Universitat Autònoma de Barcelona

GENOMIC APPROACHES FOR THE ASSESSMENT OF BOAR SPERM QUALITY

Marta Gòdia Perelló

UNIVERSITAT AUTÒNOMA DE BARCELONA

Departament de Ciència Animal i dels Aliments, Facultat de Veterinària

CENTRE DE RECERCA EN AGRIGENÒMICA

Grup de Genòmica Animal

GENOMIC APPROACHES FOR THE ASSESSMENT OF BOAR SPERM QUALITY

Marta Gòdia Perelló

Doctoral thesis to obtain the PhD degree in Animal Production of the Universitat Autònoma de Barcelona, September 2019

Supervisors

Dr. Àlex Clop Ponte

Dr. Armand Sánchez Bonastre

El Dr. Àlex Clop Ponte, investigador distingit del CSIC (Consejo Superior de Investigaciones Científicas) i el Dr. Armand Sánchez Bonastre, professor titular del Departament de Ciència Animal i dels Aliments de la Universitat Autònoma de Barcelona,

fan constar

que el treball de recerca i la redacció de la memòria de la tesi doctoral titulada "*Genomic approaches for the assessment of boar sperm quality*" han estat realitzats sota la seva direcció per
MARTA GÒDIA PERELLÓ

I **certifiquen**

que aquest treball s'ha dut a terme al Departament de Ciència Animal i del Aliments de la Facultat de Veterinària de la Universitat Autònoma de Barcelona i al Grup de Genòmica Animal del Centre de Recerca en Agrigenòmica,

considerant

que la memòria resultant es apta per optar al grau de Doctor en Producció Animal per la Universitat Autònoma de Barcelona.

I perquè quedi constància, signen aquest document a Bellaterra,
a Setembre del 2019.

Dr. Àlex Clop Ponte Dr. Armand Sánchez Bonastre Marta Gòdia Perelló

This work was funded by the Spanish Ministry of Economy and Competitiveness (MINECO) under grants AGL2013-44978-R and AGL2017-86946-R.

Marta Gòdia Perelló acknowledges a PhD FPI fellowship provided by MINECO (2015-2019) under grant BES-2014-070560. The Short-Stays were financed by MINECO grants. The Short-Stays were carried at University of Michigan (Michigan, USA) (EEBB-I-2016-11528), at Wayne State University (Michigan, USA) (EEBB-I-17-12229) and at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) (Brisbane, Australia) (EEBB-I-18-12860).

Per als meus pares

i al meu germà,

per treure sempre el millor de mi

CONTENT

SUMMARY

Improving the efficiency of food production, reducing its environmental impact and strengthening health and welfare is crucial for the sustainability of the animal breeding sector, including the pig industry. The sector is demanding for higher selection pressure, which will result in a lower number of elite boars in the breeding nucleus. This will bring alongside the need to increase the number of artificial insemination sperm doses from these elite boars. Thus, subfertility issues not detected now in these ejaculates may arise. Consequently, semen quality has gained interest for the scientific and animal breeding sector. Sperm quality is an intricate set of phenotypes confounded by several factors and molecular processes that remain to be resolved. The aim of this thesis was to investigate the genetic basis of porcine sperm quality using high-throughput sequencing and genotyping technologies.

First, we evaluated and assessed the protocol to purify and extract RNA from porcine sperm cells and found that the sperm recovery rate and RNA yield was low when compared to other mammalian species.

By using RNA-seq technology, we found a complex suite of RNAs in porcine sperm including messenger, long noncoding, micro-, piwi-interacting- and transfer RNAs. We detected 4,436 protein coding genes with moderate to high abundances. Most of the transcripts were highly fragmented but this was less obvious in genes related to spermatogenesis, fertility and chromatin compaction. Moreover, we identified 37 mRNA transcripts and 7 microRNAs with a seasonal payload, which could relate to the decrease of semen quality typically occurring in the warm summer months. In keeping with results obtained in other studies, these genes were mostly related to oxidative stress and DNA damage. We also sought to characterize for the first time in mammals by RNA-seq, the circular RNAome of mature sperm. We identified nearly 1,600 circular RNAs, 56 of which were potential sponges for 31 microRNAs. The

abundance of a small fraction of circular RNAs (11%) correlated with sperm motility parameters and few were validated by qPCR and Sanger sequencing. The association between sperm motility and 2 circular RNAs were validated by RT-qPCR. These circular RNAs hold potential as biomarkers for boar sperm motility traits.

We carried a Genome-Wide Association Study (GWAS) in 288 boars for 25 sperm related traits and identified 71 significant SNPs distributed in 12 QTL regions for percentage of head and neck abnormalities, abnormal acrosomes and motility. We also studied the sperm transcriptome of phenotypically divergent ejaculates and identified circa 6,100 significant correlations involving 3,007 genes and the 25 sperm-related traits. A total of 831 genes showed correlations with 3 or more traits. To study the complexity and interconnectivity of the molecular processes controlling sperm quality we used a holistic approach. We applied a systems biology approach to integrate the GWAS and RNA-seq datasets into a network and identified key SNPs and genes with potential effects across several sperm traits. We ultimately suggest a panel of 74 SNPs that together explain between 5 to 36% of the phenotypic variance from the sperm quality traits recorded in our study.

Finally, we identified by whole-genome sequencing of sperm and matched blood samples from 3 boars, genetic variants in Allelic Ratio Distortion in ejaculated mature sperm. These variants may be highlighting additional genes affecting semen quality and even boar fertility, not identified by GWAS or RNA-seq.

RESUMEN

Mejorar la eficiencia de la producción de alimentos, reducir su impacto ambiental y fortalecer la salud y el bienestar es crucial para la sostenibilidad del sector de mejora animal, incluyendo la industria porcina. El sector exige una mayor presión de selección, lo que dará como resultado un menor número de cerdos de élite en el núcleo reproductor. Esto traerá consigo la necesidad de aumentar el número de dosis de esperma para la inseminación artificial de estos cerdos de élite. Por lo tanto, pueden surgir problemas de subfertilidad que no se detectan en este momento en estos eyaculados. En consecuencia, la calidad del semen ha ganado interés para el sector científico y de mejora animal. La calidad del semen es un conjunto complejo de fenotipos formado por varios factores y procesos moleculares que aún no se han resuelto. El objetivo de esta tesis fue investigar la base genética de la calidad del esperma porcino utilizando tecnologías de secuenciación y genotipado de alto rendimiento.

Primero, evaluamos y analizamos el protocolo para purificar y extraer ARN de las células espermáticas de porcino y encontramos que la tasa de recuperación de espermatozoides y el rendimiento de extracción de ARN son bajos en comparación con otras especies de mamíferos.

Mediante el uso de la tecnología de secuenciación de ARN, encontramos un conjunto complejo de ARN en los espermatozoides porcinos, incluyendo ARN mensajeros, ARN no codificantes largos, micro-, piwi- y ARN de transferencia. Detectamos 4,436 genes codificantes de proteínas con abundancias moderadas a altas. La mayoría de los transcritos estaban muy fragmentados, pero esto era menos obvio en los genes relacionados con espermatogénesis, fertilidad y compactación de cromatina. Además, identificamos 37 transcritos de ARN mensajero y 7 micro-ARN asociados a la estacionalidad, lo que podría relacionarse con la disminución de la calidad del semen que ocurre típicamente en los cálidos meses de verano. De acuerdo con los resultados obtenidos en otros estudios, estos genes estaban relacionados principalmente con el estrés oxidativo y el daño en el ADN. También buscamos caracterizar por primera vez en mamíferos mediante secuenciación de ARN, el ARNoma circular de esperma

maduro. Identificamos casi 1,600 ARN circulares, 56 de los cuales estaban actuando como posibles esponjas para 31 micro-ARN. La abundancia de una pequeña fracción de ARN circulares (11%) se correlacionó con los parámetros de motilidad de los espermatozoides y unos pocos fueron validados por amplificación PCR y secuenciación Sanger. Dos ARN circulares fueron validados por RT-PCR cuantitativa como potenciales biomarcadores para caracteres de motilidad en esperma de cerdo.

Llevamos a cabo un estudio de asociación de genoma completo (GWAS) en 288 porcinos para 25 caracteres de calidad seminal e identificamos 71 SNPs significativos distribuidos en 12 regiones QTL asociadas al porcentaje de anormalidades de cabeza y cuello, acrosomas anormales y motilidad. También estudiamos el transcriptoma de esperma de eyaculados fenotípicamente divergentes e identificamos alrededor de 6,100 correlaciones significativas que involucraban 3,007 genes y los 25 caracteres seminales. Un total de 831 genes mostraron correlaciones con 3 o más caracteres. Para estudiar la complejidad y la interconectividad de los procesos moleculares que controlan la calidad del esperma, utilizamos un enfoque holístico. Aplicamos un análisis de biología de sistemas para integrar los de datos GWAS y secuenciación de ARN para construir una red de interacciones e identificamos SNP y genes clave con efectos potenciales en varios caracteres de calidad seminal. En última instancia, sugerimos un panel de 74 SNP que juntos explican entre el 5 y el 36% de la variación fenotípica de los caracteres de calidad espermática registrados en nuestro estudio.

Finalmente, identificamos mediante secuenciación de genoma completo de muestras de esperma y sangre de 3 porcinos, variantes genéticas con distorsión del ratio alélico en el esperma maduro eyaculado. Estas variantes pueden estar destacando genes adicionales que afectan la calidad del semen e incluso la fertilidad del porcino no identificados mediante GWAS y secuenciación de ARN.

RESUM

Millorar l'eficiència de la producció d'aliments, reduir l'impacte ambiental i reforçar la salut i el benestar són crucials per a la sostenibilitat del sector de producció animal, inclosa la indústria porcina. El sector exigeix una pressió de selecció més elevada, cosa que es traduirà en un menor nombre de porcins d'elit al nucli reproductor. D'aquesta manera s'incrementarà la necessitat d'augmentar el nombre de dosis espermàtiques per a l'inseminació artificial d'aquests porcins d'elit. Això, podria fer aparèixer problemes de subfertilitat no detectats fins ara en aquests ejaculats. En conseqüència, la qualitat seminal ha guanyat interès pel sector científic i de producció animal. La qualitat seminal és un conjunt complex de fenotips format per diversos factors i processos moleculars que encara queden per resoldre. L'objectiu d'aquesta tesi era investigar les bases genètiques de la qualitat seminal en porcí mitjançant tecnologies de seqüenciació i genotipat d'alt rendiment.

Primer es va analitzar i avaluar el protocol per purificar i extreure ARN d'espermatozoides porcins i vam trobar que la taxa de recuperació d'espermatozoides i el rendiment d'extracció d'ARN era baixa en comparació amb altres espècies de mamífers.

Mitjançant l'ús de la tecnologia de seqüenciació d'ARN, es va trobar un conjunt complex d'ARNs en els espermatozoides porcins incloent-hi l'ARN missatger, ARN llarg no codificant, micro-, piwi- i ARN de transferència. Es van detectar 4,436 gens codificants amb abundàncies de moderades a altes. La majoria dels transcripts estaven molt fragmentats, però menys en gens relacionats amb espermatogènesi, fertilitat i compactació de cromatina. A més, es van identificar 37 transcripts i 7 microARN amb una abundància diferencialment estacional, que podrien relacionar-se amb la disminució de la qualitat seminal que es produeix en els mesos calorosos d'estiu. Tal i com passa en altres estudis, aquests gens es relacionaven principalment amb estrès oxidatiu i danys en l'ADN. També es caracteritzar per primera vegada en mamífers mitjançant

seqüenciació d'ARN, el ARNoma circular de l'espermatozou madur. Es van identificar prop de 1,600 ARN circulars, 56 dels quals actuaven potencialment com a esponges per 31 miRNAs. L'abundància d'una petita fracció d'ARN circulars (11%) correlacionava amb paràmetres de motilitat seminal i alguns van ser validats per amplificació PCR i seqüenciació Sanger. L'associació entre la motilitat dels espermatozoides i 2 ARN circulars es van validar mitjançant quantificació RT-PCR. Aquests 2 ARN circulars presenten potencial com a biomarcadors per a la motilitat seminal del porcí.

Vam realitzar un estudi d'associació del genoma complet (GWAS) en 288 porcins per a 25 caràcters de qualitat seminal i vam identificar 71 SNPs significatius distribuïts en 12 regions QTL associades al percentatge d'anomalies al cap i al coll, anormalitats en acrosomes i motilitat. També es va estudiar el transcriptoma d'esperma d'ejaculats fenotípicament divergents i es van identificar al voltant de 6,100 correlacions significatives que involucren 3,007 gens i els 25 caràcters seminals. Un total de 831 gens van mostrar correlacions amb 3 o més trets. Per estudiar la complexitat i la interconectivitat dels processos moleculars que controlen la qualitat de l'espermatozoide es va utilitzar un enfocament holístic. Es va aplicar un anàlisis de biologia de sistemes per integrar les dades de GWAS i seqüenciació d'ARN i construir una xarxa d'interaccions i identificar SNPs i gens clau amb efectes potencials a varis caràcters de qualitat seminal. En última instància, suggerim un panell de 74 SNPs que expliquen entre un 5 i un 36% de la variació fenotípica dels trets de qualitat seminal registrats al nostre estudi.

Finalment, mitjançant seqüenciació del genoma sencer en mostres d'esperma i sang de 3 porcins, es va identificar variants genètiques amb proporció de distorsió al·lèlica en els espermatozoides madurs ejaculats. Aquestes variants poden estar destacant gens addicionals que afecten la qualitat seminal i fins i tot la fertilitat del porcí, no identificats pel GWAS o seqüenciació d'ARN.

LIST OF TABLES

LIST OF FIGURES

GENERAL INTRODUCTION

REVIEW

PAPER I

PAPER II

ANNEXES: Paper III

ANNEXES: Paper IV

ANNEXES: Paper V

LIST OF PUBLICATIONS

The present thesis is based on the work contained in the list of articles below:

- Review: **Gòdia M**, Swanson G, Krawtez SA. A History of Why Fathers' RNA Matters. *Biology of Reproduction.* 2018, 99(1), 147–159. doi:10.1093/biolre/ioy007.

- Paper I: **Gòdia M**, Quoos Mayer F, Nafissi J, Castelló A, Rodríguez-Gil JE, Sánchez A, Clop A. A technical assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-seq analysis. *Systems Biology in Reproductive Medicine*. 2018, 64(4):291-303. doi: 10.1080/19396368.2018.1464610

- Paper II: **Gòdia M,** Estill M, Castelló A, Balasch S, Rodríguez-Gil JE, Krawetz SA, Sánchez A, Clop A. A RNA-Seq Analysis to Describe the Boar Sperm Transcriptome and Its Seasonal Changes. *Frontiers in Genetics*. 2019, 10:299. doi: 10.3389/fgene.2019.00299

- Paper III: **Gòdia M**, Castelló A, Rocco M, Cabrera B, Rodríguez-Gil JE, Sánchez A, Clop A. Identification of circular RNAs in porcine sperm and their relation to sperm motility. *Submitted.* bioRxiv 2019, doi: https://doi.org/10.1101/608026

- Paper IV: **Gòdia M**, Reverter A, González-Prendes R, Ramayo-Caldas Y, Castelló A, Rodríguez-Gil JE, Sánchez A and Alex Clop. An integrative systems biology approach to identify the molecular basis of sperm quality in swine. *In preparation*

- Paper V: **Gòdia M**, Casellas Q, Rodríguez-Gil JE, Castelló A, , Sánchez A and Clop A. Whole genome sequencing of porcine sperm identifies Allelic Ratio Distortion in genes related to spermatogenesis. *In preparation*

OTHER PUBLICATIONS BY THE AUTHOR

(Not included in the thesis)

- Mäkeläinen S, **Gòdia M**, Hellsand M, Viluma A, Hahn D, Makdoumi K, Zeiss CJ, Mellersh C, Ricketts SL, Narfström K, Hallböök F, Ekesten B, Andersson G, Bergström T. An ABCA4 loss-of-function mutation causes a canine form of Stargardt disease. *PLoS Genetics*. 2019 15(3): e1007873. https://doi.org/10.1371/journal.pgen.1007873
- **Gòdia M**, López S, Rodríguez-Gil JE, Sánchez A and Clop A. The microbiome of the porcine sperm is related to semen quality traits. *In preparation*

ABBREVIATIONS

AI: Artificial insemination

ALH: Amplitude of lateral head displacement

ATAC-seq: Assay for transposase-accessible chromatin with sequencing

ATP: Adenosine triphosphate

ARD: Allelic ratio distortion

AWM: Association weight matrix

BAC: Bacterial artificial chromosome

Ca^{2+}: Calcium

CASA: Computer assisted semen analysis

CatSper: Cation channels of sperm

CTCF: CCCTC-binding factor

ChIP-seq: Chromatin immunoprecipitation with sequencing

circRNA: circular RNA

Cq: Quantification cycles

CV: Coefficient of variation

DMC: Differentially methylated cytosines

DP: Read depth

eGWAS: Expression GWAS

eSNP: Expression SNP

FAANG: Functional annotation of animal genomes

GWAS: Genome-wide association study

indel: Insertion or deletion

LD: Linkage disequilibrium

miRNA: micro RNA

MN: Mononucleosomes

MNAse-seq: Micrococcal nuclease with sequencing

NGS: Next-generation sequencing

ORT: Osmotic resistance test

ODF: Outer dense fibers

OXPHOS: Oxidative phosphorylation

PCA: Principal component analysis

PCIT: Partial correlation coefficient information theory

piRNA: piwi-interacting RNA

PTM: Post-translational modifications

QTL: Quantitative trait loci

RFLP: Restriction fragment length polymorphism

RNAse R: Ribonuclease R

RT-qPCR: Quantitative real time PCR

SERCA: sarco/ER Ca^{2+} ATPase

SGSC: Swine genome sequencing consortium

siRNA: small interfering RNA

SMRT: Single-molecule real time

SN: Subnucleosomes

SNP: Single nucleotide polymorphism

SOLiD: Sequencing by oligo detection

SRE: Sperm RNA element

SRR: Sperm recovery rate

TF: Transcription factor

TRD: Transmission ratio distortion

tRNA: transfer RNA

VAP: Average path velocity

VCL: Curvilinear velocity

VSL: Straight-line velocity

WES: Whole exome sequening

WGCNA: Weighted correlation network analysis

WGS: Whole genome sequencing

General Introduction

Chapter 1

1.1 Generalities on swine

1.1.1 The domestic pig

The domestic pig (*Sus scrofa*) is an *Eutherian* omnivorous mammal from the genus *Sus* within the *Suidae* family. Molecular genetics and zooarcheological evidence suggests that *S.scrofa* emerged in South East Asia during climatic fluctuations 5.3-3.5 M years ago and became independently domesticated ~10,000 years ago in western Eurasia and East Asia (Groenen et al., 2012; Larson et al., 2005). After domestication and trough the years, Europe and China became the 2 major pig-domestication centers of the Old World, developing a wide variety of local types adapted to the environmental conditions and selected mostly for behavioral and morphological traits (reviewed in: Amills et al., 2010). From these 2 primary sites of domestication, pigs spread across Europe, North Africa and Asia. Within each continent, different degrees of relatedness between the breeds and their geographically respective wild boars have been observed, thereby suggesting a long history of genetic exchange (Groenen et al., 2012).

1.1.2 Pig production

Pork is the most consumed meat and the main source of protein in many countries. The latest FAO report stated that pork meat accounted for 35.9% of the 334 M tons of total meat produced worldwide (FAO, 2017). Improvements in pig breeding strategies have resulted in a 2.3-fold increase in pig stocks since the 1960s to reach nearly 1,000 million stocks in 2017 (FAO, 2017). Strikingly, meat production has undergone a higher increment (4.8-fold) in this period, as thanks to genetic improvement and management, the animals now generate more meat than they did decades ago (Figure 1.1). Europe is responsible of 24.2% of the worldwide pig production, with Germany as the top producer (18.9% of the European Union production), followed by Spain (14.8%) (FAO, 2017). Catalonia is the first contributor of the Spanish share (41.8% of the Spanish total) with 1,796,810 tones generated in 2017 (IDESCAT, 2017).

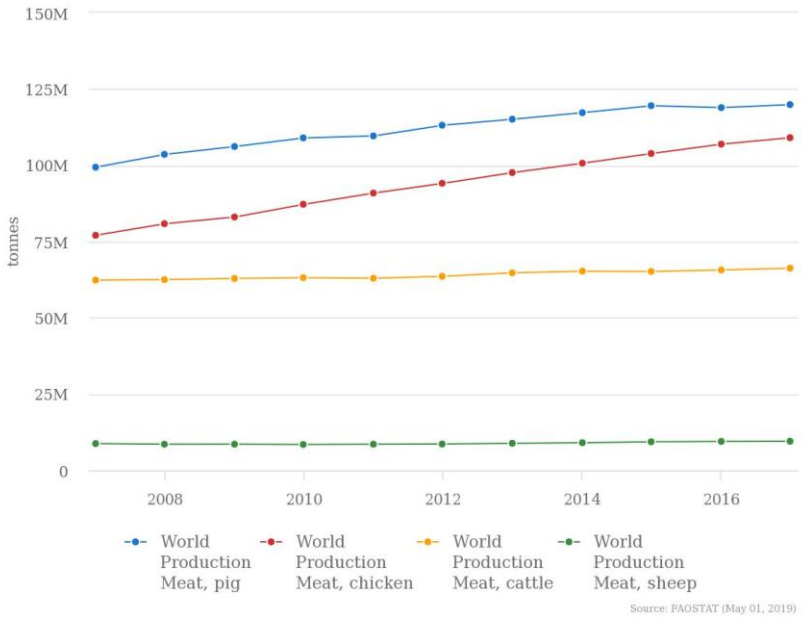


Figure 1.1 World meat production in tones of meat for pig, chicken, cattle and sheep. World meat production trendline from 2007 to 2017 (FAO 2017).

With an expected increase of human population, the pork industry will have to make even bigger efforts to fulfill the consumers' demands, improving carcass and meat quantity and quality, feed efficiency, reproduction, health, animal welfare and other traits that contribute in making the sector more sustainable (Neeteson-van Nieuwenhoven et al., 2013).

1.2 Porcine reproduction

Reproductive efficiency is crucial for the economy of the livestock production sectors. In swine, selection of reproductive traits has been largely centered in sows. These traits mostly included litter size, number of piglets born alive, litter birth weight, number of weaned piglets and age at puberty (Rothschild, 1996; Zak et al., 2017). Nowadays, the industry is beginning to be interested in male reproductive phenotypes such as sexual behavior, libido, testes size and semen quality (Rothschild, 1996; Zak et al., 2017) and even some companies have started recording the boar's conception rate and litter size.

For more than 2 decades, Artificial Insemination (AI) has been widely practiced in countries with intensive boar production. In Europe, this reproductive technique is used in 90% of the sows (Gadea, 2003; Knox, 2016). When

compared to natural mating, AI has greatly contributed in disseminating the genetic material of the boars with the highest genetic merit, reduced the risk of transmitting infectious diseases and lead to better profitability of each ejaculate (Maes et al., 2008).

Sows can be used for reproductive purposes once they reach puberty (5 months of age) (FAO, 1994). In practice, fresh diluted sperm is used for intracervical insemination (Maes et al., 2011) and if successful, pregnancy lasts 114 days (3 months, 3 weeks and 3 days) (FAO, 1994). A sow can have 2 litters per year and 10 piglets per litter, a number that highly depends on the breed (FAO, 1994). In males, a healthy boar reaches puberty at the age of 8 months old and produces spermatozoa until the end of its fertile life (Banaszewska and Kondracki, 2012). Although the boar life span is 10-15 years (Meyerholz, 2016), boars are replaced, in general, before their 4th anniversary unless they are highly exceptional (Robinson and Buhr, 2005).

1.2.1 Biology of sperm: from production to ejaculation

The male reproductive organs conform a complex and intricate system that produces the spermatozoa that will carry the paternal genetic contribution to fertilize the egg and produce offspring. Starting from a pool of self-renewing stem cells, the male germ cells complete their development into spermatozoa in the seminiferous tubules of the testes in a process called spermatogenesis.

1.2.1.1 Histological organization of the testis and epididymis

The testis is composed of 2 structurally and functionally distinct compartments: the seminiferous tubule and the interstitial compartment.

The seminiferous tubules constitute 80 to 90% of the testis volume and are bond among them forming a network known as rete testis (Figure 1.2.A) (Matsumoto and Bremner, 2016). The seminiferous tubules are composed by Sertoli cells and developing germ cells, from spermatogonium (stem cell) to spermatozoa

(Figure 1.2.B and C). Sertoli cells extend from the basement membrane to the lumen of tubules and envelop and support germ cells undergoing progressive differentiation and development into mature spermatozoa (Figure 1.2.C). Sertoli cells form tight junctions that constitute the blood-testis barrier to impede the passage of large molecules, steroids and ions into the seminiferous tubule. Developing germ cells are organized sequentially along the longitudinal axis of the tubules according to their stage of maturity. Immature sperm cells (spermatogonium) are located close to the basement membrane, the most external part of the tube. Located between adjacent Sertoli cells, the other cell types from the spermatogenesis lineage sequentially occupy different positions approaching the lumen. In this order, these cells are spermatocytes (primary and secondary), spermatids (round and elongated) and finally, mature spermatozoa. Spermatozoa will be eventually released into the lumen of the seminiferous tubule and leave the testis through the rete testis and the epididymis (Figure 1.2.A and C) (Garcia-Gil et al., 2002; Matsumoto and Bremner, 2016). There, the post-testicular maturation takes place and the spermatozoa is stored during the passage from testis to vas deferens, in a process that can last up to 12 days in boars (Garner and Hafez, 2000). The epididymis is divided in 3 segments (Figure 1.2.A): the caput epididymis, where spermatozoa concentrate; the corpus epididymis, where spermatozoa mature, and the cauda epididymis, where spermatozoa are stored. During maturation, the cells acquire motility, fertilizing ability and the small amount of excess cytoplasm (cytoplasmic droplet) is lost. In fact, the presence of cytoplasmatic droplets remaining in the ejaculated spermatozoa is a sign of immaturity and low semen quality (Garner and Hafez, 2000).

The interstitial compartment of the testis contains clusters of Leydig cells (Figure 1.2.B) which produce testosterone, a hormone that acts as a paracrine regulator to stimulate spermatogenesis. The interstitial compartment also

contains macrophages that play a role in the phagocytosis of degenerating cells (Matsumoto and Bremner, 2016).

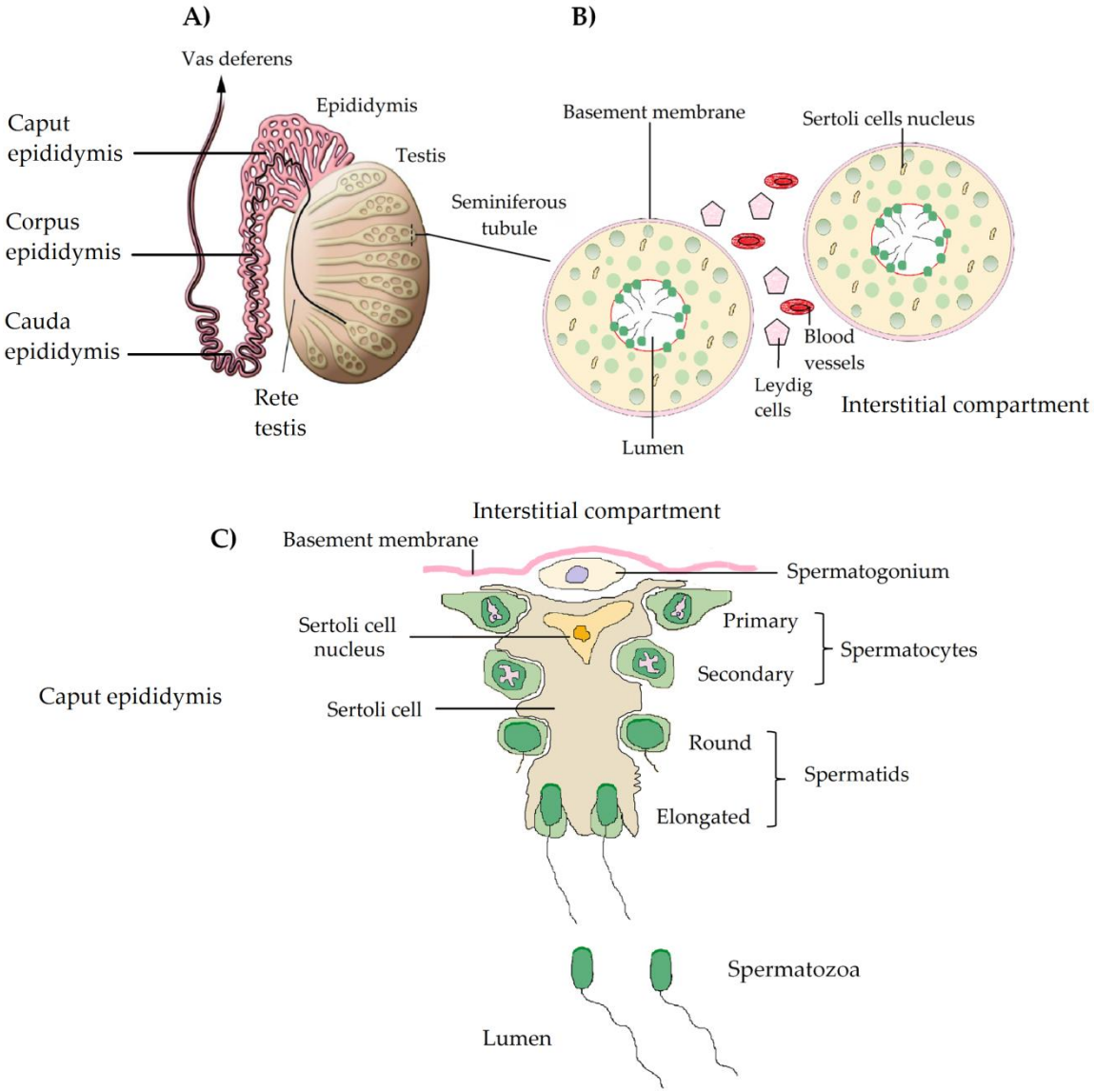


Figure 1.2 Organization of the testis and seminiferous tubules. A) Cross-section of the testis including the vas deferens, epididymis and rete testis. **B)** Diagrammatic cross-section of seminiferous tubules with the different stages of sperm developmental cells (green) between Sertoli cells (yellow). Mature sperm are released in the lumen of the tubes. Leydig cells are found in groups in the interstitial compartment. **C)** Developing germ cells and a single Sertoli cell. From the outer membrane to the lumen: pre-meiotic (spermatogonium), meiotic (primary and secondary spermatocytes) and post-meiotic (round and elongated spermatids) cells. Mature spermatozoa are released into the lumen. Modified from: (Cooke and Saunders, 2002; Hogarth and Griswold, 2010; Krawetz, 2005).

1.2.1.2 Spermatogenesis

Spermatogenesis is a sophisticate process in which stem cells (spermatogonium) differentiate into mature spermatozoa. In pigs, spermatogenesis lasts 34-36 days

(Bonet et al., 2013). During this process, the cells undergo several extensive cellular and functional changes as well as chromatin remodeling events (Krawetz, 2005). Spermatogenesis proceeds in 3 functionally distinct phases (Figure 1.3):

- **Mitotic phase:** stem cells (spermatogonium) divide mitotically to renew the stem cell pool and a minority commit to further differentiate to produce spermatocytes (Jones and Lopez, 2014). Spermatocytes differ from spermatogonium in that they enter meiosis and engage in the genetic recombination process. During this proliferative stage, there is active transcription and translation (Figure 1.3).
- **Meiotic phase:** a primary spermatocyte undergoes 2 rounds of meiotic division to become 4 spermatids (Figure 1.3). In the first cellular division, the diploid primary spermatocyte (2n and 4c; where n is the number of chromosomes in a haploid set and c the haploid amount of DNA) divides to form 2 haploid secondary spermatocytes (1n and 2c), each with 1 set of the sister chromatids. During this process, the homologous chromosomes interchange their genetic material through homologous recombination to generate genetic variation in the offspring as a mechanism of evolution. In the second meiotic division, each secondary spermatocyte undergo the second meiotic division and forms 2 spermatids (1n and 1c) (Jones and Lopez, 2014).
- **Spermiogenesis:** spermatids differentiate into elongated spermatids (Figure 1.3). The cell reduces its size and the vast majority of the cytoplasm (residual body) is lost by fagocytosis mediated by Sertoli cells (Holstein et al., 2003). The nucleolus is displaced to the cell's periphery and the acrosome and flagellum structures form. In this stage, the cell suffers several remodeling changes in the chromatin structure and transcriptional and translational activities cease. These early chromatin changes start by the incorporation of histone variants and the hiperacteylation of histones. Then, the nucleosomal structure is progressively disassembled, replaced by transition proteins and

finally associated to protamines to form a highly compact nucleoprotamine complex (Oliva, 2006) (Figure 1.3). In human and mice spermatozoa, approximately 85-90% of the DNA from the nucleus is associated to protamines but some regions accounting for 10-15% of the genome, remain associated to histones (Brykczynska et al., 2010; Hammoud et al., 2009). These studies have shown that histone retention in sperm is not stochastic. Sperm histones are preferably located in or near genes related to embryo development, miRNA clusters and imprinted gene clusters (Hammoud et al., 2009).

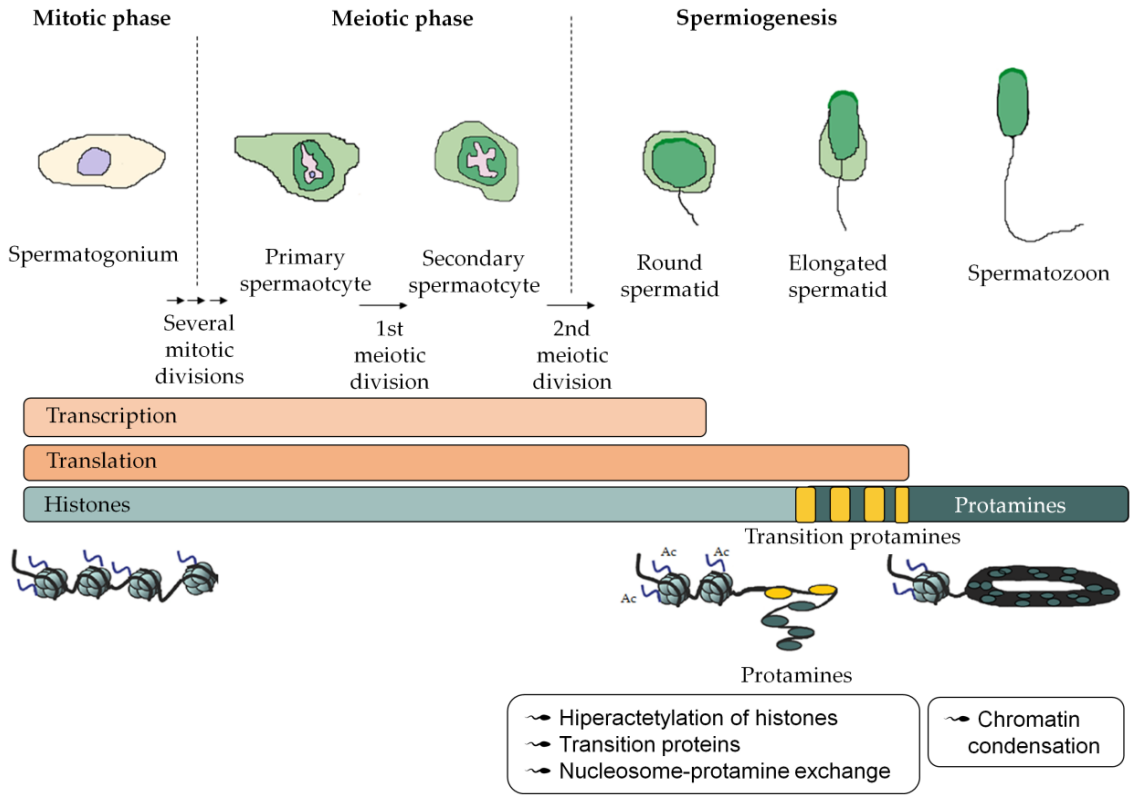


Figure 1.3 Male germ cell developmental events. Spermatogenesis can be divided into 3 major phases: mitotic, meiotic and spermiogenesis. During spermiogenesis, transcription and translation cesases and sperm cells undergo complete reorganization and extensive condensation of the nuclear chromatin, including the replacement of the majority of the nucleosomes by protamines. Modified from: (Oliva, 2006; Schagdarsurengin and Steger, 2016).

1.2.1.3 The boar spermatozoon

The boar spermatozoon is a highly specialized elongated cell about 45 µm long (Briz, 1994), divided into 2 major regions, the head (7 µm) and the tail (37.4 µm), both separated by a short connecting piece or neck (0.7 µm) (Figure 1.4) (Briz and Fàbrega, 2013).

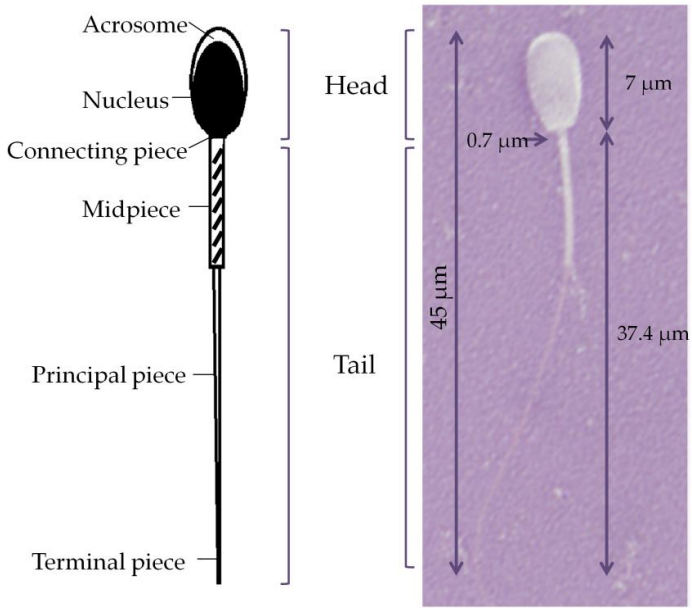


Figure 1.4 Structure of the porcine spermatozoon. Structure and dimensions of the major components of a normal porcine spermatozoon.

The **head** is bilaterally flattened and oval in shape. The nucleus constitutes the major part of the head and is extremely compacted. The apical part of the sperm head is occupied by the acrosome. It contains a large variety of enzymes involved in gamete fusion. The most important enzyme is acrosin, a protease released when the sperm gets in contact with the egg (the so-called acrosome reaction) which is essential for fertilization by allowing the penetration to the zona pellucida (Schill et al., 1988).

The **connecting piece** (or neck) is a linking segment between the sperm head and tail. It does not only act as a physical linkage, but also participates in sperm motility by starting and regulating the waveforms during swimming (Fawcett, 1975).

The **tail** (or flagellum) impulses the spermatozoon in a helical forward movement. It has 3 main distinguishable regions: the midpiece (or mitochondrial region), the principal piece and the terminal piece (Figure 1.4). The axoneme occupies the central axis of the midpiece and is covered by the mitochondrial sheath (hundreds of mitochondria in a helical shape arrangement over the underlying axoneme), responsible for providing the energy needed for the flagellar movement (Bonet et al., 2013; Briz and Fàbrega,

2013). This region extends from the connecting piece to the Jensen's ring, a ring-like structure that avoids the displacement of mitochondria located between the midpiece and the principal piece (Fawcett, 1975). The principal piece of the tail follows the axoneme. It is the longest segment of the tail and spans from the Jensen's ring to the terminal or distal piece. It presents an axoneme structure covered with fibrous sheath. The final piece of the tail is the terminal piece, formed by the axoneme covered with the plasmatic membrane or plasmalemma (Briz and Fàbrega, 2013).

1.2.1.4 Boar ejaculate

During ejaculation, spermatozoa are released from the epididymis towards the urethra where is mixed with seminal plasma. Seminal plasma is a composite secretion arising from testis, epididydimis and accessory glands (seminal vesicle, prostate gland and bulbourethral gland) (Bonet et al., 2013; Garner and Hafez, 2000) (Figure 1.5). Seminal plasma contains a wide range of proteins, lipids, fatty acids and enzymes that provide a nutritive-protective medium to the spermatozoa as well as important components for sperm metabolism, function, survival and transport (Garner and Hafez, 2000; Juyena and Stelletta, 2012). The urethra conducts the semen to the outermost part of the penis, where it is expelled.

