Tittle:

A technical assessment of the porcine ejaculated spermatozoa for a sperm specific RNA-seq analysis

Authors and affiliations:

Marta Gòdia[a,*]; email: marta.godia@cragenomica.es ORCID: 0000-0002-0439-4014

Fabiana Quoos Mayer[a,b,*]; email: bimmayer@gmail.com ORCID: 0000-0002-9324-8536

Julieta Nafissi[a,c]; email: julietanafissi@gmail.com

Anna Castelló[a,d]; email: anna.castello@uab.cat

Joan Enric Rodríguez-Gil[e]; email: juanenrique.rodriguez@uab.cat

Armand Sánchez[a,d]; email: armand.sanchez@cragenomica.es ORCID: 0000-0001-9160-1124

Alex Clop[a,#]; email: alex.clop@cragenomica.es ORCID: 0000-0001-9238-2728

[a]Animal Genomics Group, Centre for Research in Agricultural Genomics-CSIC-IRTA-UAB-UB, Campus UAB, 08193 Cerdanyola del Valles, Catalonia, Spain. Tel: +34 93 5636600

[b]Permanent adress - Instituto de Pesquisas Veterinárias Desidério Finamor, Agricultural Diagnostic and Research Departament, Secretariat of Agriculture, Livestock and Irrigation, Eldorado do Sul, Rio Grande do Sul, Brazil

[c]Current address: Technology Institute (INTEC), Argentine University of Enterprise (UADE), Buenos Aires, Argentina

[d]Unit of Animal Science, Department of Animal Science and Nutrition, Autonomous University of Barcelona, Cerdanyola del Valles, Catalonia, Spain

[e]Unit of Animal Reproduction, Department of Animal Medicine and Surgery, Autonomous University of Barcelona, Cerdanyola del Valles, Catalonia, Spain

* These authors contributed equally to this work.

[#]Corresponding author.

Keywords:

boar sperm purification; sperm recovery rate; semen quality; RNA yield; RNA-seq.

Abbreviations

FPKM - Fragments Per Kilobase of transcript per Million mapped reads

KRT1 - Keratin 1

miRNA - micro RNA

miscRNA - miscellaneous RNA

Mt rRNA - mitochondrial ribosomal RNA

Mt tRNA - mitochondrial transference RNA

OAZ3 - Ornithine Decarboxylase Antizyme 3

ORT - Osmotic Resistance Test

piRNA - Piwi-interacting RNA

PRM1 - Protamine 1

PTPRC - Protein tyrosine phosphatase receptor type C

rRNA - ribosomal RNA.

snoRNA - small nucleolar RNA

snRNA - small nuclear RNA

SRR – Sperm Recovery Rate

tRNA - transfer RNA

Abstract

The study of the boar sperm transcriptome by RNA-seq can provide relevant information on sperm quality and fertility and might contribute to animal breeding strategies. However, the analysis of the spermatozoa RNA is challenging as these cells harbor very low amounts of highly fragmented RNA, and ejaculates also contain other cell types with larger amounts of non-fragmented RNA. Here we describe a strategy for a successful boar sperm purification, RNA extraction and RNA-seq library preparation. Using these approaches our objectives were: (i) to evaluate the sperm recovery rate (SRR) after boar spermatozoa purification by density centrifugation using the non-porcine-specific commercial reagent BoviPure^{TM}; (ii) to assess the correlation between SRR and sperm quality characters; (iii) to evaluate the relationship between sperm cell RNA load and sperm quality traits and (iv) to compare different library preparation kits for both total RNA-seq (SMARTer Universal Low Input RNA and TruSeq RNA Library Prep kit) and small RNA-seq (NEBNext Small RNA and TailorMix miRNA Sample Prep v2) for high throughput sequencing. Our results show that pig SRR (~ 22 %) is lower than in other mammalian species and that it is not significantly dependent of the sperm quality parameters analyzed in our study. Moreover, no relationship between the RNA yield per sperm cell and sperm phenotypes was found. We compared a RNA-seq library preparation kit optimized for low amounts of fragmented RNA with a standard kit designed for high amount and quality of input RNA and found that for sperm, a protocol designed to work on low quality RNA is essential. We also compared two small RNA-seq kits and found not substantial differences in their performance. We propose the methodological workflow described here for the RNA-seq screening of the boar spermatozoa transcriptome.

Introduction

RNA-seq is the current gold-standard technology for the high-throughput analysis of transcriptome profiles, which is essential to understand the molecular basis of phenotypes (Wang et al. 2009). Thus, if studied in livestock species, this information could contribute to designing animal breeding strategies. This method has been applied to map the transcriptome of multiple species and tissues including spermatozoa from human (Sendler et al. 2013), mouse (Fang et al. 2014), bovine (Card et al. 2013; Selvaraju et al. 2017) and horse (Das et al. 2013). Although these cells are considered transcriptionally and translationally inactive, they contain a wide population of coding and non-coding RNA molecules (Jodar et al. 2013), which functions have been related to spermatogenesis (Ostermeier et al. 2002), sperm chromatin reorganization (Martins and Krawetz 2005; Hamatani 2012), fertility potential (Jodar et al. 2015), early embryo development (Sendler et al. 2013) and trans-generational epigenetic inheritance (Rando 2016). Hence, the study of the sperm transcriptome is crucial for understanding its biology and its role in fertility, and can be thus of interest when applied to livestock research.

One of the main challenges for the study of the spermatozoa transcriptome is the extremely low RNA yield and high fragmentation of the transcripts typically present in these cells, as the standard RNA-seq chemistry normally requires a large amount (1 µg) of good quality RNA. To overcome this problem, new protocols to prepare high quality RNA-seq libraries from samples containing only tiny amounts (200 pg) of highly degraded RNA (e.g., paraffin embedded tissues) have been developed and already tested and compared in human sperm (Mao et al. 2014). A human mature sperm cell is estimated to contain a 600-fold lower amount of RNA than a somatic cell (Zhao et al. 2006). As a typical ejaculate contains somatic cells – mainly leukocytes, keratinocytes and other type of epithelial cells – as well as germ-line cells from different stages of spermatogenesis (Patil et al. 2013), the study of the spermatozoa transcriptome requires removing these RNA-rich cells for an unbiased analysis.

