materials currently being developed, are a critical part of the measurement infrastructure needed to have confidence in clinical measurements of billions of analytes, such as a human genome sequence.

References

1 National Institute of Standards and Technology, Standard Reference Material Definitions: http://www.nist.gov/srm/definitions.cfm

2 National Institutes of Health: The Haplotype Map Project (HapMap) and Other Research on Genetic Variations http://hapmap.ncbi.nlm.nih.gov/downloads/elsi/CEPH_Reconsent_Form.pdf

3 http://www.personalgenomes.org/

4 Zook JM, Samarov D, McDaniel J, Sen SK, Salit M. Synthetic Spike-in Standards Improve Run-Specific Systematic Error Analysis for DNA and RNA Sequencing. Plos One. 2012 Jul;7(7):10.

5 Londin E, Keller M, D'Andrea M, Delgrosso K, Ertel A, Surrey S, et al. Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. BMC Genomics. 2011;12(1):464.

6 Saito S, Morita K, Kohara A, Masui T, Sasao M, Ohgushi H, et al. Use of BAC array CGH for evaluation of chromosomal stability of clinically used human mesenchymal stem cells and of cancer cell lines. Human Cell. 2011;24(1):2-8.

7 Lundin S, Gruselius J, Nystedt B, Lexow P, Kaller M, Lundeberg J. Hierarchical molecular tagging to resolve long continuous sequences by massively parallel sequencing. Scientific reports. 2013 Mar;3.

8 Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature. 2012 Jul;487(7406):190-5.

9 Fan HC, Wang JB, Potanina A, Quake SR. Whole-genome molecular haplotyping of single cells. Nature Biotechnology. 2011 Jan;29(1):51

10 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011 May;43(5):491

11 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010 Sep;20(9):1297-303.

12 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug;27(15):2156-8.

13 Ajay SS, Parker SCJ, Abaan HO, Fajardo KVF, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome Research. 2011 Sep;21(9):1498-505.

14 Howey R, Cordell HJ. PREMIM and EMIM: tools for estimation of maternal, imprinting and interaction effects using multinomial modelling. Bmc Bioinformatics. 2012 Jun;13:13.

15 Kisand V, Lettieri T. Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. BMC Genomics. 2013;14:211.

16 Farrer RA, Henk DA, MacLean D, Studholme DJ, Fisher MC. Using False Discovery Rates to Benchmark SNP-callers in next-generation sequencing projects. Scientific reports. 2013 Mar;3.

17 Thompson KL. Use of a mixed tissue RNA design for performance assessments on multiple microarray formats. Nucleic Acids Research 2005;33:e187–e187.

Figure Legends:

