Published in final edited form as:

Syst Biol Reprod Med. 2014 October ; 60(5): 308–315. doi:10.3109/19396368.2014.944318.

A Comparison of Sperm RNA-Seq Methods

Shihong Mao¹, Edward Sendler¹, Robert J. Goodrich¹, Russ Hauser^{2,3}, and Stephen A. Krawetz^{1,*}

¹Center for Molecular Medicine and Genetics, Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI 48201, USA

²Vincent Memorial Obstetrics and Gynecology Service, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts

³Departments of Environmental Health and Epidemiology, Harvard School of Public Health, Boston, Massachusetts

Abstract

A significant challenge to the effective application of RNA-seq to the complete transcript analysis of low quantity and/or degraded samples is the amplification of minimal input RNA to enable sequencing library construction. Several strategies have been commercialized in order to facilitate this goal. However, each strategy has its own specific protocols and methodology, and each may introduce unique bias and in some cases show specific preference for a collection of sequences. Our wider investigation of human spermatozoal RNAs was able to reveal their complexity despite being generally characterized by low quantity and high fragmentation. In this study, four commercially available RNA-seq library amplification protocols for the preparation of low quantity/highly fragmented samples - SMARTer™ Ultra Low RNA (SU) for Illumina® Sequencing (Clontech Laboratories, Inc.), SeqPlex RNA Amplification (SP) (Sigma-Aldrich Co.), Ovation® RNA-Seq System V2 (OR), and Ovation® RNA-Seq FFPE System from NuGEN (FFPE) (NuGEN Technologies, Inc.) - were assessed using human sperm RNAs. Further investigation analyzed the effects on the end results of two different library preparation methods - Encore NGS Multiplex System I (Enc) and Ovation Ultralow Library Systems (UL) (NuGEN Technologies, Inc.) - that appeared best suited to this type of RNA, along with other potential confounding factors such as Formalin Fixed Paraffin Embedded (FFPE) preservation. Our results indicate that for each library preparation protocol, the differences in the initial amount of input RNA and choice of RNA purification step do not generate marked differences in terms of RNA profiling. However, substantial disparity is introduced by individual amplification methods prior

*Corresponding Author: 271 C.S. Mott Center, 275 E. Hancock Ave., Detroit, MI, 48201, U.S.A., 313-577-6770, steve@compbio.med.wayne.edu.

DECLARATION of INTEREST:

The authors declare no conflicts of interest

DISCLAIMERS:

Mention of company names and/or products does not constitute endorsement by the National Institute for Occupational Safety and Health. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

AUTHOR CONTRIBUTIONS:

S.M. and E.S. analyzed the data and wrote the manuscript; R.J.G. prepared libraries for RNA-Seq and reviewed the manuscript; R.H. reviewed and edited the manuscript; and S.A.K. oversaw the project and edited the manuscript.

to library construction. These significant differences may be caused by the different priming methods or amplification strategies used in each of the four different protocols examined. The observation of intrasample variation introduced by the choice of protocol highlights the role that external factors play in planning and reliable interpretation of results of any RNA-seq experiment.

Keywords

RNA purification; reverse transcription and cDNA amplification; library preparation

INTRODUCTION

RNA-Seq enables the identification, profiling, and quantification of the complete transcript profile of a given cell. This technique has been widely used to study global gene expression, detect alternative splicing, discover novel exons, isoforms, and identify Single Nucleotide Polymorphisms [Ozsolak and Milos 2011; Wang et al. 2009]. The optimal amount of total RNA for analysis of a specific tissue using RNA-Seq is typically stated to be 1 µg or higher, although this is now being scaled to 100 ng. However, deep sequencing of low-quantity or low-quality RNA substrates such as samples fixed in paraffin (FFPE), rare cell populations, or even a single cell is often desired. Amplification protocols [Bhargava et al. 2013; Hashimshony et al. 2012; Ozsolak et al. 2010; Pan et al. 2013] introduced prior to library preparation allow sequencing libraries from low quantities of input to be constructed. The strategy for each amplification method is different in terms of the amount of RNA required, the priming methods utilized, and rRNA minimization [Tariq et al. 2011]. Comparison between RNA-Seq results generated from libraries employing various protocols, such as PCR amplification prior to library construction, have demonstrated bias towards high abundance transcripts [Sam et al. 2011]. A comprehensive and comparative analysis of RNA-Seq methods using low-quality or low-quantity RNA samples and a variety of amplification methods was recently performed [Adiconis et al. 2013]. This analysis indicated that different methods yielded different results in terms of ribosomal depletion, transcript coverage, 5' and 3' read bias and expression-level performance.

