Application Note

Genomics

Agilent

Trusted Answers

Agilent SureSelect XT HS RNA Library Preparation: A Streamlined and Improved Workflow for Direct-to-Capture Construction of RNA-Sequencing Libraries from Fresh or FFPE Samples

Authors

Carsten Carstens, Katherine Felts, and Sarah Johns

Agilent Technologies, Inc.

Abstract

In this application note, we present a condensed and improved workflow for the construction of targeted RNA-sequencing libraries. We introduced four major improvements to the Agilent SureSelect XT RNA Direct protocol:

1. Replacement of an overnight target enrichment hybridization with a 90-minute fast hybridization
2. Elimination of Uracil deglycosylase (UDG) treatment (RNA strand specificity is maintained through use of a new enzyme)
3. Replacement of older SureSelect XT adaptors with SureSelect XT HS adaptors enabling the parallel processing of DNA and RNA from the same sample
4. Providing a unique molecular barcode (MBC) for improved deduplication of PCR and fragmentation duplicates

These changes reduced what was a 2-3 day turnaround time to a 1 day process. Additionally by analyzing fusion data with the use of multiple sample input types (intact, fresh frozen, FFPE), we found that this streamlined workflow produces RNA-sequencing libraries of superior library complexity and improved sequencing performance with as low as 10 ng RNA inputs.

Introduction

The application of high-throughput sequencing to transcriptomics (RNA-sequencing, RNA-seq) has not only enabled global gene expression profiling, but also had added precise information on splice variants, fusion transcripts, post-translational editing

events, and allele-specific expression. The standard approach to generating RNA-seq libraries consists of fragmentation of RNA input by Mg^{2+}-dependent transesterification, random hexamer-primed first-strand cDNA synthesis, followed by second-strand cDNA synthesis with dUTP strand marking. To prevent concatenation, the cDNA ends are blunted and an adenine is added to the 3' termini by nontemplated DNA polymerization. This is followed by ligation of platform-specific sequencing adaptors. All steps after adaptor ligation are shared with the DNA sequencing workflow.

Despite the fact that DNA and RNA-derived sequencing libraries appear identical after adaptor ligation, there are specific differences when comparing RNA-sequencing to DNA sequencing. The first difference is the need to maintain directionality of the sequencing fragments to ascertain to which genomic DNA strand the original RNA corresponds to.This is usually achieved by including uracil in the second-strand synthesis reaction, continuing with library preparation, and at a later step using UDG to remove the second strand. The second difference is the large dynamic range of the transcripts measured in RNA sequencing, which can exceed five orders of magnitude due to the vast differences in the relative abundance of expressed coding transcripts and extremely high abundance of ribosomal RNA (rRNA) compared to coding RNA. RNA-seq therefore requires complexity reduction to avoid wasting reads on noninformative transcripts (ex: rRNA). The most common method involves removal of rRNA by either targeted depletion (ribodepletion) or specific capture of polyadenylated RNA (mRNA). In a typical sample, even after rRNA removal, the top 1% of all expressed genes will account for ≈50% of all transcripts. Therefore, if research goals require the study of low or medium expressed genes, this detection will benefit from further complexity reduction.

As an alternative to targeted depletion approaches, complexity reduction can also be achieved through targeted enrichment using biotinylated target enrichment probes (aka "baits"), a common approach in genomic sequencing. In RNA sequencing, targeted enrichment by bait selection is primarily used in conjunction with FFPE-derived samples, where rRNA depletion is notoriously inconsistent and poly(A) enrichment cannot be used due to fragmentation of the source material[1]. Target enrichment is also beneficial when only a relatively small subset of the transcriptome needs to be examined. A prominent example is the detection of fusion transcripts that are indicative of an underlying gene fusion event. However, any scenario where only a subset of transcripts is informative, such as transcriptional fingerprinting or detection of rare post-transcriptional editing, benefits from targeted enrichment[2].

Agilent offers the SureSelect XT RNA Direct library preparation kit (p/n G7564A, G7564B) to enable construction of targeted RNA-sequencing libraries. We have demonstrated successful use of this kit working with FFPE-derived samples[3]. Here we describe a streamlined and improved workflow by combining components of the SureSelect XT RNA Direct library preparation kit with the SureSelect XT HS target enrichment kit (p/n G9706A) in conjunction with a modified protocol. The workflow was simplified by eliminating lyophilization of input RNA, elimination of one SPRI bead purification step, and replacing the 24-hour bait hybridization step with a 1.5-hour fast hybridization step. We also eliminated the UDG treatment step, achieving strand specificity instead through the use of a PCR enzyme that discriminates against uracil-containing template DNA during the precapture PCR. In addition, the use of SureSelect XT HS sequencing adaptors adds a MBC for identification of fragmentation duplicates and lowers the practical barriers to split tube (aka parallel) processing of DNA and RNA from the same sample.

Experimental

RNA sources

The universal human reference RNA (UHRR) was obtained as fresh frozen material from Agilent Technologies (Santa Clara, CA, USA p/n 750500-41). A matched set of breast tumor and normal adjacent tissue was obtained as both fresh-frozen and FFPE material from CureLine human biospecimen CRO (Brisbane, CA, USA, custom part number). Seraseq FFPE tumor fusion RNA reference material v2 was purchased from SeraCare (Gaithersburg, MD, USA, cat. # 0710-0129).

RNA isolation

Where required, RNA was isolated using the RNeasy FFPE kit or RNeasy mini kit from Qiagen according to the manufacturer's instructions (Qiagen USA, Germantown, MD, USA, p/n 73504 and 74104, respectively). For a more detailed protocol, see the appendix.