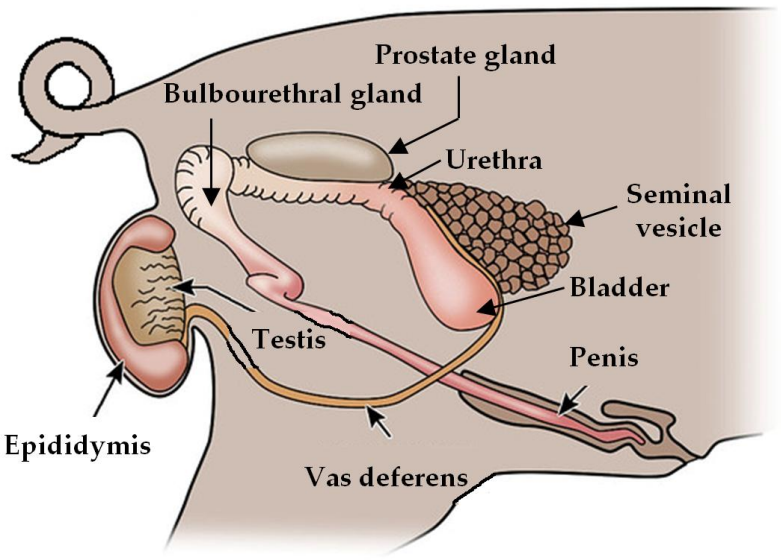


Figure 1.5 Anatomy of the boar's reproductive system. Modified from: North Carolina School of Science and Mathematics.

The ejaculate can be divided in 3 main fractions according to their composition:

- **Pre-spermatic fraction:** formed by the accessory glands, lubricates and cleans the urethra for sperm passage. It accounts for 10 to 15 ml of the ejaculate and does not contain spermatozoa (Sancho and Vilagran, 2013).
- **Sperm-rich fraction:** with 70 to 100 ml and a milky appearance, it contains high sperm concentration 0.5 to 1 x 10^9 spermatozoa/ml. It is composed of secretions from the prostate and seminal vesicle and is the only fraction used for preparing seminal doses for AI (Sancho and Vilagran, 2013).
- **Post-spermatic fraction:** with 150 to 200 ml and pale white appearance, this fraction contains low sperm concentration (1 x 10^6 spermatozoa/ml). It is mostly composed of seminal plasma and its function is to stimulate the spermatozoa. This fraction is not recommended to be included in AI doses as basal metabolism is required for the preservation of good sperm quality during storage (Sancho and Vilagran, 2013).

In AI stations, boar ejaculates are collected on a routine basis with the hand-glove method (Vyt et al., 2007) and up to 3 ejaculates per week can be collected in a boar with optimal semen quality traits (Bonet et al., 2013). The ejaculated volume varies between 150 to 300 ml depending on breed, age, season, rhythm of collection or diet, among others (Yeste et al., 2010). Average final sperm concentration is 0.2 x 10^9 spermatozoa/ml and the total sperm number per ejaculate is 10 to 100 x 10^9 spermatozoa (Bonet et al., 2013; Garner and Hafez, 2000).

After ejaculate collection, the sperm is diluted with an extender, an aquous solution used to provide the optimal condition to preserve the spermatozoa and to increase the volume of the ejaculate to the volume and concentration of the required dose. Semen extenders can be divided in short-term preservation (less than 3 days) or long-term preservation (over 4 days). To perform its action, extenders include nutrients (glucose) for the metabolic maintenance, regulation of pH (e.g. bicarbonate and Tris), control of osmotic pressure (e.g. NaCl and KCl), protection against cold shock (bovine serum albumin) and prevent

microbial growth (antibiotics) (Gadea, 2003). Dilution of ejaculates with semen extenders together with lower storage temperature (15-16ºC), enables long distance transport, health inspection, conduct diagnostic tests and generally improve the management in-farm (Gadea, 2003).

1.2.2 Semen quality

Although, young boars are mostly selected for AI based on estimated breeding values of production traits (e.g. carcass or meat quality), male fertility outcomes have become a major focus of interest for AI companies. Only the highest-merit boars are chosen for AI and a single ejaculate can inseminate 10 to 15 sows. As the pig industry is in continuous selection pressure aiming to maximize genetic progress, fewer boars will be used in the breeding nucleus and thus, the AI doses will have to be efficiently used (e.g. more doses with less sperm concentration). For this reason, the early detection of infertile or subfertile boars can have a large positive repercussion for the sustainability of AI and the pork industry (Briz and Fàbrega, 2013). Sperm quality has been proposed as a proxy of fertility outcomes (Aitken, 2006; Alm et al., 2006; Broekhuijse et al., 2012b; Gadea and Matas, 2000; Hirai et al., 2001; Juonala et al., 1998; Paston et al., 1994; Tsakmakidis et al., 2010) and AI studs routinely evaluate semen quality traits, including volume, concentration, motility, agglutination, bacterial contamination or cell morphology (Knox, 2016; Robinson and Buhr, 2005). As a matter of fact, 10 to 30% of the boars are replaced annually due to sperm quality problems (Robinson and Buhr, 2005), thereby causing substantial economic losses in AI centers.

The main traits of interest for semen quality are:

- **Sperm concentration**. Number of cells per volume of ejaculate. Previous studies have shown that increasing sperm concentration of AI doses results in higher fertility outcomes. Nowadays, the standard AI dose is set at 3 x 10^9 spermatozoa (Maes et al., 2011).

- **Sperm volume**. Varies among pig breeds and ranges between 100 and 300 ml. Together with the total cell number per ejaculate, it determines the sperm concentration (Kondracki, 2003).

- **Sperm motility**. Although spermatozoa is transported to the fertilization site through uterine contraction (Langendijk et al., 2002), sperm motility is required for penetration of the zona pellucida (Maes et al., 2011). Although it is one of the most attractive traits when purchasing semen doses, its impact in fertility remains inconclusive (Broekhuijse et al., 2012a; Hirai et al., 2001; Holt et al., 1997; Quintero-Moreno et al., 2004). Motility can be objectively assessed using the digital image analysis of the Computer Assisted Semen Analysis (CASA) instrument, which analyzes the motion properties of spermatozoa (Verstegen et al., 2002). Most AI centers consider an ejaculate as satisfactory if the percentage of motile cells is above 70% (Flowers, 2009; Shipley, 1999).

- **Sperm morphology, membrane integrity and cell viability**. The microscopic appearance of spermatozoa can provide information on morphological abnormalities and plasma membrane integrity. Morphological abnormalities can be indicative of aberrant spermatogenesis and may easily compromise the function of the cell (Maes et al., 2011). In fact, in swine, strong relationships between morphologically normal sperm and fertility have been found (Alm et al., 2006; Tsakmakidis et al., 2010). Sperm abnormalities can be assessed with Eosin-Nigrosin staining (Figure 1.6). This morphological examination can determine the percentage of altered acrosomes, morphological abnormalities (of the head, neck and tail) and cytoplasm leftovers (proximal and distal droplets) (Quintero-Moreno et al., 2004). The evaluation of the functional membrane integrity of the spermatozoa can be assessed via the Osmotic Resistance Test (ORT), which quantifies the percentage of cells that are resistant to acrosomal membrane changes when subjected to changes in the osmotic conditions of the medium (Rodríguez-Gil and Rigau, 1995) and via cell

viability, assessed by evaluating the resilience that typically living cells with intact membranes have to absorb the dye (Moskovtsev and Librach, 2013).

Figure 1.6 Porcine sperm morphology. (Left to right) Eosin-Nigrosin staining of spermatozoa with normal morphology, abnormal head, distal droplet and abnormal head and coiled tail.

Sperm morphology in swine can be affected by a wide range of genetic (e.g. breed, presence of congenital diseases), environmental (e.g. temperature, humidity, photoperiod), management (e.g. socialization, diet, sperm handling, frequency of sperm collection) and health (e.g. testicular, epididymal or sex gland pathologies), factors (Briz and Fàbrega, 2013).

1.3 Pig genomics

Classical livestock genetic selection programs target traits that impact on the cost of production and on product quality. In the recent years, traits affecting welfare and environmental sustainability are also gaining interest. In the earliest days, genetic selection involved recording genetic relationships and phenotypes of related individuals (pedigrees) to estimate the genetic merit of the animals. The recent introduction of genomics in selection goals (genomic selection) enables the evaluation of animals with higher precision and much faster thus accelerating genetic progress (Zak et al., 2017).

The pig genome is organized into 18 autosomal and 1 sexual chromosome (X and Y) pairs. In comparison to the human genome, it has similar complexity, organization and genome size. The sequencing of the swine genome was performed after the establishment of the Swine Genome Sequencing Consortium (SGSC) in 2003, which led to successful genetic (Groenen et al.,

2011) and physical maps (Raudsepp and Chowdhary, 2011). The strategy of the SGSC included Sanger sequencing of bacterial artificial chromosome (BAC) clones of the genome (Humphray et al., 2007), and was later on complemented with Illumina high throughput sequencing data (Archibald et al., 2010). These efforts resulted in the first reference genome sequence for *S. scrofa* in 2012 (Groenen et al., 2012), an assembly known as Sscrofa10.2. Since then, several hundreds of genomes from different pigs have been re-sequenced to study genetic variation, evolution and selection (Groenen, 2016). The latest update of porcine reference genome sequence was released in July 2017. This assembly (Sscrofa11.1) was obtained and curated using several sequencing platforms. The assembly was constructed mostly using Pacific Biosciences long reads and then curated with Sanger sequencing, Illumina and Oxford Nanopore sequence data. It comprises a 2.5 Gb sequence assigned to chromosomes and additional 66 Mb of sequence assembled into unplaced scaffolds. The latest Ensembl genome annotation (version 97) includes nearly 22,500 protein coding genes, 3,250 noncoding genes, over 49,000 transcripts and more than 64 M of DNA short variants, including Single Nucleotide Polymorphisms (SNPs), insertions-deletions (indels) and structural variants. This annotation is way less complete than the human counterpart (Ensembl version 97), with a similar number of annotated genes 20,454, but many more noncoding genes (nearly 24,000), transcripts (closely to 227,000) and genetic variants (over 665 M). A large international consortium (the Functional Annotation of Animal Genomes, FAANG) was recently set up with the aim to improve the reference genome sequences and annotation of livestock species. These annotations include the catalog of coding and noncoding RNAs, genetic variants and functional elements of the genomes flagged by a set of epigenetic marks (DNA methylation, chromatin accessibility, histone modifications or chromatin spatial conformation) in multiple cell types, conditions, breeds and livestock species (Andersson et al., 2015). It is thus expected that the annotation of the pig genome will improve substantially in the near future.

1.4 Genomic approaches to study sperm quality

Several studies have reported that sperm quality is heritable (Table 1.1) and that a proportion of the phenotypic variation is under genetic control. Hence, these traits can be included in genetic selection programs.

Table 1.1 Heritability estimates for semen quality traits in swine.

Semen characteristics	Heritability	References
Ejaculate volume (ml)	0.14-0.58	(Leenhouwers et al., 2008; Smital et al., 2005; Wolf, 2009; Wolf and Smital, 2009)
Motility (%)	0.05-0.38	(Leenhouwers et al., 2008; Marques et al., 2017; Smital et al., 2005; Wolf, 2009; Wolf and Smital, 2009)
Progressive Motility (%)	0.11-0.38	(Marques et al., 2017; Smital et al., 2005)
Sperm concentration (10^6/ml)	0.10-0.49	(Marques et al., 2017; Oh et al., 2006; Smital et al., 2005; Wolf, 2009; Wolf and Smital, 2009; Wolf, 2010)
Abnormal sperm (%)	0.04-0.34	(Marques et al., 2017; Wolf, 2009; Wolf and Smital, 2009)
Number of insemination doses	0.40	(Smital et al., 2005)
Number of piglets born alive	0.08	(Smital et al., 2005)
Conception rate	0.29	(Smital et al., 2005)

Genetic correlations across semen quality traits have also been studied and have been found to vary widely among trait pairs. Also, some discrepancies between studies have been observed (Smital et al., 2005; Wolf, 2009; Wolf and Smital, 2009). For example, while Smital et al. reported no correlation between sperm concentration and the percentage of morphologically abnormal cells (0.01) (Smital et al., 2005), other studies reported moderate (0.11) (Wolf and Smital, 2009) or even high (−0.60) (Wolf, 2009) correlations between these 2 traits. Discordant observations have been also observed between motility and the percentage of abnormal cells: -0.57 (Wolf and Smital, 2009), −0.87 (Wolf, 2009) or -0.34 (Smital et al., 2005).

Since the 1990's, DNA markers have been used for the selection or removal of desirable or undesirable traits, respectively (reviewed in: Dekkers, 2012). The use of DNA markers is particularly efficient for traits that:

- Cannot be measured early in life.
- Can only be assessed in one sex.
- Are difficult or expensive to measure.

Semen quality traits fulfill all these criteria as ejaculates are only produced after puberty and only in males and the phenotypes are complex and expensive to measure. Semen quality is a set of polygenic traits determined by multiple genes, molecular mechanisms and regulatory pathways (Zak et al., 2017). Although the genetic basis of boar sperm quality remains to be elucidated, several efforts employing different technological methodologies have been made. Several approaches can be used to elucidate the molecular basis of phenotypes at the nucleic acids level. These are briefly described below.

1.4.1 Genetic variation in candidate genes

One of the original approaches to identify DNA polymorphisms associated with traits of interest in livestock focused on the identification of these variants in candidate genes known to play a role in that given phenotype. Two main approaches based on genomic PCR amplification have been used to study the boar sperm quality: the restriction fragment length polymorphism (RFLP) (Botstein et al., 1980) and the Sanger sequencing methods (Sanger, 1975). In RFLP, the DNA is cut with a restriction enzyme which restriction site is altered by the DNA polymorphism that wants to be genotyped. This results in a different electrophoretic DNA banding pattern that allows genotyping the samples (Botstein et al., 1980). On the other side, Sanger sequencing is a technique for DNA sequencing based on the selective incorporation of dideoxynucleotides by the DNA polymerase while the target DNA is amplified by PCR (Sanger, 1975). Sanger sequencing has been widely applied to identify and to genotype DNA polymorphisms in candidate genes. Using both technologies, several polymorphisms in candidate gene loci have been associated to boar sperm quality traits. A summary of some of these studies is highlighted in Table 1.2.

Table 1.2. Examples of candidate genes associated with sperm quality traits identified through Sanger sequencing or restriction fragment length polymorphism.

Gene name	Gene symbol	Trait(s)	Reference
Acrosin	*ACR*	Concentration and motility	(Lin et al., 2006a)
Actin Gamma 2, Smooth Muscle	*ACTG2*	Volume	(Wimmers et al., 2005)
Cluster-of-differentiation antigen 9	*CD9*	Motility, droplets and abnormalities	(Kaewmala et al., 2011)
Cytochrome P450 Family 21	*CYP21*	Concentration and abnormalities	(Kmiec et al., 2002)
Estrogen Receptor 2	*ESR2*	Motility and droplets	(Gunawan et al., 2012)
Gonadotropin Releasing Hormone Receptor	*GNRHR*	Motility, droplets and abnormalities	(Lin et al., 2006b)
Heat Shock Protein Family A (Hsp70) Member 2	*HSP70.2*	Volume, abnormalities and proximal droplets	(Huang et al., 2002)
Inhibin Subunit Beta A	*INHBA*	Droplets	(Lin et al., 2006b)
Inhibin Subunit Beta B	*INHBB*	Concentration	(Lin et al., 2006b)
Phospholipase C-Zeta	*PLCz*	Concentration	(Kaewmala et al., 2012)
Ryanodine Receptor 1	*RYR1*	Volume, motility, abnormalities and semen doses	(Urban and Kuciel, 2001)
Secreted Phosphoprotein 1	*OPN*	Motility and abnormalities	(Lin et al., 2006a)
Sperm Flagellar 2	*KPL2*	Abnormalities	(Sironen et al., 2006) (Cheng et al., 2016)

1.4.2 Genome-wide association studies

SNPs are single nucleotide substitutions that are highly ubiquitous in the genome. Their abundance and ease to genotype using current microarray technologies has made SNPs the polymorphisms of choice in genetic association studies to link DNA variants with traits of interest. The first commercial SNP microarray panel for high throughput genotyping in swine was released in 2009 (Ramos et al., 2009). This panel (PorcineSNP60v2 BeadChip) was commercialized by Illumina and contained 64,232 markers across all autosomal and X chromosomes (Ramos et al., 2009). Two additional genotyping panels were later developed by GeneSeek/Neogen. The first was a SNP chip panel of

low density with 10,241 SNPs (GeneSeek Genomic Profiler for Porcine LD BeadChip), and the second was a panel of higher density that included 70,231 markers (GeneSeek Genomic Profiler for Porcine HD BeadChip). More recently, Affymetryx released a high-density genotyping panel Axiom Porcine Genotyping Array (Axiom_PigHDv1) containing 658,692 SNPs in the autosomal and sex chromosomes including most of the markers from the previous PorcineSNP60v2 BeadChip (Groenen, 2015).

The SNP genotyping panels have enabled the detection of genomic regions associated to quantitative traits, known as quantitative trait loci (QTL) (Geldermann, 1975). The Pig QTL database (release 38) (Hu et al., 2019) includes information on 29,045 QTLs for a wide variety of traits including reproductive performance traits (448 QTLs). A small fraction of these reproductive QTLs (81) belongs to sperm quality parameters (e.g., sperm concentration, motility, volume and concentration).

Between the 1990's and 2010, QTLs would be identified typically by linkage analyses using F_2 or back-cross experimental populations or by association studies using commercial populations, genotyped with hundreds of ultra-polymorphic microsatellite markers. This approach had several limitations due to the economic cost of the technologies at that time. In the second decade of the 21st Century, commercial SNP platforms became available and rapidly replaced microsatellite markers.

The genome-wide association study (GWAS) is an approach to identify genomic regions associated to target phenotypes, either Mendelian or complex. This scan method exploits the high throughput and cost-effective genotyping of tens or hundreds of thousands or even millions of SNP variants probed into microarrays which, allow the genotyping of a large number of samples. It is based on a statistical approach to detect the significant associations between each genetic marker and the trait of interest. GWAS relies on linkage

disequilibrium (LD) between nearby DNA variants and the concept that a portion of the genotyped variants will be in LD with - and thus tag - causal functional variants impacting on the trait under study. The statistical power to detect association depends on the sample size, the effect sizes of the causal genetic variants, the allelic frequencies of these variants and the LD between the genotyped marker and the unobserved causal variant for the trait of interest (Visscher et al., 2017). In comparison with the candidate gene approach, it scans the whole genome and does not rely on previous information. Genotyping of thousands of variants distributed along the genome using the commercial genotyping microarrays or other tools developed in-house by the breeding companies is already applied for genomic selection and it has considerably accelerated the rate of genetic progress.

In pigs, there are less than a handful of GWAS studies for sperm quality traits using SNP panels. Sironen et al. (Sironen et al., 2010) identified the Ubiquitin-protein ligase E3 *(HECW2)* as a candidate gene for the knobbed acrosome defect, a Mendelian sperm head abnormality identified only in Finnish Yorkshire boars. Diniz et al. (Diniz et al., 2014) identified a single QTL for sperm motility and suggested the mitochondrial methionyl-tRNA formyltransferase (*MTFMT)* as a candidate gene. Zhao et al. (Zhao et al., 2016) found 3 QTLs and 7 candidate SNPs associated to sperm quality traits in a White Duroc x Erhualian F_2 population. A QTL for the total number of sperm per ejaculate and sperm concentration harbored the telomerase-associated protein 1 (*TEP1*) and poly (ADP-ribose) polymerase 2 (*PARP2*) candidate genes. Zhao and co-authors also identified significant SNPs related to semen temperature and sperm motility close to the Serine Peptidase Inhibitor, Kazal Type 1 (*SPINK1)* gene and to ejaculation times within the Phosphodiesterase 1C (*PDE1C*) gene. More recently, Marques et al. (Marques et al., 2018) detected 16 and 6 semen quality QTLs in Large White and Landrace, respectively. These

QTLs were associated to sperm motility and progressive motility, number of cells per ejaculate and sperm morphological abnormalities. Reported candidate genes included the Sodium Voltage-Gated Channel Alpha Subunit 8 (*SCN8A)*, Prostaglandin-Endoperoxide Synthase 2 *(PTGS2)*, Phospholipase A2 Group IVA (*PLA2G4A),* Dynein Axonemal Intermediate Chain 2 *(DNAI2),* IQ Motif Containing G *(IQCG)* and Spermatogenesis Associated 7 (S*PATA7*).

The genetic relationship between DNA variants and gene expression can be also evaluated using the same SNP platforms and approach as used in the standard GWAS. This analysis is known as expression GWAS (eGWAS) and tests the association between genetic markers and the genes' expression levels. This approach allows the identification of genomic regions harboring functional variants regulating gene expression without previous knowledge on the genome's functional cartography (Schadt et al., 2009). Typically, eGWAS effects are classified as *cis*- (nearby) or *trans*- (distant) acting depending on the physical distance between the SNP (eSNP) and the regulated gene. The distances distinguishing *cis*- and *trans*- are normally subjective and differ between studies. To our knowledge, no eGWAS has been carried to study boar sperm quality traits.

1.4.2.1 Transmission ratio distortion

Data generated from SNP panels have also been used to study transmission ratio distortion (TRD). TRD occurs when 1 of the 2 alleles from either parent is preferentially transmitted to the offspring, causing a statistical departure from the Mendelian inheritance ratio of 0.5 (reviewed in: Huang et al., 2013). Thus, TRD can have a direct impact on reproduction. There are several mechanisms that have been suggested to induce TRD such as (i) germline selection: during mitosis, mechanisms as mutation or recombination cause cells with a certain genotypes to be preferentially produced. Hence, germ cells entering to the next stage (meiosis) have an imbalanced genotype ratio; (ii) meiotic drive: when during spermatogenesis, a sex chromosome drive occurs which leads to

unequal production of X- or Y-bearing gametes; (iii) gametic competition: spermatozoa compete to achieve fertilization and this may be allele-dependent; (iv) imprinting errors: after fertilization, errors during imprinting resetting can result in the lethality of the embryos with a specific genotype (reviewed in: Huang et al., 2013). TRD has been weakly studied in mammals (reviewed in: Huang et al., 2013). Of these, 3 independent publications reported a relationship between TRD and sperm motility (Bauer et al., 2007; Bauer et al., 2012; Véron et al., 2009). In swine, Casellas et al. (Casellas et al., 2014) identified 84 SNPs in strong evidence of TRD using genotype data from 5 boars and their 352 offspring. TRD SNPs mapped within or near regions enriched for biological processes such as chromatin assembly and organization or DNA packaging (Casellas et al., 2014).

1.4.3 Next-generation sequencing technologies

Next-generation sequencing (NGS) technologies succeed the automated Sanger sequencing method (Sanger, 1975) and a new era in the genomics field begun. NGS is able to produce a large amount of data at low cost and different methodologies and solutions have been developed to characterize the various -omics (genomics, transcriptomics, epigenomics) levels. Since their introduction in 2005 (Margulies et al., 2005), 6 companies have controlled the market using different platforms that can be divided into the second and third generation groups. Second NGS platforms rely on PCR amplification of a given DNA template and perform parallel and cyclic sequencing and imaging (Metzker, 2010). At the beginning, the 3 leading companies and their instruments were: Roche (e.g. 454 Genome Sequencer FLX+), Illumina (e.g. MiniSeq, NextSeq, HiSeq, Novaseq) and Thermo Fisher (e.g. SOLiD –Sequencing by Oligo Detection-, Ion Proton). However, Illumina quickly dominated the market and became the first choice for most NGS sequencing experiments of the last decade. Third generation platforms interrogate single molecules of DNA and can produce longer read length. The 3 main companies are: SeqLL (Helicos) (Heliscope SMS –Single Molecule Sequencing-), Pacific Biosciences (e.g. PacBio RS II SMRT – Single-Molecule Real Time-) and Oxford Nanopore (e.g. MinION

SMRT) (Pareek et al., 2011; Yeh et al., 2018). Third generation technologies such as Pacific Biosciences just recently started to become commercially available and many livestock genome assemblies are being updated with these long reads.

1.4.3.1 Transcriptome analysis

Transcriptomic studies characterize and quantify the RNA content of a cell, tissue or organism at a genomic scale. This can help identifying, for a given trait of interest, the underlying key genes and their associated changes. Transcriptome analysis can be achieved by screening RNA samples through either oligoarray technologies with probes targeting a set of pre-determined genes or transcripts or by RNA-seq. While microarrays can only interrogate the expression of the target RNA molecules, RNA-seq, a NGS technique, has the potential to analyze the expression of all the RNA molecules (long and short fractions) present in a sample, regardless of whether they have been previously annotated or not. Remarkably, RNA-seq yields a myriad of additional research opportunities including the detection of novel genes (both coding and noncoding) and splicing-isoforms, the identification of transcribed genetic variants, the inference of allele specific expression or the annotation and quantification of microbial transcripts (Wang et al., 2009). Moreover, RNA-seq overcomes the microarray based methods with higher dynamic range, specificity and sensivity (Wang et al., 2009). The studies provided in this thesis assess for the first time, with RNA-seq, the role of RNAs on boar sperm quality. Several studies have provided solid evidences of the presence of RNAs in sperm and their role in semen quality and fertility in different species. In the following section a discussion review on sperm RNAs that was developed within the work frame of this thesis is included.

1.4.3.1.1 Review: A history why father's RNA matter

A history of why fathers' RNA matters

Marta Gòdia[1], Grace Swanson[2,3], Stephen A Krawetz[2,3,4*]

[1] Animal Genomics Group, Center for Research in Agricultural Genomics (CRAG) (CSIC-IRTA-UAB-UB), Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[2] Department of Obstetrics and Gynecology, Wayne State University, Detroit, Michigan, USA

[3] Department of Obstetrics and Gynecology, Wayne State University, Detroit, Michigan, USA

[4] C.S. Mott Center for Human Growth and Development, Wayne State University, Detroit, Michigan, USA

*Corresponding author

Biology of Reproduction, Volume 99, Issue 1, 1 July 2018, Pages 147–159,
https://doi.org/10.1093/biolre/ioy007

Awarded as most cited review of Biology of Reproduction in 2018

Abstract:

Having been debated for many years, the presence and role of spermatozoal RNAs is resolving, and their contribution to development is now appreciated. Data from different species continues to show that sperm contain a complex suite of coding and non-coding RNAs that play a role in an individual's life course. Mature sperm RNAs provide a retrospective of spermatogenesis, with their presence and abundance reflecting sperm maturation, fertility potential, and the paternal contribution to the developmental path the offspring may follow.

Sperm RNAs delivered upon fertilization provide some to the initial contacts with the oocyte, directly confronting the maternal with the paternal contribution as a prelude to genome consolidation. Following syngamy, early embryo development may in part be modulated by paternal RNAs that can include epididymal passengers. This provides a direct path to relay an experience and then initialize a paternal response to the environment to the oocyte and beyond. Their epigenetic impact is likely felt prior to embryonic genome activation when the population of sperm delivered transcripts markedly changes. Here, we review the insights gained from sperm RNAs over the years, the subtypes, and the caveats of the RNAs described. We discuss the role of sperm RNAs in fertilization and embryo development, and their possible mechanism(s) influencing the offspring's phenotype. Approaches to meet the future challenges as the study of sperm RNAs continues, will include, among others, elucidating the potential mechanisms underlying how paternal allostatic load, the constant adaptation of health to external conditions, may be relayed by sperm RNAs to affect a child's health.

Introduction

Although spermatozoa were thought for many years to merely contribute their half of the genome to the offspring, it is now appreciated that sperm delivers its entire structure, from which selective components are used to build a healthy child [1]. These include, phospholipase C zeta (PLCZ) and other factors that yield a pulsatile Ca2+ response, and in humans, the sperm centriole organizing center [reviewed in 2]. In contrast, certain structures including the mitochondria are ubiquitinated and targeted for degradation [reviewed in 2]. The suite of RNAs that a sperm carries is also part of the package. The RNAs can reflect the fidelity of spermatogenesis that impacts embryo development [3, 4], that, in turn, may affect offspring phenotype [5-8]. This has prompted the field to begin to dissect their role in the fetal origins of adult disease, a concept laid out in the Barker hypothesis [9] that continues to be tested as our environment changes. Health is maintained and reflects a biological endpoint termed allostatic load [10], that is both directed by, and responds to, past and present events [11, 12]. A system overload, even by one too many inputs, leaves the potential for adverse health effects to arise.

The relationship between the allostatic load and offspring health has recently been implicated in epidemiological studies and recapitulated in mouse models [13-15], with sperm RNAs as potential messengers between generations.

In this review we will provide a historical perspective of sperm RNAs, and discuss the characteristics of this unique population and the caveats of studying sperm RNAs. Their contribution to the oocyte and their role in embryo development is considered. We conclude with a discussion of how sperm RNAs may respond to their environment epigenetically relaying the father's experience to the offspring.

History of sperm RNAs

Sperm RNAs had been considered an artifact and if present, as having no role in fertilization and obviously not embryo development. The hypothesis was founded upon the extrusion of RNAs from the cell as part of the residual body during spermatogenesis and their lack of intact ribosomal RNA (rRNAs) [16]. This led to the conclusion that if any RNA remained, it was residual, possessing no function in the absence of a complete transcript. This has not stood the test of time. With marked technological advances that the field has experienced over the years, the presence and some of the roles of sperm RNAs (Figure 1) are now established with possibly others on the horizon.

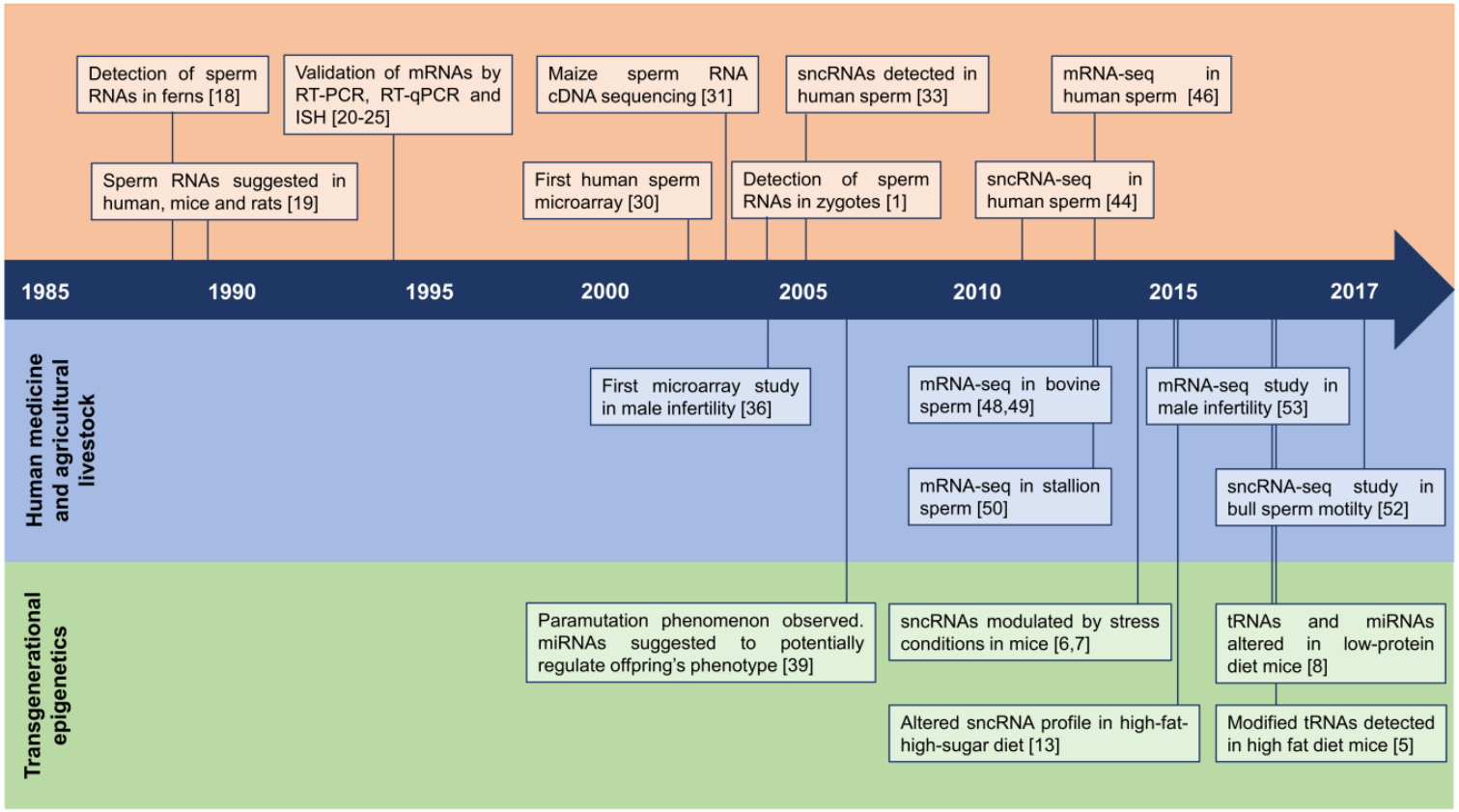


Figure 1. Three decades of sperm RNAs summarized with key studies. Sperm RNAs were first detected in ferns [18], and subsequently in mammals [19]. In the 1990s, several independent studies [20–25] validated the presence of mRNAs in sperm with different approaches including RT-PCR, RT-qPCR, and ISH. In the 2000s, with novel technologies, microarray allowed the first human sperm transcriptome profile [30], a technique posteriorly used for studying male infertility [36]. The first species to be sequenced (by sperm RNA cloning and cDNA sequencing) was in maize [31]. In 2004, sperm RNAs were shown to be delivered to the zygote [1]. Small noncoding RNAs were firstly detected in human sperm by microarray [33] and corroborated, a few years later, with sncRNA-seq [44]. With the boost of NGS techniques, mRNA-seq studies were carried in sperm RNAs in human [46], and agricultural species as bovine [48, 49] and stallion [50]. NGS has also been used to identify biomarkers related to male infertility in humans [53] or sperm quality (motility) in bulls [52]. Relevant transgenerational epigenetics studies start with the work from Rassoulzadegan and colleagues [39], where miRNAs were suggested to modify offspring phenotype. Several studies in mice have followed, where parents have been set in extreme environmental conditions, including stress [6, 7] and diet [5, 8, 13], and sncRNAs profiles have been studied. ISH (in situ hybridization); sncRNAs (small noncoding RNAs); miRNAs (micro RNAs); tRNAs (transfer RNAs).

Early studies suggested that spermatozoal RNAs were present in the epididymis and ductus deferens from mouse sperm, but were dismissed as reflective of mitochondrial contamination [17]. It was not until the late 1980's, that the presence of sperm RNAs was beginning to be considered by various independent techniques e.g., using immunogold staining in ferns [18], and finally in humans and rats by RNase colloidal gold. This provided an estimate of 70 and 100 fg RNA per cell respectively, with the majority primarily localized within the nucleus [19]. RT-PCR and In Situ Hybridization (ISH) reconfirmed these observations in human sperm with the detection of MYC [20]. However, the field was still not convinced attributing these observations to spurious hybridization or priming events. Several independent studies unknown to each other were undertaken in the 1990s that addressed the issue of the presence of RNAs in sperm using various techniques. These included RT-PCR, ISH and RT-qPCR, following the exclusion of samples with genomic DNA or contamination by somatic cell RNAs. Detection of *PRM2*, *E2R*, and *ACTB* by RT-PCR [21], *ITGB1* via PCR [22], and *PRM1*, *PRM2*, *TNP2* in humans [23] and mice [24] by ISH, as well as *LGC1* by ISH and RT-PCR in lily plants was observed [25]. It was not until 1999, when Miller, Krawetz and colleagues attempted to characterize human sperm RNA transcripts at the sequence level [26]. A group of translationally quiescent RNAs, the majority of which were derived from repetitive elements with the near or complete absence of 28S and 18S rRNAs were revealed [26]. This was a curious observation in a transcriptionally inactive sperm cell [16, 27]. Translation of nuclear encoded RNAs by mitochondrial ribosomes in mature sperm during sperm capacitation has been proposed [28, 29], but these observations still await independent confirmation.

Microarray analysis that followed provided the first characterization of sperm mRNAs from healthy human donors [30]. A set of transcripts shared between sperm samples functionally enriched for spermatogenesis and early

development were resolved, suggesting that the retention and presence of certain RNAs was not stochastic [30]. In maize (Zea mays), a set of sperm RNA transcripts were identified by cDNA sequencing with some derived by selective partitioning [31]. Delivery of human sperm RNAs to the oocyte was validated in the hamster penetration assay opening the door to a role in early embryonic development, and the male assuming a greater role in the birth of a healthy child [1]. Antisense and micro RNAs (miRNAs) were also identified amongst this population, and thus inferred to be delivered to the oocyte upon fertilization. This cemented the foundation for the beginnings of mechanistic proposals regulating parental gene activity through targeting paternal or maternal RNAs, thereby regulating early genomic events in the embryo [32, 33]. Characterization of the human sperm transcriptome by Serial Analysis of Gene Expression (SAGE), i.e., tag sequencing [34], identified a series of highly abundant transcripts in fertile donors, enriched for roles related to spermatogenesis, sperm function, fertility, and conception [35], in accordance with previous microarray studies [30]. This further encouraged interest to assign potential roles for sperm RNAs. Microarray analysis was beginning to lay a path towards developing diagnostic strategies to understand male infertility [3, 4, 36].

Advances continued over the years and in 2005, shortly after the identification of miRNAs in human sperm by microarrays [33], miRNAs were observed in mice using a microarray and RT-qPCR approach [37]. Other small non-coding (snc) RNAs like Piwi-interacting RNAs (piRNAs) were subsequently detected in mouse testis by pyrosequencing [38] but left their description in sperm wanting. The low abundance of paternal miRNAs in zygotes led to the assumption they possessed a limited role, if any, in fertilization [37]. Yet, Rassoulzadegan and colleagues provided the first study that suggested that miRNAs could influence offspring phenotype, a phenomenon known as

paramutation [39]. As with most studies in this field, this initial foray into sperm RNA mediated transgenerational epigenetics was controversial and met with skepticism by others, but this concept has also withstood the test of time.

In recent years, several compelling independent studies have corroborated the initial findings of Rassoulzadegan and colleagues that epigenetic modifications can be transmitted from father to offspring via paternal RNAs [reviewed in 13, 40, 41, 42]. Various components or stressors including mental and physical stress, induced by the physical environment, toxins, or diet that together constitute the allostatic load experienced by the father have been individually well-studied in mice and rats [5, 40, 43]. Both miRNAs [7] and piRNAs [6] have been implicated in paternal stress studies and have provided a framework. Similarly, paternal diet can coincide with alterations in the abundance of miRNAs and transfer RNAs (tRNAs) leading to transgenerational effects whether they are subject to a high-fat-high-sugar diet [13], a high fat diet [5], or a low protein diet [8]. Still, the results of these studies, especially in humans require further study (see: "RNAs and Transgenerational Epigenetic Inheritance").

Next Generation Sequencing (NGS) provided a boost to RNA-seq studies early in the first decade of the millennium. The first sncRNA-seq study performed in human sperm showed the presence of several sncRNAs including miRNAs, piRNAs, and repeat-associated small RNAs [44]. Spurred on by this work, sncRNA-seq in mouse sperm followed, that also reported a highly enriched fraction of tRNAs [45]. Several sperm RNA-seq studies have now characterized the population of transcripts found in sperm including human [46], mouse [47], and agricultural livestock species such as bovine [48, 49] and stallion [50]. Moreover, an in-depth study in mature human sperm has shown that the ribosomal RNAs that remain, while abundant, are fragmented [51]. The continually decreasing cost of sequencing has enabled the RNA-seq approach to

become widely adopted and is now a cornerstone of medicine and agricultural research [52, 53] and is becoming the diagnostic standard. Nevertheless, sperm NGS techniques present several intricacies in comparison to somatic or other germinal cells. Hence, experimental design and analyses are crucial for reliable interpretation that yields meaningful results.

