4104–4117 *Nucleic Acids Research, 2013, Vol. 41, No. 7* *Published online 6 March 2013*
doi:10.1093/nar/gkt132

Stability, delivery and functions of human sperm RNAs at fertilization

Edward Sendler[1,2], Graham D. Johnson[1], Shihong Mao[1,2], Robert J. Goodrich[1,2], Michael P. Diamond[2], Russ Hauser[3] and Stephen A. Krawetz[1,2,*]

[1]Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit, MI 48201, USA, [2]Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, MI 48201, USA, [3]Vincent Memorial Obstetrics and Gynecology Service, Massachusetts General Hospital, Harvard Medical School, Department of Environmental Health, Harvard School of Public Health and Department of Epidemiology, Harvard School of Public Health, Boston, MA, 02115, USA

Received October 31, 2012; Revised February 6, 2013; Accepted February 7, 2013

ABSTRACT

Increasing attention has focused on the significance of RNA in sperm, in light of its contribution to the birth and long-term health of a child, role in sperm function and diagnostic potential. As the composition of sperm RNA is in flux, assigning specific roles to individual RNAs presents a significant challenge. For the first time RNA-seq was used to characterize the population of coding and non-coding transcripts in human sperm. Examining RNA representation as a function of multiple methods of library preparation revealed unique features indicative of very specific and stage-dependent maturation and regulation of sperm RNA, illuminating their various transitional roles. Correlation of sperm transcript abundance with epigenetic marks suggested roles for these elements in the pre- and post-fertilization genome. Several classes of non-coding RNAs including lncRNAs, CARs, pri-miRNAs, novel elements and mRNAs have been identified which, based on factors including relative abundance, integrity in sperm, available knockout data of embryonic effect and presence or absence in the unfertilized human oocyte, are likely to be essential male factors critical to early post-fertilization development. The diverse and unique attributes of sperm transcripts that were revealed provides the first detailed analysis of the biology and anticipated clinical significance of spermatozoal RNAs.

INTRODUCTION

Sperm structure and function appear fine-tuned towards a single purpose—the delivery of the paternal content to the oocyte in the most compact and accurate form. Our view of the presence and role of RNAs in human sperm has dramatically evolved (1). As sperm are generally considered transcriptionally inert, RNAs detected in the paternal gamete were initially assumed to be either content left behind after degradation and expulsion of the residual body, or simply contaminants from other surrounding cells (2,3). However, sperm retain specific coding (4–8) and non-coding RNAs (9,10). Irregularities in the levels of sperm RNAs have been recognized as markers and potential effectors of human male infertility (11–13). Their functional role on delivery to the oocyte has been suggested (14–16).

Previous studies have used microarray approaches to identify transcripts retained in sperm (12,17,18). Although informative, these analyses were unable to provide the enhanced view of this suite of RNAs afforded by RNA-seq [reviewed in (19)]. Apart from being recognized as a more accurate assessment of transcript levels than microarrays (20,21), RNA-seq has the added advantage that novel tissue isoforms and variants can be identified and quantitatively evaluated. Further, previously unrecognized coding and non-coding transcripts can be discovered and potential function ascribed (22–25). The population of sperm RNAs as revealed by RNA-seq provides a window into the developmental history, functional viability and potential elements delivered by sperm that are likely of significance upon fertilization. Together with the previous sncRNA analysis (10), this study provides a comprehensive snapshot of the transcript composition of human sperm among several individuals. Comparison of the many genomic regions encoding sperm RNAs with elements of the sperm epigenome has provided insight into the function and temporal action of these regulatory features. Highlighted in the analysis below is the use of exon profiles to identify full-length transcripts amongst a larger population of

*To whom correspondence should be addressed. Tel: +1 313 577 6770; Fax: +1 313 577 8554; Email: steve@compbio.med.wayne.edu

fragmented RNAs. Global analysis of this unique suite of paternal transcripts echoes past transcriptional history and late-stage spermatogenic function, yet projects its role on fertilization.

MATERIALS AND METHODS

RNA isolation and sequencing

To characterize various classes and features of sperm RNA, multiple sample preparation methods were used. The first set of samples was prepared as described (10,26,27) using RNA from a total of six sperm samples from three random and three proven fertile donors and two testes samples. Briefly, four of these sperm samples (henceforth referred to as T-Sperm samples)—Sp_D62[Tt], Sp_D64[Tt], Sp_D66[Tt] (individual donors) and Sp_P1[Tt] (mixed pool of three random donors)—were not subject to any of the commonly used RNA-selection methods to fully resolve the RNA population and reduce transcript expression profile bias. With the exception of the Pooled sample Sp_P1[Tt], all samples were subject to 50% PureSperm (Nidacon, Mölndal, Sweden) gradient to purify spermatozoa away from other contaminating cell-types. RNA from donor D62 was additionally separated into Poly(A^+) (Sp_D62[A^+]) and Poly(A^-) (Sp_D62[A^-]) fractions by oligo(dT) selection, with each fraction sequenced separately. Sperm samples were all sequenced in two lanes each on Illumina GAIIx Genome Analyzer. Two samples of commercially obtained pooled testes RNA—Te_PAm[A^+] (Applied Biosystems/Ambion, Austin, TX, USA, Lot 054P010702031A) and Te_PCl[A^+] (ClonTech, Mountain View, CA, USA, Lot 3090051)—were each sequenced in a single lane after Poly(A^+) selection.

Single Primer Isothermal Amplification (SPIA) was used to prepare a second set of RNA-seq libraries from 20 ng of RNA from each of the three individual donors (Sp_D62[SP], Sp_D64[SP], Sp_D66[SP]), a single pooled sperm sample (Sp_P2[SP]) and a single pooled testes (Te_PAm[SP]) sample using the Nugen Ovation kit (Nugen Inc., San Carlos, CA, USA) for cDNA and initial linear amplification. The Nugen Encore system was used for adaptor ligation, end-repair and amplification. Samples were multiplexed at four samples per lane.

All samples were subjected to 2×36 cycles of paired-end sequencing on the Illumina GAIIx platform. Paired-end mapping was performed equivalently for all samples using Novoalign (V2.07.09, Novocraft Technologies, Selangor, Malaysia) with alignment to Human Genome build hg19, and also including ribosomal 18S and 28S and mitochondrial 12S and 16S indices for specific alignment of reads corresponding to these sequences. Reads mapping to multiple genomic locations were examined separately. Sequencing data are available at the NCBI GEO repository.

An SPIA-prepared RNA-seq Universal Human Reference (UHR; Nugen Ovation library, Nugen Inc.) sample composed of RNAs from pooled somatic tissues was used for comparison of the parallel SPIA-prepared sperm and testes RNAs with a wide representation of somatic RNAs. Data from this sequencing are publically viewable by including the text 'track type = bigWig name = "rep123.uhr.7102" description = "rep123.uhr.7102" color = 5 20 198 autoScale = on visibility = 2 viewLimits = 1:500 bigDataUrl = http://ucscdatahosting.com/labuser/Tech_Support/rep123.uhr.7102.bw' in a user track window of the UCSC genome browser at http://genome.ucsc.edu/cgi-bin/hgCustom?hgsid = 322010991:

RT-PCR-transcript validation

Remaining total RNA from Sp_D62[Tt], Sp_D64[Tt] and Sp_D66 were reverse-transcribed using SuperScriptIII (Invitrogen, Carlsbad, CA, USA) with either oligo(dT) (Sp_D62 and Sp_D64 [Tt]) or gene-specific primers (Sp_D66 [Tt]) then amplified with HotStarTaq (Qiagen, Valencia, CA, USA), as described (28). PCR primer pairs were designed to query the entire transcript length (Supplementary Table S1). Products were visualized by agarose gel electrophoresis and images captured on the Typhoon 9210 scanner (Molecular Dynamics).

Bioinformatic analysis

Total transcript read counts from each sample were measured for 37 973 different isoforms (RefSeq annotation, HG19) from a total of 22 302 genes. Sperm RNAs contain a host of transcripts exhibiting multiple polyadenylation sites, intronic elements, or reflect differential processing (29). Accordingly to minimize this effect and to quantify transcript levels, a custom algorithm was developed to count exonic reads from all sequenced RNA samples across protein-coding regions only for all annotated transcript isoforms (See Supplementary Methods). The RPKM (Reads per Kilobase per Million reads) values were then calculated for both Total RNA and SPIA samples as summarized in Supplementary Tables S2 and S3. The aggregate of four T-sperm and SPIA-sperm samples was obtained by pooling all samples and renormalizing RPKM values based on total counts. The most abundant isoform of each gene (RPKM) was then used to rank each gene in each individual sample, allowing for intersample comparison by Spearman rank correlation. Transcript levels in sperm and testes were additionally assessed using the Genomatix RegionMiner (v2.41207), as this data source also includes a series of non-coding transcripts (e.g., RNU) in its assessment of transcript abundance. Ontological analysis of selected transcript groups was performed using GeneRanker (Genomatix v2.41103). Post-fertilization significance of sperm-delivered RNAs was examined using expression data from microarray analysis of human oocyte at MII, 2-, 4-, 6-, 10-cell, morula, blastocyte and stem-cell stage (30). Relative differential expression for each transcript across these developmental stages was assessed by calculating the change in reported expression (log 2) at each stage relative to MII baseline. Predicted targets of notable sperm-delivered miRNAs were calculated using Diana-microT algorithm (v3.0; 31,32) as described (10). High confidence targets (>0.95%) were selected for downstream analysis based on strict predicted binding score (miTG >25).

An intactness score was developed to assess the stability of total sperm RNAs with a length of at least 200 nt. To portray the relative level of coverage over transcript length, RNA-seq expression profiles for selected transcripts were obtained by computationally splicing read coverage corresponding to the transcript body (Supplementary Figure S1). Analysis was based on the annotated exon map of the most abundant isoform for each sperm transcript. Bias due to lack of read coverage at exon ends was minimized by excluding the last 15 nt of each exon. Normalized coverage profiles were calculated across transcripts in 100 bins. The sum of squares of actual deviation from the expected coverage of an ideal transcript (1% of total coverage/bin) was used as the intactness score. Transcripts were additionally and separately categorized and ordered based on positioning of underrepresented or absent regions as a function of the relative degree of 5′ versus 3′ coverage. A linear scaling function from the 5′-end maximum positive value to the maximum 3′-end negative value was applied to the normalized profiles, with subsequent ranked degree of variation from a level profile now indicative of the degree and site of end degradation.

