Nature Reviews Genetics | Published online 22 Aug 2016; doi:10.1038/nrg.2016.113

PhotoDisc/Getty

SEQUENCING

A sparkling standard

DNA and RNA sequencing are transformative technologies, but the reliability of the conclusions arising from them is highly dependent on the accuracy of the various practical and bioinformatics steps in the pipelines. Two new studies from the Mercer laboratory describe designed DNA and RNA reference standards — sequencing spike-ins ('sequins') — that could serve as valuable quality-control reagents for diverse sequencing applications.

The teams sought to design reference standards for DNA (Deveson *et al.*) and RNA (Hardwick *et al.*) sequencing experiments, particularly for human studies. The rationale is that standards of predesigned composition (known DNA or RNA sequences at chosen stoichiometries) would be spiked into a test sample and would accompany it through library preparation, sequencing and bioinformatics analysis. Biases occurring at any of these steps can be identified by how robustly the expected features of the sequin pool emerge from the analysis.

A major challenge with the design of such standards is that they must be close mimics of DNA or RNA in the test sample (such as in length, complexity and GC content), yet have minimal sequence that is alignable with the test sample so that they are unambiguously distinguishable for parallel bioinformatics processing. To achieve this, both teams designed an *in silico* chromosome of ~11 Mb as the virtual source of their standards, which is derived from human genome sequences that have been inverted and often further rearranged so as to no longer align with human reference sequence.

For their DNA standards, Deveson *et al.* aimed for a quantitative resource to facilitate the detection of human genetic variation. From their reference *in silico* chromosome (*chrIS_D*) they designed custom sets of sequins that contained defined genetic variation relative to *chrIS_D*. The sequins were synthesized, cloned and purified.

One set of sequins consisted of 36 pairs of ~1 kb sequins containing single-nucleotide variants (SNVs) and small insertions and deletions (indels). Each sequin in a pair represents a chromosomal homologue; hence, variants in one or both of the sequin pairs can mimic heterozygous or homozygous genomic variation. To test performance, an equimolar mixture of sequin pairs was initially spiked into a sample of NA12878 reference human genomic DNA, for which the genetic variation is well characterized. Following sequencing and analysis, SNVs and indels were detected equivalently in the sequins and NA12878 genome, validating their use as an internal control. In a follow-up test, spiking in sequin pairs at staggered concentrations served as a quantitative control for detecting subclonal somatic mutations, such as those in a heterogeneous tumour sample against an NA12878 background. This helped to calibrate the sequencing read depths that are required for rare mutation detection and indicated that SNVs could be confidently detected only at allele frequencies of greater than 1/128.

Other sequin sets were generated for detecting larger variants. Sequins containing large deletions, insertions or inversions facilitated the detection of structural variation, and sequins containing defined copy numbers of elements served as a scale for quantifying the copy number of multi-copy elements and copy-number variants (CNVs) in NA12878.

In the accompanying paper, Hardwick *et al.* devised RNA sequins as standards for RNA sequencing (RNA-seq) experiments. Their *in silico* chromosome (*chrIS_R*) was designed to encode 78 artificial gene loci of varied length and exon number, and RNA sequins were generated for each of these genes (including splicing isoforms) by cloning and *in vitro* transcription. These sequins are a closer mimic of human transcripts than the commonly used External RNA Control Consortium (ERCC) set of 92 synthetic RNAs that are single-exon transcripts based on bacterial sequences.

The use of RNA sequins at staggered concentrations facilitated the quantitation of numerous transcriptomic features in human RNA samples, such as expression levels (including differential expression between samples) and splicing isoform use. Furthermore, a custom set of rearranged sequins aided the detection of fusion transcripts. Importantly, the sequins allowed the assessment of the most suitable bioinformatics tools for detecting these various features, and reported expression thresholds below which these features cannot be reliably called.

Overall, sequins are a versatile resource for diverse sequencing applications, and may serve the most value in inter-sample normalization (such as strengthening reproducibility between clinical laboratories) and for setting thresholds for clinical diagnoses.

Darren J. Burgess

ORIGINAL ARTICLES Deveson, I. W. *et al.* Representing genetic variation with synthetic DNA standards. *Nat. Methods* http://dx.doi.org/10.1038/nmeth.3957 (2016) | Hardwick, S. A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods* http://dx.doi.org/10.1038/nmeth.3958 (2016)
FURTHER READING Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015)