Methods, techniques and caveats when considering sperm RNA

The unique characteristics of sperm cells compared to somatic cells necessitate careful consideration as a new study is conceived or secondary analysis of a publicly available dataset is undertaken. The low concentration of sperm RNAs within each cell must be considered along with the technology that has been employed. Human sperm carries approximately 50 fg of long RNA (> 200 nt) and 0.3 fg of sncRNAs (< 200 nt) per sperm cell, which is very similar to early estimates [19]. However, this is dwarfed in comparison to somatic cells which contain 10 pg of long RNA and 1-3 pg of sncRNAs [54-56], and necessitates the use of highly purified samples to eliminate maturing cells, somatic cells, and genomic DNA that can obscure the results. Purification protocols have been developed to separate mature spermatozoa from seminal plasma, as well as immature sperm cells, leukocytes, epithelial cells, and bacteria. Three different approaches have generally been employed. The first, swim-up or sperm migration, in which 0.5 – 1 ml of semen is placed in a 45º angled centrifuge tube under a medium salt solution and incubated at 37ºC for 60 minutes [57]. The sperm cells swim out of the semen to the medium, where they are aspirated using a sterile pipette. This approach selects motile sperm, however, the number of spermatozoa recovered can be low [57]. The second, density gradient centrifugation, uses an isotonic salt solution with saline-coated silica particles, such as PureSperm® (Nidacon), to separate spermatozoa according to density [58]. In most mammalian studies, the starting concentration used is 50% [54]. As with the swim-up method, this approach favors motile and morphologically

normal spermatozoa [58], but the number of spermatozoa recovered is significantly higher by this method [59]. The third method employs a somatic cell lysis buffer that typically contains 0.1% SDS and 0.5% Triton (X-100). This effectively lyses somatic cells, and leaves the sperm head intact [30, 60]. While this method is effective in eliminating somatic cells, this treatment has been proven to compromise the midpiece, and can solubilize sperm-membrane structures, with preferential loss of mitochondrial RNAs [30, 58, 60]. Although a higher recovery rate was observed, RNA yield was significantly lower than density gradient purification [58]. Optical microscopy to visually confirm the lack of immature sperm, somatic or bacterial cells is typically used as an initial screen.

Species specific protocol optimization of sperm RNA extraction has proven useful once RNase free reagents have been confirmed as RNase free. Extraction methods generally include a homogenization and cell disruption using a chaotropic and reducing agent, e.g., TRIzol and TCEP (Tris (2-carboxyethyl) phosphine hydrochloride), typically followed by commercial RNA isolation protocols [58, 61, 62]. A final DNAse treatment step is requisite for eliminating residual genomic DNA. Several strategies to confirm spermatozoal recovery have been employed. For example, following sperm RNA extraction, the absence of intact rRNAs 18S and 28S as assessed by a Bioanalyzer trace would be consistent with their targeted removal during spermatogenesis [51]. The presence of transcripts from *CD45*/*PTPRC* (expressed in most somatic cells) is typically used as a somatic marker. RT-PCR and/or RT-qPCR quantitation can simultaneously assess the presence of genomic DNA when intron spanning primers are used. The ultimate test for sperm RNA purity is by comparative sequence analysis [63].

The construction of high throughput sperm RNA sequencing libraries is confounded by several factors including selective fragmentation of coding

transcripts [46] as well as residual fragmented rRNAs [51] and low total RNA yield [54]. While typical RNA-seq library construction protocols require 1 μg of total RNA, sperm only yields between 30 - 80 ng of total RNA per million cells [54-56]. In most tissue samples, poly(A+) selection has been used to selectively enrich the population of mRNAs as compared to mitochondrial and rRNAs. RNA-seq library construction using poly(A+) selection can result in 3′ bias due to the exclusion of all short polyadenylated tailed transcripts, those lacking this modification, and non-polyadenylated transcripts or the majority of the other segments from fragmented transcripts [55]. A random amplification strategy can, somewhat correct for potential 3′ bias. Ribosomal RNA depletion or primer design can also be used to reduce the number of sequencing reads assigned to this biotype. Different commercial library preparation kits and methods have been tested in various species, including human [64], bovine [48], and porcine (Gòdia M, Quoos Mayer F, Nafissi J, Catelló A, Rodríguez-Gil JE, Sánchez A, and Clop A, unpublished results), resulting in different outcomes with preferred protocols being developed. When comparing two studies, it is important to remember that the data obtained will vary in terms of the library preparation protocol [64]. In addition, quantification of sperm RNAs from sequencing data is dependent on the approach used to analyze the data and this can vary widely. Bioinformatics analysis can follow two different paths. The most commonly used method is "directed analysis". After read mapping, bioinformatic analyses are used to quantify the abundance of each of the annotated genes from the genome studied. This can be refined to quantify RNA levels for each of the annotated transcripts, enabling the detection of alternatively spliced isoforms, but this still remains a challenge when multiple isoforms are simultaneously expressed. Transcript quantitation is measured by FPKM (Fragments Per Kilobase of transcript per Million), when a paired-end sequencing approach is employed or RPKM (Reads Per Kilobase of transcript per Million), when single-end sequenced. These measurements consider the

total number of reads mapped to the transcript after adjusting for the full length of the gene/transcript and applying a library size (sequencing depth) correction. This analytical strategy has been widely used in sperm RNA profiling as it provides a basic way forward to direct transcript comparison with other tissues and/or between species [46-50, 64]. However, as substantial portion of the sperm transcripts are fragmented [46], all exons may not be represented in an equivalent manner and depending on the question posed, this might obscure the results. "Discovery" strategies that utilize a biologically meaningful minimal unit of detection can provide an unbiased alternative. This approach considers sperm RNAs as a collection of sperm RNA elements (REs) [sperm REs: 53]. REs are short sized sequences, formed by a set number of reads and joined as a contig, independent of their annotation. Resultant REs are considered independent of each other and are annotated by genomic location as exonic, intronic, intergenic, close to exonic regions, or novel as existing within an unannotated region (Figure 2). This approach permits a comprehensive examination [53] and addresses low quality genome annotations in some species.

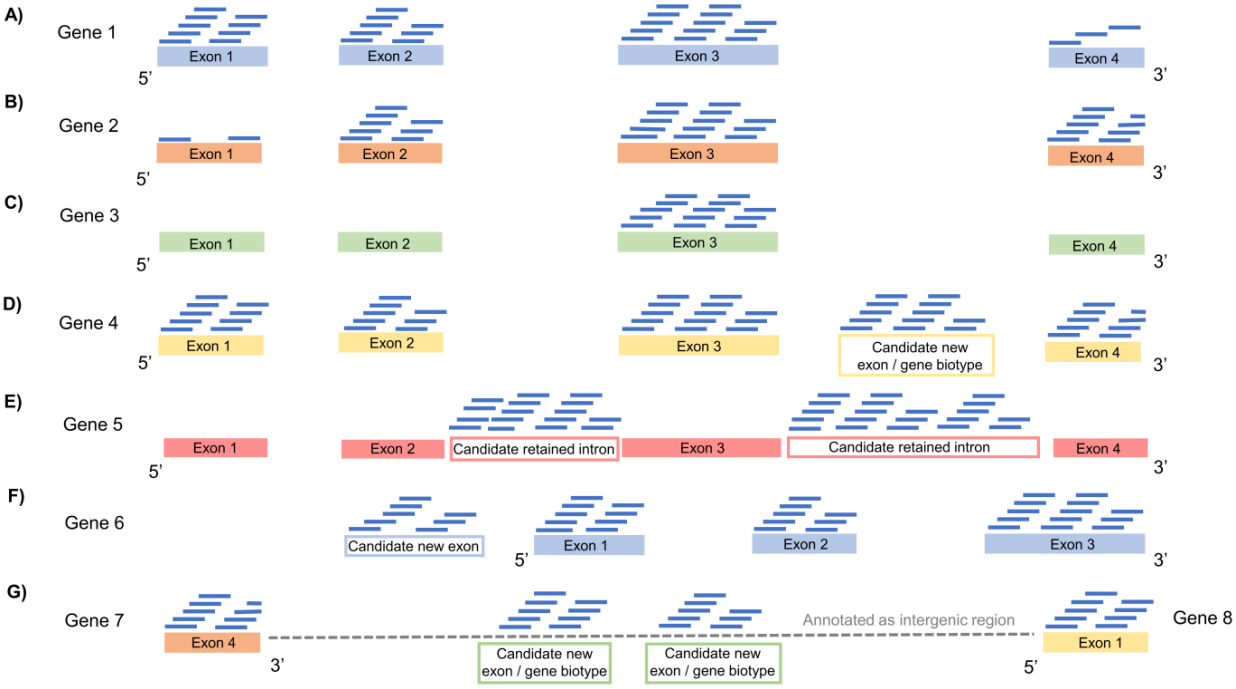


Figure 2. Defining sperm RNA Elements (REs) with the discovery analysis approach. Dark blue lines represent mapped sequencing reads. Dashed gray line corresponds to an annotated intergenic region. The approach also provides a reliable quantitation method of isoforms, where read abundance would vary as a function of transcript isoform. Since mature sperm are transcriptionally and translationally silent, the term "abundance" is preferred rather than "expression." (**A)** Four REs are detected, corresponding to the four annotated exons of Gene 1. Similar read abundance is observed in RE 1, 2, and 3, but lower abundance in RE 4, suggesting possible 3′ degradation of the transcript. (**B)** 5′—end degradation. (**C)** Only one RE is detected, corresponding to the annotated exon 3. (**D)** RE analysis detects all the annotated exons as well as an intronic RE that may correspond to an unannotated exon or other gene biotype. (**E)** Two REs are detected in intronic regions and may correspond to intronic retained elements. (**F)** RE-detected upstream of the 5′ UTR, but could also exist 3′. (**G)** Two REs are detected in an intergenic region, and may correspond to unannotated exons from nearby genes, exons from a new unannotated gene or other gene biotypes.

The population of sperm RNAs

Mature spermatozoal RNAs represent a wide range of coding and non-coding RNAs, including long and short non-coding RNAs. This rich number and classes of RNAs are implicated in different roles, including spermatogenesis, regulation, fertilization, embryo development, and offspring phenotype [reviewed in 53, 55, 65]. Early microarray studies showed that a vast majority of transcripts were shared among healthy sperm donors, enriched in sequences reminiscent of past function [30], suggesting that their distribution among sperm was not stochastic. Human sperm RNA-seq enabled a more complete understanding of the unique features of the sperm transcript population [46].

Over 22,000 transcripts were detected with approximately 700 of which were moderate to high abundance [46]. These RNAs present a complex heterogeneous profile of intact and variously fragmented transcripts [46]. As shown in human [46] and other species including mice [47], horse [50], and bovine [48, 49], a substantial proportion of RNAs are fragmented. Nevertheless, like sncRNAs, they might also function in sperm maturation and the initial events as part of and following fertilization. Integrative analysis between human RNA-seq datasets of testes, mature sperm, and seminal fluid has provided additional insight that highlights the importance of precise transcriptional regulation during spermatogenesis, to maintain the integrity of gamete generation and maturation [66]. The presence or absence of shared RNAs and corresponding proteins between the different tissue types, is indicative of cross-talk between the cellular and extracellular environment from the prostate, epididymis, and seminal vesicles [66].

Long noncoding RNAs

Long noncoding RNAs (lncRNAs) can modulate both transcriptional and posttranscriptional processes [reviewed in 55], while often serving a nuclear structural role [67]. Approximately 1/3 of sperm RNAs are confined within the bounds of the perinuclear theca encapsulated nucleus [52]. The complex packaging of the paternal genome by protamines alongside nucleosome-bound DNA forms a unique chromatin structure. In part, chromatin remodeling has been attributed to lncRNAs that directly complex with chromatin [41] as a function of its affinity to the DNA or DNA-binding proteins. Interaction with target sites through long-range chromatin interactions exemplify a means of regulation often mediated by their 3D organization [reviewed in 68, 69]. Interestingly this structure appears in close association with a repertoire of RNAs [67], positioned by chromosomal regions attached to the nuclear matrix by the Matrix Attachment Regions (MARs) [70].

Several classes of lncRNAs have been observed in human [46, 53] and mouse [71] sperm. This includes members of the RNA U small nuclear family. These components of the sperm spliceosome are intact and very abundant [46], perhaps suggestive of an early post fertilization role. Mouse sperm and testes are rich in lncRNAs of which *Lnc2* and *Lnc3* are the most abundant as evidenced by fluoresce ISH, RT-PCR, and RT-qPCR [71]. Interestingly they appear enriched in sperm when compared with testes [71]. Long noncoding RNA profiles in both testes and sperm can be affected by environmental exposures, e.g., cadmium, which impacts spermatogenesis and male fertility [72]. This is consistent with the view that lncRNAs modulate target genes that play a regulatory role in spermatogenesis, fertility, and embryo development [71-74].

Other types of noncoding RNAs have been described in germinal cells. These include the intronic retained elements and the recently discovered circular RNAs (circRNAs). Intronic-retained elements up to full-length introns have been reported in sperm, with more than 200 distinct REs of this class identified [46]. They appear more abundant in sperm than in testes [reviewed in 55]. Although suggested as retained in mature sperm it has yet to be resolved as to how they evade degradation [reviewed in 55]. Circular RNA are stable 3′ and 5′ covalent linked noncoding RNAs formed from the same transcribed segment that are inherently resistant to exonuclease degradation [75] thus evading degradation. The function of circRNAs is origin dependent. Exonic circRNAs are primarily located in the cytoplasm, thought to act as "sponges," and counteract mRNA repression by miRNAs. Nuclear localized intronic circRNAs and exon-intron circRNAs are thought to regulate their parental genes through direct cis/trans interaction [reviewed in 75, 76]. They have recently been detected in testes and seminal plasma suggestive of a role in gamete generation [77]. Gene Ontology of testes circRNAs shows enrichment for genes related to

spermatogenesis, sperm motility, and fertilization [77]. Their stability and presence in seminal plasma may provide a source of infertility biomarkers. Some of the sperm REs described in previous studies [46, 53, 56] warrant further consideration as they do possess characteristic circRNA signatures.

Small noncoding RNAs

Small non-coding RNAs encompass a variety of different RNA types that play crucial roles in the maintenance and function of the germline genome [reviewed in 42, 55, 78]. Major functions may be classified by RNA type and include regulation of gene expression by miRNAs [44], defense against transposable, repetitive elements or viruses by small interfering RNAs (siRNAs) [79] or piRNAs [80], and protein synthesis and signal modulation by tRNAs and their fragments [5, 45]. Together they can play a role in genome structure and integrity. Characterization of the population of sncRNAs from human and mouse has revealed their diversity, and provided a glimpse into their roles in spermatogenesis, early embryo development, and how they may modify the genome to heritably affect offspring [44, 45].

Micro RNAs

The most well-characterized sncRNAs are miRNAs, contributing to approximately 7% of the total sncRNAs to the fertile human sperm profile [44]. They play a crucial role in spermatogenesis and fertility, and have been reported to modulate expression during the different stages of sperm maturation [reviewed in 81]. Micro RNAs are an integral part of the RNA-induced silencing complex (RISC) and together with the AGO proteins generally target the 3′ UTRs of mRNAs using sequence complementary to repress mRNA expression through degradation or activation [reviewed in 55]. On one hand, miR-140, miR-21, miR-152, and miR-148a have been shown to repress expression of RNAs encoding epigenetic modifiers, e.g., including *DMNT3* and *RASGRP1*. This is consistent with their absence in mature sperm

[46]. On the other hand, *DNMT1* transcripts are present in sperm [46], concordant with their epigenetic role suggesting they escape this form of repression. Specific paternal origin miRNAs such as miR-34c have also been described [44]. Attempts to study its role have led to different conclusions. While in vitro they appear required for the first cleavage following fertilization [82], in vivo, it was not essential for fertilization or embryo development but was crucial for spermatogenesis, as its absence disrupted spermatogenesis leading to murine infertility [83].

Small interfering RNAs

Small interfering RNAs, also known as RNA-mediated interference RNAs, are active mediators of transcriptional and posttranscriptional gene-silencing [reviewed in 84]. As with miRNAs, siRNAs act in concert with RISC and AGO proteins to target complementary RNAs for translational inhibition or degradation [reviewed in 84]. Small interfering RNAs are primarily known to function in the host defense against transposable elements (TE) and RNA viruses, and aid in the maintenance of heterochromatic DNA [reviewed in 84]. As shown in plants, somatic cell TE-derived siRNAs migrate into sperm cells contributing to TE silencing prior to fertilization [reviewed in 85]. Preliminary work on the role of siRNAs in fertilization using Dicer knockout mice presented an altered profile of siRNAs in spermatozoa (and miRNAs), where sperm microinjection in wild-type oocytes resulted in embryos with reduced developmental potential [79]. Other independent studies have attempted to discern the role between siRNAs and miRNAs in spermatogenesis. While miRNAs require DICER and DGCR8 proteins for maturation, siRNAs only require DICER [reviewed in 86]. Results have shown that a conditional *Dicer* knockout presents severe sperm morphological defects as compared to the conditional *Dgcr8* mice mutant, implicating siRNAs in mammalian spermatogenesis [86, 87].

piRNAs and Transposable Elements

Piwi-interacting RNAs are specialized RNAs that interact with Piwi proteins, a gonad type of AGO proteins, which mediate RNA silencing of TEs [reviewed in 88, 89]. Murine Piwi proteins MIWI, MILI, and MIWI2 are essential to spermatogenesis, and their absence is associated with male infertility [reviewed in 78, 88]. In humans, ~17% of the sncRNAs correspond to piRNAs [44]. Their presence still remains controversial, although piRNAs are now beginning to be considered by others [90]. Altered levels of PIWI-like 1 and 2 mRNAs have now been detected in men with decreased sperm count and motility by RT-qPCR [91]. The PIWI-like 1 RNA remains essentially intact in mature spermatozoa as determined by RNA-Seq [53, data not shown]. Given these independent observations, their presence will likely become accepted.

In murine spermatocytes and spermatids, the vast majority of piRNAs map to specific genomic regions [reviewed in 78], and ~17% map to repeat sequences such as DNA transposons, short interspersed nuclear elements (SINEs), and long interspersed nuclear elements (LINEs) [reviewed in 89]. This is not surprising since TEs constitute a large fraction of the eukaryotic genome. SINES, LINES, and long terminal repeats (LTRs) are the most abundant [44]. LINE1 is the most common and well-studied retrotransposon in germ cells [reviewed in 88]. Piwi-interacting RNA biogenesis is regulated by the PARN family proteins [92]. The role of one of its members, PNLDC1 has been examined in vivo [93]. In the corresponding mouse knockout model, LINE1 retrotransposon silencing was disrupted leading to aberrant piRNA biogenesis and spermatogenesis [93]. Studies in mutant Piwi proteins have shed light on their roles in spermatogenesis. Miwi2 knockout mice showed spermatogenic arrest, increased LINE1 retrotransposon expression, and loss of methylation in testes, suggesting possible regulatory roles of Piwi proteins and piRNAs in methylation [80]. Moreover, Mili mice knockout mutants exhibited loss of DNA

methylation of the LINE1 element and an increased expression of both *LINE1* and *IAP* [94]. Interestingly, the piRNAs produced from lncRNAs have the capability to act as histone modifiers. For example, the piRNA sno75, derived from the lncRNA *GAS5*, has been shown to increase transcription of *TRAIL* by guiding H3K4 methylation and H3K27 demethylation [95]. Transposable elements can be regulated by DNA methylation [reviewed in 96], which may also be guided and modulated by RNAs [reviewed in 68]. This DNA methylation is maintained by DNMT1 [reviewed in 96], whose transcript is present in sperm [44]. It has also been suggested that piRNAs play a central role in the confrontation and consolidation when the sperm and oocyte meet to initially assess genetic compatibility through their interaction with repetitive elements [reviewed in 55] (see "Contributions from sperm to the oocyte and early development").

Transfer RNAs

Transfer RNAs and their fragments are some of the more abundant sperm sncRNAs [reviewed in 97]. Their abundance is directly linked to the general translational needs of a given cell type rather than to a specific gene, reflective of their metabolic state. The original description of tRNAs in mouse sperm showed an enrichment of 5′ end of tRNAs [45]. While generally scarce in testis, their abundance increases as the maturing sperm exits from testis as it passes through the caput, corpus, and cauda [8]. This has supported the view that the majority of tRNA fragments in mature spermatozoa arise from trafficking via epididymosomes from the epididymis to the sperm [8], with some upregulated in mice fed a high-fat diet [5].

In addition to fragmentation that yields 5′ enrichment of mouse sperm tRNA fragments [45], modifications have also been reported [5]. These include the incorporation of 5-methylcytidine (m5C) in response to high- or low-fat diets [5], which may act to increase their stability. This chemical post-transcriptional

modification may provide a mechanism to signal current metabolic state reflecting a change in environment [97]. It has also been argued that tRNA fragments may act to repress embryonic Mouse Endogenous RetroeLements (MERVLs). This is particularly attractive since some are in close proximity to genes expressed during the early stages of embryo development. It has thus been proposed that they regulate a significant proportion of the transcriptome in development [reviewed in 98], perhaps affecting placental size by altering metabolic pathways [8].

Contributions from sperm to the oocyte and early development

Upon fertilization, the spermatozoon and oocyte confront one another for the first time towards consolidation [2, 99]. During this initial event when the ooplasm is exposed to the paternal contribution, it is also potentially exposed to a host of retroviruses and retrotransposons that have the potential to alter the zygotic genome [99]. They can insert within active regions of the genome as it is rearranged and/or affect the expression of neighboring genes by functioning as an enhancer, altering splicing or polyadenylation [100]. LINE1 transcription affects embryo chromatin accessibility in mice, and may play a role in coordinating the activity of multiple genes throughout the genome adding to genomic stability for the embryo [98]. Yet, in contrast to MERVLs, sperm LINE1 elements do not contain genes implicated in embryonic genome activation (EGA) [98]. By a mechanism of RNA activity through maternal and paternal sncRNAs, such as miRNAs and piRNAs [2, 44, 55], parasite RNAs are degraded, leading to a consolidated state of genetic compatibility [99]. Embryo development goes forward if these checkpoints are passed. While the potential contribution of repetitive elements in early development remains unclear, embryonic arrest at the 2- or 4-cell stage in mice coincides with the disruption of LINE1-encoded reverse transcriptase [reviewed in 101].

In humans, the transition to an embryo requires approximately 3 days and comprises the migration and fusion of the germ cell pronuclei and several cleavage divisions [reviewed in 102]. The human embryo is relatively transcriptionally silent until day 3. Having reached the 4–8-cell stage, the major wave of human EGA occurs [reviewed in 102, 103]. Following EGA, the embryo continues development, including implantation of the blastocyst into the uterine wall at day 7 [reviewed in 102]. Although humans and mice morphologically share embryo similarities, key differences necessitate that cross-species extrapolations be considered with caution. In mouse embryos, EGA occurs 26–29 h postfertilization at or around the first cellular division so the transcriptionally silent period is substantially reduced [reviewed in 102]. Similarly, the timing of compaction is shorter in mice, and human embryos hold at least one extra round of cell division before implantation [reviewed in 102]. It is not surprising that marked differences between gene expression, genome instability, and epigenetic modifications between human and mouse embryos are apparent [reviewed in 102]. While the quantity of paternal RNAs delivered by sperm to the oocyte may seem small compared to the oocyte, it is sufficient to play a role in transgenerational inheritance altering offspring phenotype [99, 104].

The role of sperm RNAs in EGA in mice has recently been examined in silico providing key elements for testing [105]. This provided an RNA-seq survey of sperm, MII eggs, and zygotes, supporting a potential role of paternal RNA or proteins to interact with maternal cofactors contributing to EGA [105]. A sperm RNA corresponding to an intragenic LTR of *Hdac11*, present in sperm but absent in MII eggs and zygotes, suggested that the paternally derived *Hdac11* LTR or others might complement maternal cofactors or pathways and participate together in EGA [105], not unlike the well-known oocyte activator factor PLCZ [106].

The first days of embryo development are critical as specification begins. Perturbation may resolve as the habitual first trimester pregnancy loss, which occurs in 15–25% of pregnancies [107]. Although most of the efforts to understand possible causes have focused on the mother reflecting the direct relationship with the fetus, the possible role of the male contribution has only begun to be discussed [108]. The factors from which sperm may contribute to miscarriage have been the focus of various studies. The influence of DNA fragmentation, aneuploidy, and integrity, as well as sperm morphology have been considered to be correlated with Recurrent Miscarriage (RM), characterized as two or more consecutive failed pregnancies [107], yet, results remain controversial and inconsistent [109–112]. A recent study has focused on the possible role of sperm RNAs in RM etiology [108]. The ratio of protamine RNAs (*PRM1*/*PRM2*) was significantly different between spermatozoa when RM couples were compared to healthy donors suggesting that abnormal protamine packaging can negatively affect embryo development [108]. Whether this reflects histone retention and/or abnormal epigenetic marks [108, 113] remains to be resolved. Interestingly, alterations in the protamine transcript ratio have also been associated with reduced sperm quality, including low concentration, reduced motility, abnormalities, and increased DNA fragmentation [114].

RNAs and transgenerational epigenetic inheritance

The study of Rassoulzadegan and colleagues was the first report providing evidence of mammalian RNA-mediated epigenetic inheritance. Here, it was observed that wild type mouse offspring of *Kit* mutant parents, exhibited white patches characteristic of the *Kit* mutant phenotype coupled with altered RNA levels of *Kit* in sperm [39]. In the testis, a relatively dispersed population of RNAs was observed, including the identification of abnormal short mRNAs, derived from the wild type allele responsible for the *Kit* mutant phenotype.

These abundant RNAs were only found in the mature sperm of heterozygotic offspring [39], suggesting that the increase was responsible for the disruption of wild type phenotype. To test this hypothesis, two miRNAs thought to target *Kit* mRNA, miR-221 and miR-222, were injected into one-cell embryos. The resultant mice exhibited the same *Kit* mutant phenotype, leading to the conclusion that the presence of certain miRNAs early in embryogenesis will result in a stable, heritable change in gene expression and correspondingly, phenotype [39]. This study also highlighted miRNAs as possible modifiers of the epigenome [39], that perhaps directed in some manner DNA methylation, the most widely recognized mechanism of epigenetic heritability [reviewed in 40].

Several sncRNAs can regulate DNA methylation and histone modifications [reviewed in 68] and, as above miRNAs, are known to influence DNA methyltransferases. CARM1 is an embryonic stem cell pluripotency factor involved in the H3 promoter methylation of two transcription factors, POU5F1 and SOX2, providing an active chromatin mark upon induction [reviewed in 55]. miR-181c is known to target *CARM1*, and its immature form, pri-miR-181c, is abundant in human sperm [46]. Upon delivery to the oocyte, pri-miR-181c is processed to its mature form that is coupled with a 70% decrease in these and 27 other *CARM1*-associated target genes by the morula stage [46]. Recent results from a paternal chronic stress study identified that zygotic microinjection of nine abundant miRNAs resulted in targeted degradation of maternal mRNA transcripts, including genes involved in chromatin remodeling [7]. These results infer a role for sperm miRNAs in ensuring or modifying cell specification.

Both multigenerational and transgenerational epigenetic inheritance imposed upon future generations have important ramifications for medicine and agriculture. The concept of fetal origins of adult health and disease [9] has given rise to numerous studies that illustrate the ability for the effects of an

exacerbating input environment (allostatic overload [10]) in the father, to be inherited by later generations. While, male mediated multigenerational inheritance only passes from the affected father to child in a transient manner, transgenerational inheritance extends, affecting the father's grandchildren and/or beyond. Unlike multigenerational transmission, transgenerational inheritance can represent a permanent change to the gene pool [115]. Mechanistically, these responses to environmental factors, including stress, toxins, and diet may initially be mediated by epigenetic effectors, including RNAs. The ability of nucleic acids, including sncRNAs, to distribute throughout the body encapsulated in vesicles, provides the opportunity to influence sperm [reviewed in 115] that is normally protected from external effects by the blood-testes barrier. Data has also suggested that reverse transcription, by the LINE1 retrotransposon reverse transcriptase, might be a route for passing non-Mendelian traits to the offspring [101, 116]. The transmission of information from somatic to germ cells by exogenous RNA was suggested from a series of experiments in which mice xenografted with human A-375 melanoma cells stably expressed EGFP that was then distributed through the bloodstream. Interestingly, transcripts were also detected in mouse spermatozoa, with exosomes as the suggested mechanism of transport [117]. In this paradigm, spermatozoa would take up exogenous DNA or transcribe exogenous RNA for delivery to the oocyte at fertilization [94]. This sperm mediated reverse transcriptase may act to expand the exogenous DNA or RNAs once in the sperm [116] providing a means to directly impact early embryo development.

Several compelling studies have exhibited the capacity for the environment experienced by the father to influence the offspring without the offspring ever being in contact with the environment [reviewed in 40, 42]. For example, both chronic and acute stress models have been used to explore how the paternal environment affects progeny. For example, chronic stress mouse models have

identified increased levels of a group of sperm miRNAs [7] that act to specifically mark the response. Offspring developed a depressed hypothalamic-pituitary-adrenal (HPA) response in the absence of their father or shared paternal environment. This depressed HPA response was modulated by altered expression in the paraventricular nucleus of the hypothalamus [7]. Microinjection of the nine miRNAs into the wild-type oocyte resulted in zygotes expressing the same altered pattern of gene expression [7]. An acute stress mouse model resulted in depressive-like behaviors, exhibiting altered levels of both sperm miRNAs and piRNAs compared to controls [6]. Male offspring presented a similar depressive-like behavior as the father, in addition to altered levels of miRNAs within different brain regions and in serum. Interestingly, male offspring also exhibited altered glucose and insulin metabolic traits [6]. These results suggested that behavioral phenotypes could be transmitted down the male line, in the absence of paternal rearing, using a RNA mechanism [6, 7].

The impact of paternal diet has also drawn attention. For example, male mice subjected to a high-fat-high-sugar diet presented up-regulated levels of miR-19b [13]. When miR-19b was microinjected into zygotes, offspring presented both obesity and reduced glucose tolerance, which persisted through several generations [13]. A low-protein diet in mice has been found to alter levels of sperm miRNAs, piRNAs, and several tRNAs [8]. In mice, alterations in miRNA levels as well as 5-methylcytidine (m5C) and N2-methylguanosine (m2G) modified tRNAs in sperm have also been observed in response to a high-fat diet [5]. Microinjection of total sperm RNA from high-fat diet mice at fertilization resulted in offspring exhibiting impaired glucose tolerance, although insulin sensitivity was not observed [5]. This inferred that while modified sperm miRNA and tRNA may result in inherited metabolic traits, other layers of regulation may be involved in the inheritance of metabolic diseases such as

diabetes. In humans, epidemiological studies on male-line transmission as exemplified in those of a Swedish Överkalix cohort [15, 118] span at least three generations. The impact of a grandfathers access to excess food during his childhood correlated with a higher incidence of both diabetes related mortality and mortality risk ratios in his grandsons [15, 118], while fathers with low food access as a child correlated with a decreased mortality by cardiovascular disease in their sons [118]. These results suggested a role of paternal influence in transgenerational responses in humans. As measured by BMI, male obesity has been shown to alter the human sperm methylome, with its reversal following bariatric surgery [119]. While alterations of sperm sncRNAs were also identified [119], their association was not well defined. Nevertheless, these results identify potential alterations in epigenetic effectors that may be transferred to the oocyte at fertilization, to influence embryo development and potentially be inherited by offspring.

Challenges and questions that remain

The low viability of parthenogenic mice [120] from two maternal genomes, sheds light on the importance of sperm beyond providing genomic information for successful zygote and embryo formation. The role of sperm RNAs is indicative of the intricate regulatory mechanisms of spermatogenesis, fertilization, and embryogenesis. Different classes of sperm RNAs have been found to act with different modus operandi to maintain the integrity of the genome, regulate gene expression, and chromatin state (Figure 3).

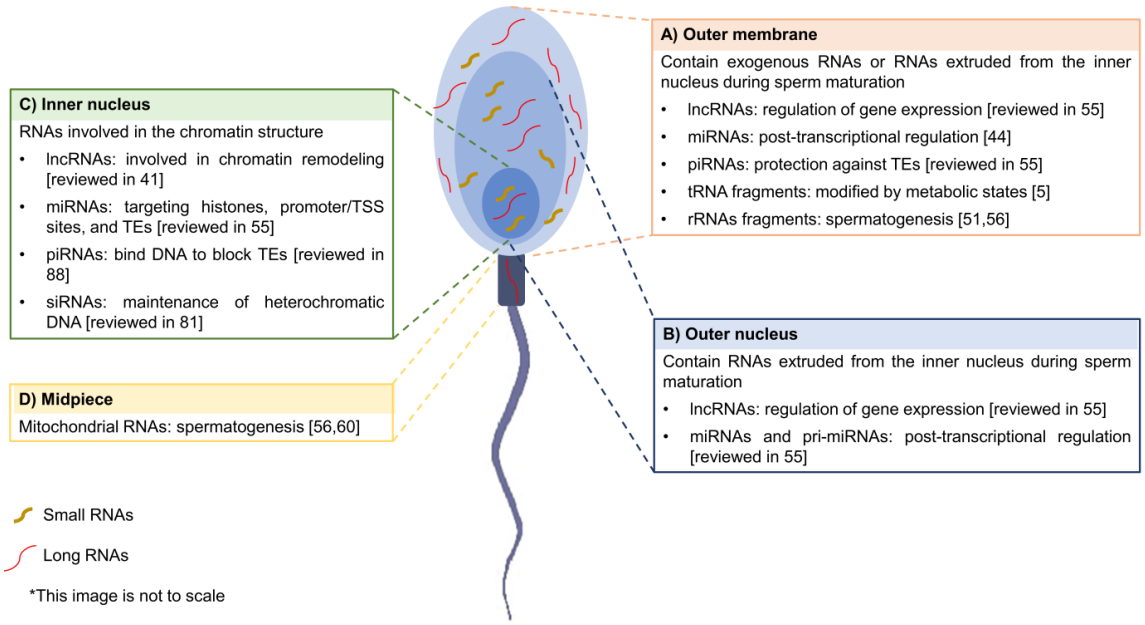


Figure 3. Distribution of sperm RNAs and their roles. (**A–B)** The outer nuclear layers indicated in light and medium blue contain ~2/3 of all sperm RNAs (including coding and noncoding), the majority of which are long RNAs (>200 nt). These RNAs include exogenous RNAs packaged within vesicles and RNAs extruded from the inner nucleus either bound to associated proteins or within vesicles during cytoplasmic extrusion. The remaining ~1/3 of total sperm RNAs reside within the dark blue inner nuclear fraction. (**C)** RNAs found within the inner nuclear fraction are likely directly associated with the DNA or chromatin bound proteins to influence chromatin structure and may provide an epigenetic signature. Examples include the chromatin associated RNAs, such as lncRNAs, piRNAs, miRNAs, and siRNAs. (**D)** Mitochondrial RNAs have been observed in the midpiece of the spermatozoa. lncRNAs (long noncoding RNAs); miRNAs (microRNAs); piRNAs (Piwi-interacting RNAs); pri-miRNAs (primary miRNAs); siRNAs (small interfering RNAs); TEs (transposable elements); tRNAs (transfer RNAs); TSS (transcript start site); rRNAs (ribosomal RNAs).

Currently, 10 - 15% of reproductive aged couples are affected by infertility [reviewed in 55]. Existing diagnostics rely on observable semen parameters [reviewed in 65] leaving many cases of male infertility unexplained. Differential RNA profiles exhibited by infertile men have been identified [reviewed in 4, 53, 55], and altered presence and/or abundance of RNAs are being used as molecular biomarkers to address reproductive health concerns. This includes the 648 sperm REs that appear essential for natural conception [53]. Circular RNAs which can be stable at room temperature for several hours [reviewed in 75] may offer yet another avenue. Factors such as advanced paternal age and increased BMI have been amplified in recent years [reviewed in 121, 122], and are increasingly linked to reproductive success. Standard semen parameters decline with advanced age [123] and increased BMI [124]. Similarly, both

paternal age and BMI impact the potential for successful live birth [125, 126]. It is likely that these factors, among others, correlate with reproductive potential through the alteration of sperm RNAs. While diet modulates RNA abundance in sperm [5], these alterations have not been assessed in terms of reproductive potential, although work on this subject should be forthcoming.

Sperm RNAs used as biomarkers are also being developed in livestock, both as a model system and for enhanced breeding. In agriculture, identification of genetic markers of sperm quality, fertility, and those minimizing frozen-thawed cryo-damage is presently being pursued to optimize animal selection [50, 127-129]. Bovine is a well-studied agricultural species due to its high economic impact. The corresponding sperm microarray studies have focused on sperm quality [127], fertility [130], and cryo-damage [131], while NGS technology is beginning to be applied to sperm motility [52]. Other agricultural species including stallion [50], porcine [128], and chicken [129] are still focused on the fundamental profiling of sperm RNAs. As the cost of sequencing decreases their potential impact for agriculture and novel studies continues to grow.

Research focusing on the concept of health maintenance and lifestyle continues to move forward. Given the comparatively long-generation times between humans as compared to mice [132], the majority of human studies addressing epigenetic inheritance are epidemiologically based [15, 118]. Their significance is highlighted by the observation that access to food during a males' slow growth period correlates with the risk of diabetes and cardiovascular-related mortality, as well as over-all mortality risk, in sons and grandsons. The age of onset of smoking of a father has also been suggested to be correlated with the son's BMI [15]. These studies while limited have identified that responses to the environment appear to reflect sex-specific transfer between generations. One must also consider that unlike females, males experience continual gamete (sperm) renewal, which can act to "wash out", or re-set the effect of a given

exposure (e.g., diet, exercise, stress level). This capability affords males the unique possibility to marginalize the transfer of a potentially harmful effect to their children [reviewed in 121]. In this model, alterations in sperm methylation or sncRNAs resultant from a lifestyle choice like diet may be at least partially restored by a simple change in caloric intake.

The relationship between the father's diet and smoking status during his prepubescent, slow growth period, on the health of his children [15, 118] suggests there are critical times of exposure. Epidemiological studies in human suggest that a critical time of exposure for males occurs prior to the onset of puberty, between 9 and 12 years of age [118]. Mouse models and human studies [reviewed in 133] have highlighted the impact of in utero environment as a time of relative phenotypic plasticity [reviewed in 134], and suggest this as a critical time of exposure for male-line epigenetic inheritance [134, 135].

The focus of most studies surrounds a single system input, but this does not reflect the intricate network of inputs that constitute allostatic load. For example, the common functional measure of obesity, BMI, will be influenced by multiple factors, including diet and exercise. While animal models provide a means to better control for these cofounding factors [132], this is rarely possible in human studies. A limited number of studies have begun to examine the relationship between age and diet with respect to this paradigm. On one hand, mouse model studies have identified that undernourishment of males in utero can lead to inheritance of altered DNA methylation via sperm [134], yet this effect was not identified if the period of undernourishment followed birth [135]. On the other hand, human epidemiological studies [15, 118] noted that the age at which a male experiences malnutrition impacts the health of multiple generations. This may be in part due to age of exposure impacting the ability for the alteration to be "washed out".

The fertilizing spermatozoon is essentially a dynamic single cell system that contains half of the information for the next generation in which a single cell is selected in some manner from millions. The field continues to seek to understand how specific paternal components and their respective mechanisms influence reproductive health. The potential importance of sperm RNAs in the maintenance and transference of biological information between generations underscores the necessity of achieving an understanding. Perhaps this provides a mechanism to assess specific fitness of a new trait prior to adaptive selection, and in part, provide one of the components of genetic resilience [11, 136]. For now, we are hindered by the lack of a comprehensive understanding of how these epigenetic effectors are modulated.

Four fundamental questions remain with respect to paternal RNAs: (1) Do sperm RNAs directly interact with other epigenetic effectors, perhaps directly with the genome (chromatin) or other RNAs or structures? (2) What is the role of RNAs in transmitting specific phenotypes? (3) What is the corresponding mechanism by which this is achieved? (4) How may we modify the future-past in a controlled manner? Addressing each of these questions in this unique human single-cell system will require the development of new molecular and computational tools. The answers will likely revolutionize our understanding of reproduction and health, as the totality of the male contribution reflective of his past experiences that impacts the future birth and life course of his child, becomes appreciated.

Acknowledgments

The authors would like to thank Dr. Àlex Clop and Dr. Meritxell Jodar for their critical review and comments. We apologize to those authors whom we were unable to cite given the constraints of the manuscript.