Somatic cells can be removed from sperm by the swim-up method (Jameel 2008), somatic cell lysis or by gradient centrifugation (Mao et al. 2013). Cell lysis approaches are efficient in eliminating somatic cells, but they also cause cell membrane damage and loss of mitochondrial sequences, thus risking to lose the sperm transcripts present in the cell's midpiece (Mao et al. 2013). Gradient centrifugation has been employed in the purification of sperm cells from several mammalian species using different commercial solutions, such as Percoll® (Ostermeier et al. 2002), PureSperm® (Sendler et al. 2013), EquiPure™ (Das et al. 2013) and BoviPure™ (Samardzija et al. 2006; Selvaraju et al. 2017). These gradients allow the motile mature spermatozoa to separate from somatic cells along with immature sperm cells (Mao et al. 2013). Typically, these commercial reagents are primarily used to improve sperm quality for artificial insemination, since they select progressive motile and morphologically normal spermatozoa (Samardzija et al. 2006). Although gradient centrifugation is convenient for these purposes, it significantly decreases the final number of recovered spermatozoa (Samardzija et al. 2006), adding yet another layer of complexity for the experimental analysis of the spermatozoa transcriptome. The sperm recovery rate (SRR) in gradient-based methods is mainly related to sperm motility (Samardzija et al. 2006), even though additional factors are likely to be involved since the number of recovered cells is lower than the expected based solely on initial motility values. The boar sperm is particularly sensitive to a wide spectrum of manipulations (Feugang 2017) and the use of a reagent not optimized for the porcine sperm may have detrimental effects on the SRR. Taking all this information into account, one of the main aims of this work was to evaluate the influence of different boar sperm quality traits on SRR after gradient density purification.

The levels of several RNA transcripts in sperm have been associated to semen quality traits and male fertility in many mammalian species including human (Jodar et al. 2012), cattle (Bissonnette et al. 2009) and pigs (Curry et al. 2011), among others. Likewise, abnormal levels of histone or protamine chromatin proteins in sperm have also been linked to spermatozoa defects (Carrell et al. 2007;

Hammoud et al. 2011) and it has been suggested that it could be related to spermatogenesis defects and alterations in RNA amounts in sperm cells (Carrell et al. 2007) and even sperm quality (Aoki et al. 2005). Thus, we searched for statistical relationship between RNA yield extracted per sperm cell and semen quality traits in swine. To determine the purity (lack of DNA or somatic RNA) of this sperm RNA, we developed three real-time quantitative PCRs (qPCRs) and tested its efficiency. Finally, we performed high throughput sequencing of a selection of these RNAs using two library preparation kits for total RNA-seq analysis (N = 6) and two kits for small RNA-seq analysis (N = 3) to compare their performances.

Results

Spermatozoa Recovery Rate

SRR was calculated in 285 samples and was in general low and with high variability between samples. The average SRR was 21.76 % with a standard deviation of 15.07 %. To shed light into the biological causes of this variance, we tested the dependence between SRR and sperm phenotypes. Significant covariates were adjusted for the given parameters: head abnormalities, tail abnormalities and distal droplets were adjusted for farm; motility 90 min for age; viability 0 min, viability 90 min, acrosomes 0 min and ORT for batch; acrosomes 90 min and neck abnormalities for farm and batch, and distal droplets for farm, age and batch. SRR and the sperm quality characters did not present normal distribution nor a linear relationship. Thus, a multivariate nonparametric test of independence was applied (Székely and Rizzo 2009). When considering the Bonferroni corrected *P*-value, SRR was found independent of all the sperm quality parameters (Table 1).

[Table 1 near here].

RNA yield

Total RNA was extracted from 190 samples. The RNA yields averaged 1.6 fg per sperm cell, with ranges from 0.4 to 4.8 fg. The RNA Integrity Number (RIN) values, measured on 70 samples, was low (RIN < 2.6) and with undetectable ribosomal RNA profiles, which indicates the absence of RNA of somatic cell origin. The amount of RNA extracted per sperm cell was not significantly associated to the covariates farm, age or batch. The test of independence indicated null relationship between the total RNA extracted per sperm cell and the sperm quality phenotypes studied (Table 2).

[Table 2 near here].

qPCR controls

The standard curves of the qPCR assays showed a good efficiency (97-97.9%). The three qPCR control assays displayed single peaks after the dissociation curve analysis, thus confirming that a single amplicon was generated in each reaction. The minus reverse transcription controls showed no amplification of *PRM1* and *PTPRC*. 70 RNA samples were subjected to qPCR, and all presented quantification cycles (Cq) ranging between 14.6 and 21.3 for the sperm-specific gene *PRM1*. In contrast, the average Cq for *PTPRC* was 35.4 in 49 sperm samples and undetectable in the other 21 samples. The ΔCq $_{\text{PTPRC-PRM1,}}$ calculated as the Cq for *PTPRC* minus the Cq for *PRM1* in the sperm samples, ranged from 14.3 to 21.3. The intergenic region was undetectable in 66 samples and had Cqs > 36 in the other 4 and the ΔCq $_{\text{Genomic-PRM1}}$ ranged from 18.4 to 21. As a comparison, the liver RNA showed a *PRM1* and *PTPRC* Cqs of 38 and 24, respectively.

RNA-seq library preparation, sequencing and mapping statistics

Four of the six samples that were chosen for total RNA-seq analysis (Sample_1 to Sample_6) presented ΔCq $_{\text{PTPRC-PRM1}}$ ranging from 17.4 to 19.1 and undetectable levels of *PTPRC* in the other two samples. Likewise, the ΔCq $_{\text{Genomic-PRM1}}$ ranged from 19.4 to 21 in three of the six samples and was undetectable in the other three.

The SMARTer and the TruSeq kits produced libraries with significantly different concentrations, which ranged between 53 and 120.7 nM (total RNA yield between 0.8 and 1.8 pmol) and between 0.5 and 2.9 nM (0.01 – 0.09 pmol), respectively (*P*-value = 0.03) (Table 3). All the libraries generated a similar percentage of high quality RNA-seq reads which mapped unambiguously to the swine reference genome, (SMARTer : 74.9 – 85.8 % and TruSeq: 70.8 – 82.3 %) (*P*-value = 0.13) (Table 3). Likewise, SMARTer yielded a higher percentage of reads uniquely mapping to annotated genes (37.8 - 48.4 %) when compared to the TruSeq libraries (28.8 - 38.5 %) (*P*-value = 0.02) (Table 3). The proportion of PCR duplicates was significantly higher for the TruSeq (89.3 - 97.9%) than for the SMARTer samples (75.9 - 80.3 %) (*P*-value = 0.03) (Table 3). We identified on average, 8,562 and 2,522 transcripts for the SMARTer and the TruSeq, respectively (*P*-value = 1.89 x 10^{-4}). The SMARTer datasets presented a mean FPKM of 363 and median FPKM of 4.8, whereas the TruSeq libraries showed a mean FPKM of 3,410 and median FPKM of 12.6. 32.5% and 46.1% of the genes were identified at intermediate or high abundance levels (FPKM ≥ 10) for both the SMARTer and the TruSeq, respectively.

[Table 3 near here].

Short RNA samples (Sample_7 to Sample_9) presented undetectable RNA levels for *PTPRC* and for the intergenic region with the qPCR assay. Sequencing and mapping of short RNAs with the NEBNext and the TailorMix displayed similar results. The proportion of reads mapping to annotated features was similar for both protocols (77.4 - 82.9 %) (Table 4). Most of these reads mapped to miRNAs (27.0 – 32.4 %) (Table 4), followed by mitochondrial tRNAs (22.1 – 27.3 %) and protein coding genes (12.6 – 15.5 %) (Table 4). The remaining mapped reads corresponded to snRNAs, piRNAs and tRNAs, among others (Table 4). Some of the most abundant miRNAs have been already identified in swine sperm or in other mammalian species and include miR10b and miR34c, among others (Capra et al. 2017; Chen et al. 2017; Jodar et al. 2013).

