External RNA Controls Consortium Beta Version Update

Hangnoh Lee1*, P. Scott Pine2*, Jennifer McDaniel2, Marc Salit2, and Brian Oliver1

1. Section of Developmental Genomics, Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
2. Joint Initiative for Metrology in Biology, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

*These authors contributed equally to this work.

Corresponding author: Hangnoh Lee, Ph.D. Email: hangnoh.lee@nih.gov Phone: 301-594-1716/ Fax: 301-496-5239.

© Ivyspring International Publisher. Reproduction is permitted for personal, noncommercial use, provided that the article is in whole, unmodified, and properly cited. See http://ivyspring.com/terms for terms and conditions.

Published: 2016.07.26

Abstract

Spike-in RNAs are valuable controls for a variety of gene expression measurements. The External RNA Controls Consortium developed test sets that were used in a number of published reports. Here we provide an authoritative table that summarizes, updates, and corrects errors in the test version that ultimately resulted in the certified Standard Reference Material 2374. We have noted existence of anti-sense RNA controls in the material, corrected sub-pool memberships, and commented on control RNAs that displayed inconsistent behavior.

Key words: ERCC, spike-in controls, external RNA controls, NIST standard reference materials.

Advances in gene expression profiling technologies not only make it possible for individual groups to ask genome-wide questions, but properly controlled experiments with well-described metadata can be used over and over to make discoveries not envisaged by the data producers. Making these data robust and durable is greatly augmented by standard reference materials. The National Institute of Standards and Technology (NIST) as a part of the External RNA Controls Consortium (ERCC) developed 176 DNA plasmids that can be used as templates for RNA controls [1-3]. NIST Standard Reference Material (SRM) 2374 is a library composed of a subset of 96 plasmids. These same materials were used for commercially available ERCC RNA spike-in mixtures (Ambion/Thermo Fisher Scientific, Waltham, MA), which are formulations of 92 RNA molecules derived from the plasmids. The Commercial collection does not include ERCC-00007, -00018, -00023, and, -00128. One of the test versions that led to the SRM contained 96 RNA sequences transcribed from the plasmids, quantified, and mixed to form defined pools to be added to unknowns in transcription profiling experiments by array, sequencing, PCRs, or other assays. These test pools were widely distributed and were used by the human and model organisms Encyclopedia of DNA Elements projects [4, 5].

When “spiked” into an individual RNA sample, the readout from a single pool of ERCC controls can be used as a ruler. Each pool is designed to have dynamic range of 2^20. It is noteworthy that the actual linear range of their measurement depends on experimental platforms. Distribution of spike-in measurement fits to straight linear line in RNA-Seq and a monotonic sigmoidal pattern against actual abundance in microarrays or bead-arrays [6], consistent with data compression in hybridization-based techniques [7].

Addition of a single pool of ERCC controls
generates useful information, but their use can be enhanced when different pools of spike-in controls from different samples are directly compared. The “pools” of ERCC controls were mixed from multiple “subpools”, such that comparisons between “subpools” that belong to different “pools” generate abundance ratios that can be used as differential expression standards. There were two distinct sets of pools in the test version. Pools 12-15 follow a modified Latin-square design, using 5 different subpools (A-E). The numbers of RNA molecules in subpool A are equal in pools 12-15, and thus subpool A molecules generate a constant 1-to-1 proportion between the pools. Subpools B-E have differing molar concentrations that produce a trend in relative abundance across the pools of 1, 1.5, 2.5, 4-fold. For example, if pool 12 and pool 13 were used for two different samples, the log2-transformed ratios between different subpools will be 0, -0.585, -0.687, -0.737, and 2 [6]. The second set of pools, 78A and 78B, provide a pair of samples with reciprocal changes in relative abundance, i.e. 1.5-fold up and down, producing log2 transformed ratios of 0, 0.585, and -0.585.

While production of the spike-in control RNAs was tightly controlled, it was a test set, and there are multiple cases where measurements of spike-in molecules do not match the original description and/or expectations. In this short note, we summarize data outlining problematic ERCC spike-ins. This information should be used in re-evaluating datasets using the test version, as well as any future work that may use remaining aliquots in circulation (Table 1, and Supplementary Material for more details).