Grant support:

M.G. was funded with a FPI Ph.D. grant from the Spanish Ministry of Education (BES-2014-070560) and a Short-Stay fellowship from the Spanish Ministry of Education (EEBB-I-2017-12229) at S. A. K. laboratory. G.S. is supported in part by a Postdoctoral Recruiting Fellowship from Wayne State University. Support from the Charlotte B. Failing Fetal Therapy and Diagnosis endowed Professorship (along with the EMD Serono 2016 Grant for Fertility Innovation) to S.A.K. is gratefully acknowledge.

Abbreviations

ACTB Actin beta
AGO Argonaute
BMI Body mass index
CARM1 Histone-arginine methyltransferase CARM1
CD45 Leukocyte common antigen
circRNA circular RNA
DGCR8 DiGeorge Syndrome Critical Region Gene 8
Dicer Endoribonuclease Dicer
DNMT1 DNA methyltransferase 1
DMNT3 DNA methyltransferase 3
E2R Ubiquitin Conjugating Enzyme E2 R2
EGA Embryonic genome activation
EGFP Enhanced green fluorescent protein
FISH Fluoresce in-situ hybridization
FPKM Fragments Per Kilobase of transcript per Million
GAS5 Growth arrest specific 5
H3K27 Histone 3 lysine 27
H3K4 Histone 3 lysine 4
Hdac11 Histone deacetylase 11
HPA Hypothalamic-pituitary-adrenal
IAP Intracisternal A Protein
ISH In Situ Hybridization
ITGB Integrin subunit beta 1
Kit KIT proto-oncogene receptor tyrosine kinase
LGC1 Low glutelin content 1
LINE Long interspersed nuclear element
LINE1 Long interspersed element 1
Lnc2 lncRNA transcript 2
Lnc3 lncRNA transcript 3

lncRNA long non-coding RNA
LTR Long terminal repeat
MII egg Meiosis II egg
m5C 5-methylcytidine
m2G N2-methylguanosine
MAR Matrix attachment region
MERVL Mouse Endogenous retroeLement
MILI Piwi like RNA-mediated gene silencing 2
MIWI Piwi like RNA-mediated gene silencing 1
MIWI2 Piwi like RNA-mediated gene silencing 4
miRNA microRNA
mRNA messenger RNA
MYC MYC Proto-Oncogene, BHLH Transcription Factor
NGS Next Generation Sequencing
PARN poly(A)-specific ribonuclease
piRNA Piwi-interacting RNAs
Piwi P element-induced wimpy testes
PLCZ Phospholipase C zeta
PNLDC1 PARN like, ribonuclease domain containing 1
POU5F1 POU class 5 homeobox 1
PRM1 Protamine 1
PRM2 Protamine 2
PTPRC Protein tyrosine phosphatase, receptor type, C
RASGRP1 RAS guanyl releasing protein 1
RISC RNA-induced silencing complex
RM Recurrent miscarriage
RNU RNA U small nuclear
RPKM Reads Per Kilobase of transcript per Million
rRNA ribosomal RNA
SAGE Serial analysis of gene expression
SINE Short interspersed nuclear element
siRNA small interfering RNA
sncRNA small non-coding RNA
sno75 piwi RNA 75
SOX2 Sex-determining region Y box 2
RE RNA element
TCEP Tris (2-carboxyethyl) phosphine hydrochloride
TE Transposable element
TNP2 Nuclear transition protein 2
tRNA transfer RNA

TRAIL Tumor necrosis factor (TNF) superfamily member 10
3′ UTR 3′ untranslated region

References

1. Ostermeier GC, Miller D, Huntriss JD, Diamond MP, Krawetz SA. Reproductive biology: delivering spermatozoan RNA to the oocyte. Nature 2004; 429:154.
2. Krawetz SA. Paternal contribution: new insights and future challenges. Nat Rev Genet 2005; 6:633-642.
3. Garrido N, Martinez-Conejero JA, Jauregui J, Horcajadas JA, Simon C, Remohi J, Meseguer M. Microarray analysis in sperm from fertile and infertile men without basic sperm analysis abnormalities reveals a significantly different transcriptome. Fertil Steril 2009; 91:1307-1310.
4. Platts AE, Dix DJ, Chemes HE, Thompson KE, Goodrich R, Rockett JC, Rawe VY, Quintana S, Diamond MP, Strader LF, Krawetz SA. Success and failure in human spermatogenesis as revealed by teratozoospermic RNAs. Hum Mol Genet 2007; 16:763-773.
5. Chen Q, Yan M, Cao Z, Li X, Zhang Y, Shi J, Feng GH, Peng H, Zhang X, Zhang Y, Qian J, Duan E, et al. Sperm tsRNAs contribute to intergenerational inheritance of an acquired metabolic disorder. Science 2016; 351:397-400.
6. Gapp K, Jawaid A, Sarkies P, Bohacek J, Pelczar P, Prados J, Farinelli L, Miska E, Mansuy IM. Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. Nat Neurosci 2014; 17:667-669.
7. Rodgers AB, Morgan CP, Leu NA, Bale TL. Transgenerational epigenetic programming via sperm microRNA recapitulates effects of paternal stress. Proc Natl Acad Sci U S A 2015; 112:13699-13704.
8. Sharma U, Conine CC, Shea JM, Boskovic A, Derr AG, Bing XY, Belleannee C, Kucukural A, Serra RW, Sun F, Song L, Carone BR, et al. Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. Science 2016; 351:391-396.
9. Skogen JC, Overland S. The fetal origins of adult disease: a narrative review of the epidemiological literature. JRSM Short Rep 2012; 3:59.
10. McEwen BS, Wingfield JC. What is in a name? Integrating homeostasis, allostasis and stress. Horm Behav 2010; 57:105-111.
11. Lundberg U. Stress hormones in health and illness: the roles of work and gender. Psychoneuroendocrinology 2005; 30:1017-1021.
12. McDade TW. Life history, maintenance, and the early origins of immune function. Am J Hum Biol 2005; 17:81-94.

13. Grandjean V, Fourre S, De Abreu DA, Derieppe MA, Remy JJ, Rassoulzadegan M. RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders. Sci Rep 2015; 5:18193.
14. de Castro Barbosa T, Ingerslev LR, Alm PS, Versteyhe S, Massart J, Rasmussen M, Donkin I, Sjogren R, Mudry JM, Vetterli L, Gupta S, Krook A, et al. High-fat diet reprograms the epigenome of rat spermatozoa and transgenerationally affects metabolism of the offspring. Mol Metab 2016; 5:184-197.
15. Pembrey ME, Bygren LO, Kaati G, Edvinsson S, Northstone K, Sjostrom M, Golding J, Team AS. Sex-specific, male-line transgenerational responses in humans. Eur J Hum Genet 2006; 14:159-166.
16. Kierszenbaum AL, Tres LL. Structural and transcriptional features of the mouse spermatid genome. J Cell Biol 1975; 65:258-270.
17. Betlach CJ, Erickson RP. A unique RNA species from maturing mouse spermatozoa. Nature 1973; 242:114-115.
18. Rejon E, Bajon C, Blaize A, Robert D. RNA in the nucleus of a motile plant spermatozoid: characterization by enzyme-gold cytochemistry and in situ hybridization. Mol Reprod Dev 1988; 1:49-56.
19. Pessot CA, Brito M, Figueroa J, Concha, II, Yanez A, Burzio LO. Presence of RNA in the sperm nucleus. Biochem Biophys Res Commun 1989; 158:272-278.
20. Kumar G, Patel D, Naz RK. c-MYC mRNA is present in human sperm cells. Cell Mol Biol Res 1993; 39:111-117.
21. Miller D, Tang PZ, Skinner C, Lilford R. Differential RNA fingerprinting as a tool in the analysis of spermatozoal gene expression. Hum Reprod 1994; 9:864-869.
22. Rohwedder A, Liedigk O, Schaller J, Glander HJ, Werchau H. Detection of mRNA transcripts of beta 1 integrins in ejaculated human spermatozoa by nested reverse transcription-polymerase chain reaction. Mol Hum Reprod 1996; 2:499-505.
23. Wykes SM, Visscher DW, Krawetz SA. Haploid transcripts persist in mature human spermatozoa. Mol Hum Reprod 1997; 3:15-19.
24. Wykes SM, Miller D, Krawetz SA. Mammalian spermatozoal mRNAs: tools for the functional analysis of male gametes. J Submicrosc Cytol Pathol 2000; 32:77-81.
25. Xu HL, Swoboda I, Bhalla PL, Singh MB. Male gametic cell-specific gene expression in flowering plants. Proceedings of the National Academy of Sciences of the United States of America 1999; 96:2554-2558.
26. Miller D, Briggs D, Snowden H, Hamlington J, Rollinson S, Lilford R, Krawetz SA. A complex population of RNAs exists in human ejaculate

spermatozoa: implications for understanding molecular aspects of spermiogenesis. Gene 1999; 237:385-392.
27. Grunewald S, Paasch U, Glander HJ, Anderegg U. Mature human spermatozoa do not transcribe novel RNA. Andrologia 2005; 37:69-71.
28. Gur Y, Breitbart H. Mammalian sperm translate nuclear-encoded proteins by mitochondrial-type ribosomes. Genes Dev 2006; 20:411-416.
29. Gur Y, Breitbart H. Protein synthesis in sperm: dialog between mitochondria and cytoplasm. Mol Cell Endocrinol 2008; 282:45-55.
30. Ostermeier GC, Dix DJ, Miller D, Khatri P, Krawetz SA. Spermatozoal RNA profiles of normal fertile men. Lancet 2002; 360:772-777.
31. Engel ML, Chaboud A, Dumas C, McCormick S. Sperm cells of Zea mays have a complex complement of mRNAs. Plant J 2003; 34:697-707.
32. Martins RP, Krawetz SA. Towards understanding the epigenetics of transcription by chromatin structure and the nuclear matrix. Gene Ther Mol Biol 2005; 9:229-246.
33. Ostermeier GC, Goodrich RJ, Moldenhauer JS, Diamond MP, Krawetz SA. A suite of novel human spermatozoal RNAs. J Androl 2005; 26:70-74.
34. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995; 270:484-487.
35. Zhao Y, Li Q, Yao C, Wang Z, Zhou Y, Wang Y, Liu L, Wang Y, Wang L, Qiao Z. Characterization and quantification of mRNA transcripts in ejaculated spermatozoa of fertile men by serial analysis of gene expression. Hum Reprod 2006; 21:1583-1590.
36. Wang H, Zhou Z, Xu M, Li J, Xiao J, Xu ZY, Sha J. A spermatogenesis-related gene expression profile in human spermatozoa and its potential clinical applications. J Mol Med (Berl) 2004; 82:317-324.
37. Amanai M, Brahmajosyula M, Perry AC. A restricted role for sperm-borne microRNAs in mammalian fertilization. Biol Reprod 2006; 75:877-884.
38. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 2006; 442:199-202.
39. Rassoulzadegan M, Grandjean V, Gounon P, Vincent S, Gillot I, Cuzin F. RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse. Nature 2006; 441:469-474.
40. Rando OJ. Daddy issues: paternal effects on phenotype. Cell 2012; 151:702-708.
41. Larriba E, del Mazo J. Role of Non-Coding RNAs in the Transgenerational Epigenetic Transmission of the Effects of Reprotoxicants. Int J Mol Sci 2016; 17:452.
42. Gapp K, Bohacek J. Epigenetic germline inheritance in mammals: looking to the past to understand the future. Genes Brain Behav 2017.

43. Anway MD, Cupp AS, Uzumcu M, Skinner MK. Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science 2005; 308:1466-1469.
44. Krawetz SA, Kruger A, Lalancette C, Tagett R, Anton E, Draghici S, Diamond MP. A survey of small RNAs in human sperm. Hum Reprod 2011; 26:3401-3412.
45. Peng H, Shi J, Zhang Y, Zhang H, Liao S, Li W, Lei L, Han C, Ning L, Cao Y, Zhou Q, Chen Q, et al. A novel class of tRNA-derived small RNAs extremely enriched in mature mouse sperm. Cell Res 2012; 22:1609-1612.
46. Sendler E, Johnson GD, Mao S, Goodrich RJ, Diamond MP, Hauser R, Krawetz SA. Stability, delivery and functions of human sperm RNAs at fertilization. Nucleic Acids Res 2013; 41:4104-4117.
47. Margolin G, Khil PP, Kim J, Bellani MA, Camerini-Otero RD. Integrated transcriptome analysis of mouse spermatogenesis. BMC Genomics 2014; 15:39.
48. Selvaraju S, Parthipan S, Somashekar L, Kolte AP, Krishnan Binsila B, Arangasamy A, Ravindra JP. Occurrence and functional significance of the transcriptome in bovine (Bos taurus) spermatozoa. Sci Rep 2017; 7:42392.
49. Card CJ, Anderson EJ, Zamberlan S, Krieger KE, Kaproth M, Sartini BL. Cryopreserved bovine spermatozoal transcript profile as revealed by high-throughput ribonucleic acid sequencing. Biol Reprod 2013; 88:49.
50. Das PJ, McCarthy F, Vishnoi M, Paria N, Gresham C, Li G, Kachroo P, Sudderth AK, Teague S, Love CC, Varner DD, Chowdhary BP, et al. Stallion sperm transcriptome comprises functionally coherent coding and regulatory RNAs as revealed by microarray analysis and RNA-seq. PLoS One 2013; 8:e56535.
51. Johnson GD, Sendler E, Lalancette C, Hauser R, Diamond MP, Krawetz SA. Cleavage of rRNA ensures translational cessation in sperm at fertilization. Mol Hum Reprod 2011; 17:721-726.
52. Capra E, Turri F, Lazzari B, Cremonesi P, Gliozzi TM, Fojadelli I, Stella A, Pizzi F. Small RNA sequencing of cryopreserved semen from single bull revealed altered miRNAs and piRNAs expression between High- and Low-motile sperm populations. BMC Genomics 2017; 18:14.
53. Jodar M, Sendler E, Moskovtsev SI, Librach CL, Goodrich R, Swanson S, Hauser R, Diamond MP, Krawetz SA. Absence of sperm RNA elements correlates with idiopathic male infertility. Sci Transl Med 2015; 7:295re296.
54. Goodrich RJ, Anton E, Krawetz SA. Isolating mRNA and small noncoding RNAs from human sperm. Methods Mol Biol 2013; 927:385-396.
55. Jodar M, Selvaraju S, Sendler E, Diamond MP, Krawetz SA, Reproductive Medicine N. The presence, role and clinical use of spermatozoal RNAs. Hum Reprod Update 2013; 19:604-624.

56. Johnson GD, Mackie P, Jodar M, Moskovtsev S, Krawetz SA. Chromatin and extracellular vesicle associated sperm RNAs. Nucleic Acids Res 2015; 43:6847-6859.
57. Smith S, Hosid S, Scott L. Use of postseparation sperm parameters to determine the method of choice for sperm preparation for assisted reproductive technology. Fertil Steril 1995; 63:591-597.
58. Mao S, Goodrich RJ, Hauser R, Schrader SM, Chen Z, Krawetz SA. Evaluation of the effectiveness of semen storage and sperm purification methods for spermatozoa transcript profiling. Syst Biol Reprod Med 2013; 59:287-295.
59. Allamaneni SS, Agarwal A, Rama S, Ranganathan P, Sharma RK. Comparative study on density gradients and swim-up preparation techniques utilizing neat and cryopreserved spermatozoa. Asian J Androl 2005; 7:86-92.
60. Lalancette C, Platts AE, Johnson GD, Emery BR, Carrell DT, Krawetz SA. Identification of human sperm transcripts as candidate markers of male fertility. J Mol Med (Berl) 2009; 87:735-748.
61. Georgiadis AP, Kishore A, Zorrilla M, Jaffe TM, Sanfilippo JS, Volk E, Rajkovic A, Yatsenko AN. High quality RNA in semen and sperm: isolation, analysis and potential application in clinical testing. J Urol 2015; 193:352-359.
62. Goodrich R, Johnson G, Krawetz SA. The preparation of human spermatozoal RNA for clinical analysis. Arch Androl 2007; 53:161-167.
63. Jodar M, Sendler E, Moskovtsev SI, Librach CL, Goodrich R, Swanson S, Hauser R, Diamond MP, Krawetz SA. Response to Comment on "Absence of sperm RNA elements correlates with idiopathic male infertility". Sci Transl Med 2016; 8:353tr351.
64. Mao S, Sendler E, Goodrich RJ, Hauser R, Krawetz SA. A comparison of sperm RNA-seq methods. Syst Biol Reprod Med 2014; 60:308-315.
65. Jodar M, Anton E, Krawetz SA. Sperm RNA and Its Use as a Clinical marker. In: De Jonge CJ, Barratt CL (eds.), The Sperm Cell: Production, Maturation, Fertilization, Regeneration, 2 ed. Cambridge: Cambridge Univeristy Press; 2017: 59-72.
66. Jodar M, Sendler E, Krawetz SA. The protein and transcript profiles of human semen. Cell Tissue Res 2016; 363:85-96.
67. Johnson GD, Lalancette C, Linnemann AK, Leduc F, Boissonneault G, Krawetz SA. The sperm nucleus: chromatin, RNA, and the nuclear matrix. Reproduction 2011; 141:21-36.
68. Chen Q, Yan W, Duan E. Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. Nat Rev Genet 2016; 17:733-743.

69. Engreitz JM, Ollikainen N, Guttman M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. Nat Rev Mol Cell Biol 2016; 17:756-770.
70. Ward WS. Function of sperm chromatin structural elements in fertilization and development. Mol Hum Reprod 2010; 16:30-36.
71. Zhang X, Gao F, Fu J, Zhang P, Wang Y, Zeng X. Systematic identification and characterization of long non-coding RNAs in mouse mature sperm. PLoS One 2017; 12:e0173402.
72. Gao F, Zhang P, Zhang H, Zhang Y, Zhang Y, Hao Q, Zhang X. Dysregulation of long noncoding RNAs in mouse testes and spermatozoa after exposure to cadmium. Biochem Biophys Res Commun 2017; 484:8-14.
73. Schmitz SU, Grote P, Herrmann BG. Mechanisms of long noncoding RNA function in development and disease. Cell Mol Life Sci 2016; 73:2491-2509.
74. Zhang C, Gao L, Xu EY. LncRNA, a new component of expanding RNA-protein regulatory network important for animal sperm development. Semin Cell Dev Biol 2016; 59:110-117.
75. Cortes-Lopez M, Miura P. Emerging Functions of Circular RNAs. Yale J Biol Med 2016; 89:527-537.
76. Dong Y, He D, Peng Z, Peng W, Shi W, Wang J, Li B, Zhang C, Duan C. Circular RNAs in cancer: an emerging key player. J Hematol Oncol 2017; 10:2.
77. Dong WW, Li HM, Qing XR, Huang DH, Li HG. Identification and characterization of human testis derived circular RNAs and their existence in seminal plasma. Sci Rep 2016; 6:39080.
78. Lau NC. Small RNAs in the animal gonad: guarding genomes and guiding development. Int J Biochem Cell Biol 2010; 42:1334-1347.
79. Yuan S, Schuster A, Tang C, Yu T, Ortogero N, Bao J, Zheng H, Yan W. Sperm-borne miRNAs and endo-siRNAs are important for fertilization and preimplantation embryonic development. Development 2016; 143:635-647.
80. Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell 2007; 12:503-514.
81. Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature 2009; 457:413-420.
82. Liu WM, Pang RT, Chiu PC, Wong BP, Lao K, Lee KF, Yeung WS. Sperm-borne microRNA-34c is required for the first cleavage division in mouse. Proc Natl Acad Sci U S A 2012; 109:490-494.
83. Yuan S, Tang C, Zhang Y, Wu J, Bao J, Zheng H, Xu C, Yan W. mir-34b/c and mir-449a/b/c are required for spermatogenesis, but not for the first cleavage division in mice. Biol Open 2015; 4:212-223.

84. Lee RC, Hammell CM, Ambros V. Interacting endogenous and exogenous RNAi pathways in Caenorhabditis elegans. RNA 2006; 12:589-597.
85. Nodine MD. Mobile small RNAs: Sperm-companion communication. Nat Plants 2016; 2:16041.
86. Modzelewski AJ, Hilz S, Crate EA, Schweidenback CT, Fogarty EA, Grenier JK, Freire R, Cohen PE, Grimson A. Dgcr8 and Dicer are essential for sex chromosome integrity during meiosis in males. J Cell Sci 2015; 128:2314-2327.
87. Zimmermann C, Romero Y, Warnefors M, Bilican A, Borel C, Smith LB, Kotaja N, Kaessmann H, Nef S. Germ cell-specific targeting of DICER or DGCR8 reveals a novel role for endo-siRNAs in the progression of mammalian spermatogenesis and male fertility. PLoS One 2014; 9:e107023.
88. Yang F, Wang PJ. Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. Semin Cell Dev Biol 2016; 59:118-125.
89. O'Donnell KA, Boeke JD. Mighty Piwis defend the germline against genome intruders. Cell 2007; 129:37-44.
90. Pantano L, Jodar M, Bak M, Ballesca JL, Tommerup N, Oliva R, Vavouri T. The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. RNA 2015; 21:1085-1095.
91. Giebler M, Greither T, Muller L, Mosinger C, Behre HM. Altered PIWI-LIKE 1 and PIWI-LIKE 2 mRNA expression in ejaculated spermatozoa of men with impaired sperm characteristics. Asian J Androl 2017.
92. Tang W, Tu S, Lee HC, Weng Z, Mello CC. The RNase PARN-1 Trims piRNA 3' Ends to Promote Transcriptome Surveillance in C. elegans. Cell 2016; 164:974-984.
93. Ding DQ, Liu JL, Dong KZ, Midic U, Hess RA, Xie HR, Demireva EY, Chen C. PNLDC1 is essential for piRNA 3 ' end trimming and transposon silencing during spermatogenesis in mice. Nature Communications 2017; 8.
94. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 2007; 316:744-747.
95. He X, Chen X, Zhang X, Duan X, Pan T, Hu Q, Zhang Y, Zhong F, Liu J, Zhang H, Luo J, Wu K, et al. An Lnc RNA (GAS5)/SnoRNA-derived piRNA induces activation of TRAIL gene by site-specifically recruiting MLL/COMPASS-like complexes. Nucleic Acids Res 2015; 43:3712-3725.
96. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 2010; 11:204-220.
97. Kirchner S, Ignatova Z. Emerging roles of tRNA in adaptive translation, signalling dynamics and disease. Nat Rev Genet 2015; 16:98-112.

98. Jachowicz JW, Bing X, Pontabry J, Boskovic A, Rando OJ, Torres-Padilla ME. LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. Nat Genet 2017; 49:1502-1510.
99. Miller D. Confrontation, Consolidation, and Recognition: The Oocyte's Perspective on the Incoming Sperm. Cold Spring Harb Perspect Med 2015; 5:a023408.
100. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 2007; 8:272-285.
101. Spadafora C. Sperm-mediated 'reverse' gene transfer: a role of reverse transcriptase in the generation of new genetic information. Hum Reprod 2008; 23:735-740.
102. Niakan KK, Han J, Pedersen RA, Simon C, Pera RA. Human pre-implantation embryo development. Development 2012; 139:829-841.
103. Vassena R, Boue S, Gonzalez-Roca E, Aran B, Auer H, Veiga A, Izpisua Belmonte JC. Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development. Development 2011; 138:3699-3709.
104. Rando OJ. Intergenerational Transfer of Epigenetic Information in Sperm. Cold Spring Harb Perspect Med 2016; 6.
105. Ntostis P, Carter D, Iles D, Huntriss JD, Tzetis M, Miller D. Potential sperm contributions to the murine zygote predicted by in silico analysis. Reproduction 2017.
106. Saunders CM, Larman MG, Parrington J, Cox LJ, Royse J, Blayney LM, Swann K, Lai FA. PLC zeta: a sperm-specific trigger of Ca(2+) oscillations in eggs and embryo development. Development 2002; 129:3533-3544.
107. Practice Committee of the American Society for Reproductive M. Evaluation and treatment of recurrent pregnancy loss: a committee opinion. Fertil Steril 2012; 98:1103-1111.
108. Rogenhofer N, Ott J, Pilatz A, Wolf J, Thaler CJ, Windischbauer L, Schagdarsurengin U, Steger K, von Schonfeldt V. Unexplained recurrent miscarriages are associated with an aberrant sperm protamine mRNA content. Hum Reprod 2017; 32:1574-1582.
109. Brahem S, Mehdi M, Landolsi H, Mougou S, Elghezal H, Saad A. Semen parameters and sperm DNA fragmentation as causes of recurrent pregnancy loss. Urology 2011; 78:792-796.
110. Coughlan C, Clarke H, Cutting R, Saxton J, Waite S, Ledger W, Li T, Pacey AA. Sperm DNA fragmentation, recurrent implantation failure and recurrent miscarriage. Asian J Androl 2015; 17:681-685.
111. Eisenberg ML, Sapra KJ, Kim SD, Chen Z, Buck Louis GM. Semen quality and pregnancy loss in a contemporary cohort of couples recruited before

conception: data from the Longitudinal Investigation of Fertility and the Environment (LIFE) Study. Fertil Steril 2017; 108:613-619.

112. Gil-Villa AM, Cardona-Maya W, Agarwal A, Sharma R, Cadavid A. Assessment of sperm factors possibly involved in early recurrent pregnancy loss. Fertil Steril 2010; 94:1465-1472.
113. Nanassy L, Carrell DT. Paternal effects on early embryogenesis. J Exp Clin Assist Reprod 2008; 5:2.
114. Aoki VW, Moskovtsev SI, Willis J, Liu L, Mullen JB, Carrell DT. DNA integrity is compromised in protamine-deficient human sperm. J Androl 2005; 26:741-748.
115. Szyf M. Nongenetic inheritance and transgenerational epigenetics. Trends Mol Med 2015; 21:134-144.
116. Spadafora C. Soma to germline inheritance of extrachromosomal genetic information via a LINE-1 reverse transcriptase-based mechanism. Bioessays 2016; 38:726-733.
117. Cossetti C, Lugini L, Astrologo L, Saggio I, Fais S, Spadafora C. Soma-to-germline transmission of RNA in mice xenografted with human tumour cells: possible transport by exosomes. PLoS One 2014; 9:e101629.
118. Kaati G, Bygren LO, Edvinsson S. Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. Eur J Hum Genet 2002; 10:682-688.
119. Donkin I, Versteyhe S, Ingerslev LR, Qian K, Mechta M, Nordkap L, Mortensen B, Appel EV, Jorgensen N, Kristiansen VB, Hansen T, Workman CT, et al. Obesity and Bariatric Surgery Drive Epigenetic Variation of Spermatozoa in Humans. Cell Metab 2016; 23:369-378.
120. Kono T, Obata Y, Wu Q, Niwa K, Ono Y, Yamamoto Y, Park ES, Seo JS, Ogawa H. Birth of parthenogenetic mice that can develop to adulthood. Nature 2004; 428:860-864.
121. Palmer NO, Bakos HW, Fullston T, Lane M. Impact of obesity on male fertility, sperm function and molecular composition. Spermatogenesis 2012; 2:253-263.
122. Herati AS, Zhelyazkova BH, Butler PR, Lamb DJ. Age-related alterations in the genetics and genomics of the male germ line. Fertil Steril 2017; 107:319-323.
123. Stone BA, Alex A, Werlin LB, Marrs RP. Age thresholds for changes in semen parameters in men. Fertil Steril 2013; 100:952-958.
124. Chavarro JE, Toth TL, Wright DL, Meeker JD, Hauser R. Body mass index in relation to semen quality, sperm DNA integrity, and serum reproductive hormone levels among men attending an infertility clinic. Fertil Steril 2010; 93:2222-2231.

125. Campbell JM, Lane M, Owens JA, Bakos HW. Paternal obesity negatively affects male fertility and assisted reproduction outcomes: a systematic review and meta-analysis. Reprod Biomed Online 2015; 31:593-604.
126. Alio AP, Salihu HM, McIntosh C, August EM, Weldeselasse H, Sanchez E, Mbah AK. The effect of paternal age on fetal birth outcomes. Am J Mens Health 2012; 6:427-435.
127. Bissonnette N, Levesque-Sergerie JP, Thibault C, Boissonneault G. Spermatozoal transcriptome profiling for bull sperm motility: a potential tool to evaluate semen quality. Reproduction 2009; 138:65-80.
128. Chen C, Wu H, Shen D, Wang S, Zhang L, Wang X, Gao B, Wu T, Li B, Li K, Song C. Comparative profiling of small RNAs of pig seminal plasma and ejaculated and epididymal sperm. Reproduction 2017; 153:785-796.
129. Singh RP, Shafeeque CM, Sharma SK, Singh R, Mohan J, Sastry KV, Saxena VK, Azeez PA. Chicken sperm transcriptome profiling by microarray analysis. Genome 2016; 59:185-196.
130. Govindaraju A, Uzun A, Robertson L, Atli MO, Kaya A, Topper E, Crate EA, Padbury J, Perkins A, Memili E. Dynamics of microRNAs in bull spermatozoa. Reprod Biol Endocrinol 2012; 10:82.
131. Chen X, Wang Y, Zhu H, Hao H, Zhao X, Qin T, Wang D. Comparative transcript profiling of gene expression of fresh and frozen-thawed bull sperm. Theriogenology 2015; 83:504-511.
132. Uhl EW, Warner NJ. Mouse Models as Predictors of Human Responses: Evolutionary Medicine. Curr Pathobiol Rep 2015; 3:219-223.
133. Sutton EF, Gilmore LA, Dunger DB, Heijmans BT, Hivert MF, Ling C, Martinez JA, Ozanne SE, Simmons RA, Szyf M, Waterland RA, Redman LM, et al. Developmental programming: State-of-the-science and future directions-Summary from a Pennington Biomedical symposium. Obesity (Silver Spring) 2016; 24:1018-1026.
134. Radford EJ, Ito M, Shi H, Corish JA, Yamazawa K, Isganaitis E, Seisenberger S, Hore TA, Reik W, Erkek S, Peters A, Patti ME, et al. In utero effects. In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. Science 2014; 345:1255903.
135. Shea JM, Serra RW, Carone BR, Shulha HP, Kucukural A, Ziller MJ, Vallaster MP, Gu H, Tapper AR, Gardner PD, Meissner A, Garber M, et al. Genetic and Epigenetic Variation, but Not Diet, Shape the Sperm Methylome. Dev Cell 2015; 35:750-758.
136. Gillespie CF, Phifer J, Bradley B, Ressler KJ. Risk and resilience: genetic and environmental influences on development of the stress response. Depress Anxiety 2009; 26:984-992.

1.4.3.2 Whole Genome Sequencing

High-throughput sequencing cost has been dropping quickly and at an unstopping pace since its development. The milestone of the $1,000 genomes (sequencing whole mammalian genomes for $1,000) is now a reality and this price will most certainly continue to drop. This is fostering the genome sequencing of a large number of individuals (humans and domestic animals) for a better characterization and annotation of their genomes and their role in phenotypes as well as the generation of reference genomes for other species with unresolved sequences. One of the most remarkable application of whole genome sequencing (WGS) is the identification of genetic variants (SNPs, indels and copy number variants) and their association to phenotypes of interest in the given species or population (Koboldt et al., 2013; Ng and Kirkness, 2010). As price drops, WGS will become more present in the livestock genetics arena, but for the time being, its cost is still a limiting factor. To date, no studies have used WGS to study the porcine sperm quality.

1.4.3.3 Epigenetics

The term 'epigenetics' was first coined as the changes in the phenotype without changes in the genotype (Waddington, 1942). Later on, epigenetics was re-defined as the heritable changes in chromatin that affect gene expression and do not involve changes in the DNA sequence (Chavatte-Palmer et al., 2018). Epigenetic mechanisms include DNA methylation of cytosines and the chromatin topography and states. Chromatin can be defined as the nucleoprotein complex in which the genome is found compacted in the nucleus of eukaryotic cells. All proteins interacting with the genome in the nucleus and their post-translational modifications contribute to chromatin structure and function. These include (i) the proteins responsible of the chromatin topography and states (histones in nucleosomes and CCCTC-Binding Factor –CTCF-), (ii) the enzymes from the transcriptional machinery (polymerases and transcription factors –TFs-) and (iii) the post-translational modifications (PTMs) of histone tails (including for example methylation, acetylation or phosphorylation). Some

descriptions of epigenetics also include the short and long noncoding RNAs (Chavatte-Palmer et al., 2018). Chromatin structure is cell- and state- specific and guides gene expression at a given genomic state. NGS can be applied to characterize epigenomic marks at a genomic level. DNA methylation can be mapped by Bi-sulphite-sequencing techniques (NGS after converting unmethylated cytosines to inositol with bi-sulphite). Open chromatin, devoid of nucleosomes and accessible to the transcriptional machinery, can be localized by techniques such as ATAC-seq (NGS of Transposase-accessible chromatin) or MNase-seq (NGS of chromatin after digestion with Micrococcal nuclease). Specific proteins such as histones in nucleosomes, CTCF, polymerases, TFs and PTMs, can be mapped by ChIP-seq (NGS after chromatin immunoprecipitation with antibodies targeting the queried protein).

While several epigenetic studies in human and mice sperm have been carried (reviewed in: Champroux et al., 2018), epigenetic studies in livestock are scarce. Wang et al. (Wang and Kadarmideen, 2019) interrogated the differentially methylated cytosines (DMC) in adult pig testis. They identified 12,738 DMC some of which were within genes associated to pig reproduction (*DICER1*, *PCK1*, *SS18*, and *TGFB3*) and human fertility (*ACACA*, *CYP21A2*, *CYP27A1*, *HSD17B2*, *LHB*, *PARVG* and *SERPINC1*). The few numbers of studies available might change in the next few years as the FAANG consortium is producing a large amount of data to understand genome functions, including epigenetic elements, to improve precision and sensitivity of genomic selection for animal improvement (Giuffra et al., 2019).

1.4.3.4 Systems biology to study complex traits

Systems biology is an integrative approach that aims to understand the whole biological system holistically, as a whole, instead of individually interrogating its components. It requires an inter-disciplinary work involving biology, mathematics and computation. Complex traits and their quantitative aspects do not merely arise from the sum of the proprieties of the individual system but rather depend on the dynamic interactions of this trait at various biological

levels (Woelders et al., 2011). Systems biology integrates information from different sources and biological organization. Information from genomics (e.g. DNA sequencing and genotyping), epigenomics (e.g. MNase-seq and ChiP-seq), transcriptomics (e.g. microarray and RNA-seq), proteomics (eg. tandem mass spectrophotometry) and metabolomics (e.g. gas chromatography) from laboratory data generated for these studies (Figure 1.7) as well as data mining from databases (e.g. Gene Ontology or Kyoto Encyclopedia of Genes and Genomes) can shed light into the cellular regulation and dynamic interactions beyond complex traits (Mackay et al., 2009; Woelders et al., 2011).

Different predictive approaches have been developed to study the dynamic interactions between (molecular) components and the trait of interest. In this context, Fortes et al. (Fortes et al., 2010) developed the Association Weight Matrix (AWM), a network approach to exploit resulting GWAS data beyond the 'single-trait single-SNP' analyses. AWM identifies gene-gene interactions based on the assumption that correlated additive effects on a complex trait are likely to share genetic regulation acting on the respective trait (Fortes et al., 2010). The resulting SNPs (representing genes) are subjected to the Partial Correlation coefficient Information Theory (PCIT) approach (Reverter and Chan, 2008) to determine statistical significance thresholds for gene-gene interactions. Gene-gene interactions can also be studied with other programs such as Weighted correlation network analysis (WGCNA) (Zhang and Horvath, 2005). The interactions can then be used to build and analyze regulatory networks with tools such as Cytoscape. The gene network would display genes (nodes) that are joined together in pairs by edges. The edges represent the relationship between the 2 genes. A gene regulatory network can also include information from external databases to, for example, represent TF or TF co-factors.

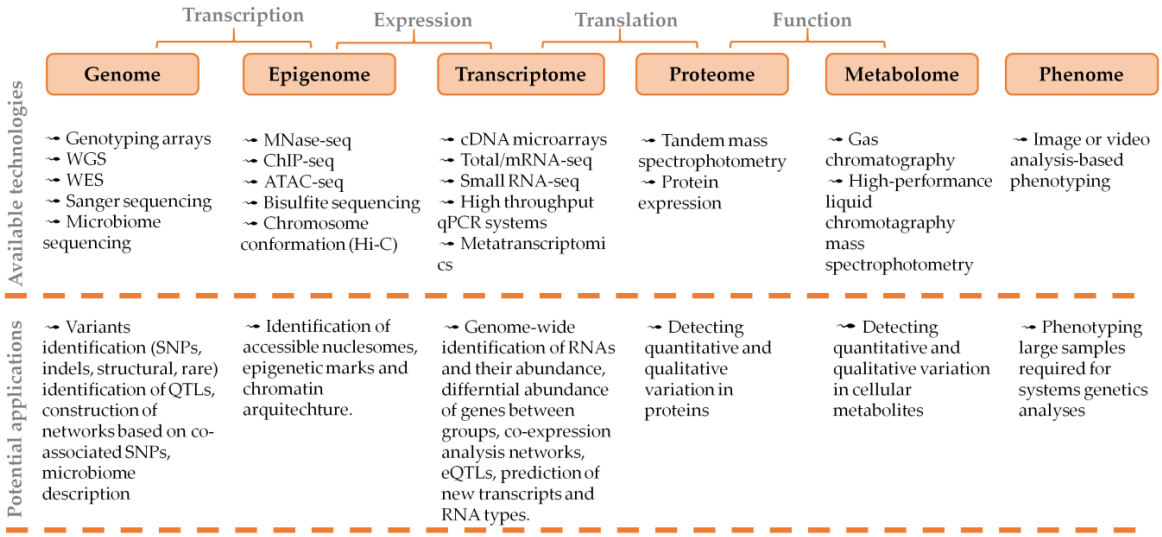


Figure 1.7 Multi-omics approaches in high-troughput technologies for genomics, epigenomics, transcriptomics, proteomics, metabolomics and phenomics. Examples of available techniques for each of the "-omics" approach and their potential applications. Systems biology enables the integration of data at the different biological levels. Whole Genome Sequencing (WGS), Whole Exome Sequening (WES), SNPs (Single Nucleotide Polymorphisms), indels (insertions/deletions), QTLs (Quantitative Traits Loci), MNAse-seq (Micrococcal nuclease sequencing), ChIP-seq (Chromatin Immunoprecipitation sequencing), ATAC-seq (Assay for Transposase-Accessible Chromatin sequencing), qPCR (quantitative PCR). Modified from: (Mackay et al., 2009; Ritchie et al., 2015).

In livestock, systems biology strategies have been implemented to study diverse complex traits such as puberty (Fortes et al., 2010), growth (Widmann et al., 2013), meat quality (Ballester et al., 2017; Diniz et al., 2019; Ramayo-Caldas et al., 2016a; Ramayo-Caldas et al., 2014) and milk production (Marete et al., 2018; Pegolo et al., 2018). To our knowledge no systems biology study has been applied to study sperm quality traits so far.

Objectives

Chapter 2

This PhD thesis was carried within the frame of the PigQSem Project funded by the Spanish Ministry of Economy and Competitiveness (grants: AGL2013-44978-R and AGL2017-86946-R). The present work has used material generated within the PigQSem Project, in coordination with the UAB's Group on Animal Reproduction and the pig companies Gepork and PIC.

The main goal of the thesis was to study the molecular mechanisms associated to boar sperm quality to ultimately identify candidate genes, pathways and DNA variants underlying these traits.

In detail, the objectives were:

1. To record semen quality phenotypic values from ejaculates from 300 Pietrain boars.
2. To purify all the ejaculates and extract DNA and RNA from all the samples.
3. To obtain a deep characterization of the porcine sperm transcriptome and identify its seasonal changes.
4. To identify SNPs and genomic regions genetically associated to sperm quality traits.
5. To identify differences in the transcriptome related to sperm quality traits.
6. To map the gene interactions, pathways and main gene regulators determining sperm quality traits using a systems biology approach.
7. To design a SNP marker panel with the potential to predict semen quality in swine.
8. To identify genetic variants in Allelic Ratio Distortion in sperm which could be related to sperm quality traits and boar fertility.

