USER GUIDE

applied biosystems® by life technologies™

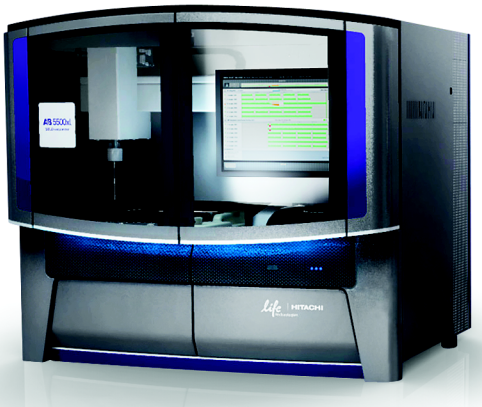
5500 Series SOLiD™ Systems

Experimental design and analysis strategies for RNA-Seq applications

RNA ANALYSIS

Publication Part Number 4460318 Rev. A
Revision Date April 2011

- **design experiment**
- prepare libraries
- prepare beads
- run sequencer
- analyze data



life technologies™

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

Part Number 4460318 Rev. A
04/2011

Contents

RNA analysis on the 5500 Series SOLiD™ System

Introduction

About RNA-Seq

Sequence-based approaches to RNA expression analysis, known as *RNA-Seq*, have been enabled by the development of massively parallel, high-throughput sequencing technology such as the SOLiD™ System. RNA-Seq methodology queries known and previously unknown RNAs in a sample, a hypothesis-neutral discovery approach that is advantageous compared to traditional microarray analysis, which interrogates only known RNAs. See Kassahn et al. (2011), Marguerat and Bahler (2010), Costa et al. (2010), Wang et al. (2009), and Wilhelm and Landry (2009) for recent reviews about applying high-throughput sequencing technology to RNA analysis.

SOLiD™ System for RNA analysis

SOLiD™ System sequencing provides several advantages over microarray platforms for RNA expression analysis:

- The SOLiD™ System enables accurate estimates of relative transcript abundance throughout a dynamic range of detection that is typically greater than that of traditional microarrays and which scales with increased sequencing depth.
- The SOLiD™ System's deep sequencing capability combined with the high accuracy of 2-base color coding enables detection of rare transcripts and transcript variants at levels below 1 copy per cell (Breu 2010).
- Library preparation with the SOLiD™ Total RNA-Seq Kit (for whole transcriptome and small RNA analysis) preserves the strandedness of the RNA during library preparation. This preparation method simplifies data analysis, allows determination of the directionality of transcription and gene orientation, and facilitates detection of opposing and overlapping transcripts.

Purpose of this guide

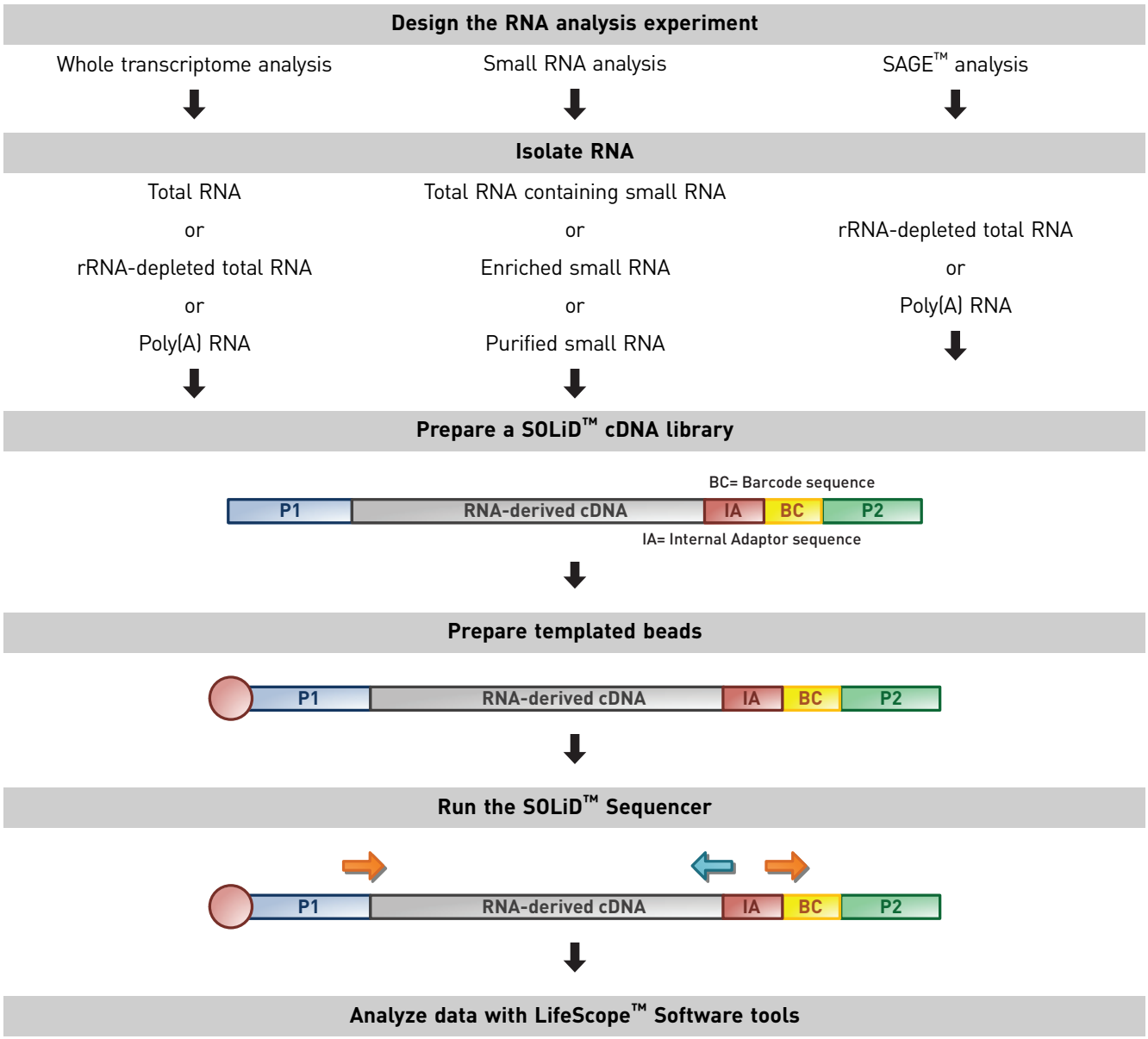
This guide discusses key parameters for RNA analysis on the 5500 Series SOLiD™ System. It also provides general guidelines for designing RNA analysis experiments using SOLiD™ System technology.

The guidelines provided here have been developed as a result of experience with analysis of human RNA expression, both in-house and in the scientific community. While it is expected that these guidelines are adaptable to any organism, they are intended only to provide a framework for discussion as you plan and implement your experiment.

RNA analysis workflow

In RNA analysis on the SOLiD™ System, a cDNA library of the RNA sample is prepared, clonally amplified onto SOLiD™ beads, and sequenced on the SOLiD™ sequencer. The sequencing reads (also known as *tags*) are mapped to one or more reference sequences, and the structure is then deduced using bioinformatic tools.

The RNA analysis workflow on the 5500 Series SOLiD™ System is illustrated in the following figure.



Overview of experimental design

Good experimental design, driven by the biological question and your experimental system, is essential to effectively applying SOLiD™ System technology to RNA analysis. The following sections cover key considerations for RNA expression analysis on the SOLiD™ System. It is always a good idea to perform pilot experiments to confirm that your plan will generate the amount of sequence data required for your experimental goal.

- Analysis strategy, RNA isolation, and library preparation (page 7).
- Sequencing strategy: read direction(s) and length (page 9).
- Number of reads required (page 11).
- Multiplex sequencing (page 13).

Analysis strategy, RNA isolation, and library preparation

Analysis strategies

Table 1 summarizes the RNA analysis strategies supported by the SOLiD™ System. The analysis strategy informs the cDNA library preparation method, the amount and type of data generated, the size/complexity of the reference sequences used for alignment, and the analysis algorithm(s).