The plasmid DNAs were sequenced and deposited in GenBank, however, the in vitro transcribed RNAs were not sequenced except during testing in RNA-Seq experiments. These experiments made it clear that seven ERCC controls had the complementary sequence indicating that the transcripts were from the other strand (ERCC-00009, -00014, -00057, -00059, -00099, -00108, and -00116). As a result, these spike-in controls would not be measurable in hybridization-based assays [6]. Similarly, they would not be aligned in a strand-specific RNA-Seq analysis unless strand specificity was “turned off” in read quantification steps, or complementary sequences were provided for alignment. Additionally, plasmids are replicated in bacteria, where errors can be introduced. Differences in the sequences of the actual RNAs and the plasmids used for transcript templates are known [4], suggesting that such mutations occurred during plasmid propagation in the test set. The certified values of SRM 2374 are the sequences of the plasmids as distributed in the final set, and were determined by exhaustive sequencing [8].

There were instances of pooling errors in the test set. From multiple experiments that used 78A and 78B, we recognized that ERCC-00085 behaves like Subpool “C”, rather than the intended Subpool “B”. Therefore, when pools 78A and 78B were compared, ERCC-00085 displayed 33.3% increased fold changes than the original description. We have not detected ERCC-00084 in our experiments and it is possible that this RNA was prepared from ERCC-00085 plasmid DNA, effectively increasing the measurement of ERCC-00085. Similarly, we have corrected pool membership of ERCC-00113 from Subpool C to Subpool D from pools 12-15. ERCC-00073 and ERCC-00144 did not provide accurate measurements [4, 6]. One reason for poor measurement may be due to the molecular properties of individual spike-in RNA species (e.g. size and secondary structure). Additionally, a previous study pointed out discrepancy in ERCC-00116 measurements between poly-dT based mRNA enrichment and rRNA depletion protocols [3, 9]. The polyA tails on the ERCC spike-ins are not optimal for PolyA+ selection, and using them prior to library production is not recommended [4]. While there could well be additional instances of unexpected behavior of ERCC spike-in measurements, the information we provide here explains the unexpected ERCC behaviors that we have encountered to date.

Table 1. Summarized information on NIST distributed ERCC spike-in control test version.

| ERCC Control | GenBank† | DNA† | Length (nt) | % GC† | MW | Subpool in pool 12 to 15 | Subpool in pool 78 |
|---------------|----------|------|-------------|-------|----|-------------------------|-------------------|
| ERCC-00002a   | DQ59430  | Syn  | 1061        | 51    | 341,162 B | B                    |
| ERCC-00003a   | DQ516784 | Mjan | 1023        | 33    | 327,530 A | A                    |
| ERCC-00004a   | DQ516752 | Mjan | 523         | 34    | 167,216 C | C                    |
| ERCC-00007a   | EF011068 | Bsub | 1135        | 46    | 362,656 D | A                    |
| ERCC-00009a   | DQ680364 | Bsub | 984         | 47    | 336,394 E | C                    |
| ERCC-00102c   | DQ883670 | Syn  | 994         | 51    | 320,263 A | A                    |
| ERCC-00103a   | EF011062 | Bsub | 808         | 43    | 261,415 B | B                    |
| ERCC-00114a   | DQ875385 | Mjan | 1957        | 44    | 631,409 C | B                    |
| ERCC-00016    | DQ883664 | Syn  | 844         | 48    | 271,684 D | A                    |