Papers and Studies

Chapter 3

Paper I

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

Marta Gòdia[a], Fabiana Quoos Mayer[a,b], Julieta Nafissi[a,c], Anna Castelló[a,d], Joan Enric Rodríguez-Gil[e], Armand Sánchez[a,d] and Àlex Clop[a*]

[a]Animal Genomics Group, Centre for Research in Agricultural Genomics-CSIC-IRTA-UAB-UB, Cerdanyola del Vallès, Catalonia, Spain;

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

[c]Department of Biotechnology and Food Technology, 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 Vallès, Catalonia, Spain;

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

*Corresponding autor

Marta Gòdia and Fabiana Quoos Mayer contributed equally to this work.

Syst Biol Reprod Med. 2018 Aug;64(4):291-303.

https://doi.org/10.1080/19396368.2018.1464610

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 the 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™; (ii) to assess the correlation between SRR and sperm quality characteristics; (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 did not find substantial differences in their performance. We propose the methodological workflow described 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 noncoding RNA molecules (Jodar et al. 2013), with functions that 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 challenge, 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 swimup 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 the loss of 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 performance.

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 basis 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 or 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 to be independent of all the sperm quality parameters (Table 1).

Table 1. Sperm quality phenotypic values and distance covariance between SRR and the semen quality parameters.

Phenotypic traits	Uncorrected mean ± SD	Distance covariance	*P*- value
Abnormal acrosomes 0 min (%)	6.4 ± 4.4	0.058	0.208
Abnormal acrosomes 90 min (%)	16.5 ± 8.1	0.072	0.317
Viability at 0 min (%)	90.4 ± 5.5	0.079	0.039
Viability at 90 min (%)	74.8 ± 14.9	0.099	0.148
ORT (%)	79.1 ± 11.8	0.098	0.108
Head abnormalities (%)	2.1 ± 6.1	0.041	0.712
Neck abnormalities (%)	2.9 ± 4.7	0.033	0.881
Tail abnormalities (%)	2.7 ± 3.4	0.084	0.029
Proximal droplets (%)	3.3 ± 4.6	0.066	0.178
Distal droplets (%)	4.5 ± 4.3	0.055	0.485
Total motility at 0 min (%)	75.4 ± 18.5	0.127	0.416
Total motility at 90 min (%)	65.4 ± 21.3	0.151	0.267

Column uncorrected mean ± SD shows the mean and standard deviation of the unadjusted semen quality phenotypic values. The adjusted phenotypic values were used to calculate its dependency with SRR. None of the traits had a significant P-value after Bonferroni correction for multiple testing ($P \leq 0.0041$); SD: standard deviation; ORT: osmotic resistance test.

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 with 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).

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*. A total of 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 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 $\Delta Cq_{\text{Genomic-}PRM1}$ ranged from 18.4 to 21. As a comparison, the liver RNA showed a *PRM1* and *PTPRC* Cqs of 38 and 24, respectively.

Table 2. Distance covariates and *P* values of the multivariate nonparametric test of independence between sperm quality phenotypes and sperm RNA extracted per cell.

Phenotypic traits	**Distance covariance**	***P*-value**
Abnormal acrosomes 0 min (%)	0.120	0.255
Abnormal acrosomes 90 min (%)	0.193	0.040
Viability at 0 min (%)	0.120	0.451
Viability at 90 min (%)	0.219	0.040
ORT (%)	0.167	0.375
Head abnormalities (%)	0.110	0.177
Neck abnormalities (%)	0.087	0.392
Tail abnormalities (%)	0.183	0.039
Proximal droplets (%)	0.162	0.020
Distal droplets (%)	0.115	0.588
Total motility at 0 min (%)	0.236	0.686
Total motility at 90 min (%)	0.262	0.667

None of the traits had a significant P-value after Bonferroni correction for multiple testing ($P \leq 0.0041$). ORT: osmotic resistance test.

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 $\Delta Cq_{PTPRC\text{-}PRM1}$ ranging from 17.4 to 19.1 and undetectable levels of *PTPRC* in the other two samples. Likewise, the $\Delta Cq_{Genomic\text{-}PRM1}$ 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 = 1.89×10^{-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.

Short RNA-seq 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 already been identified in swine sperm or in other mammalian species and include miR10b and miR34c, among others (Jodar et al. 2013; Capra et al. 2017; Chen et al. 2017).

Table 3. RNA-seq quality metrics for the SMARTer and TruSeq sperm total RNA-seq libraries.

	Sample_1		Sample_2		Sample_3		Sample_4		Sample_5		Sample_6	
	SMARTer	TruSeq	SMARTer	TruSeq	SMARTer	TruSeq	SMARTer	TruSeq	SMARTer	TruSeq	SMARTer	TruSeq
Total RNA starting amount (ng)	20	100	20	100	20	100	20	100	20	100	20	100
Library concentration (nM)	120.7	2.9	88	0.5	107.3	0.7	106.7	0.9	103.1	0.5	53	0.5
Starting number of reads (Millions)	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.3
Clean reads (%)	99.5	99.5	99.5	99.6	99.5	99.6	99.5	99.2	99.5	99.5	99.5	99.5
Reads mapped to the genome (%)	74.9	77.9	84.5	80.9	83.6	82.3	85.8	79.3	83.9	81.5	83.7	70.8
Uniquely mapped reads (%)	37.9	38.5	38.5	32.1	37.8	33.9	40.4	28.8	40.8	34.2	48.4	38.3
PCR duplicates (%)	79.8	89.3	78.2	96.4	80.3	96.3	77.1	97.9	75.9	97.2	78.1	95.7
Unmmaped reads (%)	25.1	22.1	15.5	19.1	16.4	17.7	14.2	20.7	16.1	18.4	16.3	29.2
Number of genes (FPKM > 0.1)	7,662	4,649	8,316	2,201	8,887	2,261	7,676	993	8,713	1,769	10,119	3,261

FPKM: fragments per kilobase of transcript per million mapped reads.

Table 4. RNA-seq metrics for the NEBNext and TailorMix sperm small RNA-seq libraries.

	Sample_7	Sample_8		Sample_9
	NEBNext	NEBNext	SeqMatic	SeqMatic
Total RNA starting amount (ng)	100	239	239	153
Library concentration (nM)	111.4	65.1	22.6	34.2
Starting number of reads (Millions)	0.9	0.9	0.9	0.9
Clean reads (%)	97.1	97.5	97.0	96.8
Reads mapped to the genome (%)	77.4	74.3	76.4	82.9
miRNA reads (%)	30.7	29.7	27.0	32.4
piRNA reads (%)	8.4	8	6	11
tRNA reads (%)	3.2	3	3.6	2
miscRNA reads (%)	3.7	6.5	8.9	4.7
snRNA reads (%)	9.3	9.5	5.3	4.0
snoRNA reads (%)	0.3	0.3	0.2	0.1
protein coding reads (%)	15.5	12.6	13.0	13.6
Mt tRNA reads (%)	22.7	22.1	27.3	25.8
Mt rRNA reads (%)	4.7	6.7	6.9	5.1
rRNA reads (%)	1.4	1.4	1.6	1.6
PCR duplicates (%)	89.8	90.5	92.9	92.9
Number of miRNAs detected (all depth)	205	201	174	194
Number of miRNAs dected (>10 reads)	117	121	109	119

miRNA: micro-RNA; piRNA: Piwi-interacting RNA; tRNA: transferase RNA; miscRNA: miscellaneous RNA; snRNA: small nuclear RNA; snoRNA: small nucleolar RNA; Mt tRNA: mitochondrial transference RNA; Mt rRNA: mitochondrial ribosomal RNA; rRNA: ribosomal RNA.

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).

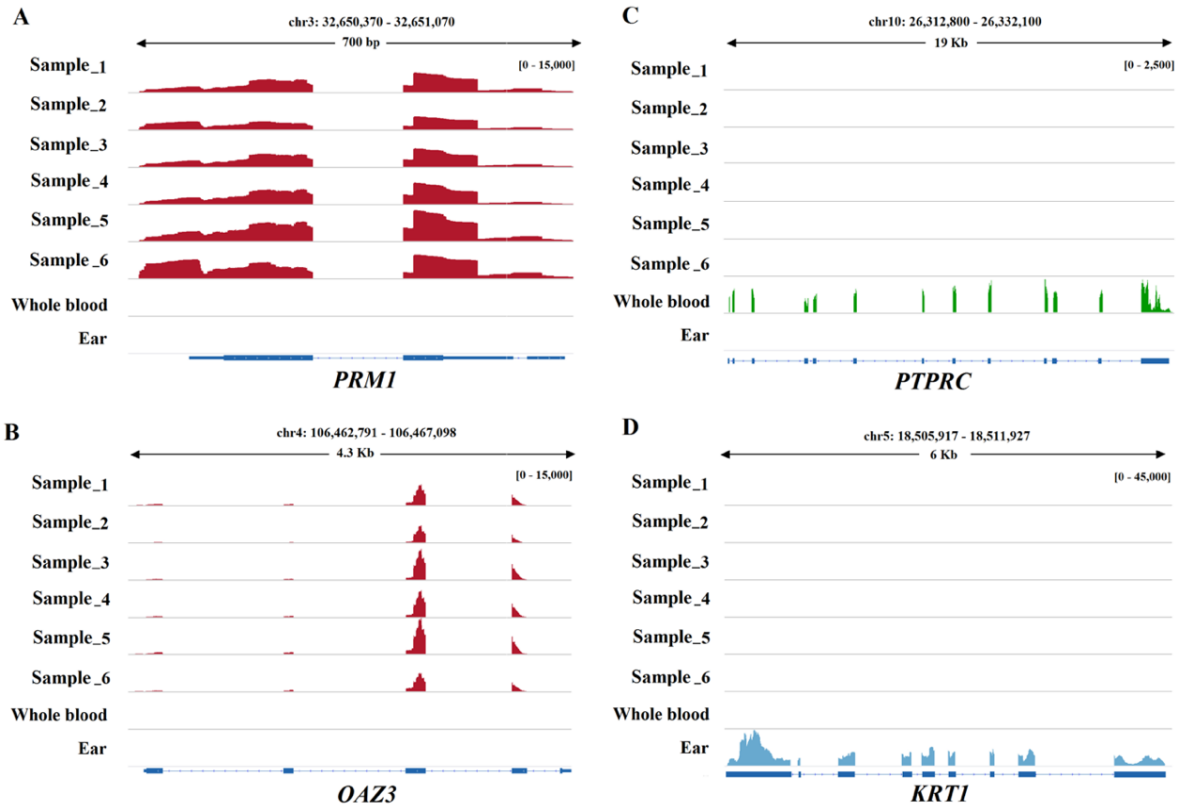


Figure 1. Read mapping depth of sperm and somatic-specific genes in the porcine sperm, whole blood and ear RNA-seq datasets. (A) Corresponds to the sperm-specific gene PRM1 (Ensembl gene ID: ENSSSCG00000021337). (B) Plot for the sperm-specific gene OAZ3 (ENSSSCG00000027091). (C) Read depth along the 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.

Discussion

Although spermatozoa are considered transcriptionally inactive, there is growing evidence that the sperm RNA population is related to spermatogenesis, fertility potential, chromatin reorganization, embryo development and transgenerational epigenetic inheritance (Jodar et al. 2013; Bohacek and Mansuy 2015). 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™ yielded highly purified spermatozoa as demonstrated by qPCR ($\Delta Cq_{PTPRC\text{-}PRM1} > 16$) 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: 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 somewhat 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™ 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™ 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 abstinence 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 (Pessot et al. 1989; Goodrich et al. 2013) 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 mechanism 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 $\Delta Cq_{PTPRC\text{-}PRM1} > 16$) and gDNA was only detected in 4 samples with $\Delta Cq_{Genomic\text{-}PRM1} > 18.4$. 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 $\Delta Cq_{PTPRC\text{-}PRM1} = 16$, the number of molecules of *PRM1* is 2^{16} = 65,536 times more abundant than the number of molecules of *PTPRC*. Likewise, when $\Delta Cq_{Genomic\text{-}PRM1} = 19$, 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. 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 2A). 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 2B).

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.

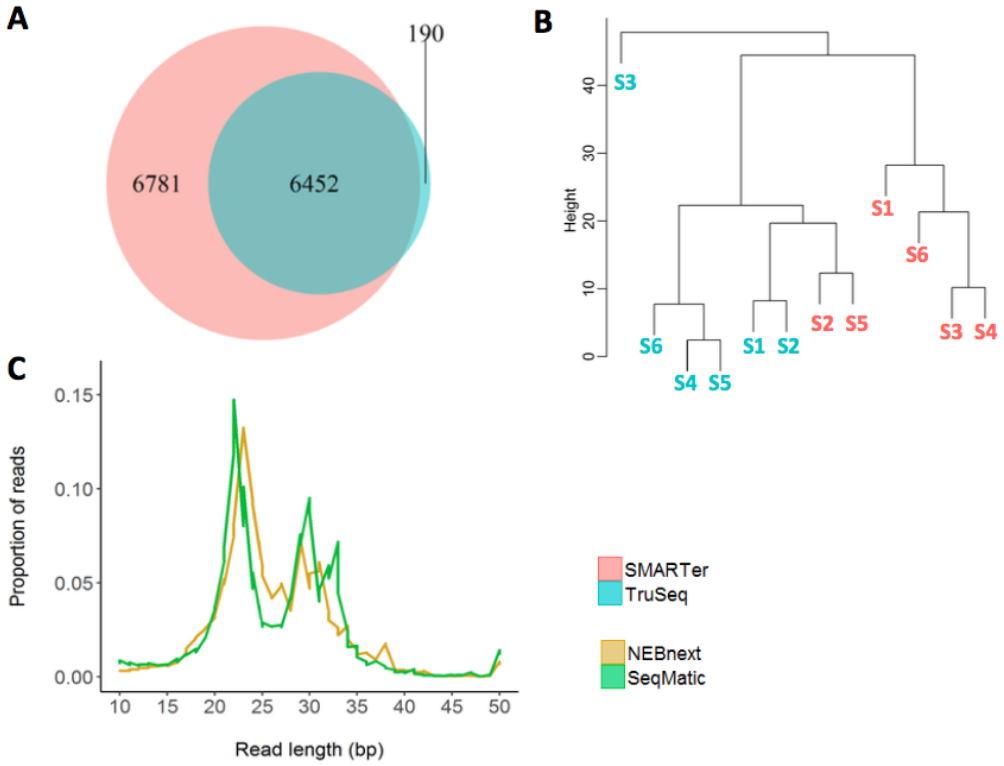


Figure 2. Comparison of the RNA-seq results from both the total and the small library preparation kits. (A) Venn diagram representing the 13,233 different transcripts detected in the SMARTer datasets and 6,642 for the TruSeq. The majority of the TruSeq transcripts detected (6,452) were detected in both 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.

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 noncoding transcriptome of the porcine sperm has been recently described (Chen et al. 2017). Chen and coauthors found a rich population of miRNAs and lower abundances of other families of small noncoding RNAs (e.g., rRNAs, tRNAs, snRNAs) but did not detect Piwi-interacting RNAs (piRNAs), a type of small noncoding RNAs with important – yet weakly explored – functions in sperm biology. In contrast, our pipeline allowed us to identify a similar catalog of small noncoding RNAs but also piRNAs (Table 4 and Figure 2C).

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 noninvasive 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™ 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°C (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°C 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 and 90 min after incubation of the samples at 37°C. 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°C 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™ (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™ (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™/semen recommended by the manufacturer was maintained. The purification was made by centrifugation at 300 × g for 20 min at 20°C 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 × g for 10 min at 20°C. 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 × g for 10 min at 20°C. The resulting pellet was stored at −80°C 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 × g for 15 min after wards. 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 × 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 × g for 5 min. The pellets were dried out at room temperature for 10 min and resuspended in 30 μL of ultra pure water. All the centrifugations were performed at 4°C.

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'-AGAACAAGGTGGATGTCTAT GGCTAT-3'; PTPRC_reverse: 5'-TGTACTGTG CCTCCACCTGAAC-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°C for 2 min, 95°C for 10 min and 40 cycles

of 95°C for 15 sec and 60°C for 1 min. Moreover, to assess the specificity of the qPCR reactions, a melting profile (95°C for 15 sec, 60°C for 15 sec and a gradual increase in temperature with a ramp rate of 1% up to 95°C) 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 ≤ 0.0041 after Bonferroni correction for multiple testing (α = 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 (Sscrofa10.2) 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).

Acknowledgments

Funding: This work was supported by the Spanish Ministerio de Economía y Competitividad (MINECO) under grant AGL2013-44978-R, (Grant Number 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 (Grant Number 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. Grant no.

BEX 6707/14-9), Brazil. Marta Gòdia acknowledges a PhD studentship from MINECO (Grant Number BES-2014-070560). Alex Clop acknowledges a MINECO's Ramon y Cajal research fellow (Grant Number 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, Marc Yeste and Isabel Serra for their suggestions on the statistical analyses.

Disclosure of interest

No potential conflict of interest was reported by the authors.

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].

Notes on contributors

Conceived and designed the experiment: ACl, AS, FQM, MG; Designed the primers and carried out the qPCR analyses: FQM, ACa; Performed sperm

purifications and RNA extrac- tions: FQM, MG, JN; Performed statistic and bioinformatics analysis: MG; Analyzed the data: FQM, MG; Carried the phenotypic analysis: JERG; Drafted the manuscript: FQM, MG, AC. All authors read and approved the final manuscript.

ORCID

Marta Gòdia Fabiana Quoos Mayer http://orcid.org/0000-0002-9324- 8536 Anna Castelló http://orcid.org/0000-0001-8497-6251 Joan Enric Rodríguez-Gil http://orcid.org/0000-0002- 1112-9884 Armand Sánchez http://orcid.org/0000-0001-9160-1124 Alex Clop http://orcid.org/0000-0001-9238-2728

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

References

Agarwal A, Gupta S, Du Plessis S, Sharma R, Esteves SC, Cirenza C, Eliwa J, Al-Najjar W, Kumaresan D, Haroun N, et al. 2016. Abstinence time and its impact on basic and advanced semen parameters. Urology. 94:102–110.

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

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

Aoki VW, Moskovtsev SI, Willis J, Liu L, Mullen JBM, Carrell DT. 2005. DNA integrity is compromised in pro- tamine-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, Sullivan R. 2013. Epididymosomes convey different repertoires of microRNAs throughout the bovine epididymis. Biol Reprod. 89(2):30.

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

Bohacek J, Mansuy IM. 2015. Molecular insights into trans- generational non-genetic inheritance of acquired behaviours. Nat Rev Genet. 16(11):641–652.

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

Byrne CJ, Fair S, English AM, Holden SA, Dick JR, Lonergan P, Kenny DA. 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 TM, Fojadelli I, Stella A, 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 CJ, Anderson EJ, Zamberlan S, Krieger KE, Kaproth M, Sartini BL. 2013. Cryopreserved bovine spermatozoal transcript profile as revealed by high-throughput ribonucleic acid sequencing. Biol Reprod. 88(2):49.

Carrell DT, Emery BR, Hammoud S. 2007. Altered protamine expression and diminished spermatogenesis: what is the link? Hum Reprod Update. 13(3):313–327.

Chan PP, Lowe TM. 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, et al. 2017. Comparative profiling of small RNAs of pig seminal plasma and ejaculated and epididymal sperm. Reproduction. 153(6):785–796.

Chen H, Boutros PC. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 12:35.

Curry E, Safranski TJ, Pratt SL. 2011. Differential expression of porcine sperm microRNAs and their association with sperm morphology and motility.

Theriogenology. 76(8):1532–1539.

Das PJ, McCarthy F, Vishnoi M, Paria N, Gresham C, Li G, Kachroo P, Sudderth AK, Teague S, Love CC, et al. 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 PJ, Paria N, Gustafson-Seabury A, Vishnoi M, Chaki SP, Love CC, Varner DD, Chowdhary BP, Raudsepp T. 2010. Total RNA isolation from stallion sperm and testis biop- sies. Theriogenology. 74(6):1099–1106.e2.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 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, et al. 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 JM. 2017. Novel agents for sperm purification, sort- ing, and imaging. Mol Reprod Dev. 84(9):832–841.

Goodrich RJ, Anton E, Krawetz SA. 2013. Isolating mRNA and small noncoding RNAs from human sperm. Methods Mol Biol. 927:385–396.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 34(Database issue):D140–D144.

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

Hammoud SS, Nix DA, Hammoud AO, Gibson M, Cairns BR, Carrell DT. 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 tech- nique of semen processing for intrauterine insemination. J Pak Med Assoc. 58(2):71–74.

Jodar M, Kalko S, Castillo J, Ballescà JL, 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 MP, Krawetz SA. 2013. The presence, role and clinical use of spermatozoal RNAs. Hum Reprod Update. 19(6):604–624.

Jodar M, Sendler E, Moskovtsev SI, Librach CL, Goodrich R, Swanson S, Hauser R, Diamond MP, Krawetz SA. 2015. Absence of sperm RNA elements correlates with idiopathic male infertility. Sci Transl Med. 7(295):295re6.

Jodar M, Sendler E, Moskovtsev SI, Librach CL, Goodrich R, Swanson S, Hauser

R, Diamond MP, Krawetz SA. 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. Curr Protoc Bioinform. Chapter 11(Unit 11):17. doi: 10.1002/0471250953.bi1107s32

Legrand JMD, Hobbs RM. 2017. RNA processing in the male germline: mechanisms and implications for fertility. Semin Cell Dev Biol. doi:10.1016/j.semcdb.2017.10.006

Li B, Dewey CN. 2011. RSEM: accurate transcript quantifica- tion from RNA-Seq data with or without a reference gen- ome. BMC Bioinformatics. 12(1):323.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence align- ment/map format and SAMtools. Bioinformatics. 25 (16):2078–2079.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2017) Cluster analysis basics and extensions. R package version 2.0.6. http://cran.r-project.org/web/packages/clus ter/index.html.

Mao S, Goodrich RJ, Hauser R, Schrader SM, Chen Z, Krawetz SA. 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 RJ, Hauser R, Krawetz SA. 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 J. 17(1):10–12.

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

Ostermeier GC, Dix DJ, Miller D, Khatri P, Krawetz SA. 2002. Spermatozoal RNA profiles of normal fertile men. Lancet. 360(9335):772–777.

Patil PS, Humbarwadi RS, Patil AD, Gune AR. 2013. Immature germ cells in semen - correlation with total sperm count and sperm motility. J Cytol. 30(3):185–189. Pessot CA, Brito M, Figueroa J, Concha II, Yañez A, Burzio LO. 1989. Presence of RNA in the sperm nucleus. Biochem Biophys Res Commun. 158(1):272–278.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841–842.

Rando OJ. 2016. Intergenerational transfer of epigenetic information in sperm.

Cold Spring Harb Perspect Med. 6 (5):a022988.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43(7):e47.

Rizzo ML, 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 JE, Bonet S. 2015. Current knowledge on boar sperm metabolism: comparison with other mammalian species. Theriogenology. 85(1):4–11.

Rodríguez-Gil JE, 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, et al. 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 AP, Krishnan Binsila B, Arangasamy A, Ravindra JP. 2017. Occurrence and functional significance of the transcriptome in bovine (Bos taurus) spermatozoa. Sci Rep. 7:42392.

Sendler E, Johnson GD, Mao S, Goodrich RJ, Diamond MP, Hauser R, Krawetz SA. 2013. Stability, delivery and func- tions of human sperm RNAs at fertilization. Nucleic Acids Res. 41(7):4104–4117.

Shafeeque CM, Singh RP, Sharma SK, Mohan J, Sastry KVH, Kolluri G, Saxena VK, Tyagi JS, Kataria JM, Azeez PA. 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, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. Plos One. 11(10):e0163962.

Siffroi JP, Alfonsi MF, Dadoune JP. 1998. Electron microscopic in situ hybridization study of simultaneous expres- sion 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 FE, Bergmann M. 2000. Expression of protamine-1 and −2 mRNA during human spermiogenesis. Mol Hum Reprod. 6(3):219–225.

Székely GJ, Rizzo ML. 2009. Brownian distance covariance. Ann Appl Stat. 3(4):1236–1265.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14(2):178–192.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolution- ary tool for transcriptomics. Nat Rev Genet. 10(1):57–63. Wykes SM, Visscher DW, Krawetz SA. 1997. Haploid tran- scripts persist in mature human spermatozoa. Mol Hum Reprod. 3(1):15–19.

Yang CC, Lin YS, Hsu CC, Wu SC, Lin EC, Cheng WTK. 2009. Identification and sequencing of remnant messen- ger 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, Qiao Z. 2006. Characterization and quantifica- tion of mRNA transcripts in ejaculated spermatozoa of fertile men by serial analysis of gene expression. Hum Reprod. 21(6):1583–1590.

Paper II

A RNA-Seq Analysis to Describe the Boar Sperm Transcriptome and Its Seasonal Changes

Marta Gòdia[1], Molly Estill[2,3], Anna Castelló[1,4], Sam Balasch[5], Joan E. Rodríguez-Gil[6], Stephen A. Krawetz[2,3,7], Armand Sánchez[1,4] and Àlex Clop[1,8*]

[1]Animal Genomics Group, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Catalonia, Spain

[2]Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, United States

[3]C.S. Mott Center for Human Growth and Development, Wayne State University, Detroit, MI, United States

[4]Unit of Animal Science, Department of Animal and Food Science, Autonomous University of Barcelona, Barcelona, Spain

[5]Grup Gepork S.A., Barcelona, Spain

[6]Unit of Animal Reproduction, Department of Animal Medicine and Surgery, Autonomous University of Barcelona, Barcelona, Spain

[7]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, United States

[8]Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, Spain

*Corresponding autor

Front Genet. 2019 Apr 16;10:299

https://doi.org/10.3389/fgene.2019.00299

Abstract

Understanding the molecular basis of cell function and ultimate phenotypes is crucial for the development of biological markers. With this aim, several RNA-seq studies have been devoted to the characterization of the transcriptome of ejaculated spermatozoa in relation to sperm quality and fertility. Semen quality follows a seasonal pattern and decays in the summer months in several animal species. The aim of this study was to deeply profile the transcriptome of the boar sperm and to evaluate its seasonal changes. We sequenced the total and the short fractions of the sperm RNA from 10 Pietrain boars, 5 collected in summer and 5 five sampled in winter, and identified a complex and rich transcriptome with 4,436 coding genes of moderate to high abundance. Transcript fragmentation was high but less obvious in genes related to spermatogenesis, chromatin compaction and fertility. Short non-coding RNAs mostly included piwi-interacting RNAs, transfer RNAs and microRNAs. We also compared the transcriptome of the summer and the winter ejaculates and identified 34 coding genes and 7 microRNAs with a significantly distinct distribution. These genes were mostly related to oxidative stress, DNA damage and autophagy. This is the deepest characterization of the boar sperm transcriptome and the first study linking the transcriptome and the seasonal variability of semen quality in animals. The annotation described here can be used as a reference for the identification of markers of sperm quality in pigs.

Keywords: sperm, sperm RNA element, RNA-seq, sperm seasonality, transcript integrity, differential gene expression

Introduction

Semen Quality Is Highly Relevant for the Sustainability of Modern Pig Breeding

Swine, together with poultry, are the most important sources of meat for human consumption (in kg) worldwide (OECD, 2018). Moreover, the global demand for animal protein is growing quickly. Thus, improving the efficiency of pork production is of paramount importance for the sustainability of the sector. Pig production relies on the genetic merit of boars kept in artificial insemination centers and the quality of their sperm to disseminate their genetic material. Hence, there is an increasing demand for molecular markers that afford early prediction of semen quality and fertility in young boars.

The Sperm Cell Contains a Complex and Functionally Relevant Transcriptome

For decades, the ejaculated mature sperm was considered a dormant cell that only carried the paternal genome to the egg. Nonetheless, in the recent years the biological complexity of sperm has become more evident, with the discovery of a rich sperm RNA population with functional roles in spermatogenesis, fertilization, early embryo development and transgenerational epigenetic transmission (Gòdia et al., 2018b). Mature sperm RNAs have been studied by NGS in several mammalian species including human (Sendler et al., 2013), horse (Das et al., 2013), mouse (Johnson et al., 2015), and cattle (Selvaraju et al., 2017). These studies have shown a sperm-specific transcriptome with a large population of transcripts most of which are present at low levels and are also highly fragmented. The sncRNA population of sperm has also been interrogated in several mammals (Krawetz et al., 2011; Das et al., 2013; Capra et al., 2017), and is composed of a large and complex repertoire of microRNAs (miRNAs), piRNAs, and tRNAs, among other RNA classes. The abundance of these transcripts has been proposed as a valuable source of bio-markers for

semen quality in animal breeding and bio-medicine (Jodar et al., 2015; Salas-Huetos et al., 2015; Capra et al., 2017).

The Boar Sperm Transcriptome

The boar sperm transcriptome has been interrogated in several studies, most employing qPCR analysis of target genes. Although qPCR is a useful tool that provides very valuable information, these studies typically assume transcript integrity and target one or two exons of only candidate genes. RNA-seq overcomes these two limitations. The first genome wide evaluation of the boar spermatozoa transcriptome was completed in 2009 by sequencing the 5′-ends of a Expressed Sequence Tag library using Sanger technology (Yang et al., 2009), which led to the identification of 514 unique sequences many of which corresponded to unknown genes. High-throughput RNA-seq was more recently applied to compare two differentially fed boars (Bruggmann et al., 2013) and to explore the short RNA component of the boar sperm (Luo et al., 2015; Pantano et al., 2015; Chen et al., 2017a; Chen et al., 2017b). These studies aimed to compare the sncRNAs at different stages of spermatogenesis or between the different components of the ejaculate, and concluded that a large proportion of these short RNAs are sperm-specific. Despite these previous studies, an in-depth analysis of the boar sperm transcriptome is still missing.

Sperm Quality Has a Seasonal Component

Sperm quality can be influenced by multi-factorial genetics (Marques et al., 2017) and environmental factors such as stress and seasonality (Wettemann et al., 1976). In pigs, a clear drop on semen quality and male fertility has been observed in the warm summer months, possibly due to heat stress (Trudeau and Sanford, 1986; Zasiadczyk et al., 2015). This seasonal effect has been linked to altered levels of some transcripts (Yang et al., 2010).

The first step toward the efficient identification of RNA markers of sperm

quality requires obtaining a profound picture of the boar sperm transcriptome. Our group has recently optimized a pipeline to extract RNA from swine mature spermatozoa and obtain a high quality and complete transcriptome profile (Gòdia et al., 2018a). In this study, we have profiled the sperm transcriptome from 10 boars, including both coding and non-coding RNAs and we have evaluated the relationship between transcript abundance and the season of collection (summer versus winter) in the northern temperate climate zone.

Material and methods

Sample Collection

Specialized professionals obtained 10 fresh ejaculates each from a different Pietrain boar from a commercial stud, with ages ranging from 9 to 28 months of age. The ejaculates were collected between July 2015 and January 2017 as previously described (Gòdia et al., 2018a). Five ejaculates were collected between December and February (winter ejaculates), and the other 5 were obtained between May and July (summer ejaculates). Fresh sperm ejaculates were obtained by the hand glove method. Spermatozoa were directly purified from the ejaculate by density gradient centrifugation (Gòdia et al., 2018a).

RNA Extraction, qPCR Validation, Library Prep and Sequencing

RNA extraction was performed as described in Gòdia et al. (2018a). The purity of the extracted RNA, defined as RNA originating exclusively from sperm cells and devoid of DNA, was determined with three qPCR assays assessing the abundance of the sperm specific *PRM1* transcript, the somatic-cell specific *PTPRC* RNA and the presence of genomic DNA (gDNA) as previously described by our group (Gòdia et al., 2018a). RNA was then quantified with Qubit™ RNA HS Assay kit (Invitrogen; Carlsbad, CA, United States) and its integrity validated with Bioanalyzer Agilent RNA 6000 Pico kit (Agilent Technologies; Santa Clara, CA, United States). Total RNA was subjected to

ribosomal RNA depletion with the Ribo-Zero Gold rRNA Removal Kit (Illumina) and RNA-seq libraries were constructed with the SMARTer Low Input Library prep kit v2 (Clontech) and sequenced to generate 75 bp paired-end reads in an Illumina's HiSeq2500 sequencing system. Short RNA-seq libraries were prepared from the same RNA aliquots (prior to rRNA depletion) with the NEBNext Small RNA (New England Biolabs) and sequenced in an Illumina Hiseq2000 to produce 50 bp single reads.

Total RNA-Seq Mapping and Analysis of the Sperm RNA Elements

The quality of the paired-end reads were evaluated with FastQC v.0.11.11, and filtered to remove low quality reads and adaptors with Trimmomatic v.0.36 (Bolger et al., 2014). Filtered reads were then mapped to the *Sus scrofa* genome (Sscrofa11.1) with HISAT2 v.2.1.0 (Kim et al., 2015) with default parameters except "–max seeds 30" and "-k 2". Duplicate mapped reads were removed using Picard Tools2 MarkDuplicates. The uniquely mapped reads were used for the detection and quantification of SREs. SREs are short-size sequences characterized by a number of RNA-seq reads clustering to a given genomic location (Jodar et al., 2015; Estill et al., 2019; Gòdia et al., 2018b). This approach enables an accurate exon-quantification (or short-size sequence quantification) instead of a whole transcript mean, which makes it useful for tissues with highly fragmented RNA such as sperm. After mapping, SREs are classified as exonic (mapping to annotated exons), intronic, upstream/downstream 10 kb (if located 10 kb upstream or downstream of annotated genes) and orphan (mapping elsewhere in the genome) (Gòdia et al., 2018b). This classification was done using the pig Ensembl genome annotation (v.91) extracted with the R package "BiomaRt" (Durinck et al., 2009). Porcine orphan SREs coordinates were converted to human (hg38) coordinates and from human to bovine (bosTau8) using the UCSC liftover tool (Kuhn et al., 2013).

All the Gene Ontology enrichment analyses described throughout the article

were performed with Cytoscape v.2.3.0 plugin ClueGO v.2.3.5 (Bindea et al., 2009) using Cytoscape's porcine dataset and the default settings. Only the significant corrected *p*-values with Bonferroni were considered.

The CV of the RNA abundance across samples was used to classify the transcripts as highly unstable (CV > 0.75), moderately stable (CV between 0.25 and 0.75) and highly stable (CV < 0.25). To carry the GO analysis we used only these genes for which all their SREs fitted within the same stability class (stable, moderately stable or unstable), to ensure that genes were robustly assigned to a specific category.

***De novo* Transcriptome Analysis**

Reads unmapped to the Sscrofa11.1 genome were screened against the porcine Transposable Elements from the Repbase database (Bao et al., 2015) using HISAT2 v.2.1.0 (Kim et al., 2015). The remaining unmatched reads were searched against bacterial and viral genomes using Kraken v.0.10.5 (Wood and Salzberg, 2014) and removed. The remaining reads were subjected to *de novo* assembly with Trinity v.2.1.0 (Grabherr et al., 2011) using default parameters and databases. The assembled contigs were quantified with RSEM and only those with identity score > 85%, abundance levels > 50 FPKM and detected in 5 samples or more were kept.

Repetitive Elements and Long Non-coding RNAs

The proportion of reads in RE was calculated with Bedtools (Quinlan and Hall, 2010) multicov using the RepeatMasker database (Bao et al., 2015). Read counts were normalized for RE length and sequencing depth. The same approach was used for lncRNAs. Only the lncRNAs annotated in Ensembl v.91 were used. The coding genes mapping less than 20 kb apart from the lncRNAs were considered as potential *cis*- regulated lncRNA targets.

Transcript Integrity

[1] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2] http://picard.sourceforge.net

RNA transcript integrity (TIN) was calculated with RseQC v.2.6.4 (Wang et al., 2012) using the Ensembl v.91 pig annotation. TIN indicates the proportion of a gene that is covered by reads. As an example, TIN = 100 indicates a fully covered transcript. Transcript abundance was calculated using expression.py from the same software. Transcript length was calculated based on CDS length, extracted with the R package "BiomaRt" (Durinck et al., 2009).

Analysis of the Short Non-coding RNAs

Trimming of adaptors and low quality bases were performed with Cutadapt v1.0 (Martin, 2011) and evaluated with FastQC v.0.11.11. The mapping of sncRNAs was performed with the sRNAtoolbox v.6.17 (Rueda et al., 2015) with default settings and giving as library datasets: tRNA database (Chan and Lowe, 2016), miRBase (Kozomara and Griffiths-Jones, 2011) release 21, piRNA database (Rosenkranz, 2016) and Mt tRNA, Mt rRNA, snRNA, snoRNA, lincRNA, CDS, and ncRNAs from Ensembl v.91. Multi- adjusted read counts were then normalized by sequencing depth. We only considered the miRNAs that were detected in all the samples processed. To determine if piRNAs were located in REs, the overlap between REs and the piRNA clusters that were shared in at least 3 samples was checked with Bedtools (Quinlan and Hall, 2010) multicov using the RepeatMasker database (Bao et al., 2015). The short RNA-seq reads that did not align to any of the datasets provided were used for the *de novo* piRNA annotation using ProTRAC v.2.4.0 (Rosenkranz and Zischler, 2012) and forcing a piRNA length between 26 and 33 bp and a default minimum cluster length of 5 kb. We then kept only these putative novel clusters that were shared in at least 3 of the sperm samples.

Analysis of the Seasonal Variation of the Boar Sperm Transcriptome

We studied the potential seasonal effect of the sperm transcriptome by comparing the summer (N = 5) and the winter (N = 5) ejaculates. Total RNA-seq analysis was performed for the transcripts annotated in the pig genome. We

quantified RNA abundance with the software StringTie v.1.3.4 (Pertea et al., 2015). Transcript counts were then used for the differential analysis using the R package DESeq2 (Love et al., 2014) correcting for sequencing run batch. Similarly, the identification of differential miRNAs was also carried with DESeq2 (Love et al., 2014). We only considered the differentially abundant transcripts and miRNAs with adjusted FDR values < 0.05 and FC > 1.5.

Results and Discussion

Total RNA-Seq Analysis: Characterization of Sperm RNA Elements

RNA extraction yielded an average of 2.1 fg per cell (Supplementary File S1). These RNAs were devoid of intact ribosomal 18S and 28S RNA with RIN values below 2.5 and were free of gDNA and RNA from somatic cell origin (Gòdia et al., 2018a). On average, the total RNA-seq libraries yielded 23.6 M paired-end reads (Supplementary File S1). A total of 81.3% of the reads that passed the quality control filter mapped unambiguously to the pig genome (Supplementary File S1). After duplicate removal, a mean of 5.6 M reads per sample were obtained, resulting in a percentage of unique reads similar to recent data on human sperm (unpublished results). These reads were used for further analysis and yielded 185,037 SREs (Estill et al., 2019). Most SREs were present at low abundances but the 10% most abundant (top decile) SREs accounted for 65% of the read count with RNA levels ranging between 83 and 378,512 RPKM (Figure 1). Most of these top decile SREs were exonic (Supplementary File S2). Notably, the majority (65%) of the intronic and upstream/downstream 10 kb SREs mapped in or near genes that also harbored exonic SREs. The exonic, intronic and the upstream/downstream 10 kb top decile SREs mapped in or near 4,436 annotated genes, which were thus considered to be abundant in the boar sperm transcriptome (Supplementary File S2).