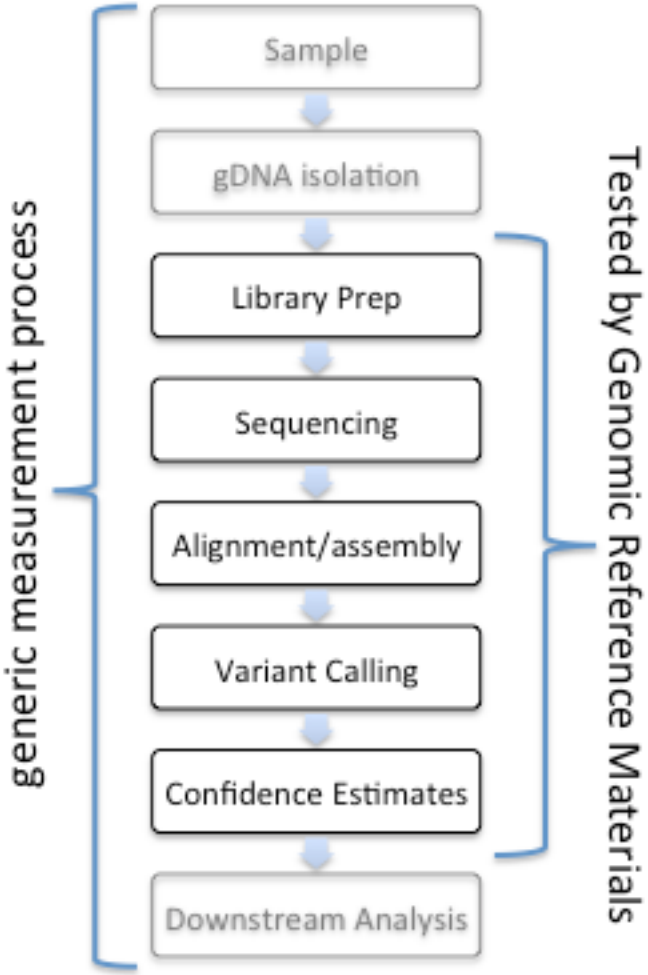

Fig. 1: Overall measurement process for sequencing DNA, with black boxes indicating the parts of the measurement process that will be assessed by candidate whole genome DNA NIST RMs.

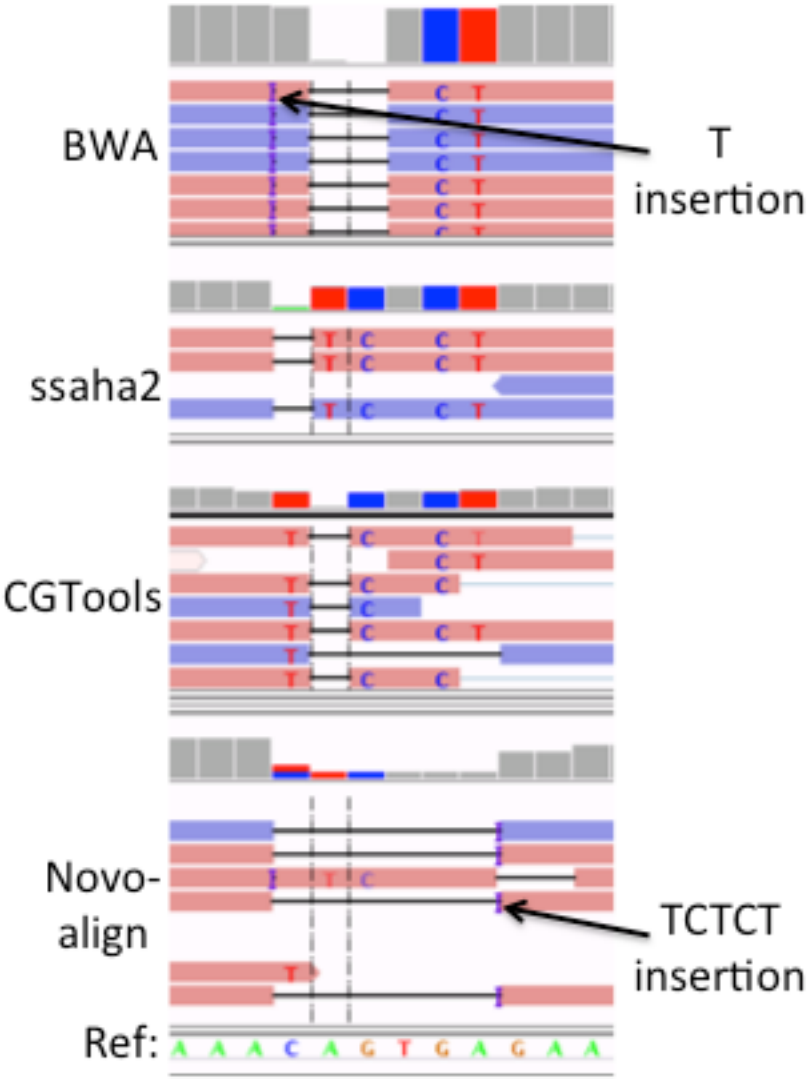

Fig. 2: Depiction of four different correct alignments around the same homozygous complex variant CAGTGA>TCTCT, which would result in 4 different sets of variant calls. BWA has a T insertion followed by a 2-base deletion follow by 2 SNPs. Ssaha2 has a 1-base deletion followed by 4 SNPs. Complete Genomics CGTools has a SNP followed by a 1-base deletion followed by 3 SNPs. Novoalign has a 6-base deletion followed by a 5-base insertion. All are correct alignments but they would result in very different variant calls, which complicates comparison of variant calls from different aligners and datasets.