To examine the association of RNA abundance in sperm with local epigenetic marks, 19 521 expression-enhanced promoter and exonic cluster regions from sperm sample Sp_D62[Tt] were identified using Genomatix RegionMiner tool (Audic-Claverie algorithm with $P < 0.05$, window size of 100 bp). External data sets were used for correlation analysis of H3K4me3 (34 912 regions), H3K27me3 (38 337 regions) and histone-enriched (25 114 regions; 33) and hypomethylated regions (79 124 regions; 34). Identified regions as reported by authors in hg18 genome build coordinates were converted to hg19 coordinates using the UCSC Genome Browser liftover tool. Minimum distance of center point of RNA clusters to closest local element of each histone region or hypomethylated DNA region set was calculated using GenomeInspector (Genomatix), yielding a correlation graph showing the distribution of clusters over a ± 10 kb region relative to the center of each element.

RESULTS AND DISCUSSION

Two separate protocols were used to construct sequencing libraries from Total and poly(A) fractionated sperm and testes RNAs. Sequencing characteristics for all sample libraries are summarized in Supplementary Table S4. Reads not aligned (total reads − aligned reads) are primarily those rejected due to a low Quality Control score. Reads that could not be assigned to a unique genomic location despite paired-end mapping were considered separately in subsequent analyses (below). Although it is assumed that sequencing libraries prepared from Total RNA can more accurately portray the complete RNA profile of a cell, alternative methods of preparation are important, as they significantly reduce the contribution of the abundant ribosomal and mitochondrial and fragmented RNAs. In this regard, it has been shown that 28S and 18S ribosomal RNAs are essentially fragmented in the mature gamete ensuring translational silencing (26,35). This precludes their depletion by available technologies and hinders efficiently achieving adequate sequencing depth. Only those sperm RNAs of subsequent biological significance are expected to escape fragmentation. Acknowledging how sequencing library construction impacts transcript representation was essential in the characterization of the unique population of RNAs retained in sperm.

Library characteristics

Both RNA transcript abundance and read profile were to some degree dependent on the protocol used for sequencing library construction. Average fragment length of the SPIA-prepared sequencing libraries was ~140 bp, while T-libraries (Total RNA) averaged between 70–90 bp. This increased the representation of transcripts of <100 nt in this set, enabling the resolution of sno and snaR RNAs. As expected, the poly(A^+) sperm and testes samples showed a significant decrease in the relative abundance of both mitochondrial and ribosomal RNAs compared with the total RNA libraries. Although SPIA did not use a poly(A^+) selection, the significant depletion of rRNA achieved by this method results from the use of a heterogeneous mixture of primers incapable of annealing to the highly structured ribosomal transcripts. The SPIA library preparation protocol provided an additional advantage over other strategies, as it required substantially less template RNA. This is of particular concern when investigating sperm RNAs (36). Briefly, total RNA was reverse transcribed using a proprietary mixture of chimeric RNA/DNA primers. Use of these hybrid oligonucleotides allows for continual priming and polymerization following second strand synthesis (37).

To assess library purity, a group of 523 transcripts were selected that were abundant in SPIA-testes (RPKM > 25, approximately 90th percentile), yet were relatively low or absent in the SPIA-sperm samples [RPKM (Aggregate) < 1; Supplementary Table S5]. RNAs that were depleted from mature sperm include those associated with Sertoli cells (22 genes, $P = 1.2$ e-04) and epididymis (22 genes, $P = 4.1$ e-06), confirming the absence of contaminating RNAs from surrounding cells. Other ontological groups of note include the absence of the large ribosomal subunit (11 genes, $P = 1.7$ e-06), and ribosome (16 genes, $P = 1.9$ e-04). This is consistent with the view that ribosomal fragmentation occurs earlier during the maturation phase of spermiogenesis, ensuring cessation of translation (26).

As exhibited by *ACSBG2* and shown in Figure 1, the majority of T- and SPIA-library prepared transcripts displayed similar sequence-read profiles. However, clear differences were apparent for several transcripts, including *PRM2* (Supplementary Figure S2), *PRM1* and several non-coding RNAs. Analyses presented below for sperm transcript profiling used the T-libraries. Comparisons between sperm and other tissues used equivalently prepared SPIA-libraries.

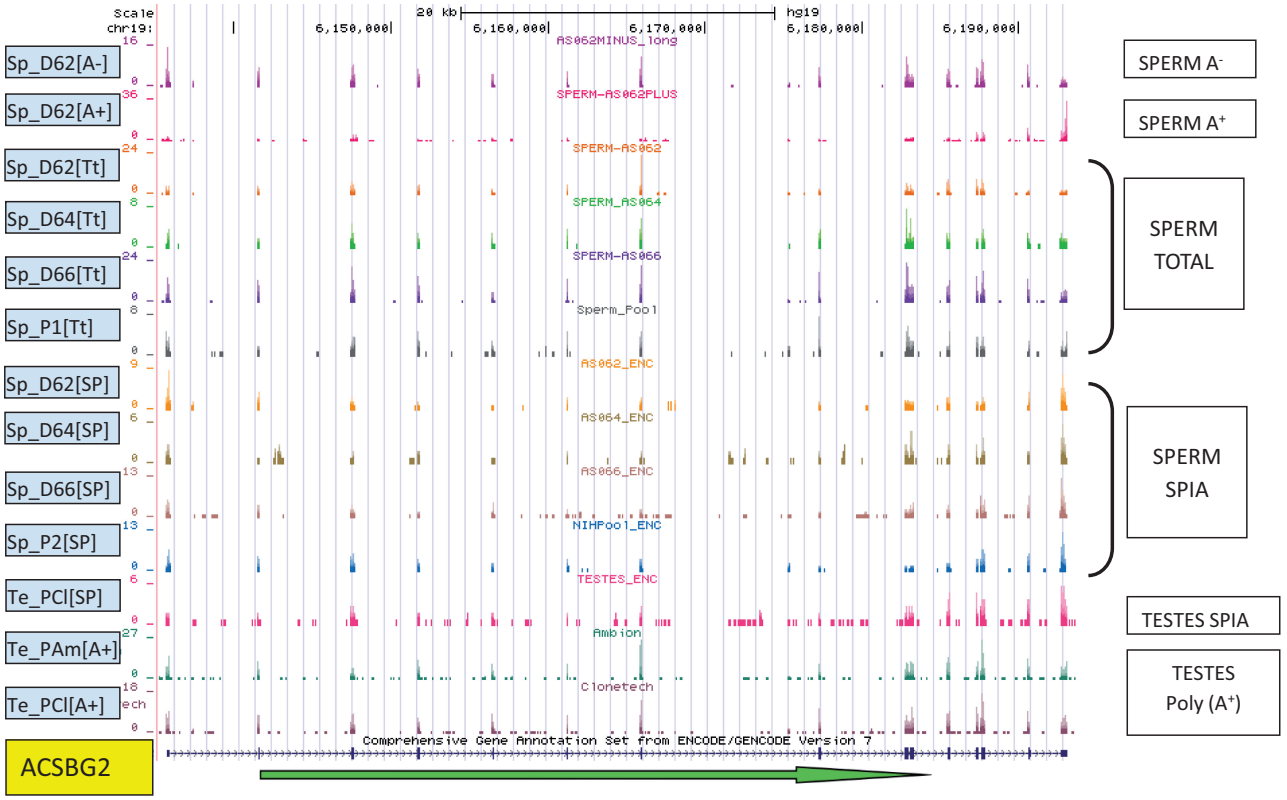

Figure 1. The distribution of sequencing reads across ACSBG2 in all libraries. The green arrow indicates transcript orientation. The number of reads corresponding to each read start at each base position is represented on the vertical axis. The scale for each sample is based on maximum read count within the displayed region. Four Total Sperm samples—Sp_D62[Tt], Sp_D64[Tt], Sp_D66[Tt] (individual donors) and Sp_P1[Tt] (mixed pool of three random donors)—were not subject to any of the commonly used RNA-selection methods. Sample D62 was additionally separated into Poly(A^+) (Sp_D62[A^+]) and Poly(A^-) (Sp_D62[A^-]) fractions by oligo(dT) selection. Commercially obtained pooled testes RNA—Te_PAm[A^+] (Applied Biosystems/Ambion, Austin, TX, USA, Lot 054P010702031A) and Te_PCl[A^+] (ClonTech, Mountain View, CA, USA, Lot 3090051)—were subject to Poly(A^+) selection. Sequences from Single Primer Isothermal Amplification (SPIA—Nugen Ovation Nugen Inc., San Carlos, CA, USA) (Sp_D62[SP], Sp_D64[SP], Sp_D66[SP]), a single pooled sperm sample (Sp_P2[SP]) and a single pooled testes (Te_PAm[SP]) prepared libraries are compared.

Comparison of rank transcript abundance between samples as measured by RNA-seq yielded an average Spearman rank correlation between the four SPIA-sperm samples of $\rho = 0.83$ (0.76–0.89). The correlation between individual SPIA-sperm samples to SPIA-testes was $\rho = 0.57$ (0.54–0.61), while the correlation of SPIA-sperm with a somatic UHR sample was $\rho = 0.38$ (0.35–0.40) and SPIA-testes with UHR was $\rho = 0.67$. These values reflect both the unique composition of sperm transcripts and their stability compared with somatic cells. Although the correlation between sperm sample transcript rank was acceptable, the absolute measure of abundance of many transcripts varied between sperm samples. This may reflect the wide-ranging degrees of fragmentation and/or the undirected expulsion of many RNAs during maturation.

Multiple-mapped reads

One issue uniquely encountered in RNA-seq as compared with microarray analysis is the treatment of reads mapping to multiple genomic locations (38,39). Many RNA-seq studies simply ignore these reads, assuming that within a coding sequence, 'repetitive' elements are of marginal significance. However, there are many genes for which identical sequences of the complete or partial transcript are annotated in multiple genomic locations. Sequence reads derived from these transcripts would map equivilently to all these locations. For example, transcripts like, *SPANXB1/B2/F1* appear replicated across multiple locations. In comparison, others exhibit limited reiteration of sequence spans, e.g., the 3′ exon of *STK19*, across multiple locations. Accordingly, if multiple alignment reads were discarded, transcript levels would be severely underestimated as highlighted by the sperm-associated *SPANX* transcript (40). Accordingly, all multiple-mapped reads were separately retained for transcripts exceeding RPKM > 2, and data are summarized in Supplementary Table S6; sample Sp_D62[Tt].