Table 1 RNA analysis with SOLiD™ System technology

Analysis strategy	Purpose†	Library preparation method and input RNA‡
Whole transcriptome analysis Global sequence analysis of coding and non-coding RNA transcripts along their entire length.	Discovery§: • Mapping all transcripts (coding and non-coding) on a genomic or regional scale. • Discovery of novel transcripts, coding and non-coding. • Discovery and mapping of translocations and fusion transcripts. • Detection of alternative splicing events. • Discovery of alternative transcription start and stop sites. Counting and discovery§: • Expression profiling of coding and non-coding RNA. • Discovery of SNPs and allele-specific expression.	Method: SOLiD™ Total RNA-Seq Kit, whole transcriptome procedure. Input RNA: • Total RNA, rRNA-depleted, for analysis of coding and non-coding transcripts. • Total RNA, *not* rRNA-depleted can be appropriate if required for a specific application or with very limited samples, to avoid incurring losses during rRNA depletion.†† • Poly(A) RNA, for analysis of polyadenylated RNAs (coding and non-coding).
Small RNA analysis Sequence analysis of small RNA species, generally 16–27 nt in length.	Counting and discovery§: • Analysis of small RNA expression patterns. • Detection of length and 3'- and 5'-sequence isoforms. • Discovery of novel small RNAs.	Method: SOLiD™ Total RNA-Seq Kit, small RNA procedure. Input RNA: • Total RNA containing small RNA. • Enriched small RNA. • Purified small RNA. Depends upon the percentage of small RNA in the total RNA sample; see the *SOLiD™ Total RNA-Seq Kit Protocol*.
SAGE™ analysis (Serial Analysis of Gene Expression) Expression profiling by counting short sequencing reads generated specifically from the 3' ends of RNA.	Counting§: • Profiling expression of known transcripts in one or more samples relative to another. • Profiling alternative poly(A) site usage in known transcripts.	Method: SOLiD™ SAGE™ Kit procedure. Input RNA: • Total RNA‡‡. • Poly(A) RNA.

† The applications in this table are not an exhaustive or all-inclusive list.

‡ Detailed information about library preparation is found in the user guide or manual for each kit.

§ The terms *counting* or *discovery* are sometimes used for applications whose main focus is relative expression levels or discovery of novel RNAs, respectively.

†† When using total RNA that includes rRNA, ≥60 % of mapped reads may be rRNA.

‡‡ The standard SOLiD™ SAGE™ Kit library preparation procedure selects poly(A) RNA from total RNA.

In many cases, the most suitable analysis strategy for your experiment is obvious. However, if your primary purpose is gene expression profiling of known transcripts, there is overlap in the suitability of whole transcriptome and SOLiD™ System SAGE™ analysis.

- Whole transcriptome analysis examines the RNA molecule along its entire length, enabling discovery of alternative splicing or start/stop sites, even if the primary objective is relative expression levels.
- In contrast, SOLiD™ System SAGE™ analysis interrogates only a short sequence at the 3′ end of each transcript, precluding analysis of transcript structure.

  SOLiD™ System SAGE™ analysis generates data that are less complex than whole transcriptome analysis and more like that generated by microarray analysis, but with the higher sensitivity and broader dynamic range of the SOLiD™ platform.

RNA isolation

The RNA isolation method is determined by your experimental system and the input requirements of the library preparation method (see Table 1 on page 7). Isolation of high quality, intact RNA is critical for preparing SOLiD™ System cDNA libraries that are representative of the cellular RNA population.

- Table 5 in Appendix A lists RNA isolation kits suggested for different sample types and applications.
- It is essential to use best practices for handling RNA, from RNA isolation through cDNA library preparation. For more information, see Ambion Technical Bulletin #159, *Working with RNA*, available at **www.invitrogen.com/workingwithrna**.
- We strongly recommend evaluating the quality of your RNA samples before proceeding with SOLiD™ System cDNA library preparation. Table 7 in Appendix B lists common criteria for evaluating RNA quality.

Library preparation

Table 1 lists kits optimized for preparation of SOLiD™ System cDNA libraries. All kits enable preparation of a library that is ready to enter the SOLiD™ System workflow at the templated bead preparation step.

- In the SOLiD™ Total RNA-Seq Kit procedures, the RNA is first ligated to SOLiD™-specific adaptor oligonucleotides (this step preserves the strandedness of the RNA in the library) and then reverse transcribed to cDNA.
- In the SOLiD™ SAGE™ Kit procedure, the RNA is first reverse transcribed and then ligated to the SOLiD™ adaptor oligonucleotides.

For multiplex sequencing, barcodes can be incorporated with barcoded 3′ PCR primers during the cDNA amplification step of either library preparation procedure. Barcoded primers are available in the SOLiD™ RNA Barcoding Kits. See "Multiplex sequencing" on page 13 and Table 9 in Appendix B for further information.

For libraries prepared using the SOLiD™ Total RNA-Seq Kit, it is important to quantitate the libraries as described in the *SOLiD™ Total RNA-Seq Kit Protocol* before proceeding with templated bead preparation. The SOLiD™ Total RNA-Seq Kit procedure, using the Agilent 2100 Bioanalyzer to determine the molar concentration of the library, gives the most reliable and consistent estimates of RNA-Seq library quantity. For the SOLiD™ SAGE™ Kit and other library preparation kits, use the SOLiD™ Library TaqMan® Quantitation Kit and real-time PCR as recommended.

Whole transcriptome libraries: using ERCC RNA Spike-In Control Mixes

We recommend including one of the ERCC RNA Spike-In Control Mixes during preparation of a whole transcriptome library.

The ERCC RNA Spike-In Control Mixes are 2 sets of 92 polyadenylated RNAs, 250–2000 nt in length, that are transcribed from a set of NIST-certified plasmids (External RNA Controls Consortium, 2005). Each Spike-In Mix is preformulated at defined quantities spanning a 10^6-fold concentration range, with defined Mix 1:Mix 2 transcript molar concentration ratios.

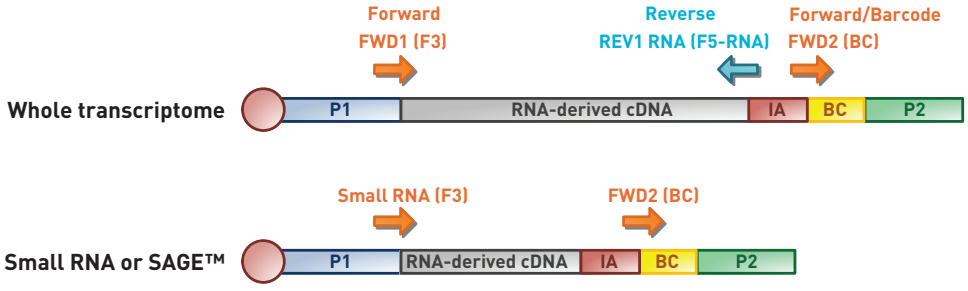
When Spike-In Mixes are added at known concentrations to RNA samples before library preparation, you can compare observed ERCC transcript amounts and ratios to known amounts and ratios, within and between samples. These comparisons can be done with the libraries themselves (using real-time PCR and TaqMan® Gene Expression Assays) and in the sequencing data. These assessments can be particularly useful when performing pilot experiments; see "Estimating the lower limit of detection using ERCC Spike-In Mixes" on page 12.

Further information about the ERCC RNA Spike-In Control Mixes and detailed instructions for use are provided in the *ERCC RNA Spike-In Control Mixes User Guide*. The ERCC RNA Spike-In Control Mixes are not recommended for small RNA analysis due to the size of the transcripts.

Sequencing strategy: read direction(s) and length

The 5500 Series SOLiD™ System supports both forward-only and forward + reverse, also called *paired-end*, sequencing reads for RNA-Seq applications. The 5500 Series SOLiD™ Sequencer allows variable read lengths (up to 75 nt forward; up to 35 nt reverse) to optimize reagent use. Figure 1 provides a conceptual overview of the primers used in SOLiD™ System sequencing of RNA-source libraries (FWD1, Small RNA, FWD2, and REV1 RNA) and the corresponding data analysis tags (F3, F5-RNA, BC).

Figure 1 5500 Series SOLiD™ System sequencing primers for RNA analysis



Generally, we recommend the following:

- Acquire the longest reads that are allowed by the insert size (see Table 2) and supported by the SOLiD™ platform.
- For whole transcriptome analysis, we recommend paired-end sequencing, which optimizes the potential for discovery of novel transcripts, fusion transcripts, and alternative splice junctions, even if that is not the primary objective. See "Supplemental data: paired-end sequencing for detection of fusion transcripts" on page 22. However, if the purpose of the experiment is only to determine relative levels of known RNAs, then only a single forward read is needed.