In contrast to the these studies where degraded or low-input RNA samples were synthetically produced for to compare methods, spermatozoal RNA is a biological sample inherently possessing both attributes of low quantity available per cell and inherent fragmentation. Over the past decade, our laboratory has described the complex population of RNAs present in mature spermatozoa using various strategies [Krawetz et al. 2011; Ostermeier et al. 2002; Ostermeier et al. 2004; Platts et al. 2007; Sendler et al. 2013]. The functions of the majority of sperm RNAs still remain to be resolved, but new studies suggest that the spermatozoon delivers functional RNAs to the oocyte during fertilization regulating early embryonic development [reviewed in Jodar et al. 2013; Krawetz 2005; Krawetz, et al. 2011]. As a snapshot of what remains after spermatogenesis, sperm RNA may provide investigational and diagnostic biomarkers of male fertility [Hamatani 2012; Krawetz 2005; Krawetz, et al. 2011; Sendler, et al. 2013; Wacławska and Kurpisz 2012]. The analysis of sperm RNA from large well-defined sample sets, like those from the Longitudinal Investigation of Fertility and the Environment (LIFE) study [Buck Louis et al. 2011], has the

potential to lead to a better understanding of the molecular basis of male fertility and how perturbations may affect this complex pathway. RNA-Seq has been applied to identify potential functions, transcript abundance and intactness, and particular characteristics of transcripts retained in spermatozoa [Jodar, et al. 2013; Johnson et al. 2011; Sendler, et al. 2013]. Compared with typical somatic cells, sperm RNA has two features: (1) low quantity of RNA per cell (~50 fg) and (2) exist in a fragmented or partially degraded state [Johnson, et al. 2011; Mao et al. 2013; Sendler et al. 2011] compared with RNA from typical somatic cells. Both characteristics make deep sequencing of sperm RNA more complicated, but also make sperm an ideal paradigm for the comparison of sequencing methodologies for low-input and fragmented samples.

In the present study, the low-quantity and partially fragmented RNAs from mature human spermatozoa were used to investigate the effectiveness of different amplification, library preparation and RNA purification strategies. This provided an assessment of the generalized characteristics of the RNA-Seq data generated by different RNA sequencing protocols, so that other independent studies may view protocol-specific data within this context.

RESULTS and DISCUSSION

A total of 16 libraries were constructed from a pool of human spermatozoa RNA samples using different methodologies designed to examine protocol-specific effects, followed by deep sequencing on Illumina HiSeq 2000. Each library was constructed by a specific protocol which was the combination of different parameters through four steps, as detailed in Table 1 and summarized in Figure 1 as follows: Step 1, RNA purification with or without column-based concentration; Step 2, Amplification evaluating four different amplification protocols including SMARTer (SU) which uses an oligo(dT) method to prime cDNA synthesis, SeqPlex (SP) which uses a random priming method, and OR and FFPE which use a combination of oligo(dT) and random primers for cDNA synthesis. SU and SP use PCR based amplification while NuGEN based protocols (OR, FFPE) use isothermal amplification (SPIA) (Table 1); Step 3, Differential amounts of initial RNA for amplification input; and Step 4, library preparation methods comparing two different protocols -Encore NGS Multiplex System I and Ovation Ultralow Library System (NuGEN Technologies, Inc.). This modular design enabled the effect of each step to be compared as a function of RNA profiling without the need to consider biological variation. The influence of each of these steps can be evaluated from several perspectives: overall RNA-Seq statistics, ribosomal and mitochondrial RNA profiling, and most importantly, the abundance and profile of RNAs at the transcript-level.

Alignment statistics

A summary of RNA-Seq statistics for each of the 16 protocol-specific libraries is provided in Table 2. Although the input to the sequencer was the same for each library (2 nmoles), the number of reads generated shows substantial variation between protocols. Overall, the samples amplified by SU and SP yielded approximately 2–3 fold more sequenced reads ($55.0 \pm 22.1\text{M}$ and $44.9 \pm 8.5\text{M}$) than the samples amplified using the FFPE or OR protocols ($25.9 \pm 21.2\text{M}$ and $15.2 \pm 0.2\text{M}$). The percentage of reads that aligned to the reference

sequences surveyed (human genome build hg19 plus ribosomal and mitochondrial sequences) also varied between different amplification and library preparation methods. The percentage of reads that aligned to reference sequences within the SU amplified library was substantially lower than the other methods - OR, FFPE and SP ($51.5 \pm 4.1\%$ vs. $83.6 \pm 7.4\%$). The choice of library preparation protocol also affected the percentage of aligned reads, with UL libraries showing markedly less alignment success than Enc ($61.8\% \pm 12.1\%$ vs $85.5\% \pm 6.0\%$, $P = 0.025$). Amplification errors, which are magnified by increasing the number of PCR cycles, are a likely factor in the final quantity of successfully mapped reads. The difference between using a combination of poorly performing steps for cDNA amplification (e.g. SU) and library construction (UL) can reduce read alignment to only 50% in comparison with 70–95% obtained using other protocols. This can affect the final RNA profile.