In comparison with the above, transcript levels determined by probe hybridization are not affected by multiple genomic origins of individual RNAs, as the level of each is estimated as an aggregate for that probe. For example, when comparing the recent microarray-based survey of human round spermatid RNAs (41) several of these transcripts are virtually absent from the RNA-Seq-processed data sets when multiple-mapped reads are ignored (Supplementary Tables S2, S3 and S6).

This apparent discordance is resolved when multiple reads are correctly assigned, e.g., *SPANX* genes, histone *HIST2HAA3/4*, *CXorf51*, *RIMBP3/3B/3C* and *HSFY1/2*. This independent validation suggests that other multiple-mapped transcripts, e.g., *snaR*s, require consideration.

Alternative polyadenylation and isoforms

Alternative polyadenylation (APA) is a transcriptional modification whereby the 3′UTR is cleaved at a secondary poly(A) recognition site, giving rise to a transcript possessing an unaltered coding sequence and a significantly shortened cell-specific 3′ UTR (42–45). *EFHD2* RNAs in the male reproductive tract offer an example of APA. This transcript is represented in all sperm RNA-seq data sets, and its coding sequence shown to be intact by RT-PCR (Supplementary Figure S3). But unlike that observed in testes, the UTR of this RNA is truncated in sperm relative to annotated transcripts (Figure 2). Approximately 20% of sperm transcripts exhibit clear APA sites relative to the full-length 3′ UTR of the primary annotated transcripts. Interestingly, more than one-half are sperm-specific and ontologically nondescript (Supplementary Table S7). This modification has been shown to influence mRNA stability and would impact relative abundance (46,47).

As determined by comparison of the T-libraries, several novel isoforms appear to be unique to sperm. These include the multi-functional pyruvate kinase *PKM2* that is closely associated with the regulation of cellular energy pathways (48). Of note, conservation of this sperm transcript extends to *Arabidopsis thaliana* sperm homologue *AT4G26390* (49) suggesting significance. In human sperm, this mRNA possesses two additional exons, which are detected at levels comparable with the remainder of the transcript (Figure 2B; boxes). Interestingly, each exon has been observed in other isoforms in testis and amygdala. Individual mapping of paired-end reads indicates that both exons are present in novel sperm-specific isoforms not observed in testes, the UHR RNAs or in unfertilized human oocytes (50). The sperm PKM2 isoforms also exhibit mutually exclusive use of exons 9 or 10. The exon 9 isoform is linked with aerobic respiration oxidative phosphorylation and plays a role in ATP production in sperm flagella (51). The M2 exon 10 isoform is a member of the aerobic glycolytic pathway often observed in growing cells (52). Delivery of these unique sperm transcripts may provide a paternal regulator which silences *OCT4* maintaining the pluripotent state (53) as required post-fertilization.

Sperm/testes-associated transcripts

Transcripts that are abundant in both testes and sperm (RPKM > 25, SPIA-samples), but are absent or at very low levels in the UHR sample (RPKM < 1) were also apparent. This set of 102 transcripts (Supplementary Table S8) comprises those that are present throughout spermatogenesis and retained through the final stages of maturation. This group was independently validated in a recent testes microarray data set (41). Within this set of 102 RNA-seq sperm transcripts, 90 transcripts were interrogated by microarray probes, of which 42 spermatid transcripts were concordant. These RNAs are likely critical to gametogenesis, e.g., structural components—flagellum (six genes, $P = 4.2$ e-09), acrosomal vesicle (five genes, $P = 1.1$ e-06) and nucleosome (three genes, $P = 2.2$ e-03). Of particular significance are the large number of genes (22 genes, $P = 1.6$ e-11) in this group that are associated with infertility. Mouse knockout data from MGI (54) shows that 13 are associated with a male factor contribution to infertility (Supplementary Table S8). This includes *PCYT2*, which is only found in testes and sperm. It is the primary regulator of the CDP–ethanolamine pathway, which when deleted, confers complete embryonic lethality (55). One can pose that like *PCYT2* those that await characterization may similarly contribute other essential male factors.

Abundant sperm transcripts

Of the 22 302 unique transcripts surveyed, 726 displayed RPKM levels greater than 50 in the aggregate of T-sperm libraries and were thus considered as representative of the highly abundant sperm transcript class (Supplementary Table S9). Of these transcripts, 565 were identified by ontological analysis as linked to testes ($P = 6.5$ e-48), 54 were spermatid related ($P = 3.2$ e-19), 44 were associated with spermatogenesis ($P = 3.0$ e-15) and 76 were linked to infertility ($P = 7.8$ e-09). This is consistent with these RNAs providing a critical male function and suggests that they may provide a resource of diagnostic markers of fitness (56). Enrichment of genes associated with sperm motility (six genes, $P = 4.6$ e-05), acrosomal vesicle (eight genes, $P = 1.4$ e-04), components of the RAN pathway (four genes, $P = 4.4$ e-04) associated with regulation of centrosomal activity during mitotic division of spermatogonia (57,58), and structural components of sperm flagella such as cytoskeleton (84 genes, $P = 1.3$ e-06) and microtubule (29 genes, $P = 1.9$ e-07), clearly reflect transcripts coding for critical functions. On one hand, it is likely that some abundant sperm transcripts including the protamine genes that are required earlier in spermatogenesis and RNAs associated with nuclear organization (10 genes, $P = 5.1$ e-08) and the chromosome (33 genes, $P = 6.3$e-04) are no longer required in fully differentiated sperm and are simply residual. On the other hand, some abundant sperm transcripts may also function post-fertilization. For example, *RANBP2*, one of the intact sperm RNAs, is known to play an important role in nucleocytoplasmic transport and mitosis, reduction of which directly effects aneuploidy (59). This raises the intriguing possibility that *RANBP2* may be an example of a sperm-delivered transcript encoding a factor specifically involved in restructuring the paternal chromatin following fertilization. Consistent with a post-fertilization function, its knockout confers embryonic lethality (60,61).

Sperm-enhanced transcripts

Sperm transcripts appear to encompass a wide spectrum of RNAs that encode proteins spanning past, current and future roles. Comparison of the sperm and testes RNAs

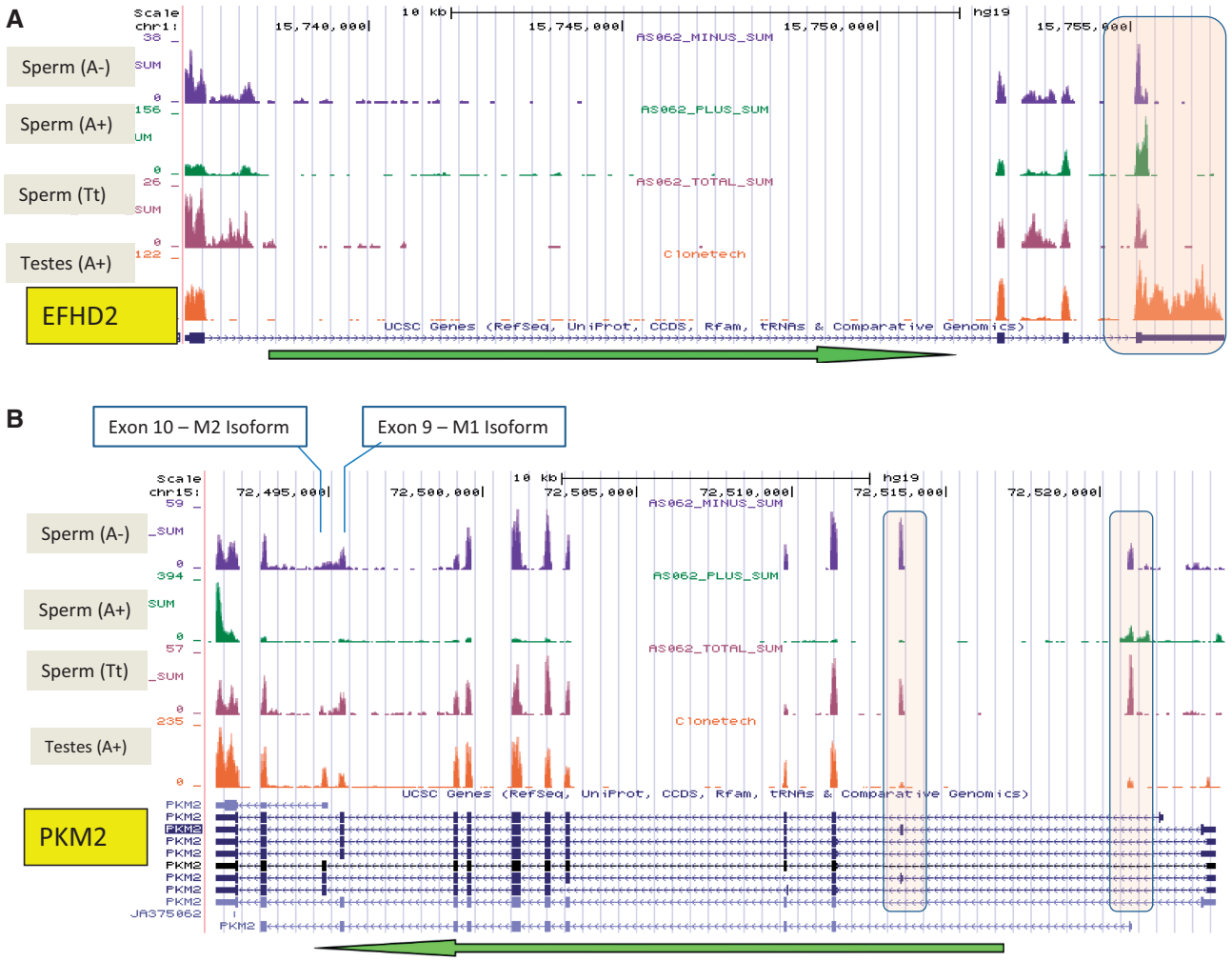

Figure 2. Novel sperm transcript expression profiles: The green arrow indicates transcript orientation. The number of reads corresponding to each base position is represented on the vertical axis. The scale for each sample is based on maximum read count within the displayed region. (**A**) The complete annotated 3′ UTR of *EFHD2* is 1.6 kb in length (pink box). In contrast to testes the 3′ UTR of this RNA in sperm is truncated to ~100 nt due to alternative polyadenylation. (**B**) An isoform of PKM2 specific to sperm contains exons (pink ovals) not observed together in other tissues. Examination of paired reads from these regions indicates splicing between these neighbouring exons and that these regions are transcribed at levels comparable with the remainder of the transcript. Previously reported isoforms containing either the M1 or M2 exon are presented below.