Quality assessment of input material and sequencing libraries

Nucleic acid samples were assessed on the Agilent 2100 Bioanalyzer system (Agilent Technologies, p/n G2939B) using either the Agilent RNA 6000 Pico kit (Agilent Technologies, p/n 5067-1513) for RNA quality scoring, or the Agilent DNA 1000 kit (Agilent Technologies, p/n 5067-1504), for assessing the quality of the sequencing libraries.

Other materials

Actinomycin D was purchased from Sigma (St. Louis, MO, USA, p/n A1410) and made up to a concentration of 4 µg/µL in DMSO stock solution. SPRI bead purifications were carried out with AMPure XP beads (Beckman Coulter, Atlanta, GA, USA, p/n A63880). Capture of the biotinylated probes was carried out using Dynabeads

MyOne streptavidin T1 beads (Thermo Fisher Scientific, Waltham, MA, USA, cat. # 65601).

SureSelect XT HS RNA library preparation

Construction of RNA-sequencing libraries was carried out using the SureSelect XT RNA Direct kit (Agilent Technologies, p/n G7564A) and the SureSelect XT HS target enrichment system for the Illumina paired-end multiplexed sequencing library (Agilent Technologies, p/n G9706A). For a detailed description, please see the appendix.

SureSelect XT RNA Direct library preparation

RNA Direct libraries were generated, enriched, and sequenced following the instructions in the SureSelect XT RNA Direct library preparation protocol (manual).

Target enrichment

Target enrichment of SureSelect XT HS RNA libraries was carried out using SureSelect human all exon V7 exome (Agilent Technologies, p/n 5191-4029) targeting the coding transcriptome. A detailed protocol for bait capture is given in the appendix.

Sequencing and data analysis

Sequencing libraries were analyzed on an Illumina HiSeq 4000 by paired-end sequencing using a 2 × 150 read format. For expression analysis (data not shown), FASTQ files were aligned to the transcriptome using the splice-aware STAR version 2.6.0a package using genome build hg38 as a reference. Expression profiles were then generated from the STAR alignment output using the RSEM tool. General library statistics (strand specificity, 5'-3' end bias, MBC-blind duplication rates, library size estimates) were generated using the Picard RNA analysis pipeline with duplicates marked using .bam files that were downsampled to 2 × 10^7 reads to generate normalized duplication rates. MBC-corrected duplication statistics and library size estimates were generated using the same pipeline but using UmiAwareMarkDuplicatesWithMateCigar for removing fragmentation duplicates. Fusion transcripts were scored using STAR-Fusion and visualized using the FusionInspector Tool, which is part of the Trinity Cancer Transcriptome Analysis Tool kit (CTAT)[4].

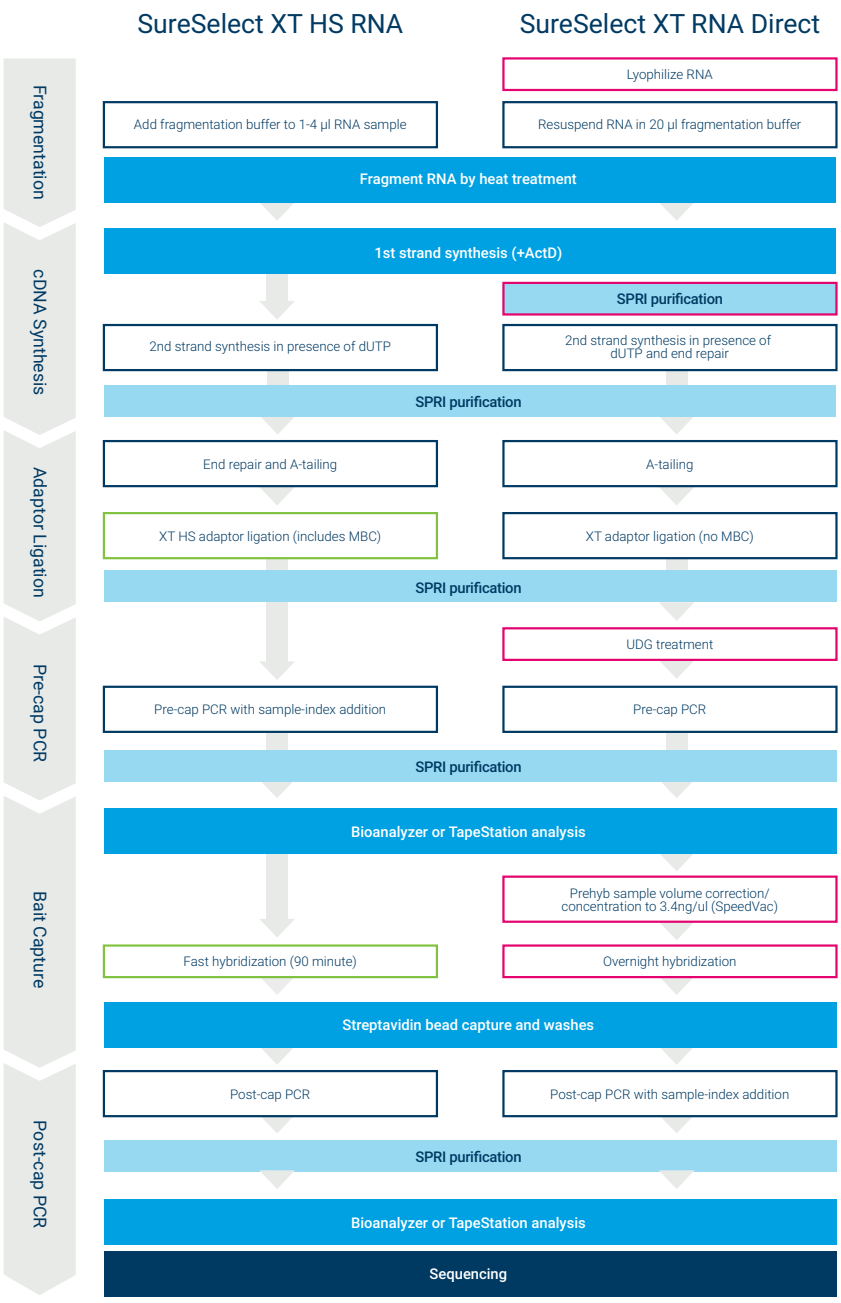

Figure 1. Comparison of the Agilent SureSelect XT HS RNA workflow (RNA XT HS) with the Agilent SureSelect XT RNA Direct library preparation kit (RNA Direct). Steps eliminated from the RNA Direct workflow are indicated by a pink border. Steps that provide improved functionality to the RNA XT HS protocol are indicated by a green border.