Effect of RNA column-concentration

Column-concentration of RNA subsequent to DNase-treatment is often employed in order to remove any remaining contaminants and also concentrate the RNA. In addition, this process tends to remove smaller RNAs and fragments. As summarized in Table 2, after this procedure, the percentage of reads that mapped to both the mitochondrial genome and ribosomal RNA shows a considerable decrease in SP samples ($P < 0.001$), with mtRNA in FFPE samples also decreasing significantly ($P = 0.027$). This may reflect a bias towards shorter length fragments within mtRNA and rRNA fragment populations, which in turn effects an increase in the relative proportion of longer length genomic RNAs after column-concentration size selection.

RNA fragment size distribution

RNA fragment size was inferred based on the average separation between aligned ends of paired sequencing reads. Their distribution is shown in Figure 2. Protocols utilizing the same amplification procedure but with different amounts of initial input show a highly similar distribution and clearly group together. Among amplification protocols, significant differences in fragment length were observed. Fragment lengths from UL (red and orange curves) are much longer ($126.8 \pm 6.3\text{bp}$) than those from Enc ($95.8 \pm 10.1\text{bp}$), with a significant difference ($P = 6.1 \times 10^{-5}$). This difference likely represents underlying variance in the favored template between amplification methods. As expected, column-based RNA concentration additionally contributes to the differences in the size distribution of fragments in SP and OR samples. For example, in SP samples without column-based concentration (dark blue), a peak with a fragment length of ~50 bp is evident that disappears after column-based concentration. In the case of OR samples, subsequent to column-based concentration, the average fragment length increases from 103 bp (light blue) to 121 bp (black). Figure 2 includes four FFPE-protocol (green) samples, with and without column concentration. The extreme similarity of these four distribution curves indicates that column-based concentration did not yield any significant difference in terms of the distribution of fragments from FFPE samples

rRNA and mtRNA

It is apparent that the number of reads mapping to ribosomal and mitochondrial sequences varies greatly despite the use of identical input RNA. In sperm, without mRNA enrichment or cDNA amplification, the 18S and 28S ribosomal RNA fragments make up approximately 80% of total RNA population, while ~ 1% – 10% of the total reads are derived from 12S and 16S mitochondrial RNAs [Johnson, et al. 2011; Sendler, et al. 2013]. All protocols examined within this study show a significant reduction of the ribosomal RNA fraction from this original amount. In some methods this appears to give rise to a resultant increase in the relative amount of mitochondrial RNA, while in others it may be repressed. Differences in the extent to which specific RNA components are removed indicates the underlying variation in each method to either intentionally or unintentionally alter the representation of these common RNAs. As SU uses oligo(dT) to select poly(A+)-enriched RNAs, the majority of poly(A-) RNAs, such as rRNA and mtRNA will be removed. This is reflected by the relative depletion of these elements in the alignment results (Table 2). Although SP uses a selective collection of a selection of random primers to synthesize and amplify cDNA, it is interesting to observe that the fraction of reads mapped to rRNA or mtRNA appears to be substantially depleted. The OR and FFPE protocols use a combination of oligo(dT) and random primers to enrich mRNAs. In these methods, substantial removal of ribosomal RNA is observed, however comparable mitochondrial RNA removal does not seem to be effected. This results in an apparent increase in the representation of mtRNAs as previously observed [Mao, et al. 2013]. The basis for this difference in the relative removal of ribosomal vs. mitochondrial RNA may at least in part reflect that RNA fragmentation appears not to be random. This yields a large number of fairly common sequences within the total RNA pool. Primers within the random or semi-random pool which correspond to these sequences will be preferentially consumed rather than the comparatively less common genomic-derived RNAs, reducing the total amount of amplified. Although this perhaps unintended behavior is indeed beneficial in terms of reducing the proportion of undesired ribosomal and perhaps mitochondrial RNA, it is also important to note that it may also skew the representation of highly abundant transcripts, such as the sperm protamines, within the total sequenced population.