(SPIA-libraries) permits delineation of the transcripts and transcript groups that appear significant in sperm by their conspicuous abundance relative to testes. As transcription is believed to halt in the latter stages of spermiogenesis, the majority of RNAs which can be classified as present in sperm also appear in testes. Of the 2200 RNAs in the top decile of sperm-abundant transcripts, >85% are also present in testes at levels that are at least 50% of that observed in sperm. Most notable, however, are a number of sperm RNAs that appear to increase markedly in the latter stages of spermatogenesis, based on comparison of relative abundance of relatively intact transcripts in parallel sperm and testes libraries. This group likely represents a final burst of transcriptional activity during spermatid differentiation responsible for producing those RNAs with functions specific to the final stages of spermiogenesis or function subsequent to delivery. For example, among this group are keratins *KRT33A*, *KRT5* and *JUP*, which are necessary structural components of sperm cytoskeleton and microtubules (62,63). Another RNA in this class is *SIX3*, a member of the *SIX* family of transcription factors, which includes the newly identified sperm-associated *SIX5*, critical to spermiogenesis (64) that is associated with DNA binding (65). Also in this group and identified by multiple-mapped reads, is the RNA encoded by the Y-chromosome gene *HSFY 1/2* that is deleted in some cases of severe male infertility (66). Potential post-fertilization effector roles also include signalling pathway members like *SOCS1* that is encoded immediately downstream of the protamine locus (67). It is both a regulator of the JAK/STAT signal-transduction pathway (68) and acts to block insulin action (69), perhaps regulating early embryonic growth. Similarly, *NRARP*, a coordinator of Notch and Lef1-dependent Wnt signalling (70) and EGR3, an early growth response regulator of genes

controlling biological rhythm, were also identified as members of this sperm-enhanced class.

Non-coding RNAs

The classes of small non-coding sperm RNAs have been previously described (10). Several classes of larger non-coding transcripts are also present at significantly higher levels within the SPIA-sperm libraries relative to testes. This includes members of the small nuclear RNU family of RNAs. Together with protein factors, these transcripts comprise both the major (nuclear) and minor (cytosolic) spliceosomes. In addition to their role as components of spliceosome, these elements have also been found to be integrally associated with polyadenylation (71). Sperm RNU transcripts above the 80th percentile of abundance yet essentially absent in SPIA-Testes include *RNU11*, *RNU4-1*, *RNU4-2*, *RNU5A*, *RNU5E* and *RNU6ATAC*. As sperm are considered to be transcriptionally and translationally silent, there is presumably no active role for a spliceosome in the mature spermatozoon even though as independently shown (e.g., *RNU11*, Supplementary Figure S3) they appear intact. The abundance of these transcripts in sperm despite the absence of spliceosome activity suggests that they may serve a role on delivery to the oocyte. In late metaphase II (MII), prior to fertilization and until zygotic genome activation, the human oocyte is transcriptionally silent (72,73). Coincident with embryonic genome activation, dormant ribosomal and spliceosomal machinery must be reactivated (74). Perhaps the sperm RNU transcripts activate dormant zygotic post-transcriptional RNA-processing pathways.

Several snaR (small NF90-associated RNAs) transcripts including *snaR-C3*, *-C4*, *-E*, *-F*, *-G1* and *-I* are also abundant in sperm compared with testes. These novel small non-coding RNAs are unique to human and chimpanzee and are embedded in genomic regions that contain many genes important to reproduction, development and regulation of male fertility (75). Of particular note is the observation that all snaR elements also overlap hypomethylated regions (HMRs) in sperm. These regions (34) occupy ~4% of the total genome, and transcripts in the vicinity of these regions are abundant in sperm (discussed below). The observation that these regions overlap snaR elements suggests that local chromatin is specifically organized with respect to these transcripts. Even though their precise role remains to be determined, the following two observations require consideration. First, *snaR-G1* is both intact as shown in Supplementary Figure S3, and one of the most abundant RNAs within the sperm-enhanced class. Second, the sequence encoding this RNA directly overlaps the predicted promoter region for the beta 1 subunit (*CGB1*) of chorionic gonadotropin glycoprotein hormone (*CGβ*), a regulator of early placental development and implantation (76).

A number of other novel classes of ncRNAs (non-coding) were additionally noted to be uniquely abundant in sperm as compared with testes and observed at relatively similar levels in all sperm samples examined. Included among these are at least 250 RNAs that appear to be a specific intron of transcripts expressed in testes, but for which the spliced mature form is no longer present in sperm (Supplementary Table S12). Recent studies have suggested that intron-encoded RNAs may function as precursors of miRNAs or as regulatory factors (77,78). Some sperm RNAs appear as relatively short regions overlapping the UTRs or individual exons of several low abundance sperm RNAs (Supplementary Table S13). Local sequence analysis suggests that these regions are expressed independently of the transcripts in which they are located and that some may possess a poly(A) consensus site (AATAAA) indicative of antisense transcription (79–81). Sense-antisense regulation is a pervasive feature of transcriptomes and contributes to gene silencing, selective transcript editing, promoter inactivation and epigenetic modification (82–86). Previous studies in other tissues have shown that such transcripts are often independently regulated and only co-expressed with a corresponding sense transcript target at specific stages of development. Additionally, many RNAs present in sperm overlap regions identified by Gencode annotation (preV13) as either lncRNAs or as antisense elements (Supplementary Table S14) or isoforms. A final class of RNAs for which there is evidence in sperm is that of chromatin-associated RNAs (CARs). These RNAs are transcribed from intergenic or intronic regions of the genome, are found to bind to DNA and may play a role in regulation of DNA architecture or transcriptional regulation (87). Several previously identified CAR regions of the genome are transcribed at significant levels in sperm (Supplementary Table S15). Further work is being done to more fully characterize and assess the functional importance of each of these ncRNA classes in sperm.

Pri-mir-RNAs

In addition to the previously characterized population of mature sperm micro-RNAs (10), the paternal gamete also harbours a number of ~200–300 nt micro-RNA precursors (pri-miRNAs) that are essentially absent in testes. Examples of pri-mirRNAs present at high levels in sperm include *pri-mir-1181*, *-miR-3648*, *-miR-3687*, *-mir-663* and *-mir-181c*. The most abundant of these, *pri-mir-181c*, is not observed in its shorter mature form in sperm (10) nor in either form in testes (Supplementary Figure S4). This strongly suggests that this transcript is delivered to the oocyte as a precursor mir-RNA requiring post-fertilization activation. The ability of fertilized mouse oocytes to process sperm-borne pri-mi-RNAs into mature forms capable of modulating target RNAs has been demonstrated (16). To evaluate the potential post-fertilization role of mir-181c, 27 high-confidence targets were computationally identified. Several of the putative targets including *FIGN*, *ONECUT2*, *ZNF197*, *POU2F1*, *ROD1*, *ACRV2A*, *ACRV2B*, *TCERG1* and *FOXP1* are associated with differentiation (77), while *TNRC6B* is linked to the regulation of miRNA pathway (78). Their average relative level as a function of the 2-, 4-, 6-, 10-post-fertilized human oocyte (30) is illustrated in Figure 3 and detailed in Supplementary Table S10. Greater than 70% of the targets decreased in expression from the oocyte to

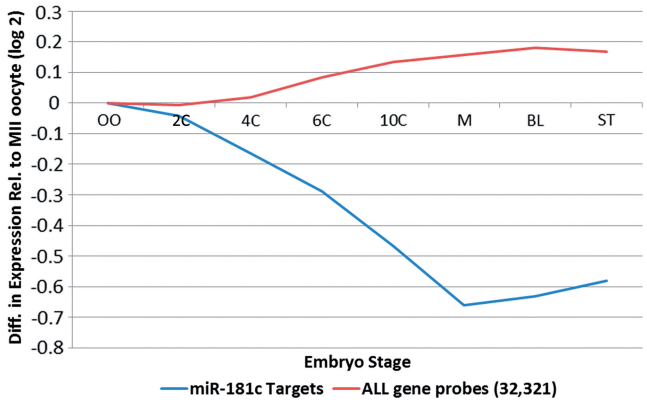

Figure 3. Average level of miR-181C targets as a function of preimplantation embryonic development. The average level of 27 high-confidence predicted miR-181C targets through preimplantation embryonic development OO: Oocyte, M:Morula, BL:blastocyst, ST:Stem cell is shown in blue. This is compared to the relative level of non-targeted genes indicated in red.

morula stage, with nine of these decreasing by a factor of 2 or more. Downregulation of targets from oocyte to morula stage is significantly greater than all genes ($P < 0.0001$) compared with the average expression of all 32 321 gene probes interrogated. This is consistent with the view that the sperm-delivered pri-*miR-181c* is processed to regulate genes of developmental consequence following delivery to the oocyte. Recent work has demonstrated that mature miR-181c modulates the expression levels of *CARM1*, a key regulator of critical pluripotency factors in human and mouse embryonic stem cells and blastomeres (79–81).