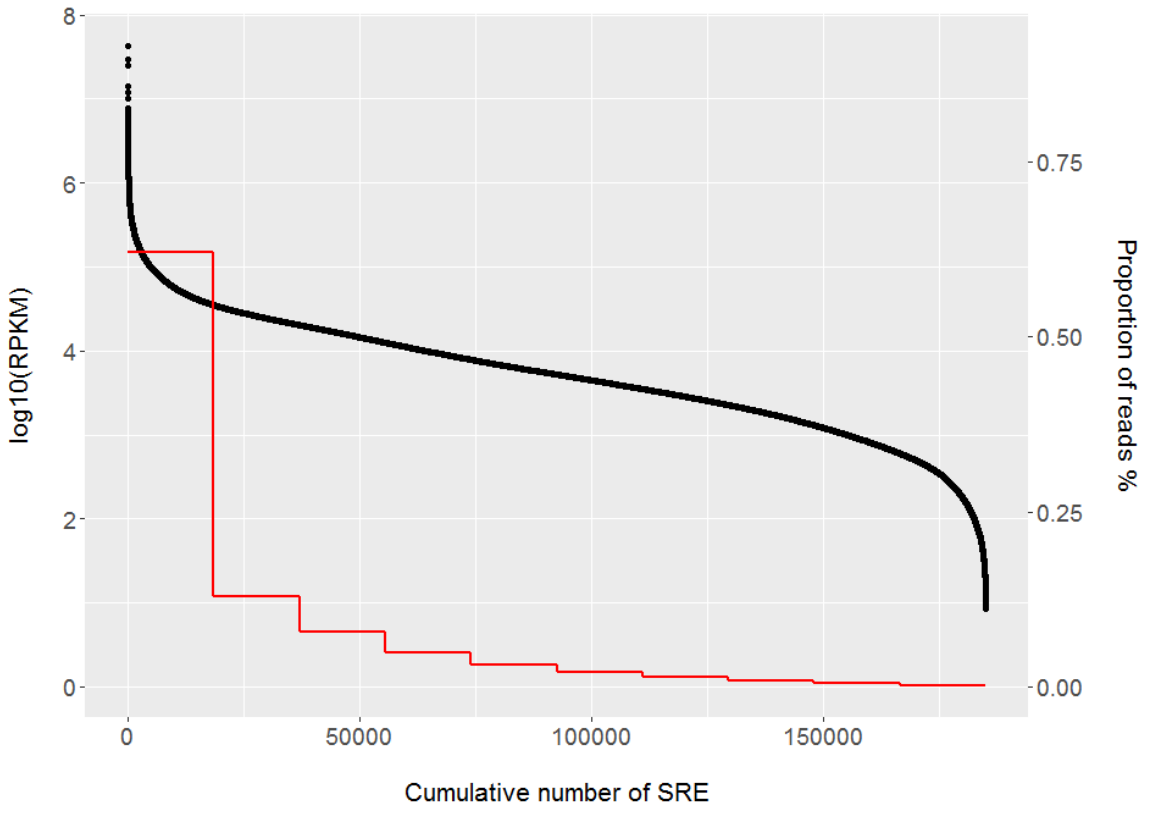


Figure 1. Cumulative abundance of the porcine SREs. The black dots indicate the log10 of the RNA abundance of each SRE. SREs are sorted in a decreasing order by their RNA abundance in the X-axis. The red line represents the total number of SREs for each abundance decile group. The first decile of the most abundant SREs accounted for 65% of the total read abundance. RPKM, reads per kilobase per million mapped reads; SRE, sperm RNA element.

The number of annotated genes with allocated SREs in our study in swine is very similar to what has been found in the human sperm (4,765) using the same SRE bioinformatics approach (Estill et al., 2019). As expected, the number of genes identified in sperm is notably low when compared to other porcine tissues. The number of expressed genes reported in porcine muscle (Chen et al., 2011), liver (Chen et al., 2011), fat (Chen et al., 2011; Corominas et al., 2013), pituitary glands (Shan et al., 2014), hypothalamus (Pérez-Montarelo et al., 2014), duodenum (Mach et al., 2014), ilium (Mach et al., 2014) and pre-pubescent male gonads (Esteve-Codina et al., 2011), ranges between 12,816 and 18,878 genes. Among these, the immature male gonads displayed the lower number of reported expressed genes. Although these values can be only taken as a guide because each study carried their own experimental pipeline, they are indicative that the boar sperm contains a way less rich and complex transcriptome when

compared to other tissues.

The top decile SREs also included 2,667 orphan SREs (SREs located more than 10 kb apart from the closest annotated gene) (Supplementary File S2). However, nearly 30% of the orphan SREs mapped within 30 kb from the closest gene, which indicates that, as the novel upstream/downstream 10 kb SREs, they may represent unannotated exons of these genes. In summary, only 10% of the top decile SREs were not linked to annotated genes. A recent study carried by Pertea et al. (2018) analyzed RNA-seq data from 9,795 human experiments from the GTEx project and concluded that the human genome annotation incorporates most of the *Homo sapiens* genes but still lacks a large proportion of the splice isoforms. While this study increased the list of coding genes by only 5%, the catalog of splice isoforms grew by 30%. Our data is in line with these recent results and does not only indicate that the novel annotation of the pig genome annotation incorporates most of the genes found in sperm but also reveals that there is still a large amount of splice isoforms to be discovered in this species. Since it is well known that the spermatozoon harbors a very specific transcriptome, a large proportion of these unannotated isoforms are likely to be sperm-specific (Sendler et al., 2013; Ma et al., 2014).

In order to dig further into the porcine sperm transcriptome, we investigated whether the porcine orphan SREs could correspond to genes not annotated in the pig but annotated in the human or cattle genomes. To this end, the location in the pig genome of the 2,667 orphan SREs were liftover onto the human and bovine genome coordinates. This resulted in 1,505 (56.4% of the 2,667 orphan SREs) human and 1,313 (49.2%) bovine syntenic regions. Forty five of the genes annotated within these regions were detected in both human and cattle (**Supplementary File S3**), including *CDYL*, a gene implicated in spermatid development and *ANXA3*, which protein levels in sperm have been found altered in men with poor semen compared to men with good sperm quality

(Netherton et al., 2018). Ontology analysis of the 4,436 most abundant genes together with the 45 orphan SRE orthologs showed an enrichment of the cellular protein metabolic process (*q*-value: 2.7 × 10− 12), macromolecular complex subunit organization (*q*-value: 2.1 × 10−9), sexual reproduction (*q*-value: 6.5 × 10−8), spermatogenesis (*q*-value: 1.2 × 10−6) and male gamete generation (*q*-value: 1.4 × 10−6), among others (**Supplementary File S4**). The transcripts detected in our study are concordant with previous results in human (Jodar et al., 2016) and bovine (Selvaraju et al., 2017) sperm and included genes related to fertilization (e.g., *HSPA1L* and *PRSS37*) or spermatogenesis (*ODF2* and *SPATA18*).

The top 30 most abundant annotated protein coding SREs mapped to 27 genes (**Table 1**), 12 from mitochondrial origin (e.g., *COX1*, *COX2*, *ATP8*, *ATP6*, and *COX3*), and 15 encoded in the nuclear genome (e.g., *PRM1*, *OAZ3*, *HSPB9*, and *NDUFS4*). The abundance of mitochondrial genes reflects the high number of mitochondria typically contained in a spermatozoa cell to provide critical functions for the cell's fertilizing ability including energy supply, regulation of molecular mechanisms involved in the development of the capacitation process, production of reactive oxygen species and calcium homeostasis (Rodriguez- Gil and Bonet, 2016). The 15 nuclear genes included members related to spermatogenesis, chromatin compaction and embryo development (Sendler et al., 2013; Selvaraju et al., 2017).

Table 1. List of the 30 most abundant SREs in the porcine sperm.

Ensembl ID	Gene ID	SRE genomic coordinates	SRE Type	Mean abundance	Abundance SD
ENSSSCG00000018075	*COX1*	MT:6511-8055	EXON	42244	14055
ENSSSCG00000018078	*COX2*	MT:8203-8890	EXON	25411	11931
ENSSSCG00000018080 ENSSSCG00000018081 ENSSSCG00000018082	*ATP8, ATP6, COX3*	MT:8959-10583	EXON	18282	10076
ENSSSCG00000021337	*PRM1*	3:31861071-31861233	EXON	14509	2711
ENSSSCG00000018094	*CYTB*	MT:15342-16481	EXON	13414	6153
ENSSSCG00000018091	*ND5*	MT:12935-14755	EXON	12285	7527
ENSSSCG00000027091	*OAZ3*	4:97442381-97442556	EXON	10492	2592
ENSSSCG00000027091	*OAZ3*	4:97441308-97441393	EXON	10441	3563
ENSSSCG00000018092	*ND6*	MT:14739-15266	EXON	8983	5521
ENSSSCG00000016203	*CFAP65*	15:121057113-121057202	NOVEL_INTRONIC	8302	7705
ENSSSCG00000018086 ENSSSCG00000018087	*ND4, LND4*	MT:11069-12736	EXON	7984	4396
ENSSSCG00000006302	*GPR161*	4:82900699-82900818	EXON	7256	1350
ENSSSCG00000018069	*ND2*	MT:5087-6128	EXON	7038	4966
ENSSSCG00000027091	*OAZ3*	4:97443314-97443450	EXON	6469	1151
ENSSSCG00000006688	*ANKRD35*	4:99454337-99454374	EXON	6130	1564
ENSSSCG00000028031	*HDAC11*	13:70866593-70866635	EXON	6012	820
ENSSSCG00000005585	*DENND1A*	1:264683712-264683755	EXON	5849	1643
ENSSSCG00000006302	*GPR161*	4:82896938-82897042	EXON	5714	839
ENSSSCG00000017609	*ANKFN1*	12:32508908-32509087	NOVEL_INTRONIC	5539	3823
ENSSSCG00000006688	*ANKRD35*	4:99459430-99459495	EXON	5483	1332
ENSSSCG00000007010	*ZMAT4*	17:9836268-9836357	NOVEL_INTRONIC	5411	4921
ENSSSCG00000017770	*PROCA1*	12:44943383-44943515	EXON	5242	1245
ENSSSCG00000017413	*HSPB9*	12:20636767-20637249	EXON	5235	1007
ENSSSCG00000000018	*KIAA0930*	5:4184013-4184090	EXON	5176	1151
ENSSSCG00000018065	*ND1*	MT:3922-4876	EXON	5155	3419
ENSSSCG00000021337	*PRM1*	3:31861339-31861529	EXON	5137	988
ENSSSCG00000016893	*NDUFS4*	16:32891178-32891257	NOVEL_INTRONIC	4843	2985
ENSSSCG00000023974	*PHF21A*	2:16386945-16386977	EXON	4792	1481
ENSSSCG00000006688	*ANKRD35*	4:99450478-99450566	EXON	4760	753
ENSSSCG00000035537	*RUNX1*	13:198392909-198392938	NOVEL_INTRONIC	4759	5935

The most abundant SREs from protein coding genes included 12 mitochondrial and 15 nuclear genes. Some genes (e.g. *PRM1, OAZ3, ANKRD35*) presented more than one highly abundant SRE. SD: Standard Deviation; SRE: Sperm RNA Element. The SRE genomic coordinates are displayed in the format chromosome:start location–end location. Mean abundance and abundance SD are indicated in RPKM: Reads Per Kilobase per Million mapped reads.

Total RNA-Seq Analysis: Variance on the SRE Abundance

We evaluated the transcripts that contained the 10% most abundant SREs across all samples and classified them as uniform (coefficient of variation or CV < 25%) or variable (CV > 75%). This identified 481 genes for which all their SREs were uniformly represented (CV < 25%) and 276 genes where each SRE was highly variable (CV > 75%). The list of 481 genes with constant abundance was enriched for several functions including the regulation of calcium, ATP generation and spermatid development and differentiation (**Supplementary File S5**). On the contrary, the highly variable genes were only enriched for the gene ontology term: single fertilization (zygote formation), which includes *SPMI*, *AQN-1* and *BSP1* among others (**Supplementary File S5**). This transcript variability is in general tolerated because it does not have severe phenotypic consequences. However, some of these transcripts may incur in a significant impact on semen quality and/or fertility and they could thus be biomarkers of the boar's reproductive ability. Thus, it would be worth exploring the relationship between these genes and reproductive phenotypes in a larger study.

Jodar et al. (2016) compared the transcriptome of testes, sperm and seminal fluid and classified the corresponding transcripts according to their relative abundance in these tissues. Subsequently, they used this classification to partition the transcripts that are present in sperm into testes-enriched, sperm–enriched and seminal fluid-enriched fractions. Testes-enriched transcripts are those that presented more than 40 FPKM in testes and less than 10 FPKM in sperm and seminal fluid. The same principle applied to the other two fractions.

According to this partition (Jodar et al., 2016), we identified in our porcine dataset, 728 testes, 448 seminal fluid and 381 sperm–enriched SREs. We compared the abundance variability of these three SRE categories and found no difference between the sperm-enriched and the other fractions (Tukey's 'Honest

Significant Difference,' *p*-values: 0.18–0.20). However, we detected a significant difference between the testes-enriched and the seminal fluid-enriched SREs (*p*-value: 3.6×10^{-4}). The seminal fluid-enriched fraction was, in average, more variable. The difference on the abundance variability between the testes-enriched and the seminal fluid-enriched fraction might have a biological explanation. Spermatogenesis is a finely orchestrated multi-step process that occurs in the testis, which may require a stable set of transcripts in each of these steps. On the contrary, the seminal fluid-enriched transcripts are likely to have been infiltrated into sperm via seminal exosomes (Vojtech et al., 2014; Jodar et al., 2016). The exosome uptake process may be relatively prone to variability as it is influenced by the concentration of exosomes in the seminal fluid, the RNA-load within these exosomes, and the efficiency in which the exosomes are merged with and release their content into the sperm cells.

Total RNA-Seq Analysis: Transcript Integrity

Sperm transcripts have been found to be highly fragmented in several mammalian species (Das et al., 2013; Sendler et al., 2013; Selvaraju et al., 2017; Gòdia et al., 2018a). We sought to investigate whether this fragmentation followed a programmatic pattern or perhaps was stochastic in the pig. For each annotated transcript, we calculated the abundance levels (in FPKM) and the TIN. In average, we found 31,287 protein coding transcripts with FPKM > 0 and TIN values > 0. Most transcripts (55%) were highly fragmented (TIN ≤ 25) whilst only 181 were almost intact (TIN > 75). Interestingly, the 10 samples showed similar TIN patterns across transcripts (Pearson correlation 0.72–0.93) (**Supplementary File S6**). The correlations between TIN and transcript length and transcript abundance were low (0.14–0.20 and 0.14–0.25, respectively) (**Supplementary File S6**).

We then searched for gene ontology enrichment using the 10% most abundant transcripts within each TIN group. The highly fragmented group (TIN < 25) was

enriched for genes related to negative regulation of JNK cascade (q-value = 1.2 × 10− 3), spindle assembly (q-value = 5.6 × 10− 3), and regulation of DNA repair (q-value = 4.5 × 10− 3), among others. These results are comparable to a previous study in human sperm (Sendler et al., 2013), where the most fragmented transcripts were not enriched for spermatogenesis or fertility functions. On the other hand, no significant pathways were found in the group of the top 10% most intact transcripts, possibly due to the low size of this group (18 transcripts), even though it contained genes related to spermatogenesis (*PRM1*, *OAZ3*, and *ACSBG2*), sperm movement (*PRM3* and *SMCP*) or heat stress response (*HSPB9*) (**Table 2**). Remarkably, the six aforementioned genes were also within the most intact transcripts in human sperm (Sendler et al., 2013), thereby indicating conservation across species and their likely basic function in supporting sperm development and/or fecundity. Altogether, this indicates that the transcript fragmentation typically found in sperm may follow a programmatic basis and possible owe to relevant functions during spermatogenesis or upon fertilization.

Table 2. List of the 10% most abundant intact transcripts (TIN > 75) in the boar sperm.

Ensembl Transcript ID	Gene ID	TIN mean	TIN SD
ENSSSCT00000018955	*ZNRF4*	97.60	0.56
ENSSSCT00000007842	*TMEM239*	93.54	2.54
ENSSSCT00000019381	*HSPB9*	92.62	3.48
ENSSSCT00000046661	*UBL4B*	91.93	2.24
ENSSSCT00000001702	*C6orf106*	89.91	4.00
ENSSSCT00000006503	*SPATC1*	86.40	2.61
ENSSSCT00000030220	*OAZ3*	84.91	2.74
ENSSSCT00000004015	*AZIN2*	83.92	3.15
ENSSSCT00000049885	*PRM3*	83.56	2.27
ENSSSCT00000029296	*DBIL5**	83.21	3.71
ENSSSCT00000014766	*ZNRF4*	82.36	2.23
ENSSSCT00000048242	*ACSBG2**	81.15	2.62
ENSSSCT00000007224	*SMCP*	79.47	5.33
ENSSSCT00000003898	*KIF17*	79.04	1.67
ENSSSCT00000007327	*ANKRD35*	78.97	2.91
ENSSSCT00000012714	*DNAJB8*	76.43	4.11
ENSSSCT00000000746	*TPI1*	75.99	3.63
ENSSSCT00000029974	*PRM1*	75.72	1.44

Transcript integrity was measured as TIN: Transcript Integrity Number; TIN mean: Average TIN. SD: Standard Deviation. * Gene symbol extracted from an orthologous gene species.

Total RNA-Seq Analysis: *De novo* Transcriptome Assembly

We sought to further exploit the RNA-seq data by performing *de novo* assembly of the reads that did not map to the porcine genome. An average of 5.1 M unmapped reads per sample were used for the analysis (**Supplementary File S1**) and assembled into a mean of 8,459 contigs per sample, with a median size (N50) of 259 bp (**Supplementary File S7**). These contigs were then contrasted by sequence homology against several protein databases and after filtering, resulted in a list of 1,060 proteins from human, cattle, mouse, pig, and other animal species with moderate to high RNA abundance (**Supplementary File S8**). Some of the proteins were detected in more than one species and accounted for a total of a non-redundant list of 768 unique genes (**Supplementary File S9**).

The majority of these genes (739) were already present in the porcine annotation whilst 29 were classified as novel genes. From the annotated genes, 699 were also detected with our initial pipeline mapping the SREs to the porcine genome but 40 were only detected by this *de novo* assembly (**Supplementary File S9**).

The unmapped reads that found a gene that is annotated in swine in the *de novo* analysis, could have remained unmapped due to two main reasons. They could have either harbored more mismatches than the maximum allowed for the mapping algorithm, or they might have corresponded to genomic segments not assembled to the current version of the porcine genome. We re-mapped the unmapped reads employing the looser mismatch penalty scores (the default is 6) 5, 4, 3, and 2 and obtained a small improvement in the read mapping percentage (92.9, 88.5, 85.8, and 77.3% of the reads remained unmapped reads, respectively). This shallow increase in the read mappeability suggests that a large proportion of the unmapped reads might have corresponded to genomic regions that are not assembled in the current version of the swine genome.

The 40 known genes detected only by the *de novo* assembly together with the 29 potential novel genes did not cluster into any GO biological process. However, some of these genes have been associated to spermatogenesis or implicated in the sperm structure such as the sperm head or flagellum (e.g., *ACSBG2*, *HSF2BP*, *CCNYL1*, *KNL1*, and *WBP2NL*). These results are in line with the recent study carried in humans by Pertea et al. (2018) as already detailed in relation to the orphan SREs. Although the number of novel protein-coding genes represents a modest increase (29 genes), our *de novo* analysis yielded a much higher number (699) of potentially novel splice variants.

Total RNA-Seq Analysis: Repetitive Elements

Repetitive elements (REs) are of particular interest as they comprise a high proportion of the porcine genome (approximately 40%) often related to genome

instability (Bzymek and Lovett, 2001). Germline cells are very sensitive to the deleterious effects of active transposable elements. For example, the disruption of LINE1 retrotransposon silencing, the most abundant RE in the pig genome, can lead to spermatogenesis aberrations (Gòdia et al., 2018a) and embryo development arrest (Beraldi et al., 2006). Due to their relevance in spermatozoa, we annotated the RE segments that were transcribed in the pig sperm. A total of 4.6% of the mapped reads overlapped with REs, which is in line with previous data in murine sperm (Johnson et al., 2015), and accounted for 42.8 Mb of the swine genome. The most enriched RE classes included simple repeats (2.58% of the total mapped reads) which could potentially correspond to porcine nuclear matrix associated RNAs (Johnson et al., 2015). The second most abundant REs were the SINEs which accounted for 0.6% of the total read abundance. SINEs are transposable elements that can be hypo-methylated and can regulate male germ cell development, sperm packaging and embryo development (Schmid et al., 2001). In pigs, LINE1 accounts for 16.8% of the genome space and in our study, 0.19% of the mapped reads overlapped with LINE1 segments and spanned 25.5 Mb of the genome. This is nearly ten times less than in mice (1.89%) (Johnson et al., 2015) even though LINE1 is just slightly more ubiquitous in the murine genome (20%) (Waterston et al., 2002). While potentially interesting, these differences may arise due to yet unknown species-specific biological particularities or technical differences in the library preparation and/or bioinformatics methods used in both studies.

Total RNA-Seq Analysis: Long Non-coding RNAs

Long non-coding RNAs are regulatory RNAs above 200 bp long implicated in a plethora of functions, including spermatogenesis and reproduction (Gòdia et al., 2018b). Sperm lncRNAs have been reported in human (Sendler et al., 2013), mice (Zhang et al., 2017), and cattle (Selvaraju et al., 2017). We identified 27 of the 361 lncRNA annotated in Ensembl v.91, and their RNA levels were clearly

below their coding SRE counterparts (**Supplementary File S10**). The predicted *cis*-regulated target genes included *ZNF217*, which is a transcriptional repressor, *DYNLRB2* which encodes for a protein belonging to the dynein family of axoneme components related to sperm motility and *YIPF5*, which caused infertility in a knock-out fruit fly model (Yu et al., 2015). The annotation of lncRNAs in the swine genome remains remarkably poor and here we provide an initial catalog that is still incomplete.

Short RNA-Seq Analysis

On average, 6.6 M reads were obtained for each short RNA-seq library. A mean of 83% of these reads aligned to the queried porcine (*S. scrofa*) databases (**Supplementary File S1**). A total of 34% of the aligned reads corresponded to sncRNAs, mainly piRNAs (37% of the sncRNA fraction), tRNAs (22.6%) and miRNAs (20.2%) (**Figure 2A** and **Supplementary File S11**). The remaining aligned reads (66%) mostly belonged to mitochondrial transfer and ribosomal RNAs (51%) but also to nuclear protein coding genes (**Supplementary File S11**).

The functional relevance of miRNAs, piRNAs, and tRNAs in sperm biology and fertility (Krawetz et al., 2011; Sharma et al., 2016; Gòdia et al., 2018b) is well known. miRNAs are a class of sncRNAs that have been found in multiple cell types and involved in a plethora of phenotypes and diseases. They post-transcriptionally repress the translation of target messenger RNAs (mRNAs) and can be ideal biomarkers for many traits including sperm quality and fertility. We detected 105 miRNAs (annotated in the pig) that were present in all the samples, with an average abundance that ranged from 4.6 to 13,192.2 CPMs. Chen et al. (2017a) carried a RNA-seq study using one pool of 3 pig sperm samples and detected a larger number of miRNAs -140- than in our study, but the overlap was remarkable, with 75 of the 140 miRNAs found in both experiments (**Supplementary File S12**). The lower number of miRNAs

described in our work compared to Chen et al. (2017a), is somewhat not surprising as we only considered those miRNAs that were present in the 10 samples and used thus more stringent parameters. The inter-species comparison also indicates a degree of conservation in the miRNA composition of the mammalian sperm with about 70% of the miRNAs shared in cattle (Capra et al., 2017) and human (Pantano et al., 2015) (**Supplementary File S12**). These results suggest a conserved functional role of these miRNAs in mammals. The most abundant miRNAs in our study, miR- 34c, miR-191, miR-30d, miR-10b and let-7a, among others (**Supplementary File S13**), are also highly abundant in cattle (Capra et al., 2017) and in human (Krawetz et al., 2011; Pantano et al., 2015) sperm. Some of these miRNAs have been linked to the male's reproductive ability. For example, miR-34c is crucial for spermatogenesis (Yuan et al., 2015) and has been related to bull fertility (Fagerlind et al., 2015) and miR-191, miR-30d, and miR-10b displayed altered levels in infertile human patients when compared to healthy controls (Salas-Huetos et al., 2015; Tian et al., 2017).

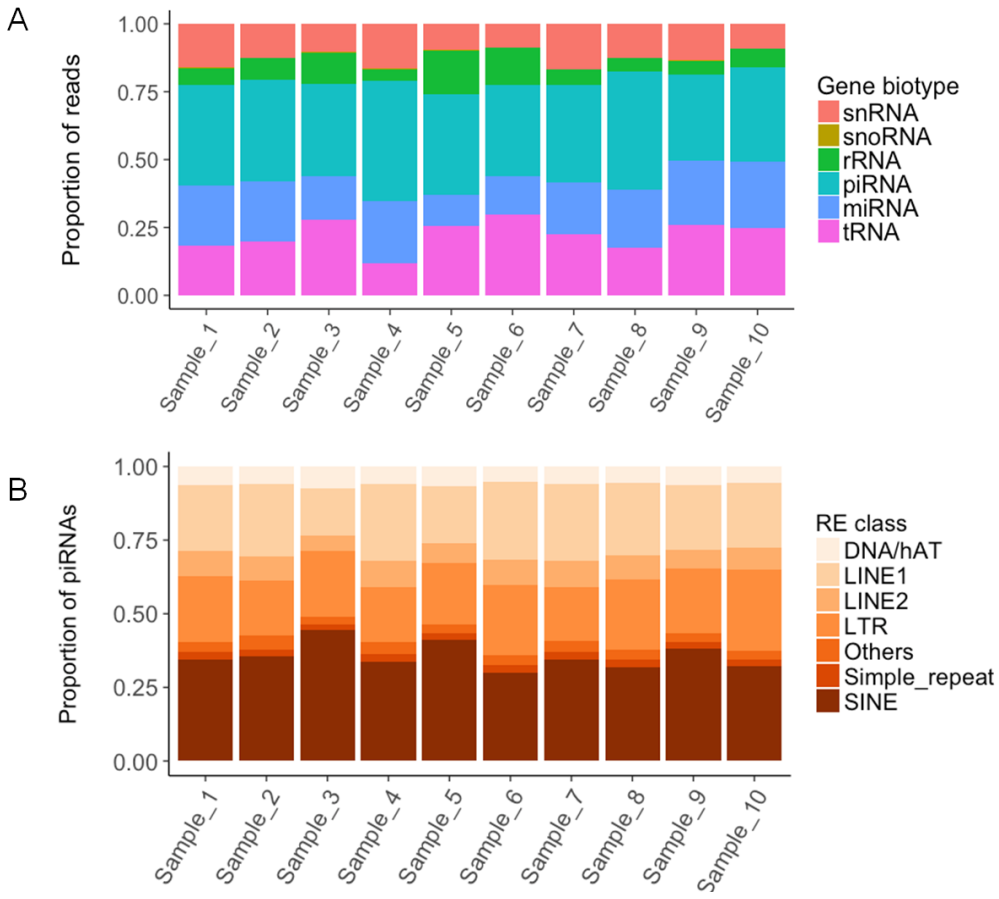


Figure 2. Read mapping distribution of the short non-coding RNA types and piRNA distribution within the Repetitive Element classes. (A). Proportion of reads mapping to each short non-coding RNA type. (B). Distribution within each Repetitive Element class of the piRNA cluster reads overlapping with Repetitive Elements.

We then assessed the CV across the sperm samples to evaluate their abundance stability (**Supplementary File S13**). Interestingly, miRNAs showed large variability, 32% of them varied markedly (CV > 75%), including the highly abundant miR-34c, miR-30c-5p, miR-186, and miR-99a, with none showing low variability. As previously mentioned, exosome vesicles may also contribute in modulating the miRNA population of recipient cells. In fact, a recent study identified altered miRNA profiles in seminal plasma exosomes from azoospermic patients (Barcelo et al., 2018). We did not measure the pairwise correlation between the abundance of miRNAs and mRNAs because in their canonical function, miRNAs inhibit translation but have a small impact on the levels of the target mRNAs.

piRNAs are a class of 26-32 bp size sncRNAs that interact with Piwi proteins to

contribute important functions to germline development, epigenetic regulation and the silencing of transposable elements (O'Donnell and Boeke, 2007). We queried a public database of 501 piRNA clusters identified in pig testes (Rosenkranz, 2016), and found that 300 were represented in boar sperm and covered 5.03 Mb (0.20%) of the Scrofa10.2 genome assembly (**Supplementary File S13**). The RNA levels ranged between 3.2 and 5,242 CPMs and the cluster length between 5,077 and 114,717 bp. piRNA clusters tend to overlap with REs, in keeping with their role in genome inactivation and transposon regulation (O'Donnell and Boeke, 2007; Krawetz et al., 2011; Pantano et al., 2015; Gòdia et al., 2018b). In our work, 25% of the piRNA clusters co-localized with REs, most of which were SINEs (**Figure 2B**). As piRNAs are tissue-specific and we queried a testes database (Rosenkranz, 2016), we also carried a *de novo* prediction of piRNA clusters with proTRAC using the remaining unaligned reads (average of 1.1 M reads) (**Supplementary File S1**). We identified 17 novel potential clusters of average abundance and length of 11.3–585 CPMs and 2,357–56,029 bp, respectively, and as a whole, they covered 159.7 kb of the Sscrofa11.1 genome. Six of the novel clusters were present in the 10 samples and are thus considered of high confidence (**Supplementary File S14**).

tRNAs were the second most abundant RNA class in porcine sperm, and their abundance is related to metabolic processes (Sharma et al., 2016). We identified 315 putative tRNAs from which 63% showed large abundance variability across samples (CV > 75%) (**Supplementary File S13**). Although the role of tRNAs in germ cells and in the offspring's health is uncertain, independent studies have shown that tRNA levels can be altered in response to certain manipulations of the paternal diet (Sharma et al., 2016; Gòdia et al., 2018b).

Seasonal Differences in the Boar Sperm Transcriptome

A seasonal variation on semen quality and fertility has been observed in several animal species including the pig. During the warm summer months, as the

scrotum is unable to thermo-regulate, spermatogenesis is negatively affected and the number of sperm cells and their motility tend to decrease alongside with an increase on morphological abnormalities (Zasiadczyk et al., 2015; Rodriguez et al., 2017). This effect on semen quality and also fertility (Suriyasomboon et al., 2006) has been related to heat stress. The molecular mechanisms underlying this phenomenon remain unclear although links to oxidative stress and the production of reactive oxidative species (ROS), with the consequent damage on sperm membrane integrity, DNA damage, apoptosis, autophagy and reduction of mitochondrial activity have been proposed (Durairajanayagam et al., 2015; Argenti et al., 2018). In a recent study, Argenti et al. (2018) identified increased superoxide dismutase anti-oxidant activity in the sperm of boars raised in sub-tropical Brazil in the summer months probably as a molecular attempt to reduce the presence of ROS and sperm damage (Argenti et al., 2018). Moreover, dietary strategies based on supplementary Zinc (Li et al., 2017) and l-arginine (Chen et al., 2018) have been related to a reduction of oxidative stress and improvement on the epididymal function and boar sperm quality in summer.

We compared the transcriptome (mRNA transcripts and miRNA) of the sperm samples collected in the summer months (May: N = 1; July: N = 4) with those collected in winter (December: N = 2; January: N = 2; February: N = 1) in a temperate climate zone (latitude 42◦ N, 800 m above sea level) with average temperatures in December–February around 2–3ºC, 12ºC in May and 19ºC in July, but which easily peaks to highs above 30ºC during this month (data from Sant Pau de Segúries weather station according to the Meteorological Service of Catalonia). The semen quality of the summer and winter groups was not significantly different when compared with a T-test, although a trend was seen for sperm cell viability (p-value = 0.05), acrosome reaction (p-value = 0.09) and neck (p-value = 0.07) and tail (p-value = 0.08) morphological abnormalities. We

detected 36 transcripts displaying a significant difference in abundance. Of these, two transcripts corresponded to the same gene and they were not taken into account due to concerns on the transcript allocation carried by the software. From the 34 remaining transcripts, each from a different gene, 14 were up-regulated and 20 were down-regulated in the summer group (**Table 3**).

The most significant difference in gene abundance between both seasonal groups (q-value = 3.13×10^{-16} , FC = 5.15) corresponded to the minichromosome maintenance 8 homologous recombination repair factor (*MCM8*) gene (**Table 3**). *MCM8* is a helicase related to the initiation of eukaryotic genome replication and may be associated with the length of the reproductive lifespan and menopause. *MCM8* plays a role in gametogenesis due to its essential functions in DNA damage repair via homologous recombination of DNA double strand breaks (Lutzmann et al., 2012).

Another gene was StAR Related Lipid Transfer Domain Containing 9 (*STARD9*), which was down-regulated in the winter group, is a lipid binding gene that has been related to asthenospermia in humans (Mao et al., 2011). Moreover, the paralog *STARD6* has been linked to spermatogenesis and spermatozoa quality (Mao et al., 2011). This is in keeping with the fact that the spermatozoon is very sensitive to oxidative damage for several reasons including the high amount of the peroxidation-prone unsaturated fatty acids that are present in its plasma membrane (Aitken and De Iuliis, 2010). Another gene that was found down-regulated in the winter group is the Oxidative Stress Induced Growth Inhibitor 1 gene (*OSGIN1*). *OSGIN1* has been related to autophagy and oxidative stress and its encoded protein regulates both cell death and apoptosis in the airway epithelium (Sukkar and Harris, 2017). Its expression is induced by DNA damage, which is one of the key sperm parameters that increase in the warm summer months (Perez-Crespo et al., 2008). Since this gene has also been identified in the sperm lineage, it could

respond with a similar anti-oxidative role in front heat stress in sperm.

Table 3. Messenger RNA transcripts showing differential abundances in the summer versus the winter ejaculates.

Transcript ID	Gene ID	Log2 (FC)	p-value	q-value (FDR)
ENSSSCT00000058763	*NSUN6*	-9.62	7.00E-16	4.35E-12
ENSSSCT00000056639	*ATG16L1*	-7.75	1.03E-08	2.14E-05
ENSSSCT00000059752	*EHBP1*	-7.49	7.95E-08	1.35E-04
ENSSSCT00000059921	*CENPC*	-6.55	8.56E-07	1.06E-03
ENSSSCT00000012060	*MTPAP*	-6.51	3.16E-07	4.21E-04
ENSSSCT00000056608	*SMARCA2*	-6.48	1.66E-05	1.15E-02
ENSSSCT00000066205	*CNOT3*	-6.22	1.75E-06	1.72E-03
ENSSSCT00000014560	*KIF18A*	-6.01	6.79E-05	3.62E-02
ENSSSCT00000057538	*ZNF24*	-5.88	4.01E-05	2.34E-02
ENSSSCT00000018135	*AOAH*	-5.47	2.54E-05	1.58E-02
ENSSSCT00000015909	*PSMD13*	-3.76	1.60E-05	1.15E-02
ENSSSCT00000037719	*STARD9*	-2.63	2.18E-05	1.45E-02
ENSSSCT00000039055	*CPEB3*	-2.32	4.98E-05	2.81E-02
ENSSSCT00000039293	*MED13L*	-2.13	1.62E-05	1.15E-02
ENSSSCT00000043522	*OSGIN1*	-1.75	9.49E-06	7.69E-03
ENSSSCT00000012151	*CUL2*	1.66	5.66E-05	3.10E-02
ENSSSCT00000001457		4.44	6.31E-14	2.94E-10
ENSSSCT00000049515	*ZMYND10*	4.72	1.26E-06	1.31E-03
ENSSSCT00000011652	*TRUB1*	4.93	2.33E-05	1.50E-02
ENSSSCT00000049377	*NUP58*	5.14	9.56E-05	4.95E-02
ENSSSCT00000007716	*MCM8*	5.15	1.68E-20	3.13E-16
ENSSSCT00000035098	*ERBIN*	5.31	9.92E-06	7.71E-03
ENSSSCT00000031111	*ANKRD6*	5.53	2.58E-06	2.40E-03
ENSSSCT00000038311	*MCPH1*	5.65	4.31E-06	3.83E-03
ENSSSCT00000018344	*WDR70*	5.72	1.04E-06	1.18E-03
ENSSSCT00000037667	*ASCC1*	5.78	2.80E-05	1.69E-02
ENSSSCT00000002542	*FUT8*	6.00	5.14E-06	4.36E-03
ENSSSCT00000032033	*TMEM230*	6.01	2.08E-07	2.98E-04
ENSSSCT00000050364	*PDE3B*	6.45	1.62E-07	2.52E-04
ENSSSCT00000015769	*FBXO38*	6.48	3.51E-08	6.54E-05
ENSSSCT00000043281	*ZNF280D*	6.49	1.08E-06	1.18E-03
ENSSSCT00000064492	*ZNF629*	6.73	2.55E-09	6.80E-06
ENSSSCT00000028805	*ZNF583*	7.34	6.81E-10	2.12E-06
ENSSSCT00000030081	*NMNAT1*	7.50	1.59E-11	5.94E-08
ENSSSCT00000039133	*ATG16L1*	7.76	4.16E-09	9.69E-06
ENSSSCT00000038377	*RUNDC3B*	8.96	3.72E-16	3.47E-12

The list includes only these transcripts with q-val < 0.05 and log2 (FC) < -1.5 or >1.5. log2 (FC) > 0 indicate up-regulation in summer when compared to winter. Empty cells in the Gene ID column correspond to transcripts without gene symbol or description. FC: Fold-Change; FDR: False Discovery Rate.

The presence of RNA differences in ejaculated sperm in summer versus winter seasons has been previously interrogated using the microarray technology (Yang et al., 2010). In that study the authors identified 33 dysregulated transcripts, none of which was differentially abundant in our dataset. This lack of concordance between works could be due to both biological and technical reasons and is somewhat expected. First, the two studies interrogated different animal populations in different geographic locations. The study by Yang et al. (2010) focused on Duroc boars breed in a sub- tropical region in Taiwan (25◦N) whilst we screened Pietrain males from a sub-Mediterranean temperate climatic zone in Catalonia with warm summers and mildly cold winters (köppen classification Cfb; latitude 42◦N). Moreover, we used a RNA-seq approach targeting the whole transcriptome whilst Yang et al. (2010) employed a custom microarray interrogating only 708 target genes and by large, ignored the vast catalog of annotated genes.

We also identified 5 miRNAs down- and 2 miRNAs up-regulated in winter (**Table 4**). This set included miR-34c, which was one of the most abundant miRNAs in our study, as well as in the sperm of other species, and was down-regulated in the winter samples. The RNA levels of miR-34c were also down-regulated in the sperm of men and mice exposed to severe early life stress events (Dickson et al., 2018), and in the testis of cynomolgus monkeys exposed to testicular hyperthermia (Sakurai et al., 2016), thus suggesting a link between the seasonality of semen quality and miR-34c. miR-1249, up-regulated in the winter group, was also found to be altered in the semen of bulls with moderate fertility (Fagerlind et al., 2015). Members of the miR-106 family were recently associated with oxidative stress in several tissues and cell types. For example, miR-106b targets the 12/15-Lipoxygenase enzymes, which are involved in the metabolism of fatty acids and oxidative stress in murine neurons (Wu et al., 2017). miR-106b has also been related to autophagy and cellular stress in

intestinal epithelial HCT116 cells (Zhai et al., 2013). A study in cattle identified a single nucleotide polymorphism in a miR-378 target site of the *INCENP* semen quality associated gene (Liu et al., 2016). In humans, miR-378 was found to also target the autophagy related protein 12 gene (*ATG12*) in cervical cancer (Tan et al., 2018). Finally, miR-221 was linked to autophagy in several tissues as well (Li et al., 2016; Qian et al., 2017) and was shown to regulate *SOD2*, which has key mitochondrial anti-oxidant functions in a murine model of ischemic skeletal muscle regeneration (Togliatto et al., 2013).

Table 4. List of the miRNAs showing distinct seasonal abundance.

miRNA ID	Log2 (FC)	p-value	q-value (FDR)
ssc-miR-221-3p	-2.70	4.19E-05	1.54E-03
ssc-miR-362	-1.81	1.63E-03	2.18E-02
ssc-miR-378	-1.71	6.16E-03	4.94E-02
ssc-miR-106a	-1.62	1.75E-05	1.29E-03
ssc-miR-34c	-1.53	5.87E-04	9.59E-03
ssc-miR-1306-5p	1.68	1.81E-04	3.81E-03
ssc-miR-1249	3.14	2.58E-08	3.79E-06

The list includes only these miRNAs with q-val < 0.05 and log2FC >1.5. FC: Fold-Change; FDR: False Discovery Rate

Our results are in consonance with previous reports suggesting that oxidative stress and autophagy are the key causes of the loss of semen quality in the warm summer periods (Suriyasomboon et al., 2006; Zasiadczyk et al., 2015). This data should be confirmed in a matched study where the winter and summer ejaculates come from the same boars using additional animals and several ejaculates per boar to account for non-genetic intra-individual variation.

Conclusion

We have identified a rich and complex sperm transcriptome with known and novel coding RNAs, lncRNAs and sncRNAs that resembles the human, mouse and cattle counterparts. Their roles are mainly related to the regulation of spermatogenesis, fertility and early embryo development. These spermatozoal transcripts are fragmented, likely in a selective manner, consistently affecting

some genes more than others across samples. This suggests that their fragmentation is not stochastic and follows an unknown deterministic pattern with potential functional implications. Similarly, the variability of the transcript abundance between samples was transcript specific. This in-depth transcriptome profile can be used as a reference to identify RNA markers for semen quality and male fertility in pigs and in other animal species.