Distribution of rRNA and mtRNA sequencing reads

Apart from the fractional difference in the proportion of rRNA and mtRNA observed across different protocols, the distribution of NGS reads along the length of transcript is dependent upon individual amplification protocol. Figure 3, shows the distribution of aligned reads across the 28S rRNA for each library protocol. Neither differing amounts of RNA input (Panels A, E and C vs. F) nor column-based concentration (Panels B and D) show any significant effect. In contrast, the choice of the amplification or library preparation protocol yields considerable differences in the relative size or even presence or absence of specific RNAs, with similar alterations in profile between methods observed in profiles of the 18S and mitochondrial RNAs (Supplemental Figures 1 and 2). Altered profiles likely reflect the protocol-specific priming strategies and primers used. For example, while the SU, OR, and FFPE protocols use oligo(dT), or at least a primer mix enhanced with oligo(dT) to prime cDNA synthesis, SP uses only random primers. Interestingly the distribution of reads even with random primers is far from uniform. This likely represents additional amplification

biases, such as differences in local secondary structure, GC content, and intrinsic sequencing artifacts [Sendler, et al. 2011].

Relative enrichment of genomic regions

As RNA-seq sequencing originates from transcribed RNA, it is expected that the majority of sequencing reads should map to exonic and promoter regions, with a relatively small proportion of reads originating in the intronic and intergenic regions of the genome. The relative enrichment of exonic, intronic, intergenic and promoter regions for each of the 16 libraries is shown in Table 3. Their relative abundance shows considerable variation, with samples prepared using the OR and FFPE protocols (without column-based concentration) exhibiting very high enrichment in exonic and promoter sequences. The samples prepared using the FFPE protocol (with column-based concentration) showed moderately high enrichment, in comparison to the SU and SP samples that showed relatively low enrichment in exonic and promoter regions respectively.

This difference is likely reflective of poly(A+) selection which the OR and FFPE protocols employ to enrich for mRNAs, where the bulk are annotated as poly(A+) containing. As expected, after amplification, the majority of the sequencing reads are derived from exons and promoters. In contrast, methods such as SP which do not specifically target the poly(A+) containing RNAs, a substantial proportion of reads originate from intergenic or intronic non-annotated regions. This reflects both the unusual nature of the sperm transcripts, and also the importance of protocol selection to reveal the complete spectrum of its RNA.

Transcript coverage

A further important criterion for comparison of protocols is the distribution of sequencing reads over the entire length of the gene. Since the comparison used identical input RNAs, the distribution of sequencing reads should be equivalent, reflecting only differences in relative abundance of local transcribed regions. To address this tenet, the distribution of reads across the top 100 sperm transcripts (ranked by relative abundance) was assessed. The average distribution across total transcript length is shown in Figure 4A, with the read distribution of one of the most abundant sperm transcripts, *PRM2*, shown in Figure 4B. It is evident that different amplification protocols introduce considerable differences to overall transcript profiling. SU uses oligo(dT) based primer amplification and as expected, shows significant 3' enrichment (black). After 3' poly(A+) enrichment, only a small fraction of the 5' region is observed. This confirms that this transcript has been subject to substantial fragmentation (as with the vast majority of others surveyed) in mature spermatozoa. Preferential retention of specific transcripts which remain in a comparatively intact state could likely be assessed by identifying transcripts which do not show such a significant difference in 5' vs. 3' read bias as typically observed using this method. The site-specific read preference observed in SP samples illustrated by *PRM2* is typical of that observed in all transcripts. This likely reflects a combination of high amplification and some degree of primer bias to specific RNA start sequences. It is important to note that while different amplification protocols do significantly affect the local distribution of sequencing reads, the choice of the library preparation method did not show any similar effect on read profile.

Transcripts concordance between protocols

The reproducibility of RNA-seq results is critical to the integrity and reliability of the results. Correlations between samples as a function of the protocols employed are summarized in Figure 5. Each point represents the relative abundance (RPKM) of one transcript as measured in two matched samples. Both visual representation and the statistical measure of the Pearson correlation illustrate the degree of influence of each of these factors on reproducibility. Comparison of the technical replicates (Pearson correlation $r = 0.927$) reflects the most basic level of sequencing resolution, and likely depends the degree of amplification that is necessary for low-input samples (Figure 5A). While column-based RNA concentration has significant effect on OR sample transcript abundance (Figure 5B), other protocols are not subject to this effect (data not shown). The library preparation protocol used (Figure 5C) and the differing amounts of input RNA (Figure 5D) only moderately affect the population of transcripts that are resolved. The most considerable difference between measures of the RNA population is clearly introduced by the amplification strategy (Figure 5E, 5F). As shown in Supplemental Table 1, the correlation between protocols illustrates that this single factor is by far the primary determinant of rank measure of individual transcripts within the overall transcript population.