[Table 4 near here].

Further analysis of the transcriptome profile was carried with the SMARTer datasets using the totality of the reads generated in each library (between 18.5 and 26.9 million reads per sample). Genes related to somatic cell contamination, *PTPRC* and *KRT1* were absent (mean FPKM = 0.3 and 0, respectively) in these samples (Supplementary Table 1). On the contrary, the sperm specific *PRM1* and *OAZ3* were among the most abundant transcripts with mean FPKMs of 15,368 and 22,670, respectively (Supplementary Table 1). The pattern of relative expression of these four genes in porcine white blood cells and in ear tissue was inverted when compared to sperm. Whilst *PRM1* and *OAZ3* were absent, *PTPRC* and *KRT1* were abundant in the white blood cells and in the ear RNA-seq datasets, respectively (Figure 1). We also quantified the amount of other previously reported somatic and sperm specific gene biomarkers (Jodar et al. 2016). The abundance of the epithelial *CDH1*, keratinocyte *KRT10*, leukocyte *IL8*, whole blood *HBB* and prostate *KLK3* genes ranged between 0 and 9 FPKM. On the contrary, the sperm-specific genes *PRM2*, *TNP1*, *ODF1* and *SMCP* showed average FPKMs ranging between 779 and 7,742 (Supplementary Table 1).

[Figure 1 near here].

Discussion

Although spermatozoa are considered transcriptionally inactive, there is growing evidence that the sperm RNA populations are related to spermatogenesis, fertility potential, chromatin reorganization, embryo development and transgenerational epigenetic inheritance (Bohacek and Mansuy 2015; Jodar et al. 2013). Since RNA load in spermatozoa is considerably lower than in somatic cells, an adequate separation of these populations is imperative to study the spermatozoa transcriptome. The application of purification methods decreases the final number of recovered sperm cells, and consequently the cell availability for RNA extraction.

The present study is the first to analyze the performance of porcine SRR. The purification of the boar sperm using gradient centrifugation with the non-porcine-specific reagent BoviPure^{TM} yielded highly purified spermatozoa as demonstrated by qPCR ($\Delta Cq_{PTPRC-PRM1} > 16$) for the vast majority (97 %) of the 70 samples. Nonetheless, the SRR was not only much lower but also more variable (21.76 ± 15.07 %) than that described in other species such as cattle (mean SRR = 31 %), human (69 %) and horse (63 %) (Allamaneni et al. 2005; Samardzija et al. 2006; Das et al. 2010). These differences may be due to the unique characteristics of the boar sperm. For example, the motility of the pig sperm after ejaculation is slower than in other species, (e.g., horse and cattle), while it is also very prone to be altered by a myriad of environmental incidences (Rodríguez-Gil and Bonet 2015). In light of these singularities, we addressed the question of which sperm quality factors are influencing SRR. The multivariate non-parametric test of independence revealed that none of the studied sperm traits were related to SRR. This is somehow unexpected, particularly for motility and cell viability, since a positive effect between these two traits and SRR have been previously described in cattle

(Samardzija et al. 2006). The differences in the physico-chemical properties between the ejaculate and the extender media in which the semen quality phenotypes are measured, and the BoviPure^{TM} reagent during centrifugation, may divergently affect semen quality. This would explain the lack of dependency between the semen quality measures and SRR. A complementary hypothesis is that the time and speed of the density gradient centrifugation step enables all the boar's motile sperm, either fast or slow, to end up reaching the bottom of the tube, and be thus recovered. This would imply that the sperm recovery with BoviPure^{TM} is not preferentially biased toward specific sperm sub-populations and therefore the molecular analysis of the recovered sperm robustly reflects that of the whole ejaculated mature sperm. Finally, SRR may be also affected by the composition and the physico-chemical characteristics of the ejaculates, which have been shown to be affected by diet (Byrne et al. 2017), or abstinency in humans (Agarwal et al. 2016), which is related to the frequency of ejaculates or the time from prior ejaculate in pigs.

Two determinant parameters for a successful transcriptome analysis are both the RNA quality and yield. The RNA extraction method becomes a critical step when working with spermatozoa, since these cells have low amount of highly fragmented RNA. In the present work, we chose the Trizol method for RNA extraction after having tested other protocols involving commercial kits, which yielded even lower RNA yields (data not shown). The average amount of RNA extracted per sperm cell was 1.6 fg, a similar value to previously reported data in domestic swine (Yang et al. 2009), but lower than human (Goodrich et al. 2013; Pessot et al. 1989) and mice (Pessot et al. 1989). The low amount of RNA recovered and low RIN value is in fact an indication that the removal of somatic cells, with their large amount of non-fragmented RNA, during the cell purification steps, was highly efficient. The observed variability in RNA yields between samples could be due to inter-sample differences in the epididymosomes secreted by epididymal epithelial cells, which have been involved in post-testicular spermatogenesis and are known to contain a repertoire of RNAs (Belleannée et al. 2013), yet this mechanisms remains to be elucidated.

Spermatogenesis is a highly regulated process with many genes tightly controlling the different maturation steps (Legrand and Hobbs 2017) and playing a role in the sperm's fertility potential (Jodar et al. 2015). Our study in 190 samples suggests that the sperm quality parameters that we assessed are independent of the amount of RNA recovered - as a proxy of RNA load - per sperm cell. qPCR assays were also developed with the aim to determine the presence of RNA from somatic origin and gDNA contamination in our samples. Most of our samples showed at most, only traces of *PTPRC* (68 samples displayed ΔCq $_{\text{PTPRC-PRM1}}$ > 16) and gDNA was only detected in 4 samples with ΔCq $_{\text{Genomic-PRM1}}$ > 18.4. In qPCR, the amplification curve is exponential and the template doubles at every cycle. This amplification follows this formula: $X_N = X_1 * 2^N$, where N is the number of amplification cycles, X_1 is the number of molecules prior amplification and X_N is the number of molecules after N PCR cycles. If we assume similar assay sensitivities, we can conclude that for a ΔCq $_{\text{PTPRC-PRM1}}$ = 16, the number of molecules of *PRM1* is 2^{16} = 65,536 times more abundant than the number of molecules of *PTPRC*. Likewise, when ΔCq $_{\text{Genomic-PRM1}}$ = 19, the number of *PRM1* RNA molecules is 2^{19} = 524,288 more abundant than the number of gDNA template. Hence, the majority of the RNA samples we processed were considered of sufficient spermatozoa purity. These qPCR assays can be used to determine sperm purity in porcine RNA samples and help selecting the purest RNAs for further analysis to obtain a reliable an accurate spermatozoa transcriptome. We must bear in mind that *PRM1* is also expressed in round spermatids (Siffroi et al. 1998; Steger et al. 2000) but we did not find any round-shaped cells in our samples following visual inspection of smear tissue under the microscope (Supplementary Figure 1). Thus, the presence of these cells is unlikely.