Results and discussion

Overview of RNA XT HS and RNA Direct workflows

We wanted to develop a simplified and improved workflow for generation of high-quality RNA-sequencing libraries by combining components from the SureSelect XT RNA Direct library preparation kit (RNA Direct) with the SureSelect XT HS target enrichment kit. This resulting workflow was named SureSelect XT HS RNA (RNA XT HS). An overview comparing these workflows is shown in Figure 1. The RNA XT HS workflow has obvious advantages over the RNA Direct workflow, including elimination of UDG treatment and replacement of traditional overnight hybridization with a 90-minute fast hybridization, greatly reducing turnaround time. We also wanted to determine if it would be possible to remove the SPRI purification step between the first and second strand synthesis and avoid lyophilization of RNA input in the RNA XT HS workflow (see Appendix for details).

Comparing workflow performance

To compare the performance of both library preparation processes, we first generated libraries using varying amounts of universal human reference RNA (UHRR) and SeraCare FFPE tumor fusion RNA reference material v2 (SeraCare) as input material representing an intact and an idealized FFPE sample, respectively. The RNA XT HS and RNA Direct libraries were enriched with the SureSelect human all exon V7. Lastly, these enriched libraries were sequenced using an Illumina sequencer and data analyzed using customized data analysis pipelines (See the Experimental section for details).

Table 1 shows a summary of the global sequencing statistics resulting from analysis of the RNA XT and RNA Direct sequencing data. We found that sequenced RNA XT HS libraries are indistinguishable from RNA Direct libraries with regards to several metrics, most noticeably high strand specificity (>98%) and low rRNA contamination (~0.1%). The high strand specificity of RNA XT HS libraries demonstrates the effectiveness of the RNA XT HS workflow's approach of eliminating UDG treatment in favor of utilizing a PCR enzyme that does not amplify uracil-containing templates. Strand specificity tends to be slightly better for intact versus FFPE samples, which is expected due to the compromised quality of the FFPE input material. Regardless, the strand specificities observed with FFPE input by either protocol are still very high (> 98% for FFPE material). The consistently low percentage of rRNA contamination in both library preparation methods has been previously shown in targeted enrichment approaches[3].

When comparing exonic mapping rates between RNA XT HS and RNA Direct libraries, we again find comparable performance (Figure 2). A critical step in shortening the new RNA XT HS workflow is replacement of the traditional overnight hybridization with a fast hybridization step. As shown in Figure 2, the accelerated capture does not affect the mapping rates, and we consistently observed exonic rates of 90% with very few reads mapping to intergenic regions, regardless of input material. Transcripts targeted by the V7 capture library accounted for 92.9–94.1% of all

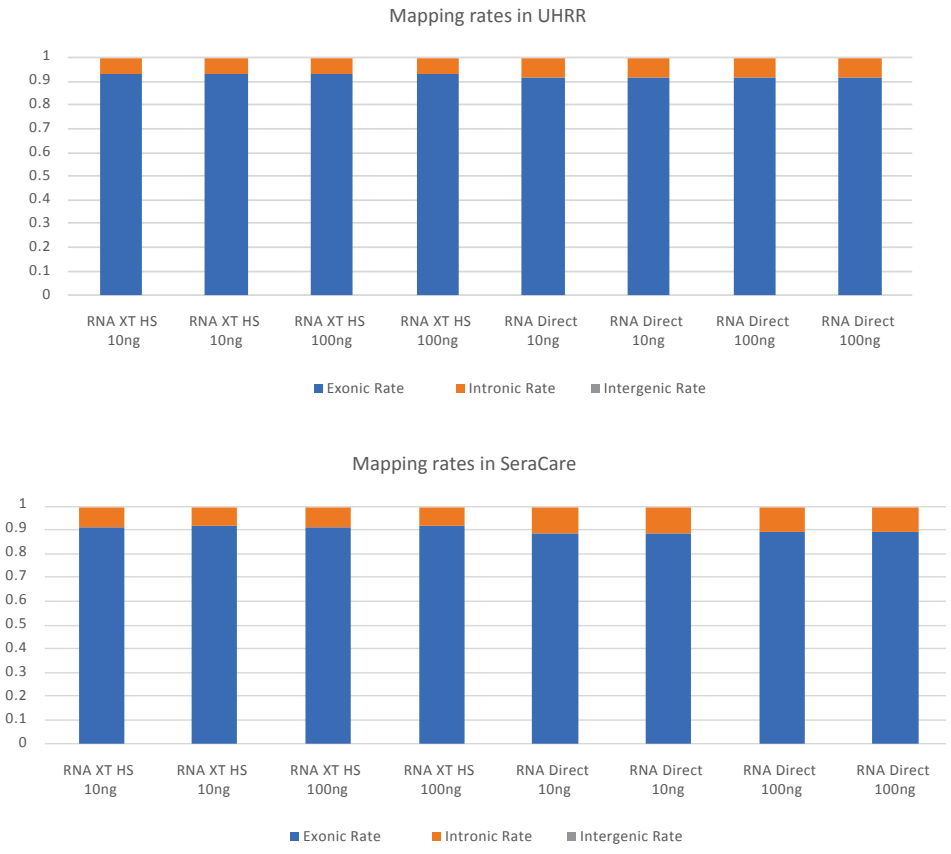

Figure 2. Mapping rate comparison of RNA XT HS and RNA Direct sequenced libraries. RNA-seq libraries were prepared from UHRR (A) and SeraCare samples (B) using the RNA XT HS or RNA Direct workflow. Mapping rates of the Agilent SureSelect human all exon V7 exome enriched sequencing libraries to exonic, intronic, and intergenic sequences are shown. The intronic rate is suspected to reflect the sequencing unprocessed mRNA.