This is further demonstrated by unsupervised hierarchical clustering of the levels of all transcripts as assessed between protocols (Supplemental Figure 3). Clusters are most clearly defined by the protocol used. The relatively high correlation between the OR and FFPE prepared samples likely reflects the use of similar priming methods. This notion is supported by the observation that SU also utilizes oligo(dT) priming, resulting in a closer correspondence to the OR/FFPE cluster. In comparison the SP protocol uses a random priming strategy, with the difference between these reflected in the distance of SP results from others. Although library preparation methods, column-based concentration and amount of input RNA all exhibit some degree of influence on protocol correlation, their effect can be seen as of minor significance as compared with the effect of amplification protocol. As with other protocol-specific effects observed, being mindful of the considerable influence that the choice of amplification strategy promotes is critical to the interpretation of comparative results.

In this study the effect of different RNA-Seq library preparation and amplification protocols on transcript abundance and profiles of low input and low quality fragmented RNA samples were assessed using the biologically relevant model of spermatozoal RNA. Results clearly indicate that the choice of amplification strategy plays a major role in the resultant RNA profile. Differences result primarily from exaggeration of protocol-specific biases after high amplification of low-quantity and low-quality RNAs. In comparison, factors such as library protocol, column-based RNA concentration, and the amount of initial RNA input have a minor impact on the final RNA-seq results. These findings emphasize that disparity in results arising from the use of different RNA preparation protocols may easily confuse or mask true biological effects. Recognition of specific distinctive characteristics of particular methods may aid in both understanding and optimization of RNA-seq library construction and analysis of results for low-quantity and low-quality samples

MATERIALS and METHODS

Sample collection and RNA extraction

Semen samples were collected as part of the Eunice Kennedy Shriver National Institute of Child Health and Human Development LIFE study [Buck Louis, et al. 2011]. RNA was isolated and analyzed as previously described [Mao, et al. 2013]. After isolation and quality control, a total of eight samples of RNA were mixed together as pooled RNA. Such pooled RNA allows for evaluation of different preparation protocols using identical RNA, eliminating variance caused by normal intra-sample variation. An aliquot of the pooled RNA was purified using the NucleoSpin® RNA Clean-up (Macherey-Nagel Inc.) to concentrate the RNA pool and remove any possible RT-PCR inhibitors.

cDNA synthesis and amplification

Pooled RNA was reverse transcribed and amplified using four different kit-based amplification protocols, SMARTer™ Ultra Low RNA (SU) for Illumina® Sequencing (Clontech Laboratories, Inc.), SeqPlex RNA Amplification (SP) (Sigma-Aldrich Co.), Ovation® RNA-Seq System V2 (OR) and Ovation® RNA-Seq FFPE System (FFPE) from NuGEN (NuGEN Technologies, Inc.). The general procedure for all protocols is similar: reverse transcription, followed by cDNA amplification. The differences between the kits appear in the amount of RNA input, the priming method for reverse transcription, and the amplification method. Differential amounts of RNA input (including both the before clean-up and after clean-up pools) were evaluated for: SU – 1 ng and 10 ng, SP – 2 ng and 10 ng, NuGEN (OR and FFPE) – 100 ng. (Table 1).

RNA library preparation, sequencing and demultiplexing

Two RNA library preparation kits were evaluated: Encore NGS Multiplex System I (Enc) and Ovation Ultralow Library Systems (UL). Both are manufactured by NuGEN (NuGEN Technologies, Inc.). Amplified cDNA is first subject to sonication. Then the fragment ends are repaired before ligation of sequencing adaptors (with barcodes attached if applicable), followed by PCR amplification and library purification. The main difference between the kits used in this study is the amount of input cDNA recommended for each. The Enc was designed to work with 200 ng of input material. The UL recommends 1 ng-100 ng of input material. In this study, 2 ng of input were utilized for UL (Table 1).

A total of 16 different libraries were constructed using identical RNA as initial input using inline barcodes. An equal amount of fragments (2 nmole) from each library were loaded to Illumina HiSeq 2000 sequencer and were subjected to paired-end sequencing for 50 cycles. Illumina genome analyzer pipeline software CASAVA (version 1.8.2) was utilized for image analysis, base calling and FASTQ generation. Inline demultiplexing was performed using software fastq_multx [Aronesty 2011].

Short read mapping and transcript abundance estimating

Novoalign (Novocraft Technologies v.2.08, Selangor, Malaysia) was used to map the sequencing reads against the human reference genome (hg19) plus human ribosomal 5S, 18S and 28S sequences at paired-end base default parameters. Genomatix software

(www.genomatix.de) was applied to calculate the relative abundance of each transcript (most abundant isoform) as RPKM (reads per kilobase exon per million fragments mapped). The RNA-Seq data has been deposited in the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) (GSE57503).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported in part by the Charlotte B. Failing Professorship to S.A.K., a grant to RH from Harvard School of Public Health, National Institute of Environmental Health Sciences [Grant Number ES017285] and in part by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development Contract 25PM6 in collaboration with the LIFE Study Working Group, Division of Epidemiology, Statistics, and Prevention Research who provided semen samples. We are grateful to Meritxell Jodar for the review of the manuscript.