Sperm RNA integrity

A computational approach was developed to globally identify the population of intact RNAs amongst the various fragmented forms. The RNA-seq profiles from the 1000 most abundant T-sperm RNAs (Sample Sp_D62[Tt]) were used to establish a means of measuring individual transcript integrity. This analysis was predicated on the assumption that sequencing coverage would vary less across the length of an intact transcript relative to one which was fragmented. To determine the uniformity of coverage for each, RNA transcripts were divided into 100 bins, and a 5-bin moving average was used to calculate localized variations in sequencing coverage. The squared deviation from expected coverage for each bin was summed and used as an intactness score to rank the 1000 RNAs according to their stability (Figure 4A). This highlighted a subpopulation of intact RNAs (top, yellow) while accentuating the biased representation of the 5′-end of the transcripts and variable depth of sequencing coverage (bottom left, red). It is important to note that shortened 3′ UTRs due to APA (see *Alternative Polyadenylation and Isoforms*) may contribute to decreased 3′ representation of some transcripts. Three mRNAs of interest were selected from the top 200 most intact profiles to test *in silico* predictions of stability by RT-PCR using transcript spanning primers (Supplementary Table S1 and Figure S3). These included *NDUFA13*, *IZUMO4* and *CIB1*, which respectively encode a component of complex I of the mitochondrial respiratory chain, a protein involved in sperm-egg fusion and a calcium-binding protein essential to spermatogenesis (82–84). As described, three non-coding RNAs, *snaR-G1*, *RNU11* and *lnc-ERGIC1-1*, which were not considered in the above integrity analysis due to their reduced length were also evaluated. As shown in Supplementary Figure S3, full length cDNA products corresponding to these transcripts were detected in all samples. The observed differential fragmentation and preservation of select RNAs may reflect the diverse roles of these transcripts during spermiogenesis or following fertilization. To determine if RNA stability was functionally correlated, the ranked transcripts were divided into quintiles and subjected to ontological analysis (1 most intact, 5 least intact; Supplementary Table S11). Transcripts in the first quintile were associated with mature spermatids (24 genes, $P = 2.8$ e-16) and male infertility (26 genes, $P = 3.8$ e-8). Subsequent quintiles generally highlighted ontological classes that included posttranscriptional regulation of gene expression (12 genes, $P = 4.9$ e-6) and RNA binding (20 genes, $P = 9.4$ e-5). The number of RNAs found within mature spermatids and male infertility ontologies decreased with each subsequent quintile, with not one spermatid or infertility transcript being observed in the 5th and most deteriorated.

A scaling function was then developed to highlight directionality and degree of fragmentation. As shown in Figure 4B, the average sequencing coverage profile for each ranked and ordered decile highlighted three general categories of RNA stability: flat/intact (green; deciles 1–3), 5′ truncated (orange; decile 4) and 3′ truncated (purple; deciles 5–10). The presence of a pronounced curve at either end of a profile was observed for those sets of RNAs with a low intactness score. Interestingly, this analysis uncovered a group of 5′ truncated RNAs that were not readily observed in Figure 4A. These compromised sets were only weakly, if at all, associated with the ontological categories shared between the flat intact RNA groups. Though it is unlikely that truncated sperm RNAs serve any functional role following fertilization, some may serve as clinical biomarkers of successful spermatogenesis alongside the subpopulation of full-length transcripts identified above.

In contrast, the profiles that exhibited reduced variance in Relative Sequence Coverage were comprised mainly of transcripts with a high intactness score in the prior ranking and were enriched in similar ontological classes as the top intact quintile. These RNAs appear preserved and therefore are the most likely group to contain candidates for function following fertilization. This set of intact transcripts likely remains incomplete, reflecting the computational and technical (26) challenge of identifying full-length RNAs amongst a larger pool of transcriptional remnants. However, included in this preliminary set of flat and intact sperm RNAs is *INTS1*. This transcript encodes a subunit of the integrator complex that is required to process components of spliceosome. *Inst1* knockout mice arrest at the early blastocyst stage and exhibit a

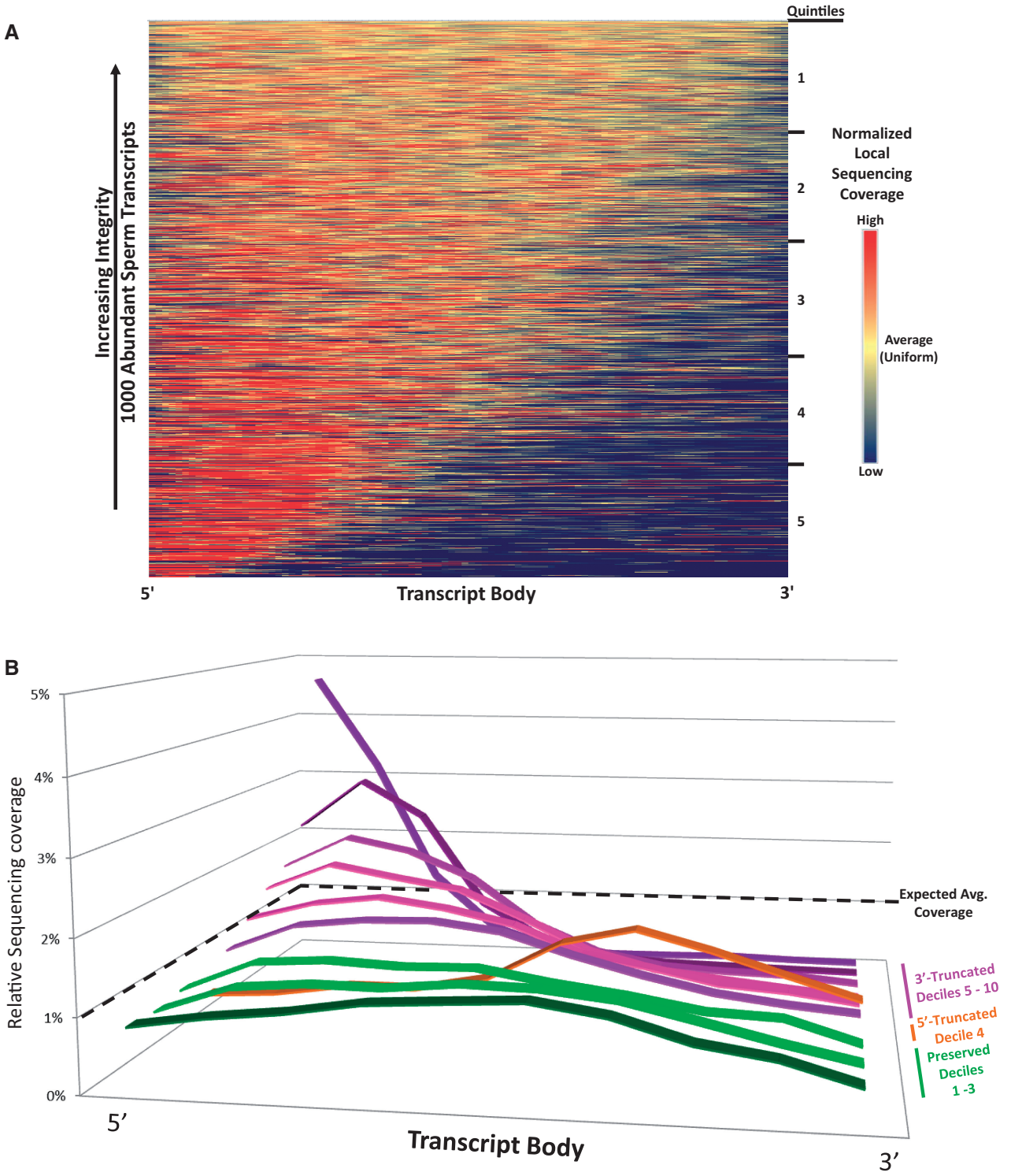

Figure 4. The most abundant sperm RNAs exhibit differential levels of stability. (**A**) Normalized RNA-seq coverage across the coding regions and UTRs of the 1000 most abundant sperm transcripts at a 100-bin resolution was used to identify intact RNAs. Transcripts are ranked in descending order of intactness and were divided into quintiles prior to ontological analysis. Profiles corresponding to stable RNAs exhibit uniform sequence coverage across the transcript (top, yellow). Profiles exhibiting increasing heterogeneity of coverage (bottom, red) are indicative of truncation. (**B**) A scaling function was applied to the above integrity scores to rank transcripts by their directionality and degree of fragmentation. The average sequencing coverage profiles for decile bins are presented. The expected level of coverage for an ideal transcript is shown as a dashed line. To highlight their likely functional importance, the flattest transcript deciles are prioritized to the foreground.

concomitant accumulation and reduction of precursor and mature U2 snRNA, respectively (85). The level of this transcript increases (fold-change 1.5) in human two-cell embryos relative to the unfertilized oocyte, suggesting a paternal contribution (30).

Epigenetic modification of sperm transcript-associated genes

During spermatogenesis, the majority of the nucleosomes are replaced with protamines, compacting chromatin to a near crystalline state (86,87). This transition to a protamine-bound genome is evident from the significant underrepresentation in sperm of the >100 transcripts encoding both canonical and variant histones present in testes and/or somatic cells. Confirming previous results, some histone transcripts, including *HIST1H1T*, *HIST1H2AA*, *HIST1H2BA (TSH2B)* and *H1FNT*, revealed by RNA-seq are more abundant in testes than in somatic cells (UHR; 88). Additionally, histone variants *H1FNT*, *H2AFJ*, *H3F3C* and *HILS1* are retained at a fairly high level in mature sperm. While *H1FNT* (*H1T2*) plays a role in spermatid elongation and DNA condensation (89,90) and *HILS1* is similarly noted to aid in chromatin remodelling during spermiogenesis (91), the specific function of the other sperm-retained histone RNAs is not yet known. They may represent the last histones used prior to replacement by the transition proteins and protamines, perhaps facilitating chromatin compaction or essential for paternal imprinting.

A comparatively small fraction of the paternal genome (5–15%) escapes protamination (92). Perhaps these regions encode transcripts required during compaction, encompass regions important early in development, or ensure protamine replacement with histones upon fertilization (8,93,94). These nucleosome-bound regions may be further differentiated based on presence or absence of specific histone modifications, DNA methylation and other genomic marks. The relationship between retained or specifically modified histones as a function of the associated sperm RNAs is illustrated in Figure 5. Sperm RNAs are associated with HMRs, H3K4me3 marks and histone-retained regions but not H3K27me3 enrichment.