Interestingly, the levels of some transcripts changed between the summer and the winter ejaculates, most likely responding to heat stress, which would in turn, cause oxidative stress, sperm membrane and DNA damage and autophagy. The biological basis of these transcriptome changes needs to be further explored. In the recent years it has become evident that the ejaculate contains different sub-populations of sperm, each with specific roles upon ejaculation. Each of these sub-populations may carry a specific transcriptome profile. Thus, the changes in transcript abundances that we identified may reflect either similar variations on the transcript's profile in all spermatozoa cells or on the contrary, may be attributed to changes in the proportion of sperm sub-populations each carrying their specific transcript profile. Discriminating both hypotheses could help defining the best strategies to mitigate this seasonal effect. Single-cell RNA-seq, a novel and powerful technology that still needs to be optimized in spermatozoa, could allow identifying the sperm sub-populations and their relevance for seasonality, semen quality and fertility. In conclusion, our results pave the way to carrying future research to understand the molecular basis of semen quality seasonality in pigs, humans and other affected species.

Ethics statement

The ejaculates obtained from pigs were privately owned for non-research purposes. The owners provided consent for the use of these samples for research. Specialized professionals at the farm obtained all the ejaculates

following standard routine monitoring procedures and relevant guidelines.

Author contributions

MG, AS, and AlC conceived and designed the experiments. SB collected the samples. JR-G carried the phenotypic analysis. MG performed sperm purifications and RNA extractions. AnC carried the qPCRs and their analysis. MG made the bioinformatics and statistic analysis. ME developed the SRE pipeline and provided bioinformatics support. MG analyzed the data, with special input from SK and AlC. MG and AlC wrote the manuscript. All authors discussed the data and read and approved the contents of the manuscript.

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under grant AGL2013- 44978-R and grant AGL2017-86946-R and by the CERCA Programme/Generalitat de Catalunya. AGL2017-86946-R was also funded by the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF). We thank the Agency for Management of University and Research Grants (AGAUR) of the Generalitat de Catalunya (Grant Number 2014 SGR 1528). We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016–2019 (Grant Number SEV-2015-0533) grant awarded to the Centre for Research in Agricultural Genomics (CRAG). MG acknowledges a Ph.D. studentship from MINECO (Grant Number BES-2014-070560) and a Short- Stay fellowship from MINECO (EEBB-I-2017-12229) at SK's laboratory. Funds through Charlotte B. Failing Professorship to SK are gratefully appreciated. AlC was recipient of a MINECO's Ramon y Cajal research fellow (Grant Number RYC-2011-07763).

Acknowledgments

We apologize for all the authors and articles that were not cited due to space

limitations.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Supplementary Material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00299/full#supplementary-material

FILE S1 | RNA-seq quality and mapping statistics. Average and Standard Deviation (SD) for the 10 boar sperm samples processed, including: amount of RNA extracted and several RNA-seq bioinformatics statistics for both total and small RNA-seq.

FILE S2 | Distribution of the top decile most abundant SREs (Sperm RNA Elements) into SRE types and gene biotypes. Number of SREs (within the top decile) for each SRE type (exonic, intronic, upstream/downstream 10 kb and orphan). Total non-redundant number of genes and their biotype for each SRE class.

FILE S3 | List of human and bovine genes identified by syntenic alignment of the orphan SREs. Orphan SRE genome coordinates were liftover to human and bovine coordinates, and the genes mapped in these regions were extracted. A total of 45 genes shared in both species were found. From these genes, 44 were already annotated in the *Sscrofa* Ensembl v.91 annotation. 17 of these genes were also detected by exonic, intronic and/or upstream/downstream 10 kb SREs. This suggests that orphan SREs could correspond to unannotated isoforms or to paralogous genes.

FILE S4 | Gene Ontology analysis of the genes including the top decile most abundant and the orphan SREs detected in the SRE pipeline. GO biological

process terms with significant Bonferroni corrected *p*-values (*p*-val < 0.05) and their associated genes.

FILE S5 | Gene Ontology analysis of the different SRE abundance variance groups. GO biological process terms with significant Bonferroni corrected *p*-values (*p*-val < 0.05) and their associated genes.

FILE S6 | Correlation between transcript integrity across samples, with transcript abundance and coding sequence length. Correlation of the TIN (Transcripts Integrity Number) between samples, with the transcript abundance and with the coding sequence length of the transcripts. This table shows the correlation of the TIN (Transcripts Integrity Number) between each pair of samples, the correlation of the TIN with the transcript average abundance in FPKM (Fragments per Kilobase per Million mapped reads) across the 10 samples, and the correlation of the TIN with the length of coding sequence of the transcripts.

FILE S7 | Summary statistics of the *de novo* transcriptome assembly. Summary statistics of the Trinity output based on the number of potential novel genes and transcripts, and size (in bp) of the contigs based on all transcripts isoforms or based only on the longest isoform for each potential gene.

FILE S8 | List of proteins identified by *de novo* analysis, with the species in which they were detected and transcript abundance. *De novo* analysis of the unmapped reads resulted in 1,060 proteins which passed the quality control filters. For each protein, we include the cognate species, the predicted RNA mean abundance in the 10 samples (in FPKM), the Standard Deviation (SD) of their RNA abundance and the gene ID symbol retrieved from Uniprot (https://www.uniprot.org/). FPKM: Fragments per Kilobase per Million mapped reads.

FILE S9 | Non-redundant list of genes identified by *de novo* analysis. 768

potentially novel genes were identified from the unmapped reads. The gene symbol IDs were retrieved with Uniprot from the Trinity output protein names. These genes were detected in at least one species (detailed in column 2 of **Supplementary File S8**). The majority of these genes were annotated in the porcine Ensembl v.91 but 29 were identified as novel genes. 40 of the genes annotated in the porcine genome were not detected with the SREs pipeline which indicates that none of their cognate reads mapped to the genome even though these genes are annotated.

FILE S10 | List of long non-coding RNAs detected in porcine sperm. Ensembl IDs of the lncRNAs identified in this study, their genome coordinates, average RNA abundance across the 10 samples and length. Most of the lncRNAs presented, as an average across all samples, low RNA abundances.

FILE S11 | Distribution of the short RNA-seq reads mapping to different RNA types. Proportion and standard deviation (SD) across the 10 samples.

FILE S12 | Concordance of miRNA identification between our dataset and other sperm RNA-seq studies. Comparison of the miRNAs identified in our study with other sperm RNA-seq experiments in pig, in human, and cattle.

FILE S13 | RNA abundance levels and coefficient of variation of miRNAs, tRNAs, and piRNAs in the porcine sperm. RNA abundance is measured in CPM (Counts Per Million) across the 10 samples. We only considered the miRNAS with >0 CPMs in all the samples. The genomic coordinates of piRNAs refer to the Sscrofa10.2 built instead of Sscrofa11.1 as provided by the piRNAs cluster database [40].

FILE S14 | Novel piRNA clusters identified in the pig sperm RNA. We detected 17 potential clusters of piRNAs that were found in at least 3 of the 10 samples analyzed in this study. Mean and standard deviation (SD) in CPM (Counts Per Million).

References

Aitken, R. J., and De Iuliis, G. N. (2010). On the possible origins of DNA damage in human spermatozoa. *Mol. Hum. Reprod.* 16, 3–13. doi: 10.1093/molehr/gap059

Argenti, L. E., Parmeggiani, B. S., Leipnitz, G., Weber, A., Pereira, G. R., and Bustamante-Filho, I. C. (2018). Effects of season on boar semen parameters and antioxidant enzymes in the south subtropical region in Brazil. *Andrologia* doi: 10.1111/and.12951

Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6:11. doi: 10.1186/ s13100- 015- 0041- 9

Barcelo, M., Mata, A., Bassas, L., and Larriba, S. (2018). Exosomal microRNAs in seminal plasma are markers of the origin of azoospermia and can predict the presence of sperm in testicular tissue. *Hum. Reprod.* 33, 1087–1098. doi: 10.1093/humrep/dey072

Beraldi, R., Pittoggi, C., Sciamanna, I., Mattei, E., and Spadafora, C. (2006). Expression of LINE-1 retroposons is essential for murine preimplantation development. *Mol. Reprod. Dev.* 73, 279–287. doi: 10.1002/mrd.20423

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10. 1093/bioinformatics/btu170

Bruggmann, R., Jagannathan, V., and Braunschweig, M. (2013). In search of epigenetic marks in testes and sperm cells of differentially fed boars. *PLoS One* 8:e78691. doi: 10.1371/journal.pone.0078691

Bzymek, M., and Lovett, S. T. (2001). Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* 98, 8319–8325. doi: 10.1073/pnas.111008398

Capra, E., Turri, F., Lazzari, B., Cremonesi, P., Gliozzi, T. M., Fojadelli, I., et al. (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* 18:14. doi: 10.1186/s12864-01 6-3394-7

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, D184–D189. doi: 10.1093/nar/gkv1309

Chen, C., Wu, H., Shen, D., Wang, S. S., Zhang, L., Wang, X. Y., et al. (2017a). Comparative profiling of small RNAs of pig seminal plasma and ejaculated and epididymal sperm. *Reproduction* 153, 785–796. doi: 10.1530/REP-17-0014

Chen, X. X., Che, D. X., Zhang, P. F., Li, X. L., Yuan, Q. Q., Liu, T. T., et al. (2017b). Profiling of miRNAs in porcine germ cells during spermatogenesis. *Reproduction* 154, 789–798. doi: 10.1530/REP-17-0441

Chen, C. Y., Ai, H. S., Ren, J., Li, W. B., Li, P. H., Qiao, R. M., et al. (2011). A global view of porcine transcriptome in three tissues from a full-sib pair with extreme phenotypes in growth and fat deposition by paired-end RNA sequencing. *BMC Genomics* 12:448. doi: 10.1186/1471-2164-12-448

Chen, J. Q., Li, Y. S., Li, Z. J., Lu, H. X., Zhu, P. Q., and Li, C. M. (2018). Dietary l-arginine supplementation improves semen quality and libido of boars under high ambient temperature. *Animal* 12, 1611–1620. doi: 10.1017/ S1751731117003147

Corominas, J., Ramayo-Caldas, Y., Puig-Oliveras, A., Estellé, J., Castelló, A., Alves, E., et al. (2013). Analysis of porcine adipose tissue transcriptome reveals differences in de novo fatty acid synthesis in pigs with divergent muscle fatty acid composition. *BMC Genomics* 14:843. doi: 10.1186/1471-2164-14-843

Das, P. J., McCarthy, F., Vishnoi, M., Paria, N., Gresham, C., Li, G., et al. (2013). Stallion sperm transcriptome comprises functionally coherent coding and regulatory RNAs as revealed by microarray analysis and RNA-seq. *PLoS One* 8:e56535. doi: 10.1371/journal.pone.0056535

Dickson, D. A., Paulus, J. K., Mensah, V., Lem, J., Saavedra-Rodriguez, L., Gentry, A., et al. (2018). Reduced levels of miRNAs 449 and 34 in sperm of mice and men exposed to early life stress. *Transl. Psychiatry* 8:101. doi: 10.1038/ s41398- 018- 0146- 2

Durairajanayagam, D., Agarwal, A., and Ong, C. (2015). Causes, effects and molecular mechanisms of testicular heat stress. *Reprod. Biomed.* 30, 14–27. doi: 10.1016/j.rbmo.2014.09.018

Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi: 10.1038/nprot.2009.97

Esteve-Codina, A., Kofler, R., Palmieri, N., Bussotti, G., Notredame, C., and Pérez- Enciso, M. (2011). Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC Genomics* 12:552. doi: 10.1186/1471-2164-12-552

Estill, M. S., Hauser, R., and Krawetz, S. A. (2019). RNA element discovery from germ cell to blastocyst. *Nucleic Acids Res.* 47, 2263–2275. doi: 10.1093/nar/ gky1223

Fagerlind, M., Stalhammar, H., Olsson, B., and Klinga-Levan, K. (2015).

Expression of miRNAs in bull spermatozoa correlates with fertility rates. *Reprod. Domest. Anim.* 50, 587–594. doi: 10.1111/rda.12531

Gòdia, M., Mayer, F. Q., Nafissi, J., Castelló, A., Rodríguez-Gil, J. E., Sánchez, A., et al. (2018a). A technical assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-seq analysis. *Syst. Biol. Reprod. Med.* 64, 291–303. doi: 10.1080/19396368.2018.1464610

Gòdia, M., Swanson, G., and Krawetz, S. A. (2018b). A history of why fathers' RNA matters. *Biol. Reprod.* 99, 147–159. doi: 10.1093/biolre/ioy007

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Jodar, M., Sendler, E., and Krawetz, S. A. (2016). The protein and transcript profiles of human semen. *Cell Tissue Res.* 363, 85–96. doi: 10.1007/s00441-01 5-2237-1

Jodar, M., Sendler, E., Moskovtsev, S. I., Librach, C. L., Goodrich, R., Swanson, S., et al. (2015). Absence of sperm RNA elements correlates with idiopathic male infertility. *Sci. Transl. Med.* 7:295re296. doi: 10.1126/scitranslmed. aab1287.

Johnson, G. D., Mackie, P., Jodar, M., Moskovtsev, S., and Krawetz, S. A. (2015). Chromatin and extracellular vesicle associated sperm RNAs. *Nucleic Acids Res.* 43, 6847–6859. doi: 10.1093/nar/gkv591

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39, D152–D157. doi: 10.1093/nar/gkq1027

Krawetz, S. A., Kruger, A., Lalancette, C., Tagett, R., Anton, E., Draghici, S., et al. (2011). A survey of small RNAs in human sperm. *Hum. Reprod.* 26, 3401–3412. doi: 10.1093/humrep/der329

Kuhn, R. M., Haussler, D., and Kent, W. J. (2013). The UCSC genome browser and associated tools. *Brief. Bioinfomatics* 14, 144–161. doi: 10.1093/bib/bbs038

Li, L., Wang, Z., Hu, X., Wan, T., Wu, H., Jiang, W., et al. (2016). Human aortic smooth muscle cell-derived exosomal miR-221/222 inhibits autophagy via a PTEN/Akt signaling pathway in human umbilical vein endothelial cells. *Biochem. Biophys. Res. Commun.* 479, 343–350. doi: 10.1016/j.bbrc.2016. 09.078

Li, Z., Li, Y., Zhou, X., Cao, Y., and Li, C. (2017). Preventive effects of supplemental dietary zinc on heat-induced damage in the epididymis of boars. *J. Therm. Biol.* 64, 58–66. doi: 10.1016/j.jtherbio.2017.01.002

Liu, J., Sun, Y., Yang, C., Zhang, Y., Jiang, Q., Huang, J., et al. (2016). Functional SNPs of INCENP affect semen quality by alternative splicing mode and binding affinity with the target bta-miR-378 in Chinese holstein bulls. *PLoS One* 11:e0162730. doi: 10.1371/journal.pone.0162730

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Luo, Z. G., Liu, Y. K., Chen, L., Ellis, M., Li, M. Z., Wang, J. Y., et al. (2015). microRNA profiling in three main stages during porcine spermatogenesis. *J. Assist. Reprod. Genet.* 32, 451–460. doi: 10.1007/s10815-014-0406-x

Lutzmann, M., Grey, C., Traver, S., Ganier, O., Maya-Mendoza, A., Ranisavljevic, N., et al. (2012). MCM8- and MCM9-deficient mice reveal gametogenesis defects and genome instability due to impaired homologous recombination. *Mol. Cell* 47, 523–534. doi: 10.1016/j.molcel.2012.05.048

Ma, X., Zhu, Y., Li, C., Xue, P., Zhao, Y., Chen, S., et al. (2014). Characterisation of *Caenorhabditis elegans* sperm transcriptome and proteome. *BMC Genomics* 15:168. doi: 10.1186/1471-2164-15-168

Mach, N., Berri, M., Esquerre, D., Chevaleyre, C., Lemonnier, G., Billon, Y., et al. (2014). Extensive expression differences along porcine small intestine evidenced by transcriptome sequencing. *PLoS One* 9:e88515. doi: 10.1371/journal.pone. 0088515

Mao, X. M., Xing, R. W., Jing, X. W., Zhou, Q. Z., Yu, Q. F., Guo, W. B., et al. (2011). [Differentially expressed genes in asthenospermia: a bioinformatics-based study]. *Zhonghua Nan Ke Xue* 17, 694–698.

Marques, D. B. D., Lopes, M. S., Broekhuijse, M., Guimaraes, S. E. F., Knol, E. F., Bastiaansen, J. W. M., et al. (2017). Genetic parameters for semen quality and quantity traits in five pig lines. *J. Anim. Sci.* 95, 4251–4259. doi: 10.2527/jas2017. 1683

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Netherton, J. K., Hetherington, L., Ogle, R. A., Velkov, T., and Baker, M. A. (2018). Proteomic analysis of good- and poor-quality human sperm demonstrates that several proteins are routinely aberrantly regulated. *Biol. Reprod.* 99, 395–408. doi: 10.1093/biolre/iox166

O'Donnell, K. A., and Boeke, J. D. (2007). Mighty Piwis defend the germline against genome intruders. *Cell* 129, 37–44. doi: 10.1016/j.cell.2007.03.028

OECD (2018). *Meat Consumption (indicator)* [Online]. Available at: https://www. oecd-ilibrary.org/content/data/fa290fd0-en (accessed June 15, 2018).

Pantano, L., Jodar, M., Bak, M., Ballesca, J. L., Tommerup, N., Oliva, R., et al. (2015). The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes. *RNA* 21, 1085–1095. doi: 10.1261/rna.046482.114

Perez-Crespo, M., Pintado, B., and Gutierrez-Adan, A. (2008). Scrotal heat stress effects on sperm viability, sperm DNA integrity, and the offspring sex ratio in mice. *Mol. Reprod. Dev.* 75, 40–47. doi: 10.1002/mrd.20759

Pérez-Montarelo, D., Madsen, O., Alves, E., Rodréguez, M. C., Folch, J. M., Noguera, J. L., et al. (2014). Identification of genes regulating growth and fatness traits in pig through hypothalamic transcriptome analysis. *Physiol. Genomics* 46, 195–206. doi: 10.1152/physiolgenomics.00151.2013

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Pertea, M., Shumate, A., Pertea, G., Varabyou, A., Chang, Y.-C., Madugundu, A. K., et al. (2018). Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* [Preprint]. doi: 10.1101/332825

Qian, L. B., Jiang, S. Z., Tang, X. Q., Zhang, J., Liang, Y. Q., Yu, H. T., et al. (2017). Exacerbation of diabetic cardiac hypertrophy in OVE26 mice by angiotensin II is associated with JNK/c-Jun/miR-221-mediated autophagy inhibition. *Oncotarget* 8, 106661–106671. doi: 10.18632/oncotarget. 21302

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rodriguez, A. L., Van Soom, A., Arsenakis, I., and Maes, D. (2017). Boar management and semen handling factors affect the quality of boar extended semen. *Porcine Health Manage* 3:15. doi: 10.1186/s40813-017-0062-5

Rodriguez-Gil, J. E., and Bonet, S. (2016). Current knowledge on boar sperm metabolism: comparison with other mammalian species. *Theriogenology* 85, 4–11. doi: 10.1016/j.theriogenology.2015.05.005

Rosenkranz, D. (2016). piRNA cluster database: a web resource for piRNA producing loci. *Nucleic Acids Res.* 44, D223–D230. doi: 10.1093/nar/gkv1265

Rosenkranz, D., and Zischler, H. (2012). proTRAC–a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* 13:5. doi: 10.1186/1471-2105-13-5

Rueda, A., Barturen, G., Lebron, R., Gomez-Martin, C., Alganza, A., Oliver, J. L., et al. (2015). sRNAtoolbox: an integrated collection of small RNA research

tools. *Nucleic Acids Res.* 43, W467–W473. doi: 10.1093/nar/gkv555

Sakurai, K., Mikamoto, K., Shirai, M., Iguchi, T., Ito, K., Takasaki, W., et al. (2016). MicroRNA profiles in a monkey testicular injury model induced by testicular hyperthermia. *J. Appl. Toxicol.* 36, 1614–1621. doi: 10.1002/jat.3326

Salas-Huetos, A., Blanco, J., Vidal, F., Godo, A., Grossmann, M., Pons, M. C., et al. (2015). Spermatozoa from patients with seminal alterations exhibit a differential micro-ribonucleic acid profile. *Fertil. Steril.* 104, 591–601. doi: 10. 1016/j.fertnstert.2015.06.015

Schmid, C., Heng, H. H. Q., Rubin, C., Ye, C. J., and Krawetz, S. A. (2001). Sperm nuclear matrix association of the PRM1 -(PRM2 -(TNP2 domain is independent of Alu methylation. *Mol. Hum. Reprod.* 7, 903–911. doi: 10.1093/ molehr/7.10.903

Selvaraju, S., Parthipan, S., Somashekar, L., Kolte, A. P., Krishnan Binsila, B., Arangasamy, A., et al. (2017). Occurrence and functional significance of the transcriptome in bovine (*Bos taurus*) spermatozoa. *Sci. Rep.* 7:42392. doi: 10. 1038/srep42392

Sendler, E., Johnson, G. D., Mao, S., Goodrich, R. J., Diamond, M. P., Hauser, R., et al. (2013). Stability, delivery and functions of human sperm RNAs at fertilization. *Nucleic Acids Res.* 41, 4104–4117. doi: 10.1093/nar/gkt132

Shan, L., Wu, Q., Li, Y. L., Shang, H. T., Guo, K. N., Wu, J. Y., et al. (2014). Transcriptome profiling identifies differentially expressed genes in postnatal developing pituitary gland of miniature pig. *DNA Res.* 21, 207–216. doi: 10. 1093/dnares/dst051

Sharma, U., Conine, C. C., Shea, J. M., Boskovic, A., Derr, A. G., Bing, X. Y., et al. (2016). Biogenesis and function of tRNA fragments during sperm maturation and fertilization in mammals. *Science* 351, 391–396. doi: 10.1126/science. aad6780

Sukkar, M. B., and Harris, J. (2017). Potential impact of oxidative stress induced growth inhibitor 1 (OSGIN1) on airway epithelial cell autophagy in chronic obstructive pulmonary disease (COPD). *J. Thorac. Dis.* 9, 4825–4827. doi: 10. 21037/jtd.2017.10.153

Suriyasomboon, A., Lundeheim, N., Kunavongkrit, A., and Einarsson, S. (2006). Effect of temperature and humidity on reproductive performance of crossbred sows in Thailand. *Theriogenology* 65, 606–628. doi: 10.1016/j.theriogenology. 2005.06.005

Tan, D., Zhou, C., Han, S., Hou, X., Kang, S., and Zhang, Y. (2018). MicroRNA-378 enhances migration and invasion in cervical cancer by directly targeting autophagy-related protein 12. *Mol. Med. Rep.* 17, 6319–6326. doi: 10.3892/mmr. 2018.8645

Tian, H., Li, Z. L., Peng, D., Bai, X. G., and Liang, W. B. (2017). Expression

difference of miR-10b and miR-135b between the fertile and infertile semen samples (p). *Forensic Sci. Int. Genet. Suppl. Ser.* 6, E257–E259. doi: 10.1016/j.fsigss.2017.09.092

Togliatto, G., Trombetta, A., Dentelli, P., Cotogni, P., Rosso, A., Tschop, M. H., et al. (2013). Unacylated ghrelin promotes skeletal muscle regeneration following hindlimb ischemia via SOD-2-mediated miR-221/222 expression. *J. Am. Heart. Assoc.* 2:e000376. doi: 10.1161/JAHA.113.000376

Trudeau, V., and Sanford, L. M. (1986). Effect of season and social environment on testis size and semen quality of the adult Landrace boar. *J. Anim. Sci.* 63, 1211–1219. doi: 10.2527/jas1986.6341211x

Vojtech, L., Woo, S., Hughes, S., Levy, C., Ballweber, L., Sauteraud, R. P., et al. (2014). Exosomes in human semen carry a distinctive repertoire of small non-coding RNAs with potential regulatory functions. *Nucleic Acids Res.* 42, 7290–7304. doi: 10.1093/nar/gku347

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28, 2184–2185. doi: 10.1093/bioinformatics/bts356

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520–562. doi: 10.1038/nature01262

Wettemann, R. P., Wells, M. E., Omtvedt, I. T., Pope, C. E., and Turman, E. J. (1976). Influence of elevated ambient temperature on reproductive performance of boars. *J. Anim. Sci.* 42, 664–669. doi: 10.2527/jas1976.423664x

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb- 2014- 15- 3- r46

Wu, Y., Xu, D., Zhu, X., Yang, G., and Ren, M. (2017). MiR-106a associated with diabetic peripheral neuropathy through the regulation of 12/15-LOX-meidiated oxidative/nitrative stress. *Curr. Neurovasc. Res.* 14, 117–124. doi: 10.2174/1567202614666170404115912

Yang, C. C., Lin, Y. S., Hsu, C. C., Tsai, M. H., Wu, S. C., and Cheng, W. T. (2010). Seasonal effect on sperm messenger RNA profile of domestic swine (*Sus Scrofa*). *Anim. Reprod. Sci.* 119, 76–84. doi: 10.1016/j.anireprosci.2009.12.002

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, 143–155. doi: 10.1016/j.anireprosci.2008.08.012

Yu, J., Wu, H., Wen, Y., Liu, Y. J., Zhou, T., Ni, B. X., et al. (2015). Identification of seven genes essential for male fertility through a genome-wide association study of non-obstructive azoospermia and RNA interference-mediated large-

scale functional screening in Drosophila. *Hum. Mol. Genet.* 24, 1493–1503. doi: 10.1093/hmg/ddu557

Yuan, S., Tang, C., Zhang, Y., Wu, J., Bao, J., Zheng, H., et al. (2015). mir-34b/c and mir-449a/b/c are required for spermatogenesis, but not for the first cleavage division in mice. *Biol. Open* 4, 212–223. doi: 10.1242/bio.201410959

Zasiadczyk, L., Fraser, L., Kordan, W., and Wasilewska, K. (2015). Individual and seasonal variations in the quality of fractionated boar ejaculates. *Theriogenology* 83, 1287–1303. doi: 10.1016/j.theriogenology.2015.01.015

Zhai, Z., Wu, F., Chuang, A. Y., and Kwon, J. H. (2013). miR-106b fine tunes ATG16L1 expression and autophagic activity in intestinal epithelial HCT116 cells. *Inflamm. Bowel Dis.* 19, 2295–2301. doi: 10.1097/MIB.0b013e31829e71cf

Zhang, X., Gao, F., Fu, J., Zhang, P., Wang, Y., and Zeng, X. (2017). Systematic identification and characterization of long non-coding RNAs in mouse mature sperm. *PLoS One* 12:e0173402. doi: 10.1371/journal.pone.0173402

Identification of circular RNAs in porcine sperm and their relation to sperm motility

Marta Gòdia[1], Anna Castelló[1,2], Martina Rocco[1,3], Betlem Cabrera[1,2], Joan E. Rodríguez-Gil[3], Armand Sánchez [1,2] and Alex Clop[1,4*]

[1]Animal Genomics Group, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[2]Unit of Animal Science, Department of Animal and Food Science, Autonomous University of Barcelona, Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[3]Unit of Animal Reproduction, Department of Animal Medicine and Surgery, Autonomous University of Barcelona, Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[4]Consejo Superior de Investigaciones Científicas (CSIC), Barcelona, Catalonia, Spain.

*Corresponding author:

Submitted

Available at: bioRxiv 2019

doi: https://doi.org/10.1101/608026

Abstract

Background

Circular RNAs (circRNAs) are emerging as a novel class of noncoding RNAs which potential role as gene regulators is quickly gaining interest. Although circRNAs have been studied in several tissues and cell types across several animal species, the characterization of the circRNAome in ejaculated sperm remains unexplored. In this study, we profiled the sperm circRNA catalogue in 40 boar samples.

Results

A complex population of 1,598 circRNAs was shared in at least 30 samples. The predicted circRNAs presented in general low abundances and were highly tissue-specific. Circa 80% of the circRNAs identified in the boar sperm were reported as novel. We also constructed a circRNA-miRNA interaction network based on experimental and predictive microRNA (miRNA) binding sites. Moreover, we found significant correlation between the abundance of some circRNAs and sperm motility parameters and confirmed two of these correlations by RT-qPCR in 36 samples with extreme sperm quality.

Conclusions

Our study provides a thorough characterization of a novel cell type for the circRNA encyclopedia collection and suggests that circRNAs are potential noninvasive biomarkers for male sperm quality and thus, might also hold potential to predict male fertility.

Keywords: sperm, circular RNA (circRNA), sperm motility, sperm quality, swine, male infertility.

Background

Food scarcity is a growing concern for our society due to human population growth and climate change. Swine and poultry are the main sources of meat worldwide and large efforts are being taken to increase their production in a sustainable manner. In this sense, reproductive efficiency is one of the key aspects for the sustainability of animal breeding [1]. In addition, due to their similarity in genome sequence, anatomy and physiology, swine is quickly becoming an important model for human bio-medical research [2]. In humans, infertility is an increasing problem in contemporary society, affecting one in twenty males [3], and has become a subject of bio-medical research in swine [4-6]. Unlike in humans, where fertility data is based on few records per person, the swine industry has begun to record fertility traits each time a male is used for artificial insemination (AI). In pig intensive production systems, during its productive life, an AI boar typically inseminates around 1,500 sows (S Balasch, pers. comm.). Moreover, the porcine ejaculates are routinely evaluated in the AI centers after every extraction. Thus, large datasets of the reproductive ability of these boars is becoming available. Taken all together, swine becomes a good animal model to study semen quality and male fertility.

Sperm motility and kinetic parameters provide an objective and reproducible measurement of semen quality that is automatically assessed by the computer-assisted semen analysis (CASA) system. CASA records the percentage of total motile sperm, curvilinear velocity (VCL), straight line velocity (VSL) and velocity of the sperm cells (VAP) and head frequency based on trajectories of motile sperm. This approach has been commonly used to assess semen quality in animal breeding strategies prior to AI in cattle [7], horse [8] and swine [9-11] where significant correlations between sperm motility and fertility have been found. In humans, this technique is also applied to estimate the *in vitro*

fertilizing potential of the ejaculates used in assisted reproductive treatments [12-14].

Multiple research efforts have demonstrated that sperm quality parameters and fertility outcomes are related to the presence or absence of sperm RNAs. Different studies have provided evidence that the absence or deregulation of certain RNAs is associated to infertility and/or sperm motility as in human [15], mice [16] and cattle [17]. Studies have focused their research on messenger RNAs (mRNAs) and on different classes of non-coding RNAs. In humans, several microRNAs (miRNAs) have been found to be dysregulated in infertile patients [18]. Similarly in bulls, Capra *et al.* found differential abundances of some miRNAs and transference RNAs (tRNAs) between high and low motility sperm populations [19].

Circular RNAs (circRNAs) are a novel class of non-coding RNAs with a closed loop structure, mainly formed through pre-mRNA back splicing event [20]. CircRNAs are highly stable *in vivo* in comparison to their mRNA linear counterparts, because of their circular structure, which confers protection against exonucleases, ribozymes, antisense RNAs and small-interfering RNAs [21, 22]. Expanding views on their biogenesis and function suggests that circRNAs can directly regulate the abundance of their cognate mRNA and can also act as miRNA sponges [23], sequestering these miRNAs and thus impeding their post-transcriptional inhibitory roles on target mRNAs [23]. circRNAs have been identified across several species and tissues including the fruitfly [24], human [25, 26], mice [26, 27] and swine [28, 29]. These studies revealed species-, developmental- and tissue- specific expression patterns.

During the last few years, there has been a considerable interest in the potential use of circRNAs as biomarkers for health. Several studies have identified circRNAs which abundances were associated to cancer, aneurysms, hypertension, heart failure, diabetes and arthritis, among others [reviewed in

30]. In addition, there are few reports characterizing circRNAs in reproductive organs and cell types including oocytes, embryo, placental tissue, granulosa cells, immature spermatogenic cells and testis [reviewed in 31]. Up to date, a small number of studies have assessed the circRNA predictive potential for reproduction outcomes. For example, Chang and colleagues identified circRNAs associated to embryo quality and suggested their role as potential predictors of live birth [32] and Qian *et al.* identified differentially expressed circRNAs in the placenta of pregnant woman affected with preeclampsia [33].

Our group has recently carried a thorough porcine sperm transcriptome analysis [34] and the data, in line with previous studies [reviewed in 35], showed that the majority of the transcripts are highly fragmented and with low abundances. Considering their stability and abundance, circRNAs could hold an important potential as reliable biomarkers for sperm quality and fertility traits. Here, we have characterized the circRNA repertoire of the porcine spermatozoa, assessed their potential role as miRNA sponges and investigated the relationship between their abundance levels and sperm motility.

Results

Characterization of the sperm circular RNA repertoire

1,598 potential circRNAs were present and shared in at least 30 of the 40 ejaculates (Additional file 1). The majority of the circRNA species were derived from exonic regions (CDS, 3' and 5' UTR) (82.1%), while only 13.5% and 4.4% circRNAs originated from intergenic and intronic segments, respectively (Figure 1A). Most circRNAs included less than 4 exons (81.0%), and just a few (14 circRNAs) contained 10 or more exons (Figure 1B). In addition, the majority of the exonic circRNAs (76.9%) were less than 400 bp long (Figure 1C). RNA abundance across the different circRNAs types was low, with a range between 0.19 and 136.4 CPMs and mean and median values of 2.42 and 0.89 CPMs, respectively (Additional file 1). The top 20 most abundant circRNAs (Table 1)

involved 15 coding genes and one annotated long non-coding RNA. These genes included *CEP63*, *ATP6V0A2*, *PPA2*, *PAIP2* and *PAXIP1*, which have been linked to sperm related traits and male fertility.

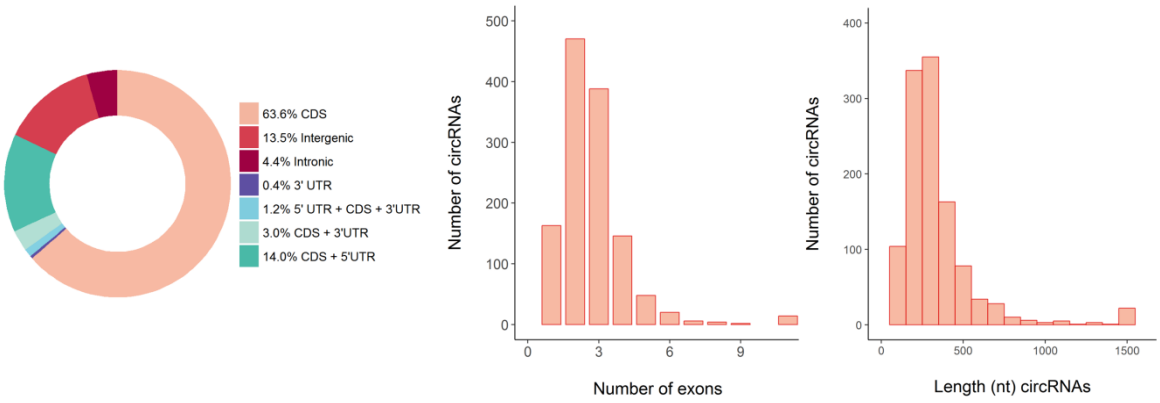


Figure 1. Characteristics of the genomic features of the circRNAs identified in the boar sperm. **A.** Distribution of the genomic location (CDS, intergenic, intronic, 3' UTR or 5'UTR) of the 1,598 circRNAs identified in the boar sperm. **B.** Distribution of the number of exons that form the exonic circRNAs. **C.** Distribution of the nucleotide length of the exonic circRNAs.

Only a small fraction of genes produce more than one circRNA. These are considered hotspot circRNAs genes. The circRNAs from hotspot genes are often produced from different back-splicing events of one exon with several others [29]. We detected 12 genes with 5 or more circRNA isoforms (Table 2). Some of these genes, namely *TESK2*, *SPATA19*, *PTK2* and *SLC5A10* are related to sperm function and fertility.

Table 1. List of the 20 most abundant circRNAs in the swine sperm. circRNA genomic coordinates are indicated as chromosome:start_position..end_position. Mean and SD: standard deviation, are in CPM (Counts Per Million).

circRNA name	circRNA genomic coordinates	Mean	SD	circRNA type	Ensembl ID	Host gene symbol
ssc_circ_0493	13:75694364..75694548	136.4	80.7	intronic	ENSSSCG00000011645	*CEP63*
ssc_circ_0220	11:17636879..17653264	128.2	52.5	intergenic		
ssc_circ_0475	13:61967873..61969753	100.2	56.8	lincRNA	ENSSSCG00000035295	
ssc_circ_1097	4:590305..590427	75.8	51.9	CDS	ENSSSCG00000005916	*WDR97*
ssc_circ_1141	5:22008783..22009750	75.0	35.4	CDS	ENSSSCG00000000406	*PTGES3*
ssc_circ_1062	4:130322062..130339222	62.2	33.1	CDS	ENSSSCG00000006939	*ZNHIT6*
ssc_circ_0537	14:29281863..29290551	59.9	34.8	CDS	ENSSSCG00000009766	*ATP6V0A2*
ssc_circ_0777	18:14243637..14259662	47.3	23.9	CDS	ENSSSCG00000016529	*AGBL3*
ssc_circ_0860	2:3096936..3098647	45.1	35.6	CDS	ENSSSCG00000037451	*PPFIA1*
ssc_circ_0805	18:54260018..54270584	40.0	52.3	CDS	ENSSSCG00000035581	*SUGCT*
ssc_circ_0132	1:641933..653683	33.2	19.1	CDS	ENSSSCG00000004008	*WDR27*
ssc_circ_1452	8:56584279..56590933	28.7	21.9	intergenic		
ssc_circ_1575	AEMK02000682.1:1719140..1720285	28.4	24.6	CDS	ENSSSCG00000005753	*CAMSAP1*
ssc_circ_1413	8:116275909..116294261	26.9	15.5	CDS	ENSSSCG00000022788	*PPA2*
ssc_circ_0895	2:64840342..64840811	25.8	17.5	5′ UTR	ENSSSCG00000013776	*DDX39A*
ssc_circ_0102	1:264102211..264103554	24.3	22.1	intronic	ENSSSCG00000005582	*STRBP*
ssc_circ_0363	13:119893907..119895097	24.2	14.9	inetergenic		
ssc_circ_1101	4:6367950..6392035	22.6	28.8	CDS	ENSSSCG00000005941	*KHDRBS3*
ssc_circ_0839	2:141254413..141254577	21.8	20.9	CDS + 5′ UTR	ENSSSCG00000026606	*PAIP2*
ssc_circ_0795	18:3101570..3102902	21.7	9.0	CDS	ENSSSCG00000025221	*PAXIP1*

Table 2. circRNA hotspot genes in swine sperm. 12 genes harbored 5 or more exonic circRNAs.

Gene symbol	Ensembl ID	Number of circRNAs
DENND1B	ENSSSCG00000010900	8
DENND1A	ENSSSCG00000005585	7
TESK2	ENSSSCG00000003917	6
UIMC1	ENSSSCG00000022508	6
ARMC9	ENSSSCG00000023994	5
CAMSAP1	ENSSSCG00000005753	5
KDM5B	ENSSSCG00000010928	5
PTK2	ENSSSCG00000038397	5
RPS6KC1	ENSSSCG00000015586	5
SPATA19	ENSSSCG00000025612	5
SLC5A10	ENSSSCG00000018049	5
WNK1	ENSSSCG00000000753	5

Sperm circRNAs might be involved in epigenetic regulation and spermatogenesis

We analyzed the potential roles of the boar sperm circRNAs under the assumption that their function is related to their known mRNA counterpart. Gene Ontology (GO) analysis of the circRNAs host genes revealed an enrichment for epigenetic functions including histone modification (q-val: 5.52 x 10^{-6}), histone H3-K36 methylation (q-val: 8.65 x 10^{-3}) and chromatin organization (q-val: 2.16 x 10^{-8}) (Additional file 2). We also identified significant ontologies in spermatogenesis (q-val: 5.81 x 10^{-4}), cilium assembly (q-val: 4.15 x 10^{-3}) and developmental process (q-val: 1.33 x 10^{-2}), among others (Additional file 2).