Abbreviations

SU	SMARTer Ultra Low RNA
SP	SeqPlex RNA Amplification
OR	Ovation RNA-Seq System V2
FFPE	Ovation RNA-Seq FFPE System
Enc	Encore NGS Multiplex System I
UL	Ovation Ultralow Library Systems
ac	after column-based purification
bc	before column-based purification
tr	technical replicates
M	million

REFERENCES

- Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods*. 2013; 10:623–629. [PubMed: 23685885]
- Aronesty E. ea-utils : "Command-line tools for processing biological sequencing data". 2011 <http://code.google.com/p/ea-utils>.
- Bhargava V, Ko P, Willems E, Mercola M, Subramaniam S. Quantitative transcriptomics using designed primer-based amplification. *Sci Rep*. 2013; 3:1740. [PubMed: 23624976]
- Buck Louis GM, Schisterman EF, Sweeney AM, Wilcosky TC, Gore-Langton RE, Lynch CD, et al. Designing prospective cohort studies for assessing reproductive and developmental toxicity during sensitive windows of human reproduction and development--the LIFE Study. *Paediatr Perinat Epidemiol*. 2011; 25:413–424. [PubMed: 21819423]
- Hamatani T. Human spermatozoal RNAs. *Fertil Steril*. 2012; 97:275–281. [PubMed: 22289287]
- Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*. 2012; 2:666–673. [PubMed: 22939981]

- Jodar M, Selvaraju S, Sendler E, Diamond MP, Krawetz SA. The presence, role and clinical use of spermatozoal RNAs. *Hum Reprod Update*. 2013; 19:604–624. [PubMed: 23856356]
- Johnson GD, Sendler E, Lalancette C, Hauser R, Diamond MP, Krawetz SA. Cleavage of rRNA ensures translational cessation in sperm at fertilization. *Mol Hum Reprod*. 2011; 17:721–726. [PubMed: 21831882]
- Krawetz SA. Paternal contribution: new insights and future challenges. *Nat Rev Genet*. 2005; 6:633–642. [PubMed: 16136654]
- Krawetz SA, Kruger A, Lalancette C, Tagett R, Anton E, Draghici S, et al. A survey of small RNAs in human sperm. *Hum Reprod*. 2011; 26:3401–3412. [PubMed: 21989093]
- Mao S, Goodrich RJ, Hauser R, Schrader SM, Chen Z, Krawetz SA. Evaluation of the effectiveness of semen storage and sperm purification methods for spermatozoa transcript profiling. *Syst Biol Reprod Med*. 2013; 59:287–295. [PubMed: 23869956]
- Ostermeier GC, Dix DJ, Miller D, Khatri P, Krawetz SA. Spermatozoal RNA profiles of normal fertile men. *Lancet*. 2002; 360:772–777. [PubMed: 12241836]
- Ostermeier GC, Miller D, Huntriss JD, Diamond MP, Krawetz SA. Reproductive biology: delivering spermatozoan RNA to the oocyte. *Nature*. 2004; 429:154. [PubMed: 15141202]
- Ozsolak F, Goren A, Gymrek M, Guttman M, Regev A, Bernstein BE, et al. Digital transcriptome profiling from attomole-level RNA samples. *Genome Res*. 2010; 20:519–525. [PubMed: 20133332]
- Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*. 2011; 12:87–98. [PubMed: 21191423]
- Pan X, Durrett RE, Zhu H, Tanaka Y, Li Y, Zi X, et al. Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc Natl Acad Sci U S A*. 2013; 110:594–599. [PubMed: 23267071]
- Platts AE, Dix DJ, Chemes HE, Thompson KE, Goodrich R, Rockett JC, et al. Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs. *Hum Mol Genet*. 2007; 16:763–773. [PubMed: 17327269]
- Sam LT, Lipson D, Raz T, Cao X, Thompson J, Milos PM, et al. A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS One*. 2011; 6:e17305. [PubMed: 21390249]
- Sendler E, Johnson GD, Krawetz SA. Local and global factors affecting RNA sequencing analysis. *Anal Biochem*. 2011; 419:317–322. [PubMed: 21889483]
- Sendler E, Johnson GD, Mao S, Goodrich RJ, Diamond MP, Hauser R, et al. Stability, delivery and functions of human sperm RNAs at fertilization. *Nucleic Acids Res*. 2013; 41:4104–4117. [PubMed: 23471003]
- Tariq MA, Kim HJ, Jejelowo O, Pourmand N. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res*. 2011; 39:e120. [PubMed: 21737426]
- Waclawska A, Kurpisz M. Key functional genes of spermatogenesis identified by microarray analysis. *Syst Biol Reprod Med*. 2012; 58:229–235. [PubMed: 22676428]
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009; 10:57–63. [PubMed: 19015660]

Figure 1.