The purification and RNA extraction from boar sperm have proven to be suitable for the sequencing of total and small RNA by RNA-seq. To test the suitability of our samples for total RNA-seq analysis, we prepared sequencing libraries from six purified boar RNAs from different pigs with the SMARTer Universal Low Input RNA kit (Clontech) and with the TruSeq RNA Library Prep (Illumina) in parallel. Despite the fact that both protocols use the preferable amplification with random primers instead of poly-dT (Mao et al. 2014), the SMARTer libraries still outperformed the TruSeq in several standard RNA-seq quality control parameters. Nevertheless, this is expected as the SMARTer protocol and chemistry is optimized for samples with low amount (10 ng) of highly fragmented RNA as for example, formalin-fixed paraffin embedded tissues. First, although the SMARTer protocol required less input RNA and it included a lower number of cycles in the amplification steps, it consistently yielded a much higher amount and concentration of cDNA library (Table 3), which is crucial to obtain optimal sequencing results. Second, even though the RNA-seq metrics of the two kits were similar (Table 3), the significantly higher proportion of PCR duplicates in the TruSeq datasets indicates a lower library complexity and number of unique transcripts. This was also indicated by the significant difference between the number of uniquely identified transcripts in both kits. The SMARTer datasets yielded twice the number of transcripts than TruSeq (13,233 versus 6,642). The vast majority of the TruSeq transcripts, 6,452, were also detected in the SMARTer dataset (Figure 2.A). This suggests that a proportion of RNAs were not captured with the TruSeq library preparation protocol. With the SMARTer kit we detected transcripts with lower abundance than with TruSeq and ultimately, a more comprehensive profile of the sperm transcriptome. We need to point out that there are other protocols from different providers, including Illumina, that have been designed for the RNA-seq analysis of samples with low amount and quality RNA, which have not been tested in this study. Furthermore, cluster analysis of the transcript levels shows a major kit effect, clustering together the libraries generated with the same kit rather than the libraries generated from matched RNAs (Figure 2.B).

To evaluate the adequacy of our samples for the sequencing of small RNAs, we used three samples and two short RNA library prep kits, the NEBNext Small RNA Library Prep Set (New England Biolabs) and the TailorMix miRNA Sample Prep v2 (SeqMatic). Both protocols showed very similar RNA-seq metrics (Table 4) but the analysis of the NEBNext libraries evidenced first, a larger number of detected miRNAs and second, slightly more similar miRNA abundance between samples with a Pearson correlation of expression of 0.95 and 0.90 for the NEBNext and TailorMix, respectively.

RNA-seq is the gold standard approach for the genomic analysis of gene expression. This technology has been already applied to ejaculated sperm of different animal species such as human (Sendler et al. 2013), mouse (Fang et al. 2014), cattle (Card et al. 2013; Selvaraju et al. 2017) or horse (Das et al. 2013). In pig so far, the spermatozoa transcriptome was explored in 2009, using medium throughput sequencing approaches (Yang et al. 2009). The authors generated an Expressed Sequence Tag (EST) library and sequenced circa 5,000 clones using Sanger sequencing chemistry. This resulted in the identification of 271 genes with known function or cellular localization. That study, of high quality at that time, yielded a low number of transcripts and did not offer a comprehensive view of the boar sperm transcriptome (Yang et al. 2009). Current RNA-seq technologies, offer higher throughput and thousands of transcripts are typically detected. This is clearly illustrated in our study with the SMARTer datasets, whereby 13,233 genes were identified. The small non-coding transcriptome of the porcine sperm has been recently described (Chen et al. 2017). Chen and co-authors found a rich population of miRNAs and lower abundances of other families of small non coding RNAs (e.g., rRNAs, tRNAs, snRNAs) but dit not detect Piwi-interacting RNAs (piRNAs), a type of small non coding RNAs with important - yet weakly explored - functions in sperm biology. In contrast, our pipeline allowed us to identify a similar catalogue of small non coding RNAs but also piRNAs (Table 4 and Figure 2.C).

[Figure 2 near here].

Sperm RNA is likely to be transcriptionally inactive and contain fragmented transcripts that are remnants from spermatogenesis (Ostermeier et al. 2002) and RNA molecules that may function in signaling for embryogenesis after fertilization (Sendler et al. 2013). Hence, the study of the sperm transcriptome and the identification of its alterations could help the scientific community to identify robust, easy and non-invasive markers for sperm defects and male fertility. The purification method and RNA extraction protocol described here, together with control qPCRs to evaluate the sperm RNA purity warrants high quality RNA-seq experiments in boar sperm.

In conclusion, we have concatenated a series of well-established protocols to first, purify spermatozoa from porcine ejaculates by gradient centrifugation, then extract RNA from the purified sperm cells and finally prepare sequencing libraries from these samples to successfully sequence the boar sperm-specific transcriptome by RNA-seq. We have also developed three qPCR assays to assess the purity of the sperm RNA and compared the quality control metrics of different total and small RNA-seq library preparation protocols. In addition, we have evaluated the boar's SRR with the BoviPure^{TM} and found that the boar's SRR is lower than in other mammalian species and not dependent on any of the sperm quality parameters measured in our study. This recovered sperm was thereafter used for RNA extraction. RNA yield per sperm cell was also lower than other species. Moreover, we found no relationship between the quantity of RNA per sperm cell and the sperm quality traits included in the analysis. Despite these caveats of the pig sperm, we obtained sufficient sperm-specific RNA for RNA-seq studies. Thus, we recommend the methodological workflow described here for the high throughput analysis of the boar spermatozoa transcriptome.

Material and Methods

Sperm phenotyping

From March 2015 to January 2017, specialized professionals at the farms obtained fresh ejaculates from 285 Pietrain boars kept in commercial farms. The ages of the animals ranged from 9 months to 5 years old. Sperm was collected with the hand glove method and immediately diluted (1:2) in freshly prepared commercial extender for storage at 16 ºC (MR-A extender; Kubus, S.L.; Majadahonda, Spain). No animal experiment has been performed in the scope of this research.

The samples were maintained at 16 ºC for a maximum time of 2 h for the phenotypic evaluation and a maximum of 4 h for the sperm cell purification. The analysis of sperm motility was performed with the commercial Computer-Aided Sperm Analysis (CASA) system (Integrated Sperm Analysis System V1.0; Proiser, Valencia, Spain) at 5 min and 90 min after incubation of the samples at 37 ºC. The percentages of sperm cell viability, structurally altered acrosomes and morphological abnormalities were measured after staining the samples with the eosin-nigrosin technique after 5 and 90 min incubation at 37 ºC as previously described (Bamba 1988). The osmotic resistance test (ORT) was performed by incubation at 37 °C for 10 min of the sperm samples on iso- and hypo-osmotic solutions, as previously described (Rodríguez-Gil and Rigau 1995). Sperm cell count was performed using a Neubauer cell chamber with not less than 200 cells examined.