Table 2 5500 Series SOLiD™ System: sequencing strategies and primers for RNA analysis

Analysis strategy, library preparation method, and insert size	Suggested read direction(s) and read length(s)	Sequencing primers†	
		F3 tag	F5-RNA tag
Whole transcriptome analysis: SOLiD™ Total RNA-Seq Kit‡, whole transcriptome procedure RNA-derived sequence§: • ~110 to ~200 nt • Inserts are derived from fragmentation of RNA along the entire transcript	Forward-only/75 nt suitable for: • Expression profiling of coding and non-coding RNA • SNP discovery	SOLiD™ FWD1 Seq. Primers	—
	Forward-only/75 nt OK, paired-end forward/75 nt + reverse/35 nt recommended for: • Discovery of novel transcripts • Mapping all transcripts (coding and non-coding) in a region • SNP discovery	SOLiD™ FWD1 Seq. Primers	SOLiD™ REV1 (RNA) Seq. Primers
	Paired-end forward/75 nt + reverse/35 nt highly recommended for: • Detection of fusion transcripts • Discovery and mapping of translocations and fusion transcripts • Detection of alternative splicing	SOLiD™ FWD1 Seq. Primers	SOLiD™ REV1 (RNA) Seq. Primers
Small RNA analysis: SOLiD™ Total RNA-Seq Kit‡, small RNA procedure RNA-derived sequence§: • 16–27 nt • Insert encompasses entire RNA molecule	Forward-only/read length as appropriate for expected size of the small RNA††	SOLiD™ Small RNA Seq. Primers	—
SAGE™ analysis: SOLiD™ SAGE™ Kit‡ RNA-derived sequence§: • 27 nt • Insert is derived from 3'-end of transcript	Forward-only/read length as appropriate for the size of the SAGE™ tags††	SOLiD™ Small RNA Seq. Primers	—

† For sequencing reads through the cDNA insert only; for all library types, use the FWD2 primers for barcode reads.

‡ See "Reagents and kits for RNA analysis" on page 17 for ordering information.

§ Does not include P1, P2, IA, or BC sequences.

†† Acquiring reads longer than the insert size allows confirmation that the entire insert has been sequenced, but may require trimming of the adaptor sequence to map the read effectively.

Exact Call Chemistry

The 5500 Series SOLiD™ Sequencers offer an optional Exact Call Chemistry (ECC) primer round that enables:

- Higher accuracy reads for enhanced mutation detection.
- Base-space sequence in the XSQ output file, without the need to map to a reference sequence.

For further information on ECC, see publication number CO31266, "SOLiD™ System accuracy with the Exact Call Chemistry module."

Number of reads required

Overview

Two complementary strategies are typically used for estimating the number of sequencing reads needed for an RNA analysis experiment.

- Estimating the number of mapped reads needed to saturate hits in the reference (such as RefSeq or miRBase), at a certain calling threshold.
- Estimating the lower limit of detection using external RNA controls, such as the ERCC RNA Spike-In Control Mixes, that are "spiked in" and sequenced along with the endogenous RNA. This estimate is based on a given total number of mapped reads at a certain calling threshold.

See "RNA sequence analysis concepts" on page 21 for information about the concepts of mapped reads and calling thresholds.

The following discussion is derived from experiments for detection of human RNAs. It is meant to serve only as a guideline for your application.

Saturation of reads mapping to a reference

Because every cell type or tissue expresses only a subset of the total transcriptome, detection of 100% of a reference is not seen using RNA from a single cell type. Instead, the following saturation characteristics are observed (see "Supplemental data: saturation of reads mapping to a reference" on page 23 for examples with human RNA analysis).

- As the number of mappable reads in a sequencing run increases, the fraction of an RNA reference that is detected increases, until a plateau is reached. After the plateau is reached, or saturation is achieved, the rate of detecting additional reference transcripts, genes, or other genomic regions of interest with additional reads is much lower.
- The number of reads needed to reach the plateau depends on a number of factors, including the RNA sample used, the size of the genome, the percent of the genome transcribed into RNA, the regions of experimental interest (for example, transcripts, exons, or splice junctions), and the threshold used for calling an element detected. In general, greater read depth is required to saturate detection of splice junctions than to saturate detection of exons, because the region spanning a splice junction is a subset of the corresponding exons. Likewise, more reads are required to detect a single exon than its corresponding transcript, if the transcript contains multiple exons.

Given these considerations, there are no strict rules. We recommend running a pilot study in your experimental system to determine the saturation curve for the reference sequence of interest.

This approach gives you an estimate of the number of reads needed to maximize detection of known RNA species in the reference. The assumption is that discovery of novel RNAs and sequence variants requires acquisition of a number of mapped reads corresponding to a location well past the plateau of the saturation curve.

Estimating the lower limit of detection using ERCC Spike-In Mixes

In general, the more reads in a sequencing run, the greater the sensitivity to detect RNA species of interest. You can estimate the lower limit of detection (LLD) of a SOLiD™ System sequencing run using the ERCC RNA Spike-In Control Mixes. See "Supplemental data: estimating the lower limit of detection with ERCC Spike-In Mixes" on page 25 for an example with human RNA.

We recommend running a pilot experiment similar to that described on page 25 to estimate the LLD for your planned sequencing run(s). This approach can give you confidence that rare events that occur at a frequency above the determined LLD should be detectable in your system.

Sequencing depth, replicates, and differential gene expression

The power to detect differential expression is affected by several factors, including:

- An accurate estimate of the population variation in your experimental system. Biological replicates of a treatment or condition give you an estimate of the variation, and they are necessary to give statistical power in experiments designed to detect differential expression caused by the treatment or condition.
- The size of the fold-change in expression that you are trying to detect.
- The number of reads. More reads (deeper sequencing) give greater detection sensitivity.

How do these factors affect experimental planning? Generally, if the estimated number of reads that is required to give adequate detection sensitivity (LLD) is well under the capacity of your sequencing run(s), consider including more biological replicates rather than deeper sequencing. This approach gives more power to your experiment for detection of differential gene expression. For further information, see Auer and Doerge (2010).

Summary

Table 3 summarizes guidelines developed at Life Technologies for human RNA analysis, based on our current understanding of the human transcriptome (see Appendix B for supporting data). It is expected that the number of reads scales with the size of the genome or transcriptome; however, we recommend empirically determining the number actually required by your experimental needs, by running pilot experiments in your experimental system similar to those presented in Appendix B. It is always a good idea to generate more rather than less data in pilot experiments.

Table 3 SOLiD™ System analysis of human RNA: suggested number of mapped reads

Type of analysis (human RNA)	Suggested number of mapped reads†
Whole transcriptome analysis	≥50 × 10^6 uniquely mapped reads (gives good coverage of human RefSeq at 1 RPKM‡): • Detection of novel transcripts • Mapping all transcripts (coding and non-coding) in a region
	≥100 × 10^6 reads to detect rare events (at 1 RPKM): • Alternative splicing • Fusion transcripts • Variant transcripts • SNPs in RNA • Allele-specific gene expression
Small RNA analysis	3–4 × 10^6 mapped reads: detection of known miRNAs
	10–30 × 10^6 mapped reads: discovery of new small RNAs or iso-miRs
SAGE™ analysis	2–5 × 10^6 mapped reads

† See Appendix B for examples of the experiments used to derive the information in this table.

‡ RPKM: reads per kilobase transcript length per million mapped reads; see Appendix B.

Multiplex sequencing

The 5500 Series SOLiD™ Sequencer flowchip has 6 lanes that can be individually programmed for a sequencing run. For certain applications, the bead capacity of the flowchip lane is several-fold higher than the number of beads required to acquire all the necessary data. In these cases, multiple barcoded libraries of the same type (for example, all small RNA libraries or all SAGE™ libraries) can be combined and sequenced in one lane for multiplex sequencing. Table 9 in Appendix B lists example calculations for determining the number of libraries that can be accommodated in a single flowchip lane.