RNA-Seq Comparison. RNA was isolated from a pooled set of sperm cells from multiple subjects. Following DNase treatment, some samples underwent column-based purification and concentration. RNA was reverse transcribed followed by one of four amplification strategies (SU: SMARTer Ultra Low RNA ; SP: SeqPlex RNA Amplification ; OR: Ovation RNA-Seq System V2; FFPE: Ovation RNA-Seq FFPE System) applied to resultant cDNA. Additionally, two different amounts of starting inputs were assessed for SU and SP strategies and technical replicates (“x2”) included. All protocols yielded sufficient template for RNA-Seq library preparation..

Figure 2.

Distribution of fragment sizes of RNA-seq libraries. The libraries were divided into several subgroups based on the amplification strategy, library preparation, with (ac) or without (bc) column-based purification and concentration. SU_UL: samples were amplified by SMARTer Ultra Low system and libraries were prepared by Ovation Ultralow Library system; FFPE_Enc: samples were amplified by Ovation RNA-Seq FFPE System and libraries were prepared by Encore NGS Multiplex System I; OR_Enc_ac and OR_Enc_bc: samples were amplified by Ovation RNA-Seq System V2 and libraries were prepared by Encore NGS Multiplex System I; SP_Enc_ac and SP_Enc_bc: samples were amplified by SeqPlex RNA Amplification and libraries were prepared by Encore NGS Multiplex System I; SP_UL_ac: samples were amplified by SeqPlex RNA Amplification and libraries were prepared by Ovation Ultralow Library Systems. Differing amounts of input RNA had little effect on fragment distribution when using same protocol, and they are shown in same color.

Figure 3.

The distribution of reads mapping across length of 28S rRNA. The X-axis is the 28S rRNA nucleotide position from the 5' to 3' end. The Y-axis is the relative abundance of the reads at any location. The 16 libraries were divided into six subgroups: SU_UL (A), OR_Enc (B), SP_Enc (C), FFPE_Enc (D), SP_UL (E) and SP_Enc (F).

Figure 4.

Transcript Profile Sequencing Characteristics of SU_UL_ac FFPE_Enc_ac, OR_Enc_ac, SP_UL_ac SP_Enc_ac at various library input concentrations. A. Average coverage of top 100 transcripts (ranked by relative abundance) across complete transcript length (ignoring intronic regions). The variance in library input concentration has no effect on read distribution. B. The varying distribution of reads as generated by different protocols across common sperm transcript *PRM2*. Arrow indicates transcript 5' to 3' orientation. Thicker rectangles indicate exons, while thinner rectangles at two ends indicate 5' UTR and 3' UTR. The thin line in the middle indicates intron. The number of reads corresponding to each base position is represented on the vertical axis.

Figure 5.

Scatter plots show the correlations of transcript abundance (log₂(RPKM + 1)) between samples prepared from protocols with any of the following treatments: A. technical replicates; B. before column-based concentration vs after column-based concentration; C. Enc library preparation vs UL library preparation; D. 2ng input RNA vs 10ng input RNA; E. FFPE amplification vs SU amplification; and F. OR amplification vs SP amplification. The Pearson correlation coefficient 'r' is noted.

Table 1

Overview of 16 libraries

Library name	Amp Kit	Amp method	RNA loads to Amp (ng)	Lib prep	cDNA loads to library prep (ng)
OR_Enc_100_bc	OR	SPIA	100	Enc	200
OR_Enc_100_ac	OR	SPIA	100	Enc	200
FFPE_Enc_100_bc	FFPE	SPIA	100	Enc	200
FFPE_Enc_100_bc_tr	FFPE	SPIA	100	Enc	200
FFPE_Enc_100_ac	FFPE	SPIA	100	Enc	200
FFPE_Enc_100_ac_tr	FFPE	SPIA	100	Enc	200
SU_UL_1_ac	SU	PCR	1	UL	2
SU_UL_10_ac	SU	PCR	10	UL	2
SP_Enc_2_bc	SP	PCR	2	Enc	200
SP_Enc_10_bc	SP	PCR	10	Enc	200
SP_Enc_10_bc_tr	SP	PCR	10	Enc	200
SP_Enc_2_ac	SP	PCR	2	Enc	200
SP_Enc_10_ac	SP	PCR	10	Enc	200
SP_Enc_10_ac_tr	SP	PCR	10	Enc	200
SP_UL_2_ac	SP	PCR	2	UL	2
SP_UL_10_ac	SP	PCR	10	UL	2

For each library name, such as SP_Enc_1_ac_tr, the differences in steps of protocol are indicated by division of underscore “_”: fields indicate amplification system: library preparation: ng of initial input RNA for amplification; plus (ac) minus (bc) column-based purification; and tr indicating technical replicate. SU: SMARTer Ultra Low RNA system from Clontech; SP: SeqPlex RNA Amplification from Sigma-Aldrich; OR: Ovation RNA-Seq System V2 from NuGEN; FFPE: Ovation RNA-Seq FFPE System from NuGEN; Enc: Encore NGS Multiplex System I from NuGEN; UL: Ovation Ultralow Library Systems from NuGEN; SPIA: Single Primer Isothermal Amplification.