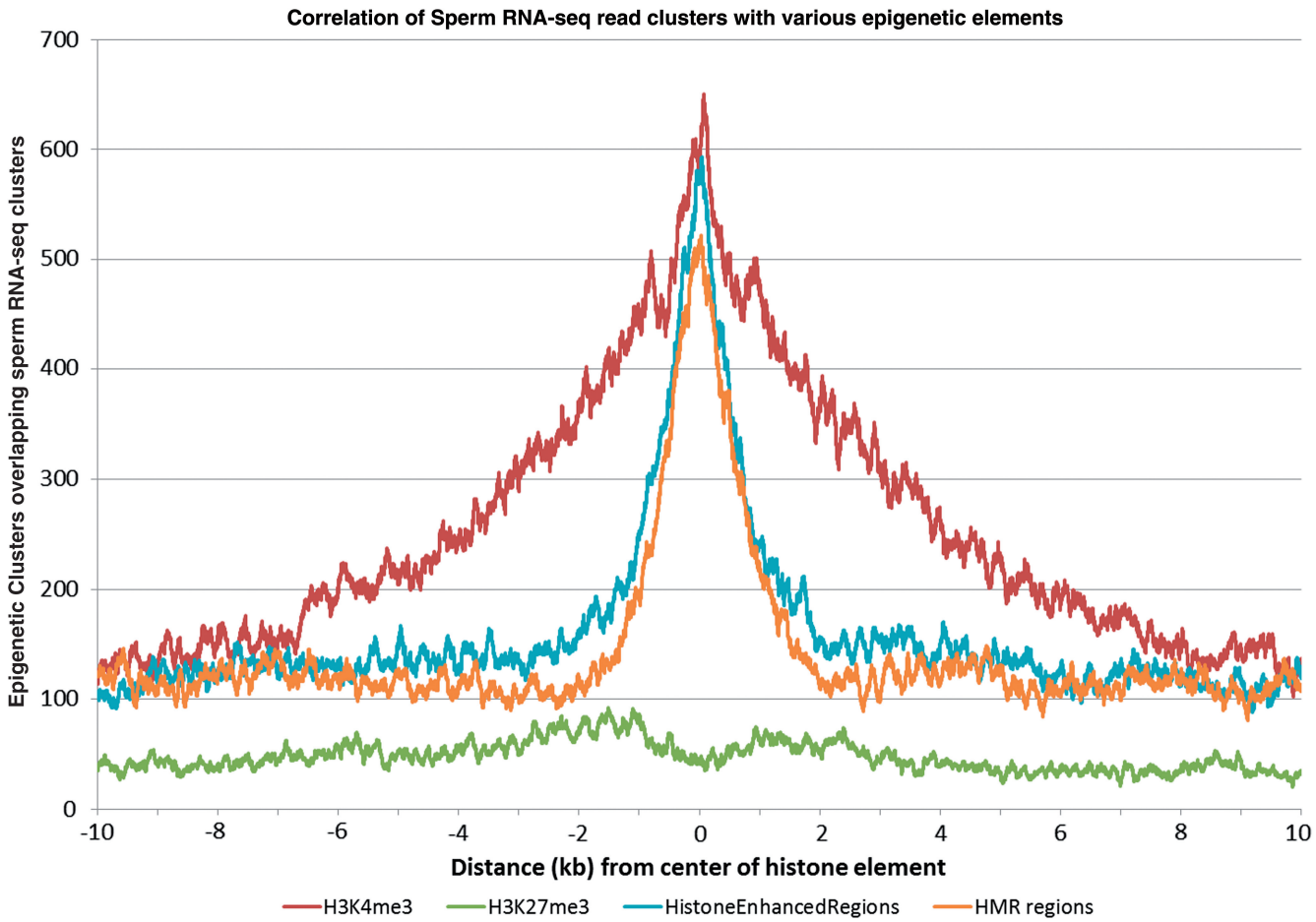

Figure 5. Sperm RNAs are correlated with the genomic positioning of histones and epigenetic elements. A total of 19 521 highly expressed RNA-seq clusters were identified in the promoter and exonic regions of sperm sample Sp_D62[Tt]. Each distance correlation is centered on the genomic coordinates of elements within one of the four following classes, H3K4me3 and H3K27me3 enriched regions (n = 34 912 and 38 337, respectively; 95), histone-enriched regions (n = 25 114; 33) and hypomethylated DNA regions (n = 79 124; 34).

The high level of H3K4me3 observed in sperm in association with developmental factors *EVX1*, *EVX2*, *ID1*, *STAT3*, *KLF5*, *FGF9*, *SOX7* and *SOX9* has previously been suggested to mark sites poised for transcription following fertilization (33). Although these genes are marked by histone modifications correlated with active transcription when delivered to the oocyte (95–97), the coincident occurrence of their corresponding transcripts in sperm must be considered. Perhaps the H3K4me3 mark correlated with the high levels of *EVX2*, *KLF5*, *SOX9* and *EVX1* RNAs retained by the mature male gamete represents a transcriptional ghost. In contrast, the developmental regulatory HOX genes are not expressed in spermatogenesis and yet they are hypomethylated and enriched in nucleosomes, some of which contain the active H3K4me3 and repressive H3K27me2 modified histones (Supplementary Figure S4A and C; 33). Accordingly, the transcriptionally silent yet epigenetically poised bipotential state of the HOX clusters in sperm infers that these genes are packaged for early embryonic use. These domains also harbour *EVX1* and *EVX2*, two regulators of early embryonic development, which are evolutionarily related to the HOX genes (98,99). Examination within the vicinity of *EVX1* and *EVX2* revealed significant retention of an RNA encoded by the last two exons of *EVX2* in addition to transcripts corresponding to sequences proximal to the 3′-ends of both genes (Supplementary Figure S5B and D). Reads from these regions appear unique to sperm compared with testes, and may be indicative of novel non-coding, potentially regulatory transcripts.

The population of RNA in sperm as revealed by RNA-seq provides a window into the developmental history, functional viability and potential elements present by sperm that may serve a role in the final stages of spermiogenesis or at fertilization. These include a number of coding and non-coding RNAs, the primary role of which in either sperm or the oocyte can now be delineated. Correlation of RNA abundance in sperm with epigenetic marks has permitted a more complete picture of how these marks may act to regulate both the pre- and post-fertilization genome. The observation that an experimentally derived mouse gynogenote may not strictly require a paternal contribution (100) remains to be reconciled. It is likely that in this case, the confrontation–consolidation pathway [reviewed in (101)], in which the paternal RNAs may participate in genome compatibility recognition, was sidestepped. Further, thought must be given to the documented role of the paternally derived mir-34c mediating first cell division (102), the most abundant human sperm micro RNA (10) or the function of other RNAs as paternal epigenetic modifiers [reviewed in (103)]. The identification of transcripts based on traits including relative abundance, integrity, knock-out phenotypes and presence or absence in the unfertilized oocyte, provides the essential foundation to elucidate the role of male factors critically important to the fertilization, birth and long-term health of a child.

ACCESSION NUMBERS

Sequencing data is deposited in GEO as:

GSM721696	AS062 (Sp_D62[Tt])
GSM721697	AS064 (Sp_D64[Tt])
GSM721698	AS066 (Sp_D66[Tt])
GSM721699	AS062PLUS (Sp_D62[A^+])
GSM721700	Testes Ambion (Te_PAm[A^+])
GSM721701	Testes Clonetech (Te_PCl[A^+])
GSM1037852	AS062 (Sp_D62[SP])
GSM1037853	AS064 (Sp_D64[SP])
GSM1037854	AS066 (Sp_D66[SP])
GSM1037855	Pool2 (Sp_P2[SP])
GSM1037856	Testes Ambion (Te_PAm[SP])

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–15, Supplementary Figures 1–5 and Supplementary Methods.

ACKNOWLEDGEMENTS

The authors are grateful to Ester Anton Martorell and Meritxell Jodar Bifet for review and comments to the manuscript and Leah Sendler for her visual review and identification of many transcript features. Samples were also prepared for sequencing by Claudia Lalancette. Support 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, which provided some of the semen samples for analysis, is appreciated. Conceived and designed the experiments: M.P.D., R.H., S.A.K.; Performed the experiments: E.S., G.D.J., R.J.G.; Analyzed the data: E.S., G.D.J., S.M., S.A.K.; Contributed reagents/materials/analysis tools: M.P.D., R.H.; Wrote the manuscript: E.S., G.D.J., S.A.K.; Reviewed the manuscript: E.S., G.D.J., S.M., R.J.G., M.P.D., R.H., S.A.K.

FUNDING

GENI pilot grant from Harvard School of Public Health (to S.A.K. and R.H.); National Institute of Environmental Health Sciences [ES017285 to R.H.]; Charlotte B. Failing Professorship (to S.A.K.); Michigan Core Technology grant from the State of Michigan's 21st Century Fund Program (to the Wayne State University Applied Genomics Technology Center). Funding for open access charge: Charlotte B. Failing Professorship (to S.A.K.).

Conflict of interest statement. None declared.