Spermatozoa purification

The purification of the spermatozoa cells was performed using 3 mL of BoviPure^{TM} (Nidacon; Mölndal, Sweden), a commercial suspension of colloidal silica particles coated with silane in an isotonic salt solution, diluted to a final ratio of 60% (v/v) with BoviDilute^{TM} (Nidacon; Mölndal, Sweden) in 15 mL RNase-free tubes. The volume of sperm that was layered on top of the cushion varied according to its concentration, with a maximum of 1 billion cells and not exceeding 11 mL. In all cases, the minimum volume ratio of 25 % diluted BoviPure^{TM}/semen recommended by the manufacturer was maintained. The purification was made by centrifugation at 300 x g for 20 min at

20 ºC with slow acceleration and deceleration rates (Allegra X-15R, Beckman Coulter; Brea, USA). After centrifugation, all the upper phases were removed and the cell pellet was transferred to a new RNase-free 15 mL tube, washed with 10 mL of RNase-free PBS and centrifuged at 1,500 x g for 10 min at 20 ºC. The supernatant was then removed and the pellet was gently resuspended in 1 mL of RNase-free PBS. Optical microscopy was used to confirm somatic cell removal of the purified spermatozoa and sperm cell number was assessed in a Neubauer cell chamber. The resuspended pellets were transferred to 1.5 mL RNase-free tubes and centrifuged at 1,500 x g for 10 min at 20 ºC. The resulting pellet was stored at -80 ºC in 1 mL of Trizol® until further use for RNA extraction. SRR was calculated as the number of cells obtained after purification divided by the initial number of cells subjected to purification.

RNA extraction

Total RNA was extracted from 190 purified sperm samples, each from a different boar. The starting number of cells ranged between 48 and 200 million (mean = 143 million) according to availability. First, the cells were pre-lysated using a 5 mL sterile syringe with a 25 G needle for 5 min on ice, followed by 2 min of vigorous vortex. Then, 200 µL of chloroform were added to the samples and these were incubated for 3 min at room temperature first, and centrifuged at 12,000 x g for 15 min afterwards. Supernatants were transferred to new RNase-free tubes and 500 µL of isopropanol were added for RNA precipitation. Samples were then centrifuged at 12,000 x g for 10 min and the supernatants were carefully removed. To wash the pellet, 500 µL of 75% (v/v) ethanol solution were added and the samples were centrifuged at 13,000 x g for 5 min. The pellets were dried out at room temperature for 10 min and resuspended in 30 µL of ultrapure water. All the centrifugations were performed at 4 ºC.

All RNA samples were subjected to DNase treatment with the Turbo DNA-free™ kit (Life Technologies, USA) following the manufacturer's instructions. The RNA samples were then quantified with Qubit™ RNA HS Assay kit (Invitrogen; Carlsbad, USA). To analyze overall RNA fragmentation, RNA integrity number (RIN) was assessed on a 2100 Bioanalyzer using the Agilent RNA 6000 Pico kit (Agilent Technologies; Santa Clara, USA). cDNA was synthesized using 2 µL of RNA (1.7 - 38 ng) and the High Capacity cDNA Reverse Transcription kit in a final volume of 20 µL (Applied Biosystems; Waltham, USA) following the manufacturer's protocol.

qPCR controls

To verify that the purified samples were free of somatic cells and genomic DNA (gDNA), three qPCR assays were developed. One assay targets *Protamine 1 (PRM1*) gene, which transcript is specific to later stages of spermatogenesis and ejaculated mature spermatozoa (Wykes et al. 1997) (PRM1_forward primer: 5'-AGTAGCAAGACCACCGCACT-3'; PRM1_reverse: 5'-AGAGGGTCTTGAAGGCTGGT-3'). The second assay targets the *Protein tyrosine phosphatase receptor type C* (*PTPRC*) gene, which is used as a marker of somatic cell contamination, since it is expressed on most somatic cells and absent in spermatozoa (Das et al. 2013; Shafeeque et al. 2014) (PTPRC_forward: 5'-AGAACAAGGTGGATGTCTATGGCTAT-3'; PTPRC_reverse: 5'-TGTACTGTGCCTCCACCTGAAC-3'). The third assay amplifies an intergenic region (Sscrofa10.2; chr18:25,459,856 - 25,459,926) and was designed to monitor the presence of gDNA contamination (Intergenic_forward: 5'-ACGCAGTCAGAAGCCTGTGA-3'; Intergenic_reverse: 5'-TGGTGTACATGCTCCGAAGGT-3').

To evaluate the performance of our qPCR assays, standard curves with serial dilutions of control cDNA were made. For *PRM1*, *PTPRC* and gDNA qPCRs, pig cDNA from spermatozoa, liver and gDNA were used as input, respectively. Liver cDNA was generated as indicated for sperm but using 1 µg of RNA starting material. The standard curve was generated with five ten-fold serial dilutions of the cDNA templates. The reactions were performed with 10 µL of SYBR® Select Master Mix (Life Technologies, USA), 0.3 µM of each primer, 5 µL of cDNA (for the serial dilutions 1/5 to 1/50,000 and for the query a dilution 1/5) or DNA (from 2 pg/uL to 20 ng/uL) and ultrapure water to a final volume of 20 µL. The thermal profile was: 50 ºC for 2 min, 95 ºC for 10 min and 40 cycles of

95 ºC for 15 sec and 60 ºC for 1 min. Moreover, to assess the specificity of the qPCR reactions, a melting profile (95 ºC for 15 sec, 60 ºC for 15 sec and a gradual increase in temperature with a ramp rate of 1% up to 95 ºC) was programmed following the thermal cycling protocol. A minus reverse transcription control was also included for the two cDNA assays (*PRM1* and *PTPRC*). The reactions for the standard curves were performed in triplicate. For the queried samples (N = 70), the reactions were performed in triplicate. Moreover, a liver cDNA was also included to monitor the expression of *PRM1* and *PTPRC* in a somatic cell type.

Statistical analysis

R v.3.3.0 was utilized for statistical analysis. The Shapiro-Wilk test was used to assess normality of the data. One-way analysis of variance (ANOVA) was used to assess the effects of farm (N = 3), age (N = 3) and batch collection day (N = 59) on SRR, fg per sperm cell and sperm quality parameters. Significantly correlated covariates were adjusted with the R package 'limma' (Ritchie et al. 2015), considering age and farm as fixed effects and batch collection day as batch effect. The R package 'energy' (Rizzo and Szekely 2016) was applied to assess a multivariate nonparametric test of independence covariates between sperm quality phenotypes and SRR and fg of RNA per sperm cell. The nominal significance threshold was set to a *P*-value $\leq$ 0.0041 after Bonferroni correction for multiple testing ($\alpha = 0.05/12 = 0.0041$). To determine whether the RNA-seq quality metrics of the SMARTer Universal Low Input RNA and the TruSeq RNA library prep kits were significantly different, we used the t-test for normally distributed data for reads mapped to the genome, number of uniquely mapped reads and the number of detected genes, and the Wilcoxon test for the non-normally distributed data, i.e. library concentration and proportion of PCR duplicates. The tests were carried with R.