Advantages of multiplex sequencing

- Multiplex sequencing optimizes use of reagents and time on the sequencer.
- Multiplex sequencing lends itself to the strategy of sequencing multiple libraries from biological replicates, which can give more statistical strength for discovery and gene expression analysis than a single library (Auer and Doerge, 2010).

  For example, sequencing 4 libraries from biological replicates instead of one library at 4× the number of reads may give higher confidence in differential gene expression experiments. See also "Sequencing depth, replicates, and differential gene expression" on page 12.

In general, multiplex sequencing enables greater flexibility in experimental design than does singleplex sequencing.

Maintaining color balance with multiplex sequencing

Multiplex sequencing must maintain color balance. Color balance is the relative proportion of beads that are called as each of the four colors in a given cycle. The SOLiD™ 3' Primers in the SOLiD™ RNA Barcoding Kit Modules are optimized to maintain color balance within consecutive groups of 4 primers. To preserve color balance during multiplex sequencing, each flowchip lane must have at least one set of

4 color-balanced barcoded libraries. Further information about constructing color-balanced, barcoded cDNA libraries is provided in the protocols for the SOLiD™ Total RNA-Seq Kit, the SOLiD™ RNA Barcoding Kit Modules, and the SOLiD™ SAGE™ Kit with Barcoding Adaptor Module.

Experimental design: putting it all together

Table 4 Examples of 5500 Series SOLiD™ System sequencing for human RNA analysis

Looking at...	Analysis strategy	RNA sample†	Sequencing strategy and mapped reads required	Libraries/ lane‡
Changes in expression levels of known transcripts during a treatment or condition	Whole transcriptome, counting	rRNA-depleted total RNA or poly(A) RNA	Forward only, ~50 nt ≥50 × 10^6 mapped reads	1
	SAGE™, counting	Total RNA or poly(A) RNA	Forward only, ~25 nt 2–5 × 10^6 mapped reads	12–33
Changes in gene expression of known or unknown transcripts during a treatment or condition, looking for alternative splicing, or promoter or terminator usage	Whole transcriptome, mostly counting with option of discovery	rRNA-depleted total RNA	Forward/75 nt + reverse/ 35 nt ≥50 × 10^6 mapped reads	1
Discovery of novel transcripts and features of transcripts, such as fusion transcripts during cancer progression, allele-specific gene expression, or SNP discovery	Whole transcriptome, discovery	rRNA-depleted total RNA or total RNA	Forward/75 nt + reverse/ 35 nt 100 × 10^6 mapped reads	1
Global expression of small RNAs in a cell line or tissue	Small RNA, discovery	Small RNA-enriched RNA or purified small RNA	Forward only, ~35 nt 10–30 × 10^6 mapped reads	2–8
Changes in expression levels of known miRNAs during a treatment or condition.	Small RNA, counting with option of discovery	Small RNA-enriched RNA or purified small RNA	Forward only, ~35 nt 5 × 10^6 mapped reads	14–16

† See Table 1 on page 7 for a summary of input RNA options for each library type. See Table 5 on page 17 for a summary of kits for high-quality RNA isolation from a variety of biological sources.

‡ Barcoded, color-balanced libraries for >1 library/lane. See Table 9 in Appendix B for example calculations.

Data analysis

Sequencing run data are automatically exported from the 5500 Series SOLiD™ Sequencer in XSQ binary file format. If an ECC primer round has been performed, the XSQ output includes the sequence information in base space.

LifeScope™ Genomic Analysis Software incorporates tools for whole transcriptome and small RNA analysis. LifeScope™ Software supports analysis of data from barcoded libraries.

The SOLiD™ SAGE™ Analysis Software v1.10 provides tools for mapping and counting SAGE™ sequencing reads with a reference database of your choice. This software requires CSFASTA and QUAL input files. A standalone program is available to convert XSQ files to CSFASTA and QUAL formats. If the XSQ file includes base space data, the conversion also exports the base space data into a FASTQ file. For further information, see the *LifeScope™ Genomic Analysis Software: Command Shell User Guide*.

The *ERCC RNA Spike-In Control Mixes User Guide* provides information about mapping ERCC Control RNA reads and assessing performance of the 5500 Series SOLiD™ Sequencer using the ERCC RNA Spike-In Control Mixes.

Information about third-party analysis software can be found at the SOLiD™ Software Development Community website: **info.appliedbiosystems.com/solidsoftwarecommunity**. If necessary, use the conversion program described above to convert XSQ output files to CSFASTA and QUAL input formats for third-party software.

Validation of results

Validate your SOLiD™ System results with TaqMan® Assays targeting RNAs of interest, available at **www.appliedbiosystems.com**:

- TaqMan® Gene Expression Assays
- TaqMan® Micro RNA Assays
- TaqMan® ncRNA Assays

Product information and support

Reagents and kits for RNA analysis

- Table 5 lists reagents and kits for high-quality RNA isolation.
- Table 6 lists reagents and kits for SOLiD™ library preparation from RNA.
- For products for templated bead preparation, see the SOLiD™ EZ Bead™ System product page at **www.appliedbiosystems.com**.
- For 5500 Series SOLiD™ System sequencing reagents, see the *5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference* (Publication no. 4465650).

Detailed instructions for use and a complete list of required materials are provided in the user guides accompanying each product.

Table 5 Key reagents and kits for RNA isolation

Kit (Part no.†)	Description
Whole transcriptome analysis	
• MagMAX™-96 Total RNA Isolation Kit (AM1830) • MagMAX™-96 for Microarrays Total RNA Isolation Kit (AM1839)	Isolation of total RNA from cells or plant/mammalian tissue samples in tubes or 96-well plates.
MagMAX™-96 Blood RNA Isolation Kit (AM1837)	Isolation of total and viral RNA from mammalian whole blood.
TRIzol® Plus RNA Purification System (12183555)	Isolation of total RNA from animal and plant cells and tissue, bacteria, and yeast.
PureLink® RNA Mini Kit (12183020, 12183018A)	Isolation of total RNA from animal, plant, yeast, bacteria, and blood.
Poly(A) selection or rRNA depletion of total RNA	
• Poly(A)Purist™ Kit (AM1916) • MicroPoly(A)Purist™ Kit (AM19190 • Poly(A)Purist™ MAG Kit (AM1922) • mRNA Catcher™ PLUS Kit (K157002)	Selection of poly(A)-containing RNA from total RNA preparations.
RiboMinus™ Eukaryote Kit for RNA-Seq (A10837)	Depletion of eukaryote 18S, 28S, 5.8S, 5S rRNA from total RNA.
RiboMinus™ Plant Kit for RNA-Seq (A10838)	Depletion of 25/26S and 17/18S rRNA, 23S and 16S chloroplast RNA, and 18S mitochondrion RNA from total RNA.
Small RNA analysis	
• *mir*Vana™ miRNA Isolation Kit (AM1560) • *mir*Vana Paris™ Kit (AM1556)	Isolation of small RNA-containing total RNA from tissues and cells; enrichment of small RNA optional.
RecoverAll™ Total Nucleic Acid Isolation Kit for FFPE (AM1975)	Isolation of total RNA, including miRNAs, from formaldehyde or paraformaldehyde-fixe, paraffin-embedded (FFPE) tissues.

Kit (Part no.†)	Description
PureLink® miRNA Isolation Kit (K157001)	Isolation of small RNA from mammalian cells and tissues, plant tissues, yeast, and bacteria.
SAGE™ analysis	
TRIzol® Plus RNA Purification System (12183555)	Isolation of total RNA from animal and plant cells and tissue, bacteria and yeast.

† Available at **www.invitrogen.com**.