Table 2

Statistical summary of RNA-seq results

Name	Total reads	reads aligned	Unique aligned	% Genome	% mtRNA	% rRNA
OR_Enc_100_bc	15.33	12.66	10.77	45.9	50.5	3.6
OR_Enc_100_ac	15.09	12.74	11.21	32.3	62.8	4.9
FFPE_Enc_100_bc	16.41	11.82	9.82	55.5	32.2	12.4
FFPE_Enc_100_bc_tr	16.77	13.34	11.23	59.7	31.4	8.9
FFPE_Enc_100_ac	57.59	47.92	41.72	67.7	21.9	10.4
FFPE_Enc_100_ac_tr	12.75	10.60	9.17	68.6	23.5	7.9
SU_UL_1_ac	39.36	19.14	16.37	97.9	1.6	0.5
SU_UL_10_ac	70.65	38.44	33.57	98.1	1.4	0.5
SP_Enc_2_bc	35.66	31.41	28.82	81.6	6.5	11.9
SP_Enc_10_bc	43.47	37.93	34.47	80.8	7.3	11.9
SP_Enc_10_bc_tr	56.81	51.93	47.42	81.8	6.7	11.5
SP_Enc_2_ac	47.26	42.67	39.62	92.6	3.3	4.1
SP_Enc_10_ac	49.63	44.66	41.22	92.1	3.6	4.4
SP_Enc_10_ac_tr	53.94	50.59	46.88	93.5	3.0	3.5
SP_UL_2_ac	38.91	27.71	24.90	85.8	7.9	6.2
SP_UL_10_ac	33.55	24.43	21.71	84.7	8.8	6.5

The read counts (first three columns) in the table are in millions of reads. "Total reads" indicates the number of reads generated from sequencer; "# reads aligned" indicates the number of reads that could be aligned to the reference genome plus ribosomal RNA plus mitochondrial; "Unique aligned" indicates the number of reads that were uniquely aligned to reference genome, rRNA and mitochondria. "% to Genome" indicates the percentage of uniquely mapped reads that were aligned to reference genome only – ignoring ribosomal and mitochondrial reads; "% align to rRNA" indicates the percentage of uniquely mapped reads that mapped to mitochondria; "% align to rRNA" indicates percentage of reads that mapped to ribosomal RNA.

Table 3

The reads enrichment on four genomic elements

Name	Exonic	Intronic	Intergenic	Promoter
OR_Enc_100_bc	10.6	0.7	0.4	7.5
OR_Enc_100_ac	8.4	0.8	0.5	5.9
FFPE_Enc_100_bc	9.3	0.7	0.4	6.8
FFPE_Enc_100_bc_tr	9.8	0.7	0.4	7.2
FFPE_Enc_100_ac	5.7	0.9	0.6	4.3
FFPE_Enc_100_ac_tr	5.6	0.9	0.6	4.2
SU_UL_1_ac	3.3	1.0	0.8	2.7
SU_UL_10_ac	3.5	1.0	0.8	2.9
SP_Enc_2_bc	2.5	1.0	0.8	2.5
SP_Enc_10_bc	2.6	1.0	0.8	2.6
SP_Enc_10_bc_tr	2.5	1.0	0.8	2.5
SP_Enc_2_ac	1.7	1.0	0.9	1.9
SP_Enc_10_ac	1.7	1.0	0.9	1.8
SP_Enc_10_ac_tr	1.6	1.0	0.9	1.8
SP_UL_2_ac	2.0	1.0	0.9	2.0
SP_UL_10_ac	1.9	1.0	0.9	2.0

Unique sequences of the human genome were segmented into exonic, intronic, intergenic and promoter elements. The percentage of these four genomic regions (exonic, intronic, intergenic and promoter) constitute 4.1%, 42.1%, 53.8% and 2.5% of the human genome respectively. In each library, the percentage of reads that mapped to these four genomic elements was quantified. The values in the table are the enrichment of reads corresponding to each of these elements relative to total genomic coverage of each element. Larger values indicate significant enrichment.