observed expression regardless of input amount or sample type. The remaining ~7% are mostly due to contaminating, highly expressed transcripts, such as mitochondrial genes, or annotation errors.

A critical indicator of the overall library preparation process efficiency is the estimated starting library complexity. Larger and hence more complex libraries are preferable, as sequencing these libraries provides greater insights and more confidence in quantification. Library complexity is calculated from the sequencing depth (in read pairs) and the number of unique variants observed, derived from the duplication rate. We used the Picard pipeline, which assumes all duplicates to be PCR duplicates, to calculate the estimated library complexities. This approach underestimates true library complexity and will be discussed in more detail in the results section. The projected resulting library complexities are listed in Table 1 and shown in Figure 3. When comparing the RNA XT HS and RNA Direct workflows, we found no significant differences between libraries constructed from high amounts (100 ng) of fresh and FFPE RNA input. However, when the input was reduced to 10 ng for either UHRR or SeraCare, we found noticeable differences in library complexity. First, these lower input libraries are smaller and less complex than the higher input libraries. This is expected, as the lower the amount of RNA input one tries to convert into a sequencing library, the smaller the final library should be. We also found that at low input, the streamlined RNA XT HS workflow produces libraries that were about 1.5–2-fold more efficient than the RNA Direct workflow. There are multiple factors in the RNA XT HS workflow that are under investigation (data not shown), which lead to this gain in efficiency.

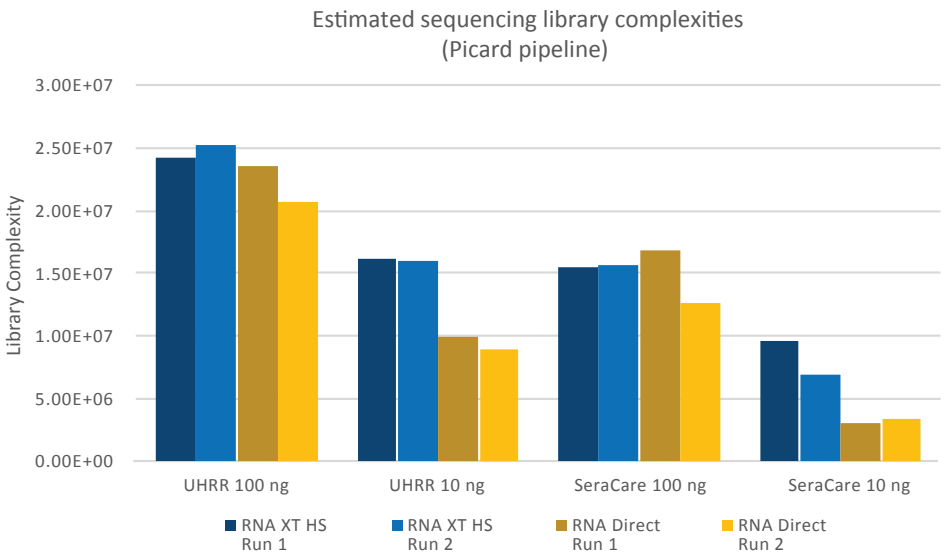

Figure 3. Library complexity differences between RNA XT HS and RNA Direct libraries at low sample input. Sequencing libraries prepared using the indicated sources and input amounts were sequenced on an Illumina HiSeq 4000 platform by paired-end sequencing. Estimated library complexities were determined based on the observed duplication rates and respective number of read pairs determined by the Picard RNA analysis tool. This pipeline does not discriminate between PCR and fragmentation duplicates.

Table 1. Global sequencing statistics of SureSelect XT HS RNA and SureSelect XT RNA Direct libraries: RNA-sequencing libraries generated from the same input material using either the RNA XT HS or RNA Direct protocol were sequenced on the Illumina HiSeq 4000 platform. After down-sampling to 2 × 107 reads, library statistics were generated using the Picard RNA analysis tool. MBC-corrected statistics were generated with the same pipeline after tagging MBCs with UmiAwareMarkDuplicatesWithMateCigar. (A) Fresh-frozen ("intact") RNA input; (B) FFPE input.

A. Universal human reference RNA (UHRR)

Protocol	SureSelect XT HS RNA				SureSelect XT RNA Direct			
Sample input (ng)	100	100	10	10	100	100	10	10
Reads analyzed (million)	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9
rRNA rate (%)	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%	0.1%
Strand specificity (%)	99.1%	99.1%	99.0%	99.0%	99.1%	99.2%	99.1%	99.1%
Duplication rate (%)	17.3%	16.7%	24.3%	24.7%	17.9%	20.0%	35.8%	38.7%
Estimated library size (x 10^6)	24.2	25.2	16.2	15.9	27.5	23.6	10.3	9.2
MBC-corrected duplication rate (%)	7.1%	6.4%	17.1%	17.8%	NA	NA	NA	NA
MBC-corrected estimated library size (x 10^6)	65.9	72.8	24.6	23.6	NA	NA	NA	NA