RNA-seq library preparation

Purified RNA from six ejaculates from different boars (Sample_1 to Sample_6) was subjected to ribosomal RNA depletion with the Ribo-Zero Gold rRNA Removal Kit (Illumina). Depleted RNA was then used to prepare long RNA-seq libraries with two different protocols in parallel. On the one hand, SMARTer Universal Low Input RNA library Prep kit (Clontech) was used, following the manufacturer's instructions. On the other hand, TruSeq RNA Library Prep kit (Illumina) was employed, adhering to the manufacturer's protocol with slight modifications adapted to a low amount of starting RNA yield (100 ng). The concentration of the 12 RNA libraries was quantified with the High Sensitivity DNA kit on a 2100 Bioanalizer (Agilent Technologies). The libraries were sequenced in a HiSeq2000 system (Illumina) to generate 75 bp long paired-end reads.

RNA from 3 additional sperm samples (Sample_7 to Sample_9) was used to compare two short RNA library Prep kits: NEBNext Small RNA Library Prep Set (New England Biolabs) and TailorMix miRNA Sample Prep v2 (SeqMatic). One sample was prepared with the NEBNext kit, another with the TailorMix, and a third sample with both kits to allow a more direct comparison of results. Libraries were prepared following the company's instructions. The three samples were quantified with the High Sensitivity DNA kit on a 2100 Bioanalizer and sequenced on a HiSeq2000 to generate 50 bp single-end reads.

Bioinformatics analysis

Read quality of the 12 long RNA-seq datasets was checked with FastQC v.0.11.2 (Andrews 2010). Reads were then filtered using Trimmomatic v.0.33 (Bolger et al. 2014) for read quality and adaptor contamination, with a minimum Phred quality score of 20 and length of over 30 bp. Trimmed reads were mapped to the pig reference genome (Sscrofa 10.2) with STAR v.2.5.3a (Dobin et al. 2013) using the default parameters and including the Ensembl v.83 pig reference annotation (ftp://ftp.ensembl.org/pub/release-83/gtf/sus_scrofa). Transcript abundance was quantified as Fragments Per Kilobase of transcript per Million mapped reads (FPKM) with RSem v.1.3.0 (Li and Dewey 2011) with default parameters. FPKM is a normalized measure of gene expression based on the number of reads mapping to a given gene corrected by the length of that gene and the sample sequencing depth. To compare the performance of the SMARTer and the TruSeq protocols we evaluated the proportion of PCR duplicates as it is a measure of the complexity of each library. To allow a fair comparison of the two protocols we analyzed the same number of reads in all the samples. More in detail, we randomly sub-selected 2,336,549 reads per sample since this number corresponds to the lowest sequencing depth obtained. The read selection was carried with seqtk v.1.2 (Shen et al. 2016). The proportion of PCR duplicates was calculated with Picard Tools v.1.110 (http://picard.sourceforge.net) MarkDuplicates. Graphs were performed with R: venn diagram with the R package 'VennDiagram' (Chen and Boutros 2011) and cluster dendrogram with the R package 'cluster' (Maechler et al. 2017).

We also evaluated the absence of RNA from somatic cell origin in our samples. For this we used the SMARTer libraries, which showed better outcomes, and the totality of the reads obtained for each of these libraries (between 18.5 and 26.9 million reads per sample). We also included two publicly available (http://www.ncbi.nlm.nih.gov/sra) boar RNA-seq datasets, one from whole blood cells (ERR1898477), which contains a large abundance of leukocytes and a second one from ear biopsy (SRR3437133), which contains a high proportion of keratinocytes, a specialized type of epithelial cells. We screened the presence of the two sperm specific genes, *PRM1* and *OAZ3* (*Ornithine Decarboxylase Antizyme 3*) (Jodar *et al.*, 2016), and two genes of somatic cell origin *PTPRC* (expressed in most somatic cells) and *KRT1* (*Keratin 1*), which is specific from keratinocytes. For data visualization, SMARTer mapped bam files were indexed with SAMtools v.1.3.0 (Li et al. 2009) and uploaded into the IGV viewer (Thorvaldsdóttir et al. 2013). We used a manual script to extract RNA levels of tissue-specific genes as described in (Jodar et al. 2016) as an ultimately control for RNA purity.

The 3 small RNA-seq datasets were analyzed for read quality with FastQC v.0.11.2 (Andrews 2010) and reads were sub-sampled to 887,406 reads per library with seqtk v.1.2 (Shen et al. 2016). Library adaptors and indexes were trimmed using Cutadapt v.1.0 (Martin 2011) and filtered for read quality, with a minimum quality score of 20, and minimum length of 10 bp with Trimmomatic v.0.33 (Bolger et al. 2014). Trimmed reads were mapped to the pig reference genome (Sscrofa 10.2) with Bowtie 1 v.1.2.0 (Langmead 2010) with default parameters but allowing 0 mismatches (-n) in ‘seed’ region of 10 bp (-l). The proportion of PCR duplicates was calculated with SAMtools v.1.3.0 (Li et al. 2009) rmdup for single-end reads. RNA levels of small non-coding RNAs were calculated with Bedtools v.2.17.0 (Quinlan and Hall 2010) intersect against the boar Ensembl v.83 ‘gtf’ annotation, miRBase database (Griffiths-Jones et al. 2006), piRNA database (Rosenkranz 2016), and tRNA v.2.0 database (Chan and Lowe 2016).

Acknowledgements

Funding: This work was supported by the Spanish Ministerio de Economía y Competitividad (MINECO) under grant AGL2013-44978-R, grant 2014 SGR 1528 from the Agency for Management of University and Research Grants of the Generalitat de Catalunya. We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016–2019 (SEV-2015-0533) grant awarded to the Centre for Research in Agricultural Genomics (CRAG). We are also thankful to the CERCA Programme of the Generalitat de Catalunya. Fabiana Quoos Mayer was recipient of a post-doctoral scholarship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Bolsista CAPES Proc. nº BEX 6707/14-9), Brazil. Marta Gòdia acknowledges a PhD studentship from MINECO (BES-2014-070560). Alex Clop acknowledges a MINECO's Ramon y Cajal research fellow (RYC-2011-07763). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors are indebted to Semen Cardona S.L., Genus, PIC, PIC Espana and Grup Gepork for providing the semen samples. We also thank Marti Bernardo and Isabel Serra for their suggestions on the statistical analyses.

Disclosure of interest:

The authors report no conflict of interest

Notes on contributors

ACl, AS, FQM and MG conceived and designed the experiment. FQM and ACa designed the primers and carried out the qPCR analyses. FQM, MG and JN performed sperm purifications and RNA extractions. MG performed statistic and bioinformatics analysis. FQM and MG analyzed the data. JERG carried the phenotypic analysis. FQM, MG and AC drafted the manuscript. All authors read and approved the final manuscript.