The boar sperm has a highly specific circRNAome

We compared our circRNA catalogue with equivalent available datasets [36] from other studies in human (including different brain sections tissues and several cell lines) [26, 37-40], mice (several brain segments, cell types and embryonic stem cells) [26, 37] and swine (lung, skeletal muscle, fat, heart, liver, spleen, kidney, ovary, testis and 5 brain sections) [28, 29]. Twenty-four % and 11.3% of the boar circRNAs had potential human and mouse orthologs, respectively (Table 3). On the other hand, 20.3% of the porcine sperm circRNAs

were also present in other porcine tissues (Table 3). The tissues showing higher overlap with sperm were testes (11.6%) and cortex (11.3%). All the other tissues were clearly less concordant (Table 3). By comparing this in the opposite direction, the proportion of the circRNAs already annotated in any pig tissue that were also present in our pig sperm list was 4.9%. This value was 12 times lower (0.4%) when evaluating the human circRNA catalog (Table 3).

Table 3. Concordance between the circRNAs catalogue of the boar sperm and the circRNA list in other tissues.

Species / Tissues	Number of circRNAs	% swine sperm	% of the total circRNAs
Human	90,067	24.0%	0.4%
Mice	15,498	11.3%	1.2%
Swine	6,663	20.3%	4.9%
Basal ganglia	456	2.6%	9.2%
Brain stem	820	5.3%	10.2%
Cerebellum	1,061	5.6%	8.4%
Cortex	2,163	11.3%	8.4%
Hippocampus	549	3.2%	9.3%
Fat	494	2.4%	7.7%
Heart	539	2.2%	6.5%
Kidney	469	1.9%	6.4%
Liver	353	2.1%	9.3%
Lung	683	2.7%	6.3%
Muscle	532	2.6%	7.9%
Ovarium	652	3.8%	9.4%
Spleen	241	1.4%	9.1%
Testes	2,685	11.6%	6.9%

Number of circRNAs after Sscrofa11.1 liftover from human, mice and swine (detailed by tissue). Column % swine sperm shows the proportion of swine sperm circRNAs that were identified in the given species and tissues. Column % of the total circRNAs lists the proportion of circRNAs from that tissue or species that found a homolog in the boar sperm.

Sperm circRNAs do not follow an age-dependent pattern

We assessed whether sperm circRNAs depicted an age-accumulating profile as has been previously observed in rat testes [41] and rat and mouse brain tissues [41, 42]. All the boars from our dataset were mature (≥ 8 months old) with ages ranging between 9 and 54 months of age. Sexual maturity in boars is a process that starts at the age of 8 months and finalizes at the age of 2 years [43]. Thus, we divided the samples in those coming from boars approaching sexual maturity with ages below 2 years old and those produced by mature pigs with

ages above 2 years old. There was no significant difference in the number of circRNAs identified (P-value: 0.68, Wilcoxon rank sum test) nor in their RNA abundance (P-value: 0.948) between the two groups. We repeated the analysis considering only extreme ages: young (N=4; between 8.6 and 9.2 month old) and mature (N=4; 29.9 to 54.6 month old). Again, there was no difference in the number of circRNAs identified (P-value: 0.89) or in their abundance (P-value: 0.2).

circRNA-miRNA interaction network

To characterize the potential role of the circRNAs as miRNA sponges, we built a co-expression network including the abundances of the 95 miRNAs and the 1,598 circRNAs identified in the short and total RNA-seq datasets. The association analysis resulted in 2,323 significant interactions between the 95 miRNAs and 564 circRNAs. On the other hand, the *in silico* prediction of miRNA targets in the circRNA sequences, based on sequence complementary, yielded 4,987 potential targets involving all the miRNAs and 1,103 circRNAs. To reduce the proportion of false-positives, only the 70 interactions (from 31 miRNAs and 56 circRNAs) that were found shared in both methods were used for network visualization (Figure 2). Most circRNAs (46) presented one miRNA target site with the exception of 10 circRNAs that harbored 2 or more potential targets (Figure 2). As the ssc_cic_08954 and ssc_circ_1454 which can potentially regulate, each, 4 distinct miRNAs (Figure 2). On the other hand, miR-26a and miR-28-5p were predicted to be regulated by 9 different circRNAs and 5 different circRNAs regulated miR140-3p and miR-423-5p (Figure 2).

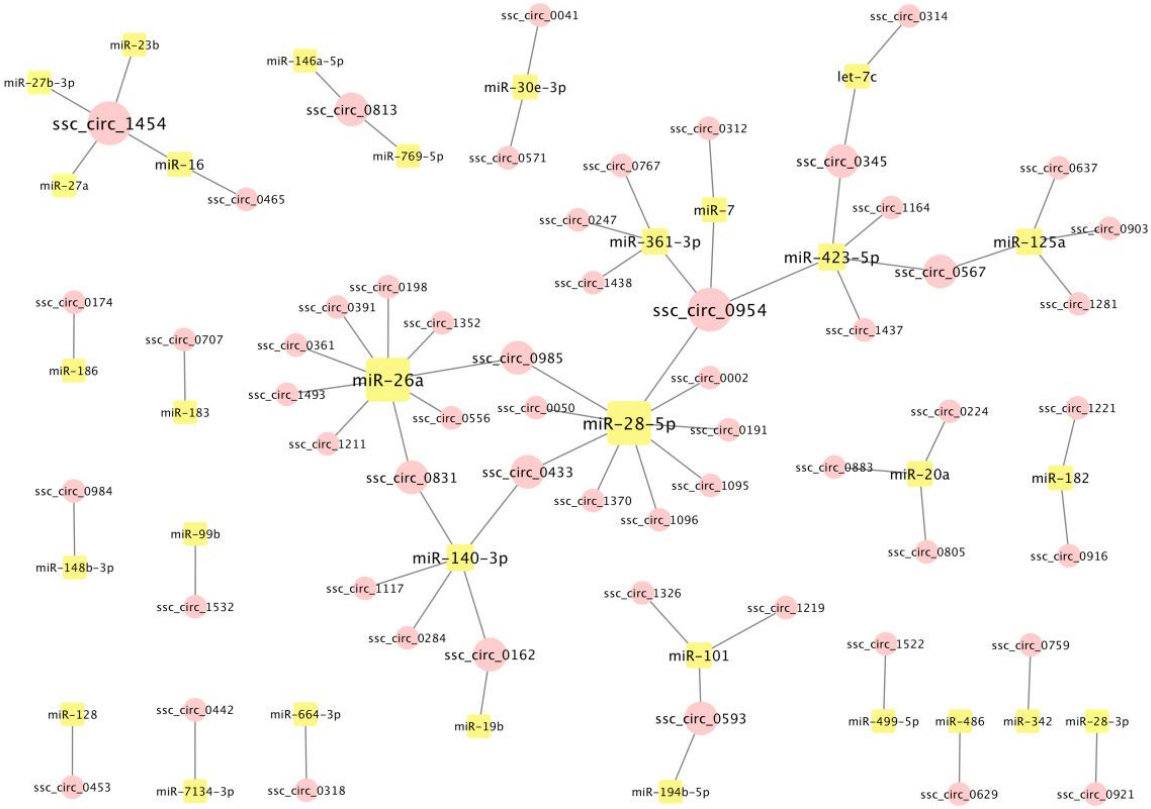


Figure 2. circRNA-miRNA interaction network. circRNA:miRNA relationships predicted by both experimental data and miRNA target sequence analysis in circRNAs. Circular and square nodes represent circRNAs and miRNAs, respectively. The node and letter sizes indicate the number of significant correlations involving the node.

Correlation of circRNAs with sperm motility and circRNA validation

179 circRNAs (from 26 intergenic regions, 10 intronic regions and 146 genes) showed significant associations between their abundance and different sperm motility parameters (Supplementary Table S3). More in detail, 28 (2 intergenic, 2 intronic and 24 protein coding genes), 94 (7 intergenic, 2 intronic and 81 genic), 35 (6 intergenic, 3 intronic and 26 genic) and 57 (11 intergenic, 3 intronic and 43 genic) circRNAs correlated with the percentage of motile cells, VCL, VAP and VSL, respectively (Additional file 3).

To confirm the existence of these circRNAs and their phenotypic correlations we undertook a Reverse Transcription quantitative PCR (RT-qPCR) and Sanger Sequencing approach. First, we randomly selected 2 circRNAs with different RNA abundance levels (75.0 and 10.6 CPM) to test whether we could confirm the bioinformatically predicted circRNAs. The chosen circRNAs were: ssc_circ_1141 (from *PTGES3*) with 75.0 CPM, and ssc_circ_0670 (from *BAZ2B*)

with 10.6 CPM, both also detected in human [36] (hsa_circ_0008137 and hsa_circ_0002463, respectively) and in swine testes [28]. The PCR amplification of the 2 circRNAs resulted in a single electrophoretic band of the expected size (Additional Figure 1.A) and Sanger Sequencing confirmed the back-splice junction (Additional Figure 2.A-B). We, therefore, confirmed the validity of the RT-PCR and Sanger Sequencing approach to validate the presence of tested circRNAs and the existence of these two circRNAs in the boar sperm.

Then, we used the same approach to confirm the existence of 8 circRNAs selected for their significant correlation with at least one motility trait (Additional file 3). These circRNAs were further selected based on the fact that (i) they were present at high abundances, or (ii) they showed significant correlation with at least 1 trait, or (iii) they had been previously identified in pig [28], human [36] or mice [36]. PCR amplification of 6 of the 8 circRNAs resulted in a band of the expected size (Supplementary Figure S1B) and the Sanger Sequencing validated the expected back-splice junction (Figure 3.A-C; Supplementary Figure 2.C-E). ssc_circ_0839 from *PAIP2* did not amplify (data not shown) and ssc_circ_0118 from *PDE10A* (also identified in pig testes and in human as hsa_circ_0078638) displayed 2 amplification bands (Additional Figure 4.B). These two circRNAs were discarded for further analysis. Thus, we also confirmed the existence of these 6 circRNAs.

The RT-qPCR levels of these 6 circRNAs were measured in 36 animals presenting extreme and opposite values of sperm motility (N=18 for each phenotypic distribution tail) from a bank of 300 phenotyped boar ejaculates. None of the 36 samples was included in the RNA-seq study. The two sample groups displayed significant phenotypic differences for all the studied traits: percentage of motile cells (P-value: 2.65×10^{-9}, Wilcoxon rank sum test), VCL (P-value: 2.20×10^{-10}), VSL (P-value: 3.22×10^{-7}) and VAP (P-value: 3.22×10^{-7}). The RT-qPCR assays presented efficiencies between 99.6% and 105.2%.

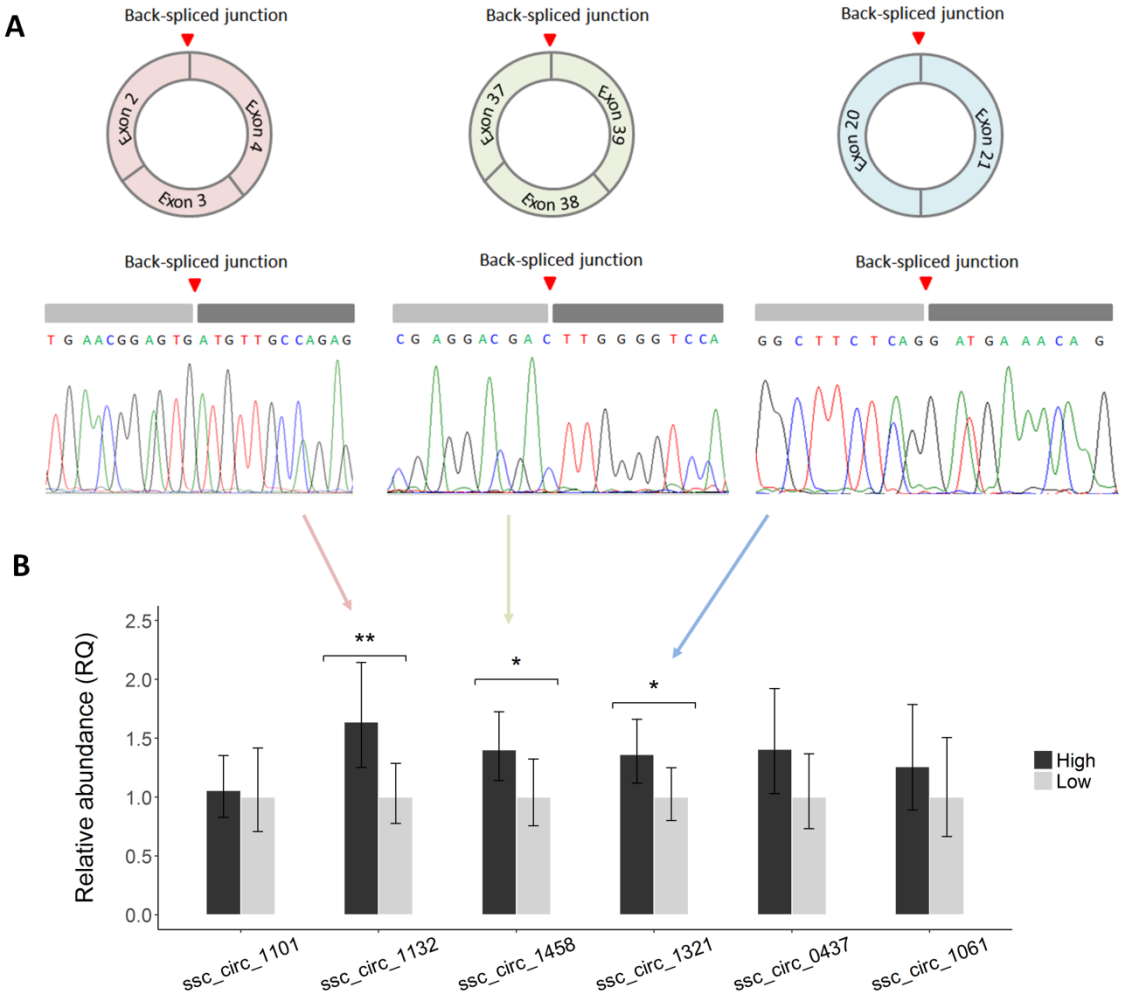


Figure 3. Validation of the circRNAs which RNA-seq based abundance correlated with sperm motility. A. Sanger sequencing validation of the circRNA black splice junction for ssc_circ_1132 from *LIN7A*, ssc_circ_1458 from *LRBA* and ssc_circ_1321 from *PAPOLA*. **B.** Relative abundance of six circRNAs in samples with extreme and divergent motility values (18 samples displaying high motility and 18 samples with low motility) obtained by RT-qPCR. The RNA-seq based association between ssc_circ_1458 and ssc_circ_1321 abundance and sperm motility was validated by RT-qPCR in the 36 samples.

Two of the 6 circRNAs showed significant differences between the two sperm motility groups (Figure 3.B). These 2 circRNAs were ssc_circ_1458 from *LRBA* (P-value: 0.049) and ssc_circ_1321 from *PAPOLA* (P-value: 0.035). A third circRNA, ssc_circ_1132, from *LIN7A*, showed significant differences between the two motility groups (P-value: 0.008) (Figure 3.B) but in the opposite direction than expected according to the RNA-seq data and was consequently considered as not validated due to inconclusive results. The other 3 circRNAs, ssc_circ_1101 from *KHDRBS3*, ssc_circ_0437 from *ULK4* and ssc_circ_1061 from

ZNHIT6 did not present significant differences between the 2 groups (Figure 3.B).

Discussion

In this study, we have profiled for the first time, the circRNA repertoire of mature spermatozoa in a mammalian species and its association with sperm motility as a parameter of semen quality. These results provide further evidence and expand the relevance of spermatoza RNAs in sperm biology and quality. circRNAs have been reported as promising potential prognostic and diagnostic biomarkers for human health including, among others, cancer, diabetes, cardiovascular diseases or pre-eclampsia [reviewed in 30]. Here, we provide novel data supporting the potential of circRNAs as biomarkers for the male's reproductive function as reflected by association with sperm motility parameters.

We have identified nearly 1,600 boar sperm circRNAs that are robustly present in most samples (at least 30). The sperm circRNA repertoire presented similar genomic circular characteristics such as the proportion of genomic overlap with protein coding or intergenic regions, among others, the number of exons and their length (Figure 1), when compared to data previously reported in other tissues from swine [28], human [44] or rat [41]. Nonetheless, the list of sperm circRNAs showed a modest overlap with other tissues and species (Table 3). Not surprisingly, the largest overlap was with porcine testes (11.6%) probably due to the fact that the spermatogenic lineage including spermatozoa is contained in the male gonads. The proportion of pig sperm circRNAs shared across swine tissues was in general low when compared to the rest of the *Sus scrofa* libraries (Additional file 4). Although this comparative analysis may be partly influenced by technical differences involving the processing of the tissues and the data between the studies, our results suggest that sperm present high tissue-specificity, with circa 80% of the circRNAs being present only in sperm,

considerably higher than testes [28], with approximately 65% testes-specific. On the opposite direction, there was a larger coincidence between the catalogue of around 6,660 porcine circRNAs [28, 29] and the swine sperm circRNAome (4.9%), than between the later and the human archive with over 90,000 circRNAs [36] (0.4%) (Table 3). This 12x fold difference indicates that circRNAs are not only tissue specific but have also a strong species-specific component.

We investigated the functional relevance of sperm circRNAs, under the hypothesis that their function is associated to the host gene. Six (*CEP63*, *ATP6V0A2*, *PPA2*, *PAIP2* and *PAXIP1*) of the 15 coding genes providing the top 20 most abundant circRNAs (Table 1) have been directly implicated, to sperm related traits and male fertility. *CEP63* is engaged in microtubule organization, and mice knockout studies revealed its essential role in male fertility [45]. *ATP6V0A2* proteins and transcripts are down-regulated in infertile men [46]. *PPA2* is located in the mitochondrial membrane and might be involved in the production of ATP, control of molecular processes linked to the launching of sperm capacitation and sperm motility [47-49]. *PAIP2* is linked to male sperm maturation and fertility [50] and *PAXIP1* is associated with developmental arrest of spermatocytes, testicular atrophy, and infertility in knockout mice [51]. We identified 12 circRNA from hotspot genes producing five or more circRNAs each (Table 2). Some of the hotspot genes were related to sperm function and fertility. *TESK2* may play a role in early stages of spermatogenesis [52]; *PTK2*, which is essential for a embryo development [53]; *SPATA19*, a gene that is critical for sperm mitochondrial function in relation to sperm motility and fertilization ability [54], and *SLC5A10*, which protein products may act as water channels in spermatozoa [55].

The ontology enrichment analysis of the genes harboring circRNAs pointed towards epigenetic related functions, which are essential in all cell types including chromatin condensation in sperm and the reprogramming of gene

expression upon egg fertilization and during embryo development (Additional file 2). Gene enrichment analysis also signaled towards spermatogenesis and developmental processes, the later also implicating the embryo development related genes, *DHX36, IPMK, RICTOR, CDC73* and *ANGPT1* which were hosting some of the circRNAs found in our analysis (Supplementary Table S1). These functions are in line with the gene ontologies highlighted in studies analyzing sperm mRNAs [34, 35, 56], thereby providing further basis for the hypothesis that circRNAs may exert their function (regulation of sperm quality and motility in our study) by controlling their cognate linear mRNA.

A previous study in rat testes identified a circRNA age-dependent dynamic pattern of expression and suggested a relation between their abundance and function with the male's sexual maturity and spermatogenesis [41]. For this reason, we sought to investigate whether, like in testes, sperm circRNAs - in boar at least - also accumulate through age. Our data highlighted that there is no association between the mature sperm circRNAs and the boar's age. Cells with high proliferation rates, seem to accumulate less circRNAs, possibly due to passive thinning out during proliferation [57]. Spermatogenesis is a process that occurs throughout the male's lifetime in which spermatogonial stem cells (SSCs) undergo continuous cell division and finally differentiate to ultimately become spermatozoon [58]. Thus, the fact SSCs undergo continuous self-renewal and keep proliferating through the lifetime of the male, may impede the accumulation of circRNAs in these cells and thus explain why no age-dependent pattern of circRNA abundance was found in the sperm of pigs with different ages. It is plausible that due to the continuous production and differentiation of the male germ cell, the sperm circRNAome does not mimic the reproductive performance of the boar, at least once the number of germ cells have stabilized which occurs at the age of 7 months old in pigs [59] and until the boar is senile. These results infer that stability of mature sperm circRNAs

might be constant for a period in normozoospermic males and provides further evidence for circRNAs as a noninvasive diagnosis tool for male reproductive diseases.

To elucidate the functional relevance of circRNAs as miRNA sponges we built an interaction network (Figure 2). We combined benchwork data of circRNA and miRNA abundances (small and total RNA-seq) and *in silico* searches of miRNA target sequences in circRNAs with the aim to increase the reliability of the circRNA-miRNA relationship predictions. This network with 70 interactions, contained some interesting circRNAs and miRNAs in relation to semen quality and male fertility. Remarkably, two circRNAs, ssc_circ_0954 and ssc_circ_1454, presented, each, 4 different miRNA target sites. The first, ssc_circ_0954 arises from *DCDC2C*, a gene identified in the human's sperm flagellum end-piece with a suggested role on microtubule dynamics by acting as a depolymerization/polymerization balancing system [60]. This circRNA regulated among others, miR-361-3p which has been found dysregulated in subfertile men [61], miR-423-5p, altered in oligozoospermic men [62] and miR-28-5p, dysregulated in normozoospermic infertile individuals [18]. The other cirRNA, ssc_circ_1454, is transcribed from *MTHFD2L*, a mitochondrial isozyme from the folate cycle metabolic pathway [63], a vitamin that has been also related to semen quality and fertility in men [64]. One of the miRNAs targeted by ssc_circ_1454 was miR-16, which was found dysregulated in subfertile men [61]. The network also showed 9 circRNAs that may be regulating miR-28, a miRNA (miR-28-5p) that is dysregulated in normozoospermic infertile individuals [18]. The potential miR-28 regulators include ssc_circ_1370, arisen from *FAM92A*, whose protein may play a role in ciliogenesis [65], implicated in the formation of the sperm flagella and ssc_circ_0002 from *WDR7*, which is associated to cattle sperm quality [66]. Likewise, 9 circRNAs were predicted to regulate miR-26a, which has been in turn, linked to VCL, VSL and VAP motility

parameters in swine [67]. These miR-26a regulatory edges implicate ssc_circ_0361 from *ACTL6A*, crucial for embryo development [68] and ssc_circ_1352 from *CAGE1*, an acrosomal protein with proposed roles in fertility [69]. Other interesting interactions included ssc_circ_0345 from the hotspot gene *SLC5A10* (sodium-dependent mannose and fructose transporter) (Table 2), which regulated miR-423-5p, a miRNA that was found to be upregulated in oligozoospermic semen [62] and let-7c, which is altered in severe aasthenozoozpermia patients [70]. Altogether, the network involves key genes and miRNAs for male fertility thereby suggesting that circRNAs play a functional role in the male reproductive ability.

The potential role of circRNAs in male reproductive traits is further substantiated by the associations between sperm motility and the abundance of some circRNAs in our study. At least 20 of the circRNA host genes implicated in the phenotypic correlations have been previously linked to sperm biology or male fertility (Additional file 3). For example, ssc_circ_0823, which correlated with VCL (p-val = 0.009), is a circRNAs hosted by *CAMK4*, a gene that has been implicated in sperm motility in humans [71]. ssc_circ_0780, correlated with the percentage of motile cells (p-val = 0.041) from *LRGUK*, a gene required for sperm assembly including the growth of the axonome, a structure that is necessary for the flagellar beating in sperm [72]. We successfully tested by RT-qPCR 6 of the 179 circRNAs that showed RNA-seq based significant correlations between their abundance and sperm motility. The results allowed us to validate the correlation with motility for 2 of these 6 circRNAs, ssc_circ_1458 from *LRBA* and ssc_circ_1321 from *PAPOLA*. Both were significantly down-regulated in the ejaculates with low sperm motility values (Figure 3.B). *LRBA* is a gene involved in coupling signal transduction and vesicle trafficking but no link with sperm function or fertility has been made thus far. This gene has been associated to immune-related disorders in humans.

ssc_circ_1321 from *PAPOLA*, has a human ortholog (human circRNA: hsa_circ_0033126) and the host gene is implicated in RNA and ATP binding. Thus, none of the two host genes have been previously related to sperm function or fertility. Moreover, these 2 circRNAs are not included in the circRNAs:miRNA interaction network. However, the association is clear and confirmed by RT-qPCR and we cannot exclude unidentified relevant functions on sperm motility for these two genes. Two other circRNAs were also of high interest and tested as they were within the top 20 most abundant (Table 1) and correlated with sperm motility (Additional file 3). They were ssc_circ_0839 from *PAIP2,* with crucial roles in spermatogenesis [50] that did not amplify, and ssc_circ_1101 from *KHDRBS3,* found highly abundant in mice testes [73], which did not present significant differences in RT-qPCR levels between motility groups (Figure 3.B).

Interestingly, some circRNAs popped up as relevant in more than one of the analysis carried through the study. For example, ssc_circ_1532 from the hotspot gene *SPATA19* (Table 2), related to sperm motility and fertility [54], was suggested to regulate miR-99a according to the network analysis (Figure 2). miR-99b has been found deregulated in low motile sperm fractions in bull [19] and in subfertile men [61]. Another circRNA, ssc_circ_1219 from *OSBPL9*, involved in male reproduction [74], displayed abundance correlation with VCL (P-value: 0.04) (Additional file 3) and was identified as a potential target of miR-101 (Figure 2), miR-101-3p was altered in asthenozoospermia men [70].

Remarkably, 4 (*DENND1B*, *PTK2*, *SLC5A10* and *CAMSAP1*) of the 12 hotspot genes hosted a circRNA with significant abundance correlation with sperm motility. Thus, one third of the hotspot genes included circRNAs correlated with motility whilst only 14.9% of the genes (147) of the 984 genes hosting the 1,598 circRNAs were correlated to motility. Noteworthy, the 12 circRNA hotspot genes were not hotspots in the other porcine tissues analyzed [28, 29]

(data not shown). Altogether, this indicates that circRNAs hotspot genes may have relevant tissue-specific functions.

CircRNAs are acknowledged to be more stable than mRNA. Our data shows, at different levels, that there is a detectable population of circRNAs in the boar sperm that is related to relevant functions in sperm and fertility. Moreover, these circRNAs are stable across age. For these reasons, our results indicate that circRNAs hold a potential as biomarkers for sperm motility and potentially, fertility outcomes [7, 8, 10-14, 75]. This potential should be further explored in swine, as well as in other domestic animals and in human fertility clinics.

Conclusions

In conclusion, our study is the first to characterize the spermatozoa circRNAs repertoire in an animal species. We have provided a comprehensive view of the boar sperm circRNAome, which is highly sperm-specific and involves genes related to sperm biology and development. Moreover, we have detected correlation between the abundance of some circRNAs with sperm motility parameters. Our findings involving sperm motility may spur novel research on male fertility in both human medicine and in animal breeding.

Methods

Sperm collection, phenotyping and library preparation

Ejaculates from 300 Pietrain boars were collected using the hand glove method by trained professionals at commercial farms. Fresh sperm motility traits were assessed with the CASA system (Integrated Sperm Analysis System V1.0; Proiser, Valencia, Spain). In this study we analyzed sperm motility parameters, including the total percentage of motile cells, VCL (μm/s), VSL (μm/s) and VAP (μm/s). The average percentage of motile cells was 75.1 with a standard deviation (SD) of 18.3, VCL (mean: 45.1; SD: 12.6), VSL (mean: 27.0; SD: 8.3) and

VAP (mean: 34.2; SD: 10.5). Phenotypes were corrected for the fixed variables: farm (1, 2, 3), age (1, 2, 3) and season and year (Autumn 2014, 2015 and 2016; Winter 2015, 2016 and 2017; Spring 2015 and 2016; Summer 2015) using the R function "lm" [76]. Ejaculates were purified to remove somatic cells and immature sperm cells and purified sperm was stored at -96°C with Trizol® as described by Gòdia et al. [77]. RNA was extracted from purified sperm cells, treated with TURBO DNA-free™ Kit (Invitrogen; Carlsbad, USA) and quantified using Qubit™ RNA HS Assay kit (Invitrogen; Carlsbad, USA). The RNA yield of these samples averaged 2.2 fg per sperm cells (the range was between 0.8 and 3.7 fg). We assessed RNA integrity with the 2100 Bioanalyzer using the Agilent RNA 6000 Pico kit (Agilent Technologies; Santa Clara, USA). All samples presented RNA Integrity Number (RIN) below 2.5, which indicates the absence of intact RNA from somatic cell origin. We then performed RT-qPCR assays for *PRM1* and *PTPRC* mRNAs as well as for intergenic/genomic DNA to verify that all the samples were free from RNA from somatic cells and from genomic DNA contamination [77].

40 sperm RNA samples were subjected to total RNA-seq. 34 of these samples were also used for small RNA-seq. For total RNA-seq libraries, ribosomal RNA (rRNA) was depleted with the Ribo-Zero Gold rRNA Removal Kit (Illumina) and libraries were constructed with the SMARTer Universal Low Input RNA library Prep kit (Clontech). Resulting libraries were sequenced in a HiSeq2500 system (Illumina) to generate 75 bp long paired-end reads. For small non coding RNA-seq libraries, extracted RNA without rRNA depletion was directly subjected to library preparation with the NEBNext Small RNA Library Prep Set kit (New England Biolabs) and sequenced on a HiSeq2000 (Illumina) to generate 50 bp single-end reads.

RNA-seq and bioinformatics analysis

For the total RNA fraction, we obtained, in average, 20.4 million paired-end reads per sample. Raw reads were then filtered by removing adaptor sequences and low-quality reads with Trimmomatic v.0.36 [78]. In average, 98.5% of these reads passed the quality control filters. The identification of circRNAs was carried on these reads with the find_circ pipeline [26] with a stricter filter stringency. To reduce the false positive rate in the discovery of circRNAs, we selected circular splice transcripts with at least two unique supporting reads in the anchor segment and with Phred quality scores of 35 or more. Moreover, only circRNAs predicted in at least 30 samples were kept. The RNA abundance of the predicted circRNAs were normalized as the number of back-splice junction spanning reads by its sequencing depth, as counts per million (CPM). The functional regions of circRNAs were identified based on their co-location with genomic features (e.g. exon, 3'UTR, 5'UTR, etc) from the Ensembl database (release 91) with BEDtools [79]. Our catalogue of boar sperm circRNAs was contrasted with other publically available porcine circRNA databases including heart, liver, spleen, lung, kidney, ovarium, testis, skeletal muscle, fat and fetal brains [28, 29]. We also queried several human tissues (including several cell lines, brain sections placenta, muscle, fat, umbilical cord, atrium, decidua and plasma) [26, 37-40] and murine (cell lines and brain sections) [26, 37] available at the circBase database [36]. Genomic coordinates from the human and mouse circRNAs were liftover to Sscrofa11.1 using the UCSC liftover tool [80].

For the small RNA-seq analysis, we obtained an average of 7.3 million reads per sample. Trimming of adaptors and low quality bases was performed with Cutadapt v1.0 [81]. 99.2% of these reads were of high quality and were thus used for downstream analysis. The mapping of sncRNAs was performed with the sRNAtoolbox v.6.17 [82] with default settings and providing miRBase [83] release 21 as library dataset. Multi-adjusted read counts were then normalized

by sequencing depth as CPM. We only considered the miRNAs that were detected > 1 CPM in all the samples.

circRNA-miRNA network visualization

For a functional annotation of the circRNA and miRNA interactions in a systems biology context, we carried a Partial Correlation with Information Theory (PCIT) analysis [84] using the combined list of circRNAs and miRNAs after normalization of abundance levels with log2. We further assessed circRNAs-miRNAs co-abundance interactions using miRanda [85] v.3.3a. Shared correlations were visualized with Cytoscape v.3.7.0 [86].

Gene Ontology analysis and correlation with sperm motility parameters

GO analysis was carried with PANTHER v.13.1 [87] with the overrepresentation test and p-values corrected with FDR. Annotation Data Set was "GO biological process complete". Pearson correlation was used to determine associations between circRNA abundance levels and sperm motility parameters. P-values < 0.05 were considered statistically significant.

Validation of circRNAs and reverse transcription quantitative PCR (RT-qPCR)

circRNAs were validated by Sanger Sequencing and quantified by RT-qPCR using divergent primers. Primers were designed as in [88] using the Primer Express software (Applied Biosystems). Primer sequences are shown in Additional file 5. For cDNA synthesis, 5 μl of RNA were reverse transcribed using the High Capacity cDNA Reverse Transcription kit in a final volume of 50 μL (Applied Biosystems; Waltham, USA) following the manufacturer's protocol. circRNAs were amplified and visualized in 3% high resolution agarose gel electrophoresis and confirmed by Sanger Sequencing.

The abundance level of 6 circRNAs, correlated to sperm motility parameters in the RNA-seq study, was analyzed by RT-qPCR in 36 samples, none of them included in the RNA-seq. These 36 samples belong to two groups with extreme and divergent values for sperm motility from a bank of 300 ejaculates with phenotypic records. Quantitative PCR reactions were performed in triplicate in 15 μL final volume including 7.5 μL SYBR Select Master Mix (Life Technologies - Thermo Fisher Scientific), 300 nM of each primer and 3.75 μL of cDNA 1:4 diluted on a QuantStudio 12K Flex Real-Time PCR System (Applied Biosystems). To evaluate the efficiency of the RT-qPCR assays, standard curves with 6 serial dilutions from a pool of sperm cDNA were generated. Thermal profile was set as follows: 50ºC for 2 min, 95ºC for 10 min and 40 cycles at 95ºC for 15 sec and 60ºC for 60 sec. Moreover, a melting profile (95°C for 15 sec, 60°C for 15 sec and a gradual increment of temperature with a ramp rate of 1% up to 95°C) was programmed at the end of the RT-qPCR to assess the specificity of the reactions. The genes *ISYNA2* and *GRP137* were selected as endogenous controls following the stability values after a GeNorm pilot experiment. Their stability was determined considering a GeNorm M value < 0.5. Relative expression values were calculated using the ThermoFisher Cloud software (Applied Biosystems) applying the $2^{-\Delta\Delta Ct}$ method. The same software was used to compare the biological groups. Significance was set at a P-value < 0.05.

List of abbreviations

AI: Artificial Insemination
circRNAs: Circular RNAs
CPM: Counts Per Million
miRNAs: microRNAs
mRNAs: messenger RNAs
PCIT: Partial Correlation with Information Theory
RT-qPCR: Reverse Transcription quantitative PCR
tRNAs: transference RNAs

VAP: velocity of the sperm cells

VCL: curvilinear velocity

VSL: straight line velocity

Declarations

Ethics approval and consent to participate

Specialized professionals at each commercial AI stud obtained all the ejaculates following their standard routine monitoring procedures and relevant guidelines. No animal experiment has been performed in the scope of this research.

Consent for publication

Not applicable

Availability of the data and material

The datasets generated and/or analysed during the current study are available at NCBI's BioProject PRJNA520978.

Competing interests

Not applicable

Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under grant AGL2013-44978-R and grant AGL2017-86946-R and by the CERCA Programme/Generalitat de Catalunya. AGL2017-86946-R was also funded by the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERDF). We thank the Agency for Management of University and Research Grants (AGAUR) of the Generalitat de Catalunya (Grant Number 2017 SGR 1060). We also acknowledge the support of the Spanish Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa 2016–2019 (Grant Number SEV-2015-0533)

grant awarded to the Centre for Research in Agricultural Genomics (CRAG). MG acknowledges a Ph.D. studentship from MINECO (Grant Number BES-2014-070560).

Author's contributions

MG, AS, and AlC conceived and designed the experiments. MR and JR-G carried the phenotypic analysis. MG performed sperm purifications. MG and MR carried the RNA extractions. AnC designed and carried the RT-qPCR and their analyses. AnC and BC performed Sanger Sequencing validation. MG made the bioinformatics, statistic analysis and analyzed the data. MG and AlC wrote the manuscript. All authors discussed the data and read and approved the contents of the manuscript.

Acknowledgments

We thank Craig Lewis from Genus PIC and Sam Balasch from Gepork for contributing the sperm samples.

Additional files

Additional file 1. List of the 1,598 circRNAs identified in sperm with their genomic coordinates, mean abundance (in CPM) and Standard Deviation (SD) in the 40 samples, and the host gene of the exonic circRNAs.

Additional file 2. Gene Ontology analysis and FDR value of the circRNA host genes.

Additional file 3. Correlation between circRNA abundance and sperm motility parameters. The table includes information on the genomic coordinates of the circRNAs, p-values of the correlation with sperm motility parameters, host gene, whether it was tested for RT-qPCR validation, and the article reference for these host genes that have previously been associated to sperm biology or male fertility. MT: total percentage of motile cells; VCL: curvilinear velocity; VSL: straight line velocity; VAP: velocity of the sperm cells; ns: not significant.

Additional file 4. Concordance on the list of circRNAs present in 15 porcine tissues.

Additional file 5. List of primers designed and used for the RT-qPCR to assess the abundance of target circRNAs and reference genes.

Additional Legends

Additional Figure 1. Figure displaying the validation of the amplified set of circRNAs by agarose-gel electrophoresis. **A.** Amplification of the 2 randomly selected circRNAs. Two different primer sets for ssc_circ_1141 from *PTGES3* were tested (primer pair a in lane 2 and b in lanes 3 and 4) and ssc_circ_0670 from *BAZ2B* (lane 5). **B.** Validation of the circRNAs correlated to sperm motility parameters: ssc_circ_1321 from *PAPOLA* (*ENSSSCG00000002505),* ssc_circ_1458 from *LRBA,* ssc_circ_0437 from *ULK4,* two different primer sets for ssc_circ_1061 from *ZNHIT6* primer pair a and b, ssc_circ_1132 from *LIN7A,* two different primer sets for ssc_circ_1101 from *KHDRBS3*, primer pair a (which resulted in amplification of two splicing forms and was excluded) and b, and ssc_circ_0118 from *PDE10A* that resulted in amplification of two splicing forms (and excluded from further analysis).

Additional Figure 2. Figure showing the Sanger sequencing based validation of the set of circRNAs. **A.** ssc_circ_1141 from *PTGES3.* **B**. ssc_circ_0670 from *BAZ2B.* **C.** ssc_circ_0437 from *ULK4.* **D.** ssc_circ_1061 from *ZNHIT6.* **E.** ssc_circ_1101 from *KHDRBS3*.

References

1. Neeteson-van Nieuwenhoven A-M, Knap P, Avendaño S. The role of sustainable commercial pig and poultry breeding for food security. Animal Frontiers. 2013;3:52-7.
2. Swindle MM, Makin A, Herron AJ, Clubb FJ, Frazier KS. Swine as Models in Biomedical Research and Toxicology Testing. Vet Pathol. 2012;49:344-56.
3. Hirsh A. Male subfertility. BMJ. 2003;327:669-72.

4. Frankenhuis MT, Wensing CJ. Induction of spermatogenesis in the naturally cryptorchid pig. Fertil Steril. 1979;31:428-33.

5. Bernabo N, Tettamanti E, Russo V, Martelli A, Turriani M, Mattoli M, Barboni B. Extremely low frequency electromagnetic field exposure affects fertilization outcome in swine animal model. Theriogenology. 2010;73:1293-305.

6. Park KE, Kaucher AV, Powell A, Waqas MS, Sandmaier SE, Oatley MJ, Park CH, Tibary A, Donovan DM, Blomberg LA, et al. Generation of germline ablated male pigs by CRISPR/Cas9 editing of the NANOS2 gene. Sci Rep. 2017;7:40176.

7. Farrell PB, Presicce GA, Brockett CC, Foote RH. Quantification of bull sperm characteristics measured by computer-assisted sperm analysis (CASA) and the relationship to fertility. Theriogenology. 1998;49:871-9.

8. Love CC. Relationship between sperm motility, morphology and the fertility of stallions. Theriogenology. 2011;76:547-57.

9. Holt C, Holt WV, Moore HD, Reed HC, Curnock RM. Objectively measured boar sperm motility parameters correlate with the outcomes of on-farm inseminations: results of two fertility trials. J Androl. 1997;18:312-23.

10. Broekhuijse MLWJ, Sostaric E, Feitsma H, Gadella BM. Application of computer-assisted semen analysis to explain variations in pig fertility. J Anim Sci. 2012;90:779-89.

11. Hirai M, Boersma A, Hoeflich A, Wolf E, Foll J, Aumuller R, Braun J. Objectively measured sperm motility and sperm head morphometry in boars (Sus scrofa): Relation to fertility and seminal plasma