REFERENCES

1. Krawetz,S.A. (2005) Paternal contribution: new insights and future challenges. *Nat. Rev.*, **6**, 633–642.

2. Pessot,C.A., Brito,M., Figueroa,J., Concha,II, Yanez,A. and Burzio,L.O. (1989) Presence of RNA in the sperm nucleus. *Biochem. Biophys. Res. Commun.*, **158**, 272–278.
3. Sutovsky,P., Moreno,R.D., Ramalho-Santos,J., Dominko,T., Simerly,C. and Schatten,G. (1999) Ubiquitin tag for sperm mitochondria. *Nature*, **402**, 371–372.
4. Miller,D. (1997) RNA in the ejaculate spermatozoon: a window into molecular events in spermatogenesis and a record of the unusual requirements of haploid gene expression and post-meiotic equilibration. *Mol. Hum. Reprod.*, **3**, 669–676.
5. Miller,D. and Ostermeier,G.C. (2006) Towards a better understanding of RNA carriage by ejaculate spermatozoa. *Hum. Reprod. Update*, **12**, 757–767.
6. Ostermeier,G.C., Dix,D.J., Miller,D., Khatri,P. and Krawetz,S.A. (2002) Spermatozoal RNA profiles of normal fertile men. *Lancet*, **360**, 772–777.
7. Wykes,S.M. and Krawetz,S.A. (2003) The structural organization of sperm chromatin. *J. Biol. Chem.*, **278**, 29471–29477.
8. Wykes,S.M., Visscher,D.W. and Krawetz,S.A. (1997) Haploid transcripts persist in mature human spermatozoa. *Mol. Hum. Reprod.*, **3**, 15–19.
9. Ostermeier,G.C., Goodrich,R.J., Moldenhauer,J.S., Diamond,M.P. and Krawetz,S.A. (2005) A suite of novel human spermatozoal RNAs. *J. Androl.*, **26**, 70–74.
10. Krawetz,S.A., Kruger,A., Lalancette,C., Tagett,R., Anton,E., Draghici,S. and Diamond,M.P. (2011) A survey of small RNAs in human sperm. *Hum. Reprod.*, **26**, 3401–3412.
11. Yatsenko,A.N., Roy,A., Chen,R., Ma,L., Murthy,L.J., Yan,W., Lamb,D.J. and Matzuk,M.M. (2006) Non-invasive genetic diagnosis of male infertility using spermatozoal RNA: KLHL10 mutations in oligozoospermic patients impair homodimerization. *Hum. Mol. Genet.*, **15**, 3411–3419.
12. Platts,A.E., Dix,D.J., Chemes,H.E., Thompson,K.E., Goodrich,R., Rockett,J.C., Rawe,V.Y., Quintana,S., Diamond,M.P., Strader,L.F. *et al.* (2007) Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs. *Hum. Mol. Genet.*, **16**, 763–773.
13. Zhao,C., Huo,R., Wang,F.Q., Lin,M., Zhou,Z.M. and Sha,J.H. (2007) Identification of several proteins involved in regulation of sperm motility by proteomic analysis. *Fertil. Steril.*, **87**, 436–438.
14. Ostermeier,G.C., Miller,D., Huntriss,J.D., Diamond,M.P. and Krawetz,S.A. (2004) Reproductive biology: delivering spermatozoan RNA to the oocyte. *Nature*, **429**, 154.
15. Rassoulzadegan,M., Grandjean,V., Gounon,P., Vincent,S., Gillot,I. and Cuzin,F. (2006) RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. *Nature*, **441**, 469–474.
16. Liu,W.M., Pang,R.T., Chiu,P.C., Wong,B.P., Lao,K., Lee,K.F. and Yeung,W.S. (2012) Sperm-borne microRNA-34c is required for the first cleavage division in mouse. *Proc. Natl Acad. Sci. USA*, **109**, 490–494.
17. Lambard,S., Galeraud-Denis,I., Martin,G., Levy,R., Chocat,A. and Carreau,S. (2004) Analysis and significance of mRNA in human ejaculated sperm from normozoospermic donors: relationship to sperm motility and capacitation. *Mol. Hum. Reprod.*, **10**, 535–541.
18. Miller,D. (2000) Analysis and significance of messenger RNA in human ejaculated spermatozoa. *Mol. Reprod. Dev.*, **56**, 259–264.
19. Waclawska,A. and Kurpisz,M. (2012) Key functional genes of spermatogenesis identified by microarray analysis. *Syst. Biol. Reprod. Med.*, **58**, 229–235.
20. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
21. Fu,X., Fu,N., Guo,S., Yan,Z., Xu,Y., Hu,H., Menzel,C., Chen,W., Li,Y., Zeng,R. *et al.* (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, **10**, 161.
22. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev.*, **10**, 57–63.
23. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
24. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
25. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
26. Johnson,G.D., Sendler,E., Lalancette,C., Hauser,R., Diamond,M.P. and Krawetz,S.A. (2011) Cleavage of rRNA ensures translational cessation in sperm at fertilization. *Mol. Hum. Reprod.*, **17**, 721–726.
27. Goodrich,R., Johnson,G. and Krawetz,S.A. (2007) The preparation of human spermatozoal RNA for clinical analysis. *Syst. Biol. Reprod. Med.*, **53**, 161–167.
28. Johnson,G.D., Platts,A.E., Lalancette,C., Goodrich,R., Heng,H.H. and Krawetz,S.A. (2011) Interrogating the transgenic genome: development of an interspecies tiling array. *Syst. Biol. Reprod. Med.*, **57**, 54–62.
29. Ramskold,D., Wang,E.T., Burge,C.B. and Sandberg,R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.*, **5**, e1000598.
30. Vassena,R., Boue,S., Gonzalez-Roca,E., Aran,B., Auer,H., Veiga,A. and Belmonte,J.C. (2011) Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. *Development*, **138**, 3699–3709.
31. Maragkakis,M., Reczko,M., Simossis,V.A., Alexiou,P., Papadopoulos,G.L., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K. *et al.* (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.*, **37**, W273–W276.
32. Maragkakis,M., Alexiou,P., Papadopoulos,G.L., Reczko,M., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K., Simossis,V.A. *et al.* (2009) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, **10**, 295.
33. Hammoud,S.S., Nix,D.A., Zhang,H., Purwar,J., Carrell,D.T. and Cairns,B.R. (2009) Distinctive chromatin in human sperm packages genes for embryo development. *Nature*, **460**, 473–478.
34. Molaro,A., Hodges,E., Fang,F., Song,Q., McCombie,W.R., Hannon,G.J. and Smith,A.D. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, **146**, 1029–1041.
35. Gilbert,I., Bissonnette,N., Boissonneault,G., Vallee,M. and Robert,C. (2007) A molecular analysis of the population of mRNA in bovine spermatozoa. *Reproduction*, **133**, 1073–1086.
36. Goodrich,R.J., Anton,E. and Krawetz,S.A. (2013) Isolating mRNA and small noncoding RNAs from human sperm. *Methods Mol. Biol.*, **927**, 385–396.
37. Kurn,N., Chen,P., Heath,J.D., Kopf-Sill,A., Stephens,K.M. and Wang,S. (2005) Novel isothermal, linear nucleic acid amplification systems for highly multiplexed applications. *Clin. Chem.*, **51**, 1973–1981.
38. Li,B., Ruotti,V., Stewart,R.M., Thomson,J.A. and Dewey,C.N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
39. Rozov,R., Halperin,E. and Shamir,R. (2012) MGMR: leveraging RNA-Seq population data to optimize expression estimation. *BMC Bioinformatics*, **13(Suppl. 6)**, S2.
40. Salemi,M., Calogero,A.E., Di Benedetto,D., Cosentino,A., Barone,N., Rappazzo,G. and Vicari,E. (2004) Expression of SPANX proteins in human-ejaculated spermatozoa and sperm precursors. *Int. J. Androl.*, **27**, 134–139.
41. Chalmel,F., Lardenois,A., Evrard,B., Mathieu,R., Feig,C., Demougin,P., Gattiker,A., Schulze,W., Jegou,B., Kirchhoff,C. *et al.* (2012) Global human tissue profiling and protein network analysis reveals distinct levels of transcriptional germline-specificity and identifies target genes for male infertility. *Hum. Reprod.*, **27**, 3233–3248.
42. Ara,T., Lopez,F., Ritchie,W., Benech,P. and Gautheret,D. (2006) Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics*, **7**, 189.