Table 6 Key kits and reagents for SOLiD™ library preparation from RNA

Application	Kit (Part no.†)	Description
• Whole transcriptome analysis • Small RNA analysis	SOLiD™ Total RNA-Seq Kit (4445374) (Optional) SOLiD™ RNA Barcoding Kits (see below)	Single kit for preparation of whole transcriptome or small RNA libraries that maintain genomic strand information for mapping.
SAGE™ analysis	SOLiD™ SAGE™ Kit with Barcoding Adaptor Module (4452811) SOLiD™ RNA Barcoding Kits (see below)	Kit for preparation of SOLiD™ libraries of cDNA tags of 3′ ends of RNAs. For use with a SOLiD™ RNA Barcoding Kit.
Barcoded cDNA libraries	SOLiD™ RNA Barcoding Kits: • SOLiD™ RNA Barcoding Kit, Modules 1–48 (4461565) • SOLiD™ RNA Barcoding Kit, Modules 49–96 (4461566) • SOLiD™ RNA Barcoding Kit, Modules 1–96 (4461567) • 16-barcode modules: – 1–16 (4427046) – 17–32 (4453189) – 33–48 (4453191) – 49–64 (4456501) – 65–80 (4456502) – 81–96 (4456503)	3′ PCR Primers with unique barcode sequences, for incorporation of barcodes into SOLiD™ cDNA libraries. For use with the SOLiD™ Total RNA-Seq Kit or SOLiD™ SAGE™ Kit with Barcoding Adaptor Module.
External RNA controls	ERCC RNA Spike-In Control Mixes • ERCC RNA Spike-In Mix (4456740) • ERCC ExFold RNA Spike-In Mixes (4456739)	Sets of synthetic RNA transcripts that are added to purified RNA samples before whole transcriptome library preparation, for SOLiD™ System performance assessment.

† Available at **www.appliedbiosystems.com**

Related documentation

For a complete list of guides for the 5500 Series SOLiD™ Systems, see the *5500 Series SOLiD™ Systems User Documentation Quick Reference* (Publication no. 4465102). Search by publication number at **www.appliedbiosystems.com**.

Document	Publication no.
SOLiD™ Total RNA-Seq Kit Protocol	4452437
SOLiD™ SAGE™ Kit with Barcoding Adaptor Module Guide	4456596
ERCC RNA Spike-In Control Mixes User Guide	4455352
5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference	4465650

Obtaining support

For the latest services and support information for all locations, go to:

www.appliedbiosystems.com

At the Applied Biosystems web site, you can:

- Access worldwide telephone and fax numbers to contact Applied Biosystems Technical Support and Sales facilities.
- Search through frequently asked questions (FAQs).
- Submit a question directly to Technical Support.
- Order Applied Biosystems user documents, MSDSs, certificates of analysis, and other related documents.
- Download PDF documents.
- Obtain information about customer training.
- Download software updates and patches.

Supplemental information

Criteria for high-quality RNA

Table 7 Criteria for high-quality RNA for SOLiD™ System library preparation

Parameter	Evaluation method	Criteria
Purity RNA samples should be free of contaminating proteins, DNA, and other cellular material, as well as phenol, ethanol, and salts associated with RNA isolation procedures.	Spectrophotometry: traditional or NanoDrop	A_{260}:A_{280} = 1.7–2.1
Integrity RNA samples should have a high proportion of full-length RNA, with little or no evidence of degradation.	Microfluidics analysis: Agilent 2100 Bioanalyzer (requires picogram to nanogram amounts of RNA)	• Discrete rRNA bands (that is, no trailing of the peaks) • 28S:18S rRNA ratio approaches 2:1† • RNA Integrity Number (RIN) >7
	Denaturing agarose gel electrophoresis and nucleic acid staining (requires microgram amounts of RNA)	• Discrete rRNA bands (that is, no significant smearing below each band) • 28S:18S rRNA ratio approaches 2:1†

† For mammalian RNA; values for RNA from other species may have different ratios.

RNA sequence analysis concepts

About mapped reads

The mappability of an RNA read depends on these parameters:

- The technical quality of the read; that is, the ability of the SOLiD™ System instrument software to align a read to a given reference sequence with a certain degree of confidence.
- The quality of the reference sequence annotation.

  Reference sequence databases such as RefSeq and miRBase are continuously changing and expanding as newly discovered transcripts are entered and the annotation of known transcripts is improved. This issue is a consideration if your experimental purpose is discovery, because by definition, the novel regions or transcripts being studied have a low probability of being in the reference sequence or of being in the reference sequence and correctly annotated.

About calling thresholds

The criteria for calling an RNA present varies according to the analysis method.

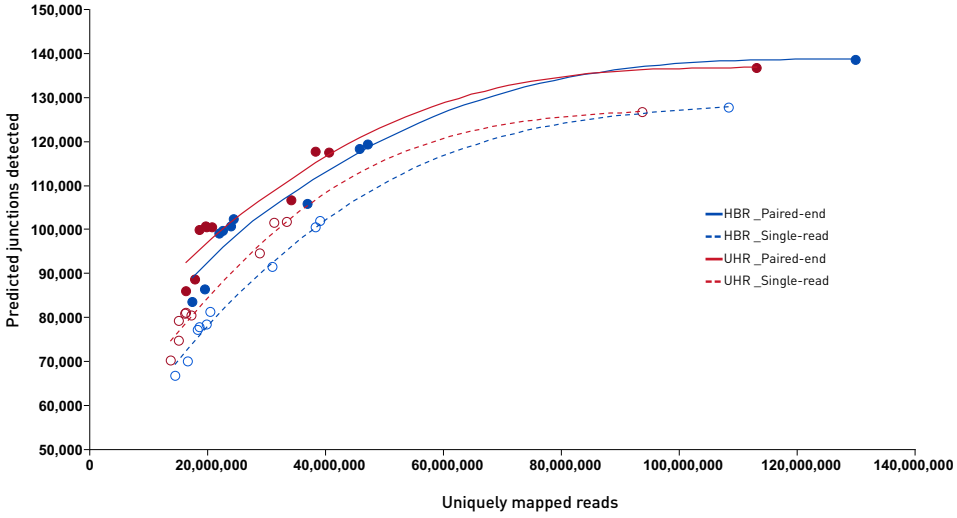
- Whole transcriptome data are often normalized for RNA length and for the total read number in the run, to facilitate comparisons between and within sequencing runs. One current method uses the concept of *reads per kilobase (kb) of exon model per million mapped reads*, or RPKM (Mortazavi et al., 2008). As an example, 1 RPKM is equivalent to 20 reads mapping to a 1-kb transcript per 20×10^6 (20 million; 20 M) mapped reads.

  The RPKM threshold for calling a transcript or new isoform present can be set more or less stringently, depending on the quality of the data and the type of application. Setting the RPKM threshold typically involves balancing the need for high accuracy with the cost associated with sequencing depth. With current technology, many researchers set the threshold at ~1 RPKM.
- SAGE™ analysis uses the concept of *hits*, also known as the *number of mapped reads per transcript*. (SAGE™ analysis generally does not use the concept of RPKM since the technique interrogates only a short read for each transcript.) With current technology, calling thresholds are typically set at 1–3 hits.
- Small RNA analysis uses the concepts of *number of reads* or, alternatively, *reads per million reads* (RPM; small RNA analysis does not use the concept of RPKM because the reads span the entire transcript length). With current technology, calling thresholds are set at ~1 RPM.

Supplemental data: paired-end sequencing for detection of fusion transcripts

Figure 2 shows a comparison of paired-end (50 nt forward/35 nt reverse) and forward-only (50 nt) sequencing on the SOLiD™ 4 System of whole transcriptome libraries prepared from Universal Human Reference RNA (UHR) and Human Brain Reference RNA (HBR) with respect to detection of known exon-exon junctions in the NCBI database. For each library, paired-end sequencing detects more exon-exon junctions for a given number of mapped reads.

Figure 2 Paired-end is superior to forward-only sequencing for detection of exon-exon junctions



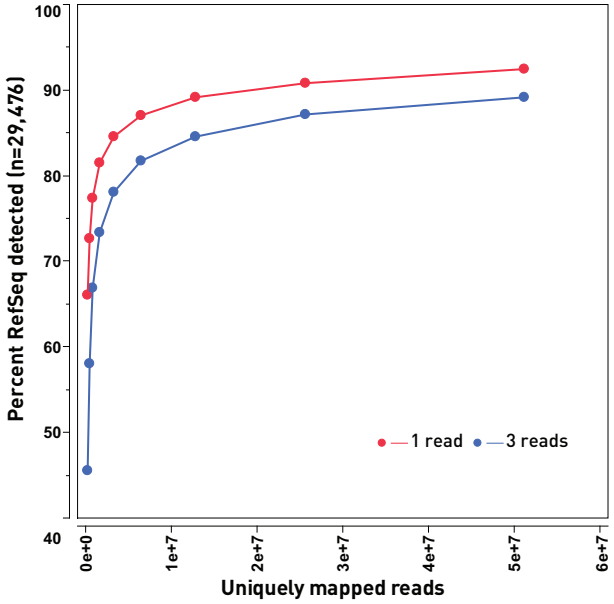
Supplemental data: saturation of reads mapping to a reference

Whole transcriptome analysis

In the example shown in Figure 3, a whole transcriptome library was prepared from poly(A) Human Brain Reference RNA using the SOLiD™ Total RNA-Seq Kit and sequenced in the SOLiD™ System. A total of 50×10^6 mapped reads was used for analysis. The curves represent thresholds of 1 (red line) and 3 (blue line) reads uniquely mapping to RefSeq.