B. Seraseq FFPE tumor fusion RNA reference material v2 (SeraCare)

Protocol	SureSelect XT HS RNA				SureSelect XT RNA Direct			
Sample input (ng)	100	100	10	10	100	100	10	10
Reads analyzed (Million)	18.6	18.6	18.7	18.6	19	18.8	18.9	18.9
rRNA rate (%)	0.1%	0.2%	0.1%	0.1%	0.1%	0.2%	0.1%	0.1%
Strand specificity (%)	98.6%	98.6%	98.6%	98.7%	98.9%	98.9%	98.9%	98.9%
Duplication rate (%)	24.8%	24.7%	36.0%	45.0%	23.5%	29.4%	69.1%	67.2%
Estimated library size (x 10^6)	15.5	15.5	9.6	6.9	18.6	13.3	3	3.3
MBC-corrected duplication rate (%)	13.2%	13.7%	29.3%	38.8%	NA	NA	NA	NA
MBC-corrected estimated library size (x 10^6)	32.5	31.2	12.7	8.66	NA	NA	NA	NA

Impact of unique molecular identifiers (MBC) on RNA-sequencing libraries

As mentioned above, accurate measurement of library complexities is critical for evaluating the efficiency of a library preparation process. The above library complexity estimates were dependent on the assumption that read pairs with the same start and stop are PCR duplicates derived from the same original library molecule. However, duplicates can also arise due to random fragmentation, resulting in two independent fragments with the same ends. Since fragmentation duplicates are true independent sequencing library members, determination of library complexities should only be made on the basis of true PCR duplicates.

Compared with DNA sequencing, RNA-sequencing may result in a higher number of fragmentation duplicates, due to the high expression of some genes increasing the random chance of fragmentation duplicates. This results in more inaccurate library complexity estimates. A potential advantage of the RNA XT HS workflow is that libraries are constructed with XT HS adaptors that contain single 10 bp MBCs. We hypothesized that this MBC could be used to distinguish PCR and fragmentation duplicates. We reanalyzed the above RNA XT HS sequencing data using a modified data analysis pipeline that takes advantage of MBC data (note: RNA Direct libraries do not include a MBC and hence were not included in this reanalysis). The output from this analysis is reported in Table 1 and Figure 4.

As shown in Figure 4, correcting for fragmentation duplicates results in higher estimates for the corresponding library complexities, especially for higher input amounts where the library complexity is underestimated by ~threefold if fragmentation duplicates are not considered. The greater bias observed for bigger libraries and the corresponding higher input amounts is anticipated due to the greater chance to observe fragmentation duplicates if more molecules derived from the same coding sequence are processed into libraries.

We next wanted to see the impact of MBC correction in a set of "real life" samples. Hence, we generated RNA XT HS libraries using a matched set of a mammary tumor and adjacent normal tissue both stored and treated as fresh-frozen as well as FFPE samples. These libraries were enriched, sequenced, and analyzed as above with the UHRR and SeraCare RNA XT HS libraries. The summary statistics of these libraries can be found in Table 2 and Figure 4. We see excellent strand specificity and rRNA rate in tumor/normal fresh-frozen and FFPE sequencing data comparable to our UHRR and SeraCare libraries. We also observed that FFPE libraries are significantly smaller than fresh-frozen