References

Agarwal, A., Gupta, S., Du Plessis, S., Sharma, R., Esteves, S. C., Cirenza, C., Eliwa, J., Al-Najjar, W., Kumaresan, D., Haroun, N., Philby, S., and Sabanegh, E. (2016) Abstinence Time and Its Impact on Basic and Advanced Semen Parameters. Urology. **94**:102–110.

Allamaneni, S. S. R., Agarwal, A., Rama, S., Ranganathan, P., and Sharma, R. K. (2005) Comparative study on density gradients and swim-up preparation techniques utilizing neat and cryopreserved spermatozoa. Asian J Androl. **7**(86):86–92.

Andrews, S. (2010) FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Aoki, V. W., Moskovtsev, S. I., Willis, J., Liu, L., Mullen, J. B. M., and Carrell, D. T. (2005) DNA integrity is compromised in protamine-deficient human sperm. J Androl. **26**(6):741–748.

Bamba, K. (1988) Evaluation of acrosomal integrity of boar spermatozoa by bright field microscopy using an eosin-nigrosin stain. Theriogenology. **29**(6):1245–1251.

Belleannée, C., Calvo, É., Caballero, J., and Sullivan, R. (2013) Epididymosomes Convey Different Repertoires of MicroRNAs Throughout the Bovine Epididymis. Biol Reprod. **89**(2):30.

Bissonnette, N., Lévesque-Sergerie, J. P., Thibault, C., and Boissonneault, G. (2009) Spermatozoal transcriptome profiling for bull sperm motility: A potential tool to evaluate semen quality. Reproduction. **138**(1):65–80.

Bohacek, J., and Mansuy, I. M. (2015) Molecular insights into transgenerational non-genetic inheritance of acquired behaviours. Nat Rev Genet **16**(11):641–652.

Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic : a flexible trimmer for Illumina sequence data. Bioinformatics. **30**(15):2114–2120.

Byrne, C. J., Fair, S., English, A. M., Holden, S. A., Dick, J. R., Lonergan, P., and Kenny, D. A. (2017) Dietary polyunsaturated fatty acid supplementation of young post-pubertal dairy bulls alters the fatty acid composition of seminal plasma and spermatozoa but has no effect on semen volume or sperm quality. Theriogenology. **90**:289–300.

Capra, E., Turri, F., Lazzari, B., Cremonesi, P., Gliozzi, T. M., Fojadelli, I., Stella, A., and Pizzi, F. (2017) Small RNA sequencing of cryopreserved semen from single bull revealed altered miRNAs and piRNAs expression between High- and Low-motile sperm populations. BMC Genomics. **18**(1):14.

Card, C. J., Anderson, E. J., Zamberlan, S., Krieger, K. E., Kaproth, M., and Sartini, B. L. (2013) Cryopreserved bovine spermatozoal transcript profile as revealed by high-throughput ribonucleic acid sequencing. Biol Reprod. **88**(2):49.

Carrell, D. T., Emery, B. R., and Hammoud, S. (2007) Altered protamine expression and diminished spermatogenesis: What is the link? Hum Reprod Update. **13**(3):313–327.

Chan, P. P., and Lowe, T. M. (2016) GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. **44**(D1):D184–D189.

Chen, C., Wu, H., Shen, D., Wang, S., Zhang, L., Wang, X., Gao, B., Wu, T., Li, B., Li, K., and Song, C. (2017) Comparative profiling of small RNAs of pig seminal plasma and ejaculated and epididymal sperm. Reproduction. **153**(6):785–796.

Chen, H., and Boutros, P. C. (2011) VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. **12**:35.

Curry, E., Safranski, T. J., and Pratt, S. L. (2011) Differential expression of porcine sperm microRNAs and their association with sperm morphology and motility. Theriogenology. **76**(8):1532–1539.

Das, P. J., McCarthy, F., Vishnoi, M., Paria, N., Gresham, C., Li, G., Kachroo, P., Sudderth, A. K., Teague, S., Love, C. C., Varner, D. D., Chowdhary, B. P., and Raudsepp, T. (2013) Stallion Sperm Transcriptome Comprises Functionally Coherent Coding and Regulatory RNAs as Revealed by Microarray Analysis and RNA-seq. PLoS ONE. **8**(2):e56535.

Das, P. J., Paria, N., Gustafson-Seabury, A., Vishnoi, M., Chaki, S. P., Love, C. C., Varner, D. D., Chowdhary, B. P., and Raudsepp, T. (2010) Total RNA isolation from stallion sperm and testis biopsies. Theriogenology. **74**(6):1099–1106.e2.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: Ultrafast universal RNA-seq aligner. Bioinformatics **29**(1):15–21.

Fang, P., Zeng, P., Wang, Z., Liu, M., Xu, W., Dai, J., Zhao, X., Zhang, D., Liang, D., Chen, X., Shi, S., Zhang, M., Wang, L., Qiao, Z., and Shi, H. (2014) Estimated Diversity of Messenger RNAs in Each Murine Spermatozoa and Their Potential Function During Early Zygotic Development. Biol Reprod. **90**(5):1–11.

Feugang, J. M. (2017) Novel agents for sperm purification, sorting, and imaging. Mol Reprod Dev. **84**(9):832–841.

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.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. **34**(Database issue):D140-4.

Hamatani, T. (2012) Human spermatozoal RNAs. Fertil Steril. **97**(2):275–281.

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**(9):2558–2569.

Jameel, T. (2008) Sperm swim-up: A simple and effective technique of semen processing for intrauterine insemination. J Pak Med Assoc. **58**(2):71–74.

Jodar, M., Kalko, S., Castillo, J., Ballescà, J. L., and Oliva, R. (2012) Differential RNAs in the sperm cells of asthenozoospermic patients. Hum Reprod. **27**(5):1431–1438.

Jodar, M., Selvaraju, S., Sendler, E., Diamond, M. P., and Krawetz, S. A. (2013) The presence, role and clinical use of spermatozoal RNAs. Hum Reprod Update. **19**(6):604–624.

Jodar, M., Sendler, E., Moskovtsev, S. I., Librach, C. L., Goodrich, R., Swanson, S., Hauser, R., Diamond, M. P., and Krawetz, S. A. (2015) Absence of sperm RNA elements correlates with idiopathic male infertility. Sci Transl Med. **7**(295):295re6.

Jodar, M., Sendler, E., Moskovtsev, S. I., Librach, C. L., Goodrich, R., Swanson, S., Hauser, R., Diamond, M. P., and Krawetz, S. A. (2016) Response to Comment on 'Absence of sperm RNA elements correlates with idiopathic male infertility'. Sci Transl Med. **8**(353):353tr1.

Langmead, B. (2010) Aligning short sequencing reads with Bowtie. Current Protocols in Bioinformatics (SUPP.32):Chapter Unit-11.7.