43. Di Giammartino,D.C., Nishida,K. and Manley,J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell*, **43**, 853–866.
44. Legendre,M., Ritchie,W., Lopez,F. and Gautheret,D. (2006) Differential repression of alternative transcripts: a screen for miRNA targets. *PLoS Comput. Biol.*, **2**, e43.
45. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
46. Ji,X., Kong,J. and Liebhaber,S.A. (2011) An RNA-protein complex links enhanced nuclear 3′ processing with cytoplasmic mRNA stabilization. *EMBO J.*, **30**, 2622–2633.
47. Richard,P. and Manley,J.L. (2009) Transcription termination by nuclear RNA polymerases. *Genes Dev.*, **23**, 1247–1269.
48. Gupta,V. and Bamezai,R.N. (2010) Human pyruvate kinase M2: a multifunctional protein. *Protein Sci.*, **19**, 2031–2044.
49. Borges,F., Gomes,G., Gardner,R., Moreno,N., McCormick,S., Feijo,J.A. and Becker,J.D. (2008) Comparative transcriptomics of Arabidopsis sperm cells. *Plant Physiol.*, **148**, 1168–1181.
50. Reich,A., Klatsky,P., Carson,S. and Wessel,G. (2011) The transcriptome of a human polar body accurately reflects its sibling oocyte. *J. Biol. Chem.*, **286**, 40743–40749.
51. Krisfalusi,M., Miki,K., Magyar,P.L. and O'Brien,D.A. (2006) Multiple glycolytic enzymes are tightly bound to the fibrous sheath of mouse spermatozoa. *Biol. Reprod.*, **75**, 270–278.
52. David,C.J., Chen,M., Assanah,M., Canoll,P. and Manley,J.L. (2010) HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature*, **463**, 364–368.
53. Lee,J., Kim,H.K., Han,Y.M. and Kim,J. (2008) Pyruvate kinase isozyme type M2 (PKM2) interacts and cooperates with Oct-4 in regulating transcription. *Int. J. Biochem. Cell Biol.*, **40**, 1043–1054.
54. Eppig,J.T., Blake,J.A., Bult,C.J., Kadin,J.A. and Richardson,J.E. (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Res.*, **40**, D881–D886.
55. Fullerton,M.D., Hakimuddin,F. and Bakovic,M. (2007) Developmental and metabolic effects of disruption of the mouse CTP:phosphoethanolamine cytidylyltransferase gene (Pcyt2). *Mol. Cell Biol.*, **27**, 3327–3336.
56. Lima-Souza,A., Anton,E., Mao,S., Ho,W.J. and Krawetz,S.A. (2012) A platform for evaluating sperm RNA biomarkers: dysplasia of the fibrous sheath–testing the concept. *Fertil. Steril.*, **97**, 1061–1066 e1061–1063.
57. Kalab,P., Pu,R.T. and Dasso,M. (1999) The ran GTPase regulates mitotic spindle assembly. *Curr. Biol.*, **9**, 481–484.
58. Carazo-Salas,R.E., Gruss,O.J., Mattaj,I.W. and Karsenti,E. (2001) Ran-GTP coordinates regulation of microtubule nucleation and dynamics during mitotic-spindle assembly. *Nat. Cell Biol.*, **3**, 228–234.
59. Dawlaty,M.M., Malureanu,L., Jeganathan,K.B., Kao,E., Sustmann,C., Tahk,S., Shuai,K., Grosschedl,R. and van Deursen,J.M. (2008) Resolution of sister centromeres requires RanBP2-mediated SUMOylation of topoisomerase IIalpha. *Cell*, **133**, 103–115.
60. Hamada,M., Haeger,A., Jeganathan,K.B., van Ree,J.H., Malureanu,L., Walde,S., Joseph,J., Kehlenbach,R.H. and van Deursen,J.M. (2011) Ran-dependent docking of importin-beta to RanBP2/Nup358 filaments is essential for protein import and cell viability. *J. Cell Biol.*, **194**, 597–612.
61. DeGregori,J., Russ,A., von Melchner,H., Rayburn,H., Priyaranjan,P., Jenkins,N.A., Copeland,N.G. and Ruley,H.E. (1994) A murine homolog of the yeast RNA1 gene is required for postimplantation development. *Genes Dev.*, **8**, 265–276.
62. Calvin,H.I., Hwang,F.H. and Wohlrab,H. (1975) Localization of zinc in a dense fiber-connecting piece fraction of rat sperm tails analogous chemically to hair keratin. *Biol. Reprod.*, **13**, 228–239.
63. Cowin,P., Kapprell,H.P., Franke,W.W., Tamkun,J. and Hynes,R.O. (1986) Plakoglobin: a protein common to different kinds of intercellular adhering junctions. *Cell*, **46**, 1063–1073.
64. Sarkar,P.S., Paul,S., Han,J. and Reddy,S. (2004) Six5 is required for spermatogenic cell survival and spermiogenesis. *Hum. Mol. Genet.*, **13**, 1421–1431.
65. Kawakami,K., Sato,S., Ozaki,H. and Ikeda,K. (2000) Six family genes–structure and function as transcription factors and their roles in development. *Bioessays*, **22**, 616–626.
66. Tessari,A., Salata,E., Ferlin,A., Bartoloni,L., Slongo,M.L. and Foresta,C. (2004) Characterization of HSFY, a novel AZFb gene on the Y chromosome with a possible role in human spermatogenesis. *Mol. Hum. Reprod.*, **10**, 253–258.
67. Kramer,J.A., Adams,M.D., Singh,G.B., Doggett,N.A. and Krawetz,S.A. (1998) A matrix associated region localizes the human SOCS-1 gene to chromosome 16p13.13. *Somat. Cell Mol. Genet.*, **24**, 131–133.
68. Komazaki,T., Nagai,H., Emi,M., Terada,Y., Yabe,A., Jin,E., Kawanami,O., Konishi,N., Moriyama,Y., Naka,T. *et al.* (2004) Hypermethylation-associated inactivation of the SOCS-1 gene, a JAK/STAT inhibitor, in human pancreatic cancers. *Jpn. J. Clin. Oncol.*, **34**, 191–194.
69. Rui,L., Yuan,M., Frantz,D., Shoelson,S. and White,M.F. (2002) SOCS-1 and SOCS-3 block insulin signaling by ubiquitin-mediated degradation of IRS1 and IRS2. *J. Biol. Chem.*, **277**, 42394–42398.
70. Phng,L.K., Potente,M., Leslie,J.D., Babbage,J., Nyqvist,D., Lobov,I., Ondr,J.K., Rao,S., Lang,R.A., Thurston,G. *et al.* (2009) Nrarp coordinates endothelial Notch and Wnt signaling to control vessel density in angiogenesis. *Dev. Cell*, **16**, 70–82.
71. Awasthi,S. and Alwine,J.C. (2003) Association of polyadenylation cleavage factor I with U1 snRNP. *RNA*, **9**, 1400–1409.
72. Tadros,W. and Lipshitz,H.D. (2009) The maternal-to-zygotic transition: a play in two acts. *Development*, **136**, 3033–3042.
73. Prioleau,M.N., Huet,J., Sentenac,A. and Mechali,M. (1994) Competition between chromatin and transcription complex assembly regulates gene-expression during early development. *Cell*, **77**, 439–449.
74. Yurttas,P., Vitale,A.M., Fitzhenry,R.J., Cohen-Gould,L., Wu,W., Gossen,J.A. and Coonrod,S.A. (2008) Role for PADI6 and the cytoplasmic lattices in ribosomal storage in oocytes and translational control in the early mouse embryo. *Development*, **135**, 2627–2636.
75. Parrott,A.M. and Mathews,M.B. (2007) Novel rapidly evolving hominid RNAs bind nuclear factor 90 and display tissue-restricted distribution. *Nucleic Acids Res.*, **35**, 6249–6258.
76. Hallast,P., Rull,K. and Laan,M. (2007) The evolution and genomic landscape of CGB1 and CGB2 genes. *Mol. Cell Endocrinol.*, **260–262**, 2–11.
77. Cox,G.A., Mahaffey,C.L., Nystuen,A., Letts,V.A. and Frankel,W.N. (2000) The mouse fidgetin gene defines a new role for AAA family proteins in mammalian development. *Nat. Genet.*, **26**, 198–202.
78. Lazzaretti,D., Tournier,I. and Izaurralde,E. (2009) The C-terminal domains of human TNRC6A, TNRC6B, and TNRC6C silence bound transcripts independently of Argonaute proteins. *RNA*, **15**, 1059–1066.
79. Xu,Z., Jiang,J., Xu,C., Wang,Y., Sun,L., Guo,X. and Liu,H. (2013) MicroRNA-181 regulates CARM1 and histone aginine methylation to promote differentiation of human embryonic stem cells. *PLoS One*, **8**, e53146.
80. Torres-Padilla,M.E., Parfitt,D.E., Kouzarides,T. and Zernicka-Goetz,M. (2007) Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature*, **445**, 214–218.
81. Wu,Q., Bruce,A.W., Jedrusik,A., Ellis,P.D., Andrews,R.M., Langford,C.F., Glover,D.M. and Zernicka-Goetz,M. (2009) CARM1 is required in embryonic stem cells to maintain pluripotency and resist differentiation. *Stem Cells*, **27**, 2637–2645.
82. Yuan,W., Leisner,T.M., McFadden,A.W., Clark,S., Hiller,S., Maeda,N., O'Brien,D.A. and Parise,L.V. (2006) CIB1 is essential for mouse spermatogenesis. *Mol. Cell Biol.*, **26**, 8507–8514.
83. Huang,G., Lu,H., Hao,A., Ng,D.C., Ponniah,S., Guo,K., Lufei,C., Zeng,Q. and Cao,X. (2004) GRIM-19, a cell death regulatory protein, is essential for assembly and function of mitochondrial complex I. *Mol. Cell Biol.*, **24**, 8447–8456.
84. Ellerman,D.A., Pei,J., Gupta,S., Snell,W.J., Myles,D. and Primakoff,P. (2009) Izumo is part of a multiprotein family whose members form large complexes on mammalian sperm. *Mol. Reprod. Dev.*, **76**, 1188–1199.

85. Hata,T. and Nakayama,M. (2007) Targeted disruption of the murine large nuclear KIAA1440/Ints1 protein causes growth arrest in early blastocyst stage embryos and eventual apoptotic cell death. *Biochim. Biophys. Acta*, **1773**, 1039–1051.
86. Oliva,R. and Ballesca,J.L. (2012) Altered histone retention and epigenetic modifications in the sperm of infertile men. *Asian J. Androl.*, **14**, 239–240.
87. Rousseaux,S., Caron,C., Govin,J., Lestrat,C., Faure,A.K. and Khochbin,S. (2005) Establishment of male-specific epigenetic information. *Gene*, **345**, 139–153.
88. Govin,J., Caron,C., Rousseaux,S. and Khochbin,S. (2005) Testis-specific histone H3 expression in somatic cells. *Trends Biochem. Sci.*, **30**, 357–359.
89. Martianov,I., Brancorsini,S., Catena,R., Gansmuller,A., Kotaja,N., Parvinen,M., Sassone-Corsi,P. and Davidson,I. (2005) Polar nuclear localization of H1T2, a histone H1 variant, required for spermatid elongation and DNA condensation during spermiogenesis. *Proc. Natl Acad. Sci. USA*, **102**, 2808–2813.
90. Singleton,S., Zalensky,A., Doncel,G.F., Morshedi,M. and Zalenskaya,I.A. (2007) Testis/sperm-specific histone 2B in the sperm of donors and subfertile patients: variability and relation to chromatin packaging. *Hum. Reprod.*, **22**, 743–750.
91. Yan,W., Ma,L., Burns,K.H. and Matzuk,M.M. (2003) HILS1 is a spermatid-specific linker histone H1-like protein implicated in chromatin remodeling during mammalian spermiogenesis. *Proc. Natl Acad. Sci. USA*, **100**, 10546–10551.
92. Johnson,G.D., Lalancette,C., Linnemann,A.K., Leduc,F., Boissonneault,G. and Krawetz,S.A. (2011) The sperm nucleus: chromatin, RNA, and the nuclear matrix. *Reproduction*, **141**, 21–36.
93. Balhorn,R. (1982) A model for the structure of chromatin in mammalian sperm. *J. Cell Biol.*, **93**, 298–305.
94. Li,Y., Lalancette,C., Miller,D. and Krawetz,S.A. (2008) Characterization of nucleohistone and nucleoprotamine components in the mature human sperm nucleus. *Asian J. Androl.*, **10**, 535–541.
95. Brykczynska,U., Hisano,M., Erkek,S., Ramos,L., Oakeley,E.J., Roloff,T.C., Beisel,C., Schubeler,D., Stadler,M.B. and Peters,A.H. (2010) Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nat. Struct. Mol. Biol.*, **17**, 679–687.
96. Hammoud,S.S., Nix,D.A., Hammoud,A.O., Gibson,M., Cairns,B.R. and Carrell,D.T. (2011) Genome-wide analysis identifies changes in histone retention and epigenetic modifications at developmental and imprinted gene loci in the sperm of infertile men. *Hum. Reprod.*, **26**, 2558–2569.
97. Vavouri,T. and Lehner,B. (2011) Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome. *PLoS Genet.*, **7**, e1002036.
98. Garcia-Fernandez,J. (2005) The genesis and evolution of homeobox gene clusters. *Nat. Rev.*, **6**, 881–892.
99. Kalisz,M., Winzi,M., Bisgaard,H.C. and Serup,P. (2012) EVEN-SKIPPED HOMEOBOX 1 controls human ES cell differentiation by directly repressing GOOSECOID expression. *Dev. Biol.*, **362**, 94–103.
100. Kawahara,M., Wu,Q., Takahashi,N., Morita,S., Yamada,K., Ito,M., Ferguson-Smith,A.C. and Kono,T. (2007) High-frequency generation of viable mice from engineered bi-maternal embryos. *Nat. Biotechnol.*, **25**, 1045–1050.
101. Bourc'his,D. and Voinnet,O. (2010) A small-RNA perspective on gametogenesis, fertilization, and early zygotic development. *Science*, **330**, 617–622.
102. Liu,W.M., Pang,R.T., Chiu,P.C., Wong,B.P., Lao,K., Lee,K.F. and Yeung,W.S. (2012) Sperm-borne microRNA-34c is required for the first cleavage division in mouse. *Proc. Natl Acad. Sci. USA*, **109**, 490–494.
103. Rando,O.J. (2012) Daddy issues: paternal effects on phenotype. *Cell*, **151**, 702–708.