In this experiment, the fraction of known RefSeq sequences detected begins to plateau at about 20×10^6 uniquely mapped reads. The more stringent the calling criteria (higher threshold), the more mapped reads are needed to achieve the same fraction of RefSeq hits.

Figure 3 SOLiD™ whole transcriptome analysis: detection of human RefSeq

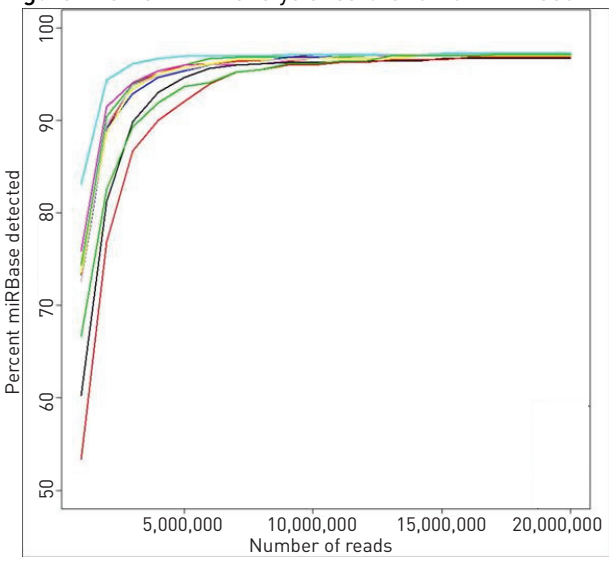


Small RNA analysis

In the example shown in Figure 4, small RNA libraries were prepared from small RNA-containing human RNA from 10 tissues using the SOLiD™ Small RNA Expression Kit (this kit has been replaced by the SOLiD™ Total RNA-Seq Kit), sequenced on the SOLiD™ System, and mapped against the Sanger miRBase reference. The calling threshold for these data was 3 reads.

For most tissues, saturation of the Sanger miRBase reference begins at about 5 million mapped reads. Based on these data, one might choose to aim for 5 million reads for detection of RNAs in miRBase and about 10 million reads for discovery of new small RNAs. More complex questions, such as discovery of iso-miRs, would require more than 10 million reads.

Figure 4 Small RNA analysis: saturation of miRBase

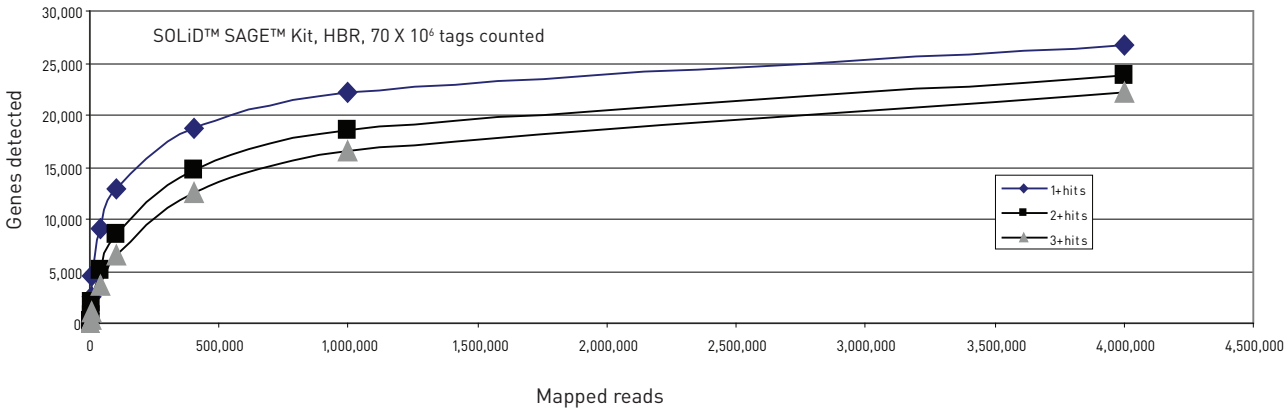


SAGE™ analysis

In the example shown in Figure 5, a SAGE™ library was prepared from human brain reference RNA (HBR) using the SOLiD™ SAGE™ Kit and sequenced on the SOLiD™ System. 70 million reads were counted. The number of unique RefSeq hits was plotted against the total of the mapped reads, using thresholds of 1, 2, and 3 hits to call a transcript present.

While the total number of RefSeq hits continues to increase out to 4 million total mapped reads and beyond, the number of newly detected unique RefSeq hits starts to plateau at ~1 million mapped reads. These data indicate that for routine expression profiling of human RNA, 2 to 5 million mapped reads per library is likely sufficient.

Figure 5 SOLiD™ SAGE™ analysis: detection of human RefSeq



Supplemental data: estimating the lower limit of detection with ERCC Spike-In Mixes

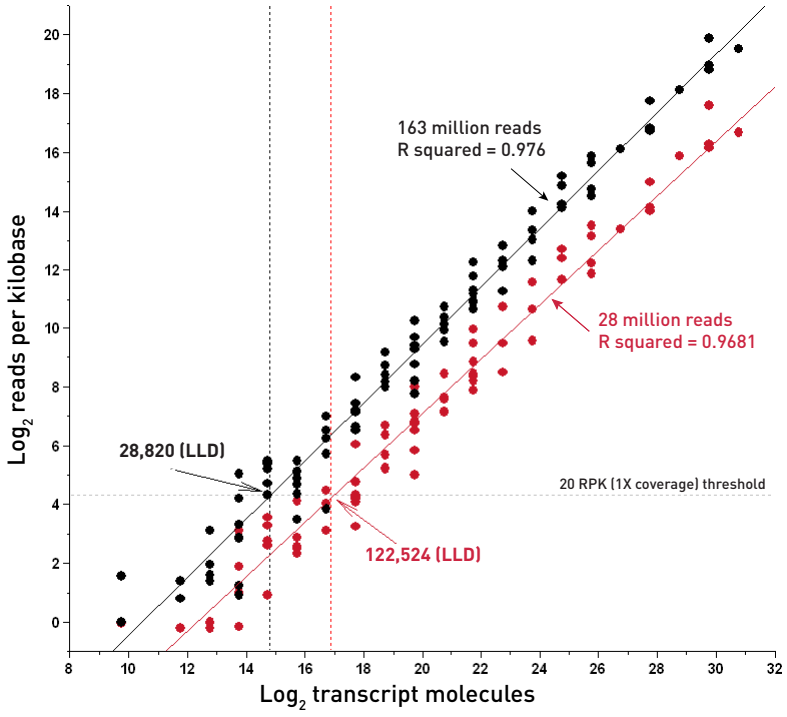
More reads = lower limit of detection (LLD)

In the experiment shown in Figure 6, ERCC Spike-In Mix 1 was added to poly(A) HeLa RNA, and whole transcriptome libraries were prepared and sequenced on the SOLiD™ System. (See "Whole transcriptome libraries: using ERCC RNA Spike-In Control Mixes" on page 9 for background information on the ERCC RNA Spike-In Control Mixes.)

- The dose-response curves for the ERCC transcripts are shown at 28 million and 163 million total reads.
- The X-axis indicates the level of each ERCC transcript ($\log_2$ transcript molecules/ 100 ng poly(A) RNA), and the Y-axis shows $\log_2$ reads per kilobase (RPK).
- The horizontal line is placed at 1× coverage as a threshold for calling a transcript present. This threshold was determined empirically and is designed to help reduce measurement noise and increase confidence in the accuracy of quantification and detection. Coverage is dependent on read length, transcript length and the number of mapped reads. For a 1000-nt transcript, 1× coverage requires 20 reads with a 50-nt read length (20 RPK for 50-nt reads); the same fold coverage could be achieved with 10 reads of 100-nt length.
- The lower limit of detection (LLD) is defined as the X value at the point where the regression line for the dose response data passes the 1× threshold.