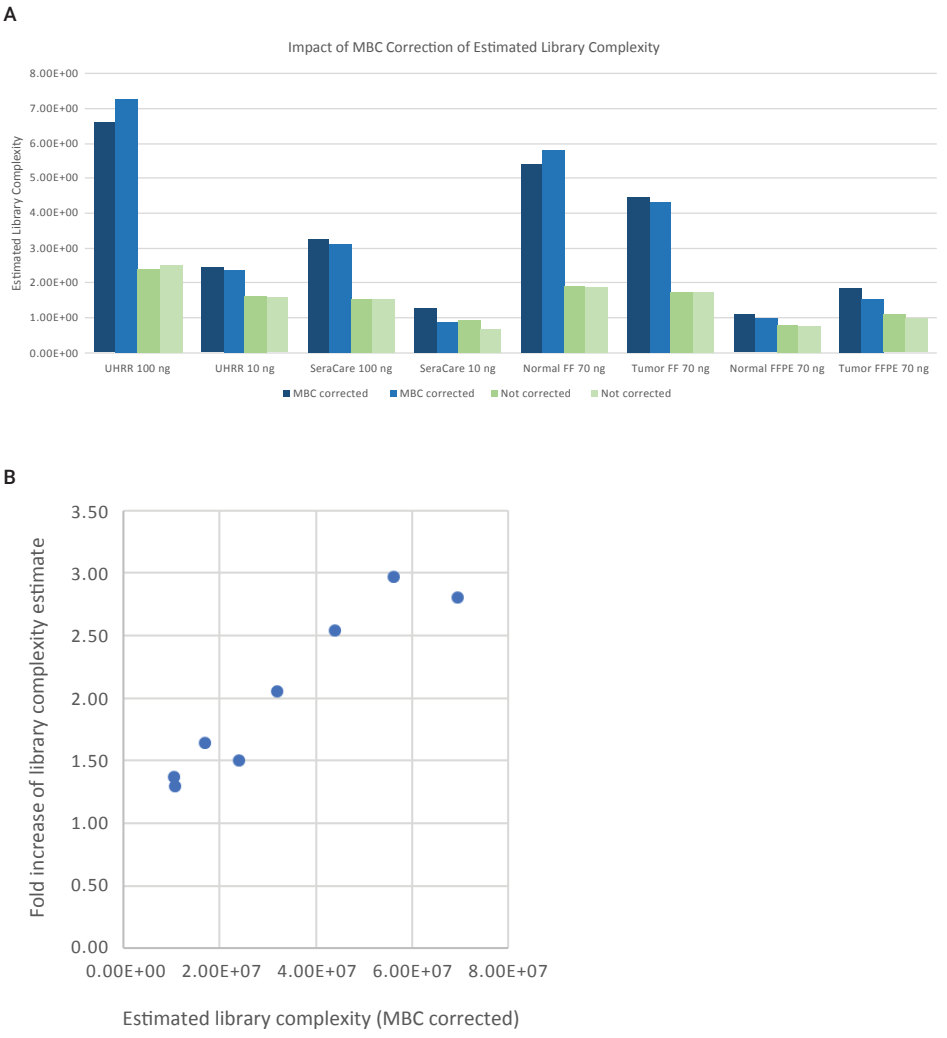

Figure 4. Impact of unique molecular identifiers (MBC) correction on library size estimates. The sizes of RNA XT HS libraries were calculated either by Picard deduplication (not corrected) or after using MBCs (MBC corrected). Picard deduplication assumes that all duplicates are PCR duplicates, whereas MBC correction distinguishes between PCR and fragmentation duplicates. **A)** Estimated library complexities for several RNA XT HS sequencing libraries derived from a range of different samples and inputs. **B)** Differences between MBC corrected and not corrected library complexities are more pronounced for larger libraries and therefore correspondingly higher input amounts.

libraries, which is a well-reported phenomenon. When we analyzed our FFPE sequencing data using our MBC corrective approach, we saw improvements of 0.3–3.5 fold in library estimation complexity and therefore recovered data that would have been lost if we had assumed all duplicates were PCR duplicates versus PCR and fragmentation duplicates. This shows that in order to maximize sequencing information delivered from FFPE RNA samples, it is critical that PCR and fragmentation duplicates are accurately identified in these samples.

Detection of gene fusions

Targeted RNA-seq has been shown to have utility in the detection of gene fusion events especially with challenging samples (FFPE). We wanted to determine if our RNA XT HS approach would work for this use case under the conditions of enriching with exome probes (versus a fusion-specific panel) and maximizing sequence length. As mentioned above, our RNA XT HS libraries from all samples (UHRR, SeraCare, fresh frozen, and FFPE) were subjected to targeted enrichment using the SureSelect human all exon V7 and enriched libraries were sequenced as a paired-end library at 2 × 150 read length. The tumor/normal fresh-frozen and FFPE sequencing data was analyzed using STAR-Fusion and visualized with FusionInspector.

The FusionInspector output detected a range of putative fusions including known false positives such as VDJ recombination events, indicating the presence of immune cells in some samples. The VDJ recombinants were removed from Table 3. Data analysis shows 15 potential gene fusions present in our samples, with 12 of 15 fusion partners previously reported as being involved in tumor-related gene fusions, albeit not in the observed combination. One of the fusions, FCHSD2-FAM168A (highlighted grey Table 3), has previously

Table 2. Global sequencing statistics of fresh-frozen and FFPE Agilent SureSelect RNA XT HS libraries. RNA-sequencing libraries generated from a matched set of tumor and normal adjacent tissue both as fresh-frozen tissue and FFPE using the RNA XT HS protocol. The libraries were sequenced on the Illumina HiSeq 4000 platform. After down-sampling to 2 × 10^7 reads, library statistics were generated using the Picard RNA analysis tool either with or without tagging the MBC.

Sample Type	Fresh Frozen				FFPE			
Source	Normal	Normal	Tumor	Tumor	Normal	Normal	Tumor	Tumor
Sample input (ng)	70	70	70	70	70	70	70	70
Reads analyzed (million)	18.8	18.8	18.7	18.8	18.6	18.6	18.7	18.7
rRNA rate (%)	0.2%	0.2%	0.2%	0.2%	0.2%	0.2%	0.1%	0.1%
Strand specificity (%)	99.0%	99.0%	99.1%	99.2%	98.4%	98.4%	98.8%	98.8%
Duplication rate (%)	21.0%	21.0%	23.0%	23.0%	41.0%	43.0%	33.0%	35.0%
Estimated library size (x 10^6)	19	18.8	17.3	17.3	7.9	7.4	10.9	9.8
MBC-corrected duplication rate (%)	8.0%	8.0%	10.0%	10.0%	33.0%	35.0%	22.0%	25.0%
MBC-corrected estimated library size (x 10^6)	53.9	58.2	44.7	43.1	11	9.93	18.6	15.4

Table 3. Detection of gene fusions in fresh-frozen and FFPE normal/tumor samples. RNA XT HS sequencing libraries were sequenced and then analyzed for gene fusions using a STAR-Fusion pipeline. Putative fusions detected in one or more samples are listed on each row along with evidence supporting the detection of this fusion. Junction reads and supporting reads are expressed as reads per million unique library read pairs. Each column represents the average of two technical replicates. The fusion highlighted in grey is identified in the TGCA database as a fusion associated with breast cancer and may be a driver mutation for this tumor. Observed VDJ recombination events indicative of immune cells present in the samples were removed from the analysis. Detection is based on analysis of 0.7–1.4 × 10^7 unique read pairs for each sample.