http://www.jgenomics.com
| Accession | Description | NCBIID | Length | E-value | C-score | A-score |
|-----------|-------------|--------|--------|---------|---------|---------|
| ERCC-00017 | DQ459420 | Syn | 1136 | 51 | 367,042 | E |
| ERCC-00018 | EF011065 | Bsub | 1026 | 43 | 330,493 | C |
| ERCC-00019 | DQ83651 | Syn | 644 | 49 | 207,543 | B |
| ERCC-00020 | DQ516433 | Syn | 712 | 47 | 241,178 | C |
| ERCC-00023 | DQ516744 | Mjan | 273 | 31 | 88,186 | D |
| ERCC-00024 | DQ516787 | Mjan | 536 | 13 | 173,128 | E |
| ERCC-00025 | DQ836589 | Syn | 1994 | 50 | 640,941 | A |
| ERCC-00028 | DQ516719 | Syn | 1330 | 51 | 364,285 | B |
| ERCC-00031 | DQ459431 | Syn | 1138 | 48 | 365,732 | E |
| ERCC-00033 | DQ516796 | Mjan | 2022 | 33 | 651,554 | D |
| ERCC-00034 | DQ516803 | Syn | 1019 | 49 | 328,139 | E |
| ERCC-00035 | DQ459413 | Syn | 1130 | 51 | 364,378 | A |
| ERCC-00039 | DQ836565 | Syn | 740 | 49 | 238,322 | B |
| ERCC-00040 | DQ836631 | Syn | 744 | 53 | 239,738 | B |
| ERCC-00041 | EF011069 | Bsub | 1123 | 45 | 363,007 | D |
| ERCC-00043 | DQ516783 | Mjan | 1023 | 39 | 325,750 | E |
| ERCC-00046 | DQ516748 | Mjan | 522 | 35 | 168,087 | C |
| ERCC-00048 | DQ836737 | Syn | 992 | 48 | 320,110 | D |
| ERCC-00051 | DQ516740 | Mjan | 274 | 34 | 88,356 | A |
| ERCC-00053 | DQ516785 | Mjan | 1023 | 31 | 327,971 | C |
| ERCC-00054 | DQ516731 | Mjan | 274 | 37 | 88,976 | B |
| ERCC-00057 | DQ516735 | Mjan | 1136 | 50 | 366,548 | D |
| ERCC-00058 | DQ516748 | Mjan | 522 | 35 | 168,195 | C |
| ERCC-00061 | DQ516747 | Mjan | 1023 | 36 | 325,442 | B |
| ERCC-00062 | DQ516733 | Mjan | 1023 | 31 | 328,505 | C |
| ERCC-00067 | DQ83653 | Syn | 644 | 47 | 207,451 | D |
| ERCC-00069 | DQ459419 | Syn | 1137 | 50 | 366,664 | E |
| ERCC-00071 | DQ83634 | Syn | 644 | 48 | 206,115 | A |
| ERCC-00073 | DQ459404 | Syn | 314 | 30 | 372,337 | B |
| ERCC-00074 | DQ656538 | Bsub | 603 | 47 | 193,958 | B |
| ERCC-00075 | DQ516778 | Mjan | 1023 | 36 | 325,680 | E |
| ERCC-00076 | DQ836532 | Syn | 644 | 49 | 206,436 | E |
| ERCC-00077 | DQ516742 | Mjan | 273 | 33 | 87,694 | A |
| ERCC-00078 | DQ587865 | Syn | 993 | 50 | 320,094 | B |
| ERCC-00079 | DQ836582 | Syn | 644 | 49 | 207,737 | A |
| ERCC-00080 | DQ516790 | Mjan | 1023 | 35 | 172,323 | D |
| ERCC-00083 | DQ516780 | Mjan | 1023 | 35 | 325,668 | B |
| ERCC-00084 | DQ83682 | Syn | 994 | 50 | 320,443 | A |
| ERCC-00085 | DQ836369 | Syn | 844 | 49 | 271,323 | A |
| ERCC-00086 | DQ516791 | Mjan | 1020 | 32 | 328,632 | B |
| ERCC-00092 | DQ516925 | Syn | 1124 | 50 | 361,716 | D |
| ERCC-00093 | DQ516799 | Mjan | 321 | 37 | 166,307 | E |
| ERCC-00094 | DQ516748 | Syn | 1130 | 51 | 356,569 | A |
| ERCC-00097 | DQ516758 | Mjan | 523 | 36 | 167,189 | B |
| ERCC-00098 | DQ516745 | Syn | 1143 | 51 | 368,970 | C |
| ERCC-00099 | DQ75387 | Bsub | 1350 | 41 | 434,408 | D |
| ERCC-00104 | DQ516815 | Mjan | 2022 | 33 | 647,370 | E |
| ERCC-00108 | DQ660365 | Bsub | 1022 | 49 | 328,424 | A |
| ERCC-00109 | DQ585498 | Syn | 536 | 46 | 172,925 | B |
| ERCC-00111 | DQ836465 | Syn | 994 | 47 | 319,339 | C |
| ERCC-00112 | DQ459422 | Syn | 1136 | 47 | 364,932 | D |
| ERCC-00113 | DQ516763 | Syn | 840 | 50 | 270,609 | A |
| ERCC-00114 | DQ660367 | Bsub | 1991 | 50 | 639,986 | B |
| ERCC-00120 | DQ516792 | Syn | 536 | 48 | 172,005 | D |
| ERCC-00123 | DQ516782 | Mjan | 1022 | 36 | 324,911 | E |
| ERCC-00126 | DQ516927 | Syn | 1119 | 54 | 359,790 | A |
| ERCC-00128 | DQ516928 | Syn | 1133 | 48 | 364,405 | B |
| ERCC-00130 | EF011072 | Bsub | 1059 | 46 | 342,268 | C |
| ERCC-00143 | DQ585903 | Syn | 771 | 47 | 248,276 | D |
| ERCC-00146 | DQ516799 | Mjan | 274 | 31 | 88,594 | C |
| ERCC-00147 | DQ516790 | Mjan | 1023 | 36 | 351,125 | C |