These dose-response data demonstrate that more reads give greater detection sensitivity and a lower limit of detection.

Figure 6 Dose-response of ERCC transcripts in HeLa RNA



Detection sensitivity: transcripts per cell equivalent

In Table 8, the LLD calculated in Figure 6 is transformed into an estimate of the number of copies of a transcript that are detectable per cell equivalent.

- The complexity ratio reflects the number of molecules that can be detected as a function of the total RNA molecules. This sensitivity metric is similar in concept to detection in parts per million.

  The complexity ratio is the number of native transcripts estimated in 100 ng HeLa poly(A) RNA (9×10^{10}; assumes mean transcript length is 2 kb) divided by the LLD.
- The detection expressed as copies of transcript per cell is estimated from the complexity ratio based on an average transcript number of 300,000/cell.

A calculation such as that shown in Table 8 is designed to help you decide whether the LLD expected in the sequencing runs is adequate for your experimental needs.

Table 8 Detection sensitivity estimates of ERCC Control transcripts in HeLa poly(A) RNA

Sample	Total reads	Uniquely mapped reads	LLD (ERCC transcripts detected in 100 ng RNA)	Complexity ratio	Detection (copies/cell)
Sample 1	28,200,852	17,485,966	122,524	1:735,000	0.41
Sample 2	163,452,796	99,165,396	28,820	1:3,123,000	0.10

Calculations for the number of libraries per flowchip lane

The following table illustrates detailed calculations to estimate the number of libraries that can be configured on a 5500 Series SOLiD™ Sequencer.

Table 9 Human RNA analysis: theoretical library configuration on 5500 Series SOLiD™ Sequencers

Type of analysis (human RNA)	For this number of mapped reads (10^6)†...	And this % expected mapped reads‡...	You need to deposit this number of P2+ beads (10^6)§...	And you can combine this number of libraries/lane††
Whole transcriptome: counting	≥50	~80‡‡	~63	1§§
Whole transcriptome: discovery	≥100	~80‡‡	~125	1§§
Small RNA: counting	5	35–40	12–14	14–16†††
Small RNA: discovery	10–30	35–40	25–85	2–8†††
SAGE™ (counting)	2–5	~30	6–17	12–33†††

† See "Supplemental data: saturation of reads mapping to a reference" in this Appendix for examples of the experiments used to derive these numbers.

‡ The percentage of all reads that map to the reference sequence. Based on in-house or published experiments with human RNA; the percentage also depends on the quality and rRNA content of the RNA sample, and the reference sequence used.

§ Total P2+ beads required to achieve required number of mapped reads; = Mapped reads/Expected fraction (for example, 30% = 0.3) mapped reads. P2+ beads are beads that have a P2-containing template DNA; for these calculations, it is assumed that >95% of the deposited beads are P2+.

†† Calculation: (Bead density)/(Total P2+ beads). For 1-micron beads, bead density is assumed to be 200×10^6 beads/lane.

‡‡ Based on in-house data using rRNA-depleted RNA samples.

§§ Assumes 75 nt X 35 nt paired-end sequencing run. No more than 1 human transcriptome per lane recommended.

††† Each flowchip lane must include at least one full set of 4 color-balanced barcodes. The last set can be incomplete.

Bibliography

Auer PL, and Doerge RW. 2010. Statistical Design and Analysis of RNA Sequencing Data. *Genetics.* 185: 405–416

Breu H. 2010. A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction. Applied Biosystems white paper. Publication no. 139WP01-02. Search by publication number at **www.appliedbiosystems.com**

Costa V, Angelini C, De Feis I, Ciccodicola A. 2010. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotech.* Article ID 853916; doi:10.1155/2010/853916.

External RNA Controls Consortium. 2005. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* 6:150. Available at **www.biomedcentral.com/1471-2164/6/150**.

Kassahn K, Waddell N, Grimmond SN. 2011. Sequencing transcriptomes *in toto*. *The Royal Society of Chemistry*. DOI: 10.1039/c0ib00062k.

Life Technologies white paper. 2011. SOLiD™ System accuracy with the Exact Call Chemistry module. Publication no. CO31266. Search for <CO31266> at **www.appliedbiosystems.com**.

Marguerat, S., and Bahler, J. 2010. RNA-Seq: from technology to biology. *Cell Mol Life Sci*. 67: 569–579.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L, and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 5: 621–628.

Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 10: 57–63.

Wilhelm, B.T., and Landry, J.-R. 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods.* 48: 249–257.

Glossary

alignment	The process of mapping sequencing reads to a reference genome or sequence.
allele	One of two or more alternative nucleotide sequences at the same location on homologous chromosomes.
alternative splicing	A process whereby exons in a primary transcript are joined in multiple ways as part of the splicing process, resulting in different mature mRNAs.
annotated gene	Within one of several reference databases, a gene sequence that has biological attributes attached that describe structure or function, such as coding regions or biochemical function.
annotation	Biological attributes or metadata that are attached to sequence data or files. Examples include: genes and protein-coding features, and verified variants.
barcode	A short, unique sequence that is incorporated into a library that enables identification of the library during multiplex sequencing.
barcoded library	A library that has a unique barcode sequence incorporated that enables identification of the library during multiplex sequencing.
BC tag	Sequencing data derived from the barcode region of the SOLiD™ templated bead, using forward ligation chemistry. The BC tag is generated using the SOLiD™ FWD2 Seq. Primers. See the *5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference* (Part no. 4465650) for an illustration.
breakpoint	In genomic or gene expression analysis, the junction(s) of structural variations such as inversions, deletions, or insertions.
calling threshold, call stringency	The criteria for calling a genetic variant or novel RNA present in the biological sample.
color balance	The relative proportion of beads in a given sequencing cycle that are called as each of the four colors.
counting	• In genomics, sequence analysis that generates read or tag counts for annotated regions of a reference sequence. • In transcriptomics, expression analysis that focuses on relative or absolute quantification of RNA molecules in a biological specimen.

coverage	• In genomic analysis, the number of aligned (or mapped) sequencing reads that span a position in the reference genome. • In RNA analysis, this term is sometimes used to describe the fraction of the reference sequence that has sequencing reads aligned (or mapped), at a certain calling threshold.
dbSNP	A Single Nucleotide Polymorphism database repository for single nucleotide polymorphisms and short insertion and deletion polymorphisms, hosted by the NCBI.
deletion	A gap in a nucleotide sequence with respect to a reference genome or sequence.
discovery	Analysis that focuses on detection of novel genetic variants or RNA species that are not already present, or are present but not annotated, in a reference database.
downstream, upstream	In transcriptomics, these terms are used with respect to the direction of transcription in a genomic segment. Upstream is to the 5' side of the mRNA, downstream is to the 3' side. In genomics analysis, these terms are used with respect to the 5' side (upstream) or 3' side (downstream) of a specific location on one DNA strand.
epigenomics	Global analysis of changes in the genome that do not involve changes to the nucleotide sequence itself that result in regulation of gene activity or expession. Examples include methylation of the DNA and changes in chromatin structure.
Exact Call Chemistry (ECC)	An optional primer round on the 5500 Series SOLiD™ Sequencer that enables higher accuracy reads and base sequence output (derived without alignment to a reference sequence) from the instrument that is included in the .xsq output file.
exon	In eukaryotic organisms, a segment of a gene that encodes part or all of a protein. Exons may be separated by introns that are spliced out of the primary transcript to produce a mature mRNA for translation into protein. See also *splicing*, *alternative splicing*.
exome	The compilation of annotated exon sequences in a genome.
F3 tag	Sequencing data derived from the P1 end of the template in the SOLiD™ templated bead, using forward ligation chemistry. The F3 tag is generated using the SOLiD™ FWD1 Seq. Primers or the SOLiD™ Small RNA Seq. Primers. See the *5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference* (Part no. 4465650) for an illustration.
F5 tag	Sequencing data derived from the P2 end of the template in the SOLiD™ templated bead, using reverse ligation chemistry. The F5 tag is generated using the SOLiD™ REV1 (DNA) Seq. Primers or the SOLiD™ REV1 (RNA) Seq. Primers. See the *5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference* (Part no. 4465650) for an illustration.
flowchip	The microfluidics chamber housing six lanes in which SOLiD™ templated beads are deposited and through which SOLiD sequencing reagents flow during a sequencing run.