	Fresh Frozen				FFPE			
	Normal		Tumor		Normal		Tumor	
Putative Fusion	Junction	Support	Junction	Support	Junction	Support	Junction	Support
MYLK--LPAR6	0.00	0.00	8.86	0.86	0.00	0.00	1.68	0.00
FCHSD2--FAM168A	0.00	0.04	5.60	0.21	0.00	0.00	2.79	0.04
FAM157A--RB1	0.00	0.00	2.59	0.04	0.00	0.00	0.91	0.00
RP4-565E6.1--HYDIN	0.00	0.00	1.47	0.00	0.00	0.00	1.33	0.00
CDR2--FRG1	0.00	0.00	0.42	0.00	0.00	0.00	0.48	0.00
SLC7A5--RP11-645C24.2	0.00	0.04	0.33	0.20	0.00	0.00	0.09	0.04
PHRF1--TXNDC5	0.00	0.00	2.78	0.17	0.00	0.00	0.00	0.00
NUDT1–AC004840.8	0.00	0.00	1.71	0.00	0.00	0.00	0.00	0.00
MPZL1–RCSD1	0.00	0.00	1.34	0.06	0.00	0.00	0.00	0.00
CCDC66–SLMAP	0.00	0.00	0.96	0.00	0.00	0.00	0.00	0.00
UBE2Q2–C15orf27	0.00	0.00	0.77	0.06	0.00	0.00	0.00	0.00
POLR2J–AC004980.9	0.00	0.00	0.00	0.00	11.53	0.00	0.00	0.00
POLR2J–UPK3B	0.00	0.00	0.00	0.00	3.96	0.00	0.00	0.00
RP11-634B7.4–TRIM58	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RB1–MYLK-AS1	0.00	0.00	0.16	0.03	0.00	0.00	0.00	0.00

been identified as associated with breast cancer and may potentially be a driver mutation for this tumor[5]. For all other tumor-specific fusions, at least one fusion partner (usually both) is listed in the TCGA database and is associated with tumor formation. We find the fusion transcript A:B. both, A and B are found in the tumor database. However, there is no listing for A:B. There are examples of A:C and D:B fusion transcripts though. This data suggests that RNA XT HS libraries enriched with an exome enable researchers to identify fusions at the whole transcriptome level, even in challenging samples.

Conclusion

In order to enable more efficient construction of targeted RNA-seq libraries, we set out to improve the current SureSelect XT RNA Direct workflow by utilizing components of the SureSelect XT RNA Direct library preparation kit with the SureSelect XT HS target enrichment kit. The new SureSelect XT HS RNA workflow incorporates several significant workflow improvements:

- RNA input lyophilization was replaced by direct addition of the fragmentation buffer to the RNA sample.
- The SPRI bead purification step after the first strand cDNA synthesis was eliminated.
- An alternative PCR enzyme was used that does not amplify uracil-containing templates, eliminating the need for UDG treatment in maintaining strand specificity.
- The probe ("bait") capture procedure was changed from a 24-hour hybridization to a 1.5-hour fast hybridization protocol.
- The XT sequencing adaptors were replaced by the XT HS adaptors, which enable the use of a MBC for more accurate library complexity estimation and "recovery" of sequencing reads.

We found that the RNA XT HS workflow reduces library and target enrichment turnaround time from 2–3 days to 1–2 days. The new, streamlined workflow yields RNA-seq libraries that are indistinguishable from libraries generated by the SureSelect XT RNA Direct protocol in terms of strand specificity, rRNA rates, and mapping rates at high input amounts. At lower input (10 ng) amounts, we found that the shortened workflow did not affect the overall performance, and in fact appears to be more efficient. We also found that the inclusion of MBCs in the improved adaptor (XT HS) design now enables the identification of fragmentation duplicates. This enhanced identification of fragmentation duplicates improves sequencing output by preventing loss of reads encountered when using standard start stop duplication methods.

When we tested "real life" samples, we found that the RNA XT HS workflow continued to produce high-quality data even with FFPE samples. Preliminary gene fusion analysis detected potential gene fusions in our fresh-frozen and FFPE tumor samples, suggesting a potential real-world use case for sequencing FFPE RNA using RNA XT HS. Although we did not show gene expression or splicing data, our preliminary analysis indicates that targeted RNA seq can be used in global gene expression analysis, splice variant detection, variant expression detection, and allele-specific expression analysis. Lastly, the development of the SureSelect RNA XT HS workflow "aligns" it with the DNA SureSelect XT HS target enrichment kit, making it easier to sequence DNA and RNA from the same sample in parallel. This could be beneficial in several applications, including general multiomics research and parallel processing of samples for TMB-MSI analysis and fusion detection.

Abbreviations

FFPE, formalin-fixed, paraffin embedded; TPM, transcripts per kilobase million; nt, nucleotide; UHRR, universal human reference RNA; MBC, molecular barcode; UDG, uracil deglycosylase;

References

1) Cieslik, M., *et al.* The Use of Exome Capture RNA-Seq for Highly Degraded RNA with Application to Clinical Cancer Sequencing. *Genome Res.* **2015**, 25, 1372–1381.

2) Mittempergher, L., *et al.,* MammaPrint and BluePrint Molecular Diagnostics Using Targeted RNA Next-Generation Sequencing Technology. *J. Mol. Diagn.* **2019**, 21, 808–823.

3) Jones, J. C.; Alex Siebold, A.; Lucas, A. B. SureSelect XT RNA Direct Protocol Provides Simultaneous Transcriptome Enrichment and Ribosomal Depletion of FFPE RNA. Agilent Technologies application note, publication number 5991-8119EN, **2017**.

4) Haas, B., *et al.* STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. bioRxiv. 120295 (**2017**).

5) Hu, X., *et al.* TumorFusions: An Integrative Resource for Cancer-Associated Transcript Fusions. *Nucleic Acids Res.* **2018**, 4, 46(D1), D1144-D1149.

Appendix

1. Detailed protocol for the construction of RNA-sequencing libraries

RNA sample preparation

Total RNA from FFPE curls was isolated using a Qiagen RNeasy FFPE kit according to the manufacturer instructions. RNA from frozen tissue was isolated using a Qiagen RNeasy mini kit. All total RNA samples were analyzed on an Agilent 2100 Bioanalyzer system using an RNA 6000 Pico kit. Samples were heated to 80 °C for 2 minutes prior to loading on the chip. RIN and DV200 values were calculated by the Bioanalyzer software. These sample quality metrics for all the total RNA samples tested are recorded in Table S1.