Legrand, J. M. D., and Hobbs, R. M. (2017) RNA processing in the male germline: Mechanisms and implications for fertility. Semin Cell Dev Biol. https://doi.org/10.1016/j.semcdb.2017.10.006.

Li, B., and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. **12**(1):323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics. **25**(16):2078–2079.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2017) Cluster Analysis Basics and Extensions. R package version 2.0.6. http://cran.r-project.org/web/packages/cluster/index.html.

Mao, S., Goodrich, R. J., Hauser, R., Schrader, S. M., Chen, Z., and Krawetz, S. A. (2013) Evaluation of the effectiveness of semen storage and sperm purification methods for spermatozoa transcript profiling. Syst Biol Reprod Med. **59**(5):287–295.

Mao, S., Sendler, E., Goodrich, R. J., Hauser, R., and Krawetz, S. A. (2014) A comparison of sperm RNA-seq methods. Syst Biol Reprod Med. **60**(5):308–315.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. **17**(1):10–12.

Martins, R. P., and Krawetz, S. A. (2005) Towards understanding the epigenetics of transcription by chromatin structure and the nuclear matrix. Gene Ther Mol Biol. **9**(B):229–246.

Ostermeier, G. C., Dix, D. J., Miller, D., Khatri, P., and Krawetz, S. A. (2002) Spermatozoal RNA profiles of normal fertile men. Lancet. **360**(9335):772–777.

Patil, P. S., Humbarwadi, R. S., Patil, A. D., and Gune, A. R. (2013) Immature germ cells in semen - correlation with total sperm count and sperm motility. J Cytol. **30**(3):185–189.

Pessot, C. A., Brito, M., Figueroa, J., Concha, I. I., Yañez, A., and Burzio, L. O. (1989) Presence of RNA in the sperm nucleus. Biochem Biophys Res Commun. **158**(1):272–278.

Quinlan, A. R., and Hall, I. M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. **26**(6):841–842.

Rando, O. J. (2016) Intergenerational transfer of epigenetic information in sperm. Cold Spring Harb Perspect Med. **6**(5):a022988.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43**(7):e47.

Rizzo ML and Szekely GJ. (2016) energy: E-Statistics: Multivariate Inference via the Energy of Data. version 1.7-2 http://cran.r-project.org/package=energy.

Rodríguez-Gil, J. E., and Bonet, S. (2015) Current knowledge on boar sperm metabolism: Comparison with other mammalian species. Theriogenology. **85**(1):4–11.

Rodríguez-Gil, J. E., and Rigau, T. (1995) Effects of slight agitation on the quality of refrigerated boar sperm. Anim Reprod Sci. **39**(2):141–146.

Rosenkranz, D. (2016) piRNA cluster database: A web resource for piRNA producing loci. Nucleic Acids Res. **44**(D1):D223–D230.

Samardzija, M., Karadjole, M., Matkovic, M., Cergolj, M., Getz, I., Dobranic, T., Tomaskovic, A., Petric, J., Surina, J., Grizelj, J., and Karadjole, T. (2006) A comparison of BoviPure and Percoll on bull sperm separation protocols for IVF. Anim Reprod Sci. **91**(3–4):237–247.

Selvaraju, S., Parthipan, S., Somashekar, L., Kolte, A. P., Krishnan Binsila, B., Arangasamy, A., and Ravindra, J. P. (2017) Occurrence and functional significance of the transcriptome in bovine (Bos taurus) spermatozoa. Sci Rep. **7**:42392.

Sendler, E., Johnson, G. D., Mao, S., Goodrich, R. J., Diamond, M. P., Hauser, R., and Krawetz, S. A. (2013) Stability, delivery and functions of human sperm RNAs at fertilization. Nucleic Acids Res. **41**(7):4104–4117.

Shafeeque, C. M., Singh, R. P., Sharma, S. K., Mohan, J., Sastry, K. V. H., Kolluri, G., Saxena, V. K., Tyagi, J. S., Kataria, J. M., and Azeez, P. A. (2014) Development of a new method for sperm RNA purification in the chicken. Anim Reprod Sci. **149**(3–4):259–265.

Shen, W., Le, S., Li, Y., and Hu, F. (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLOS ONE. **11**(10):e0163962.

Siffroi, J. P., Alfonsi, M. F., and Dadoune, J. P. (1998) Electron microscopic in situ hybridization study of simultaneous expression of TNP1 and PRM1 genes in human spermatids. Ital J Anat Embryol. **103**(4 Suppl 1):65–74.

Steger, K., Pauls, K., Klonisch, T., Franke, F. E., and Bergmann, M. (2000) Expression of protamine-1 and -2 mRNA during human spermiogenesis. Mol Hum Reprod **6**(3):219–225.

Székely, G. J., and Rizzo, M. L. (2009) Brownian distance covariance. Ann Appl Stat. **3**(4):1236–1265.

Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013) Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. **14**(2):178–192.

Wang Z, Gerstein M, Snyder M. (2013) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. **10**(1):57-63.

Wykes, S. M., Visscher, D. W., and Krawetz, S. A. (1997) Haploid transcripts persist in mature human spermatozoa. Mol Hum Reprod. **3**(1):15-19.

Yang, C. C., Lin, Y. S., Hsu, C. C., Wu, S. C., Lin, E. C., and Cheng, W. T. K. (2009) Identification and sequencing of remnant messenger RNAs found in domestic swine (Sus scrofa) fresh ejaculated spermatozoa. Anim Reprod Sci. **113**(1–4):143–155.

Zhao, Y., Li, Q., Yao, C., Wang, Z., Zhou, Y., Wang, Y., Liu, L., Wang, Y., Wang, L., and Qiao, Z. (2006) Characterization and quantification of mRNA transcripts in ejaculated spermatozoa of fertile men by serial analysis of gene expression. Hum Reprod. **21**(6):1583–1590.

Figure legends

Figure 1. Read mapping depth of sperm and somatic-specific genes in the porcine sperm, whole blood and ear RNA-seq datasets. A) Corresponds to the sperm-specific gene *PRM1* (Ensembl gene ID: ENSSSCG00000021337). B) Plot for the sperm-specific gene *OAZ3* (ENSSSCG00000027091). C) Read depth along the somatic cell specific gene *PTPRC* (ENSSSCG00000010908). D) Depth for the Keratinocyte specific gene *KRT1* (ENSSSCG00000000251). The number of reads produced for the sperm datasets are between 18.5 and 26.5 million. The white blood cells and the ear RNA-seq libraries include 18.3 and 21.7 million reads, respectively. The scale provided on the upper left axis of each graph indicates the raw number of reads mapped to the gene.

Figure 2. Comparison of the RNA-seq results from both the total and the small library preparation kits. A) Venn diagram representing the 13,233 different transcripts detected in the SMARTer datasets and 6,642 for the TruSeq. The majority of the TruSeq transcripts detected (6,452) were detected in both kits. B) Cluster dendrogram of the RNA transcript levels from both kits. S: Sample. C) Size distribution of mapped sequencing reads from small RNAs in both kits.