http://www.jgenomics.com
ERCC-00148  DQ883642  Syn  494  49  159,911  D  B
ERCC-00150  DQ883659  Syn  743  47  239,128  E  A
ERCC-00154*  DQ854997  Syn  537  50  173,317  A  C
ERCC-00156  DQ883643  Syn  494  49  159,199  B  B
ERCC-00157*  DQ899618  Syn  1019  50  328,635  C  C
ERCC-00158*  DQ516795  Mjan  1021  34  328,797  D  A
ERCC-00160  DQ883658  Syn  743  46  239,437  E  C
ERCC-00162*  DQ516750  Mjan  523  36  166,409  A  A
ERCC-00163*  DQ668359  Bsub  543  47  174,949  B  B
ERCC-00164*  DQ516779  Mjan  1022  37  324,738  C  A
ERCC-00165  DQ668363  Bsub  872  50  279,768  D  C
ERCC-00169*  DQ516776  Mjan  1024  34  326,399  E  A
ERCC-00170*  DQ516773  Mjan  1024  34  330,808  A  B
ERCC-00171  DQ854994  Syn  505  48  163,022  B  B

(a) Sequence mismatches between the GenBank entries and the resequenced RNAs (see [4]).
(b) Syn: De novo synthetic design, Mjan: Methanocaldococcus jannaschii, Bsub: Bacillus subtilis.
(c) Length and GC content include poly(A) sequence.
(d) Reversed (anti-sense) in Pools 12-15.
(e) ERCC-00084 is not detected. E.g. ERCC-00084 and ERCC-00085, may have both been prepared from ERCC-00085 plasmid. ERCC-00085 behaves as C in some batches of Pool 78A and 78B.
(f) Corrected Pool membership to D and corrected Pool concentrations accordingly.
(g) Poor performing.
(h) Consistently under-reports abundance.
(i) Consistently over-reports abundance in Pools 78A and 78B.
(j) Particularly unsuitable for polyA+ isolation.
(k) ERCC-00104 has a length of either 2202 nt or 2203 nt.
(l) Not present in current commercial collections.

Supplementary Material
Supplemental file 1.
http://www.jgenomics.com/v04p0019s1.csv

Acknowledgements
The authors would like to acknowledge the careful experimental work by Sarah Helber to prepare the RNA and complex mixtures required for the test pools. This work supported in part by the Intramural Research program of the National Institutes of Health, NIDDK.

Abbreviations
ERCC - External RNA Controls Consortium, NIST - National Institute of Standards and Technology, SRM - Standard Reference Material, ENCODE - Encyclopedia of DNA Elements.

Competing Interests
The authors declare no competing interests.

References
1. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrrill J, Causton H, Conley MP, Elsperku R, Fero M et al: The External RNA Controls Consortium: a progress report. Nature methods 2005, 2(10):731-734.
2. ERCC: Proposed methods for testing and selecting the ERCC external RNA controls. BMC genomics 2005, 6:150.
3. Munro SA, Lund SP, Pine PS, Binder H, Clevet DA, Conesa A, Dopazo J, Fasold M, Hochreiter S, Hong H et al: Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nature communications 2014, 5:5125.
4. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B: Synthetic spike-in standards for RNA-seq experiments. Genome research 2011, 21(9):1543-1551.
5. Malone JH, Oliver B: Microarrays, deep sequencing and the true measure of the transcriptome. BMC biology 2011, 9:34.
6. Pine PS, Munro SA, Parsons JR, McDaniel J, Lucas AB, Lozach J, Myers TG, Su Q, Jacobs-Helber SM, Salit M: Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. BMC Biotechnol 2016, 16(1):54.
7. Malone JH, Oliver B: Microarrays, deep sequencing and the true measure of the transcriptome. BMC biology 2011, 9:34.
8. Pine PS, Munro SA, Parsons JR, McDaniel J, Lucas AB, Lozach J, Myers TG, Su Q, Jacobs-Helber SM, Salit M: Evaluation of the External RNA Controls Consortium (ERCC) reference material using a modified Latin square design. BMC Biotechnol 2016, 16(1):54.
9. Qin T, Yu Y, Du T, Shi L: mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Science China Life sciences 2013, 56(2):134-142.