flowchip lane	A region of the flowchip upon which a single sequencing sample is loaded. Each 5500 Series SOLiD™ Sequencer flowchip contains six lanes.
forward read	Sequencing reads in the direction from the P1 to P2 sequence, using forward ligation chemistry.
gene fusion	A section of the genome that maps to an exon from one gene followed by an exon from another gene. It can occur as the result of a translocation, deletion, or chromosomal inversion. A gene fusion junction excludes exon-to-exon boundaries that arise from alternative splicing of a transcript.
genomics	Global analysis of the genome to discern elements involved in regulation of gene activity or expession, with an emphasis on genetic variation such as single nucleotide polymorphisms, small and large insertions and deletions, and other structural variants such as translocations and inversions. Some use the term genomics as an umbrella term that includes transcriptomics, epigenomics, and analysis of the genome.
fusion transcript	An RNA molecule that results from transcription of a gene fusion. See also *gene fusion*.
insert size	The physical size of the genomic DNA segments or RNA molecules represented in a SOLiD™ library. • Fragment libraries: the size of the sheared DNA fragments. • Mate-paired libraries: the length of the genomic DNA segment spanned by the corresponding mate-pair tags. • Whole transcriptome libraries: the size of the RNA fragments.
insertion	An insertion of nucleotide sequence with respect to the reference genome or sequence.
internal adaptor (IA)	The internal adaptor sequence is incorporated into the template during library construction and provides a common hybridization target for SOLiD™ sequencing primers. See the *5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference* for a schematic of sequencing primers compatible with each type of SOLiD™ library. • The IA sequence is different in DNA-source libraries and RNA-source libraries, therefore sequencing primers specific for RNA and DNA libraries must be used for reverse reads (F5 tag). • The IA-containing adaptors used during mate-paired library preparation are different from the adaptors used for fragment library preparation, but the SOLiD™ FWD2 Seq. Primers are used for all forward reads originating in the IA sequence, generating the R3 and BC tags.
intron	The genomic sequence between two exons that is spliced out of a primary transcript prior to translation. See *exon*.
junction	A place where two regions that are not contiguous in the genomic sequence are joined in a single sequenced region under consideration.
lane	See *flowchip lane*.

library	A set of DNA or cDNA molecules prepared from the same biological specimen and prepared for sequencing on the SOLiD™ System.
mapped read	A sequencing read that has been aligned to a reference sequence.
mapping	The process of aligning sequencing reads to a reference genome or sequence.
miRBase	An annotated database archive of miRNA sequences. www.mirbase.org
multiplex sequencing	Sequencing runs in which multiple barcoded libraries are simultaneously sequenced in a single flowchip lane. Each bead is assigned to the correct library after the sequencing run according to the sequence of its barcode.
ncRNA	Non-coding RNA; RNA that has not been shown to encode a protein. Many ncRNAs have been shown to have profound effects on the levels of proteins in the cell that are derived from coding RNAs.
paired-end sequencing	Sequencing runs that acquire sequence from each end of the insert in a DNA fragment or whole transcriptome library, using both forward and reverse reads.
polymorphism	A genetic variant in a population of individuals that may or not may be associated with an observable (phenotypic) trait.
pooled libraries	Barcoded libraries that are combined before templated bead preparation, to then be deposited in a single flowchip lane for multiplex sequencing.
primer set	In the SOLiD™ System, the set of primers that are used sequentially to initiate ligation sequence chemistry.
R3 tag	The R3 tag applies only to mate-paired libraries; sequencing data derived from the mate-pair tag closer to the P2 end of the SOLiD™ templated bead, using forward ligation chemistry. The R3 tag initiates in the IA sequence using the SOLiD™ FWD2 Seq. Primers. See the *5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference* (Part no. 4465650) for an illustration.
read, sequencing read	Sequencing data from a single bead with a single primer set.
reference, reference genome, reference sequence	A sequence against which sequencing reads are aligned.
RefSeq	A multi-organism database archive of DNA, RNA, and protein sequences, hosted by the NCBI.
reverse read	Sequencing reads in the P2 to P1 direction, using reverse ligation chemistry.
RNA-Seq	Gene expression analysis using sequence-based approaches. RNA-Seq can include whole transcriptome analysis, small RNA analysis, and SOLiD™ SAGE™ analysis.

RPKM	The number of reads mapping to a transcript per kilobase of transcript length per million mappable reads. RPKM is used to set a threshold for calling a transcript or new RNA species or isoform "present." 1 RPKM is equivalent to 20 reads mapping to a 1 kb transcript per 20×10^6 mappable reads.
run	Sequencing of beads on one or more flowchips at the same time.
SAGE™ analysis	(Serial Analysis of Gene Expression) Nucleotide sequence analysis seeking to find specific gene expression information using short stretches of cDNA (also known as *tags*) from the 3' ends of RNA molecules. In the SOLiD™ System, the SAGE™ tag is 25–27 bp in length.
SAGE™ library	A SOLiD™ System-compatible library that is prepared from short cDNA segments generated from the 3' ends of RNA molecules.
sample, barcoded sample	In the 5500 Series SOLiD™ ICS, the set of templated beads that will be sequenced in a single flowchip lane. A barcoded sample contains templated beads from up to 96 barcoded libraries.
sequencing primer	An oligonucleotide that is the reverse complement of a designated site on the template strand and which serves as the initation point for subsequent ligation-based sequencing cycles.
single read	Sequence data that generates a specific analysis tag.
small RNA analysis	Global sequence analysis of the small RNA population of a cellular RNA sample; small RNA includes microRNAs (miRNAs), short interfering RNAs (siRNAs), piwi-interacting RNAs (piRNAs), and repeat-associated siRNAs (rasiRNAs).
small RNA library	A SOLiD™ System-compatible library that is prepared from the small RNA fraction of a total RNA sample.
SNP	Single Nucleotide Polymorphisms (SNPs); single base pair variants in genomic DNA or the corresponding RNA transcript.
splicing	The process whereby introns are removed from a primary mRNA, resulting in a mature mRNA that is ready for translation into protein.
splice junction	Exon-to-exon boundaries that arise from splicing of a transcript. See also *gene fusion*.
strandedness	The polarity or orientation of a nucleic acid strand with respect to being sense or antisense. Libraries prepared using the SOLiD™ Total RNA-Seq Kit preserve the strandedness of the original RNA molecule such that F3 tag reads align to the sense strand and F5 tag reads align to the antisense strand.

tag	*Tag* is used in two ways: • Sequencing data from a single bead with a single primer set; sometimes used interchangeably with *read*. • A length of DNA or cDNA to be sequenced; especially, a relatively short stretch of DNA or cDNA that is used to infer information about the longer native molecule from which it is derived, such as in mate-paired library sequencing and SAGE™ analysis, respectively.
tags: BC, F3, F5, R3	Sequencing data derived from specific locations on the SOLiD™ templated bead. See the *5500 Series SOLiD™ Sequencers: Reagents and Consumables Ordering Information Quick Reference* (Part no. 4465650) for an illustration.
templated bead preparation	Process of clonally amplifying template strands on beads by emulsion PCR, enriching the beads to remove beads without template, then modifying the 3' end of the template on the beads to prepare for bead deposition and sequencing
templated bead	A single SOLiD™ P1 DNA Bead with a clonal population of templates for sequencing. Sometimes called *clonal bead*.
transcriptome	The compilation of all transcribed sequences from a genome, both coding and non-coding.
uniquely mapped read	A read that is mapped only once in a genome with a given number of mismatches.
variant	A difference in the nucleotide sequence of interest, with respect to the reference sequence.
whole transcriptome analysis	Global sequence analysis of coding and non-coding RNA transcripts along their entire length.
whole transcriptome library	A SOLiD™ System-compatible library that is prepared from total RNA, rRNA-depleted total RNA, or poly(A) RNA that enables sequence analysis of the transcripts along their entire length.
WTA	See *whole transcriptome analysis*.

Headquarters
5791 Van Allen Way | Carlsbad, CA 92008 USA | Phone +1 760 603 7200 | Toll Free in USA 800 955 6288
For support visit www.appliedbiosystems.com/support

www.lifetechnologies.com

life technologies™