Table S1. Sample quality metrics and fragmentation conditions.

Sample Description	RIN	DV200	Fragmentation
Universal human reference RNA (UHRR)	9.2	94%	94 °C, 8 min
Seraseq FFPE tumor fusion RNA reference material v2	2	54%	94 °C, 3 min, 65 °C, 2 min
Breast normal frozen	6.3	94%	94 °C, 8 min
Breast tumor frozen	5	88%	94 °C, 8 min
Breast normal FFPE	2.1	48%	65 °C, 5 min
Breast tumor FFPE	2	47%	65 °C, 5 min

Preparation of cDNA using reagents from the Agilent SureSelect XT RNA Direct kit

Note: A 4 µg/µL stock of actinomycin-D in DMSO was prepared in advance and stored frozen in single use aliquots (3 µL) at –20 °C.

Note: The fragmentation mix, first strand master mix, and second strand enzyme and oligo mixes from the SureSelect XT RNA Direct kit were thawed on ice and vortexed for 5 seconds on high speed, then spun down briefly before use.

1. Total RNA samples were prepared in a 4 µL volume of nuclease-free water. Input amounts varied between experiments and are indicated in the presentation of results.

Note: Smaller RNA sample input volumes (< 4 µL) are allowable but larger RNA sample input volumes (>4 µl) are not recommended.

2. Fragmentation mix was added to the RNA sample to achieve a final sample volume of 20 µL.
3. The RNA sample was fragmented by heating in a SureCycler 8800 (or equivalent thermal cycler) using conditions based on the quality metrics of the individual RNA sample as recommended in the SureSelect XT RNA Direct kit protocol (G9691). The fragmentation parameters used for each sample are indicated in Table S1. After fragmentation, samples were stored on ice until ready to proceed to the first-strand synthesis step.
4. A 4 µg/µL stock of actinomycin-D in DMSO was diluted in water to 120 ng/µL (3 µL actinomycin-D + 97 µL water).
5. An eight-sample bulk reaction mixture was prepared for first-strand synthesis (Table S2). The reaction mix was vortexed and kept on ice until use.
6. 8.5 µL of the first-strand reaction mix was added to each 20 µL fragmented sample on ice. Samples were mixed by vortex and spun briefly.
7. The 28.5 µL reactions were incubated in a preprogrammed SureCycler 8800 for 10 minutes at 25 °C, then 40 minutes at 37 °C, then stored at 4 °C (or on ice) until proceeding to the second-strand synthesis.
8. Both the second strand + end repair enzyme mix and the RNA-seq second strand + end repair oligo mix tubes were vortexed before use.
9. 25 µL of the second strand + end repair enzyme mix (blue cap) was added to the 28.5 µL first-strand reaction on ice.
10. Immediately following, 5 µL of the second strand + end repair oligo mix (yellow cap) was added.
11. Samples were capped, mixed by vortexing, spun briefly, and then returned to ice.

Table S2. First-strand reaction mixture.

Reagent	Volume for One Reaction	Volume for Eight Reactions + 10% Excess
RNA-seq first strand master mix	8 µL	70.4 µL
Actinomycin-D (120 ng/µL)	0.5 µL	4.4 µL
Total	**8.5 µL**	**74.8 µL**

Table S3. SPRI wash protocol parameters.

AMPure Bead Volume	105 µL (1.8X volume)
Beads Incubation Time	5 minutes
70% Ethanol Wash (Perform Twice)	200 µL
Dry time @ 37 °C	1–2 min or less
Elution Volume	50 µL nuclease free water

12. The 58.5 µL reactions were incubated in a preprogrammed SureCycler 8800 for 60 minutes at 16 °C, then stored at 4 °C (or on ice) until proceeding to SPRI purification.
13. After the second-strand synthesis, the cDNA was SPRI purified using the protocol outlined in Table S3.

Preparation of SureSelect XT HS cDNA libraries using reagents from the SureSelect XT HS library preparation kit

14. End repair, dA-tailing, and XT HS adaptor ligation of the cDNA was performed using the reagents and instructions detailed in the SureSelect XT HS target enrichment system for Illumina paired-end multiplexed sequencing kit and protocol (G9702).

Note: The SureSelect XT HS protocol was followed starting at Step 3 on page 27 with the exceptions and modifications outlined in steps 15 to 20 of this protocol.

15. At Step 6 on page 34 of the SureSelect XT HS Protocol. Precapture PCR amplification of the libraries was performed according to the instructions in that protocol, with the modified numbers of PCR cycles indicated below:
    a. UHRR, Seraseq™ v2 for 12 PCR cycles (high-quality RNA)
    b. Breast quad samples for 14 PCR cycles (low-quality RNA)
16. PreCap libraries were assessed by Bioanalyzer or TapeStation for yield and distribution of library molecule sizes.

Target enrichment and sequencing of SureSelect XT HS cDNA libraries

17. At Step 1 on page 46 of the SureSelect XT HS protocol. Target enrichment was performed using 200 ng of PreCap library as input, SureSelect XT HS fast hybridization reagents were used. 5 µL of SureSelect human all exon V7 exome probes were used for hybridization.
18. Streptavidin bead capture and subsequent washes were performed according to the protocol.
19. Postcapture PCR amplification was performed according to the protocol using 12 PCR cycles for all samples. PostCap libraries were assessed by Bioanalyzer or TapeStation for yield and distribution of library molecule sizes.
20. All libraries were sequenced on an Illumina HiSeq 4000 platform at 2 × 150 read length.

www.agilent.com

For Research Use Only. Not for use in diagnostic procedures.

This information is subject to change without notice.

PR7000-2381

Printed in the USA, January 29, 2020
5994-1644EN

Agilent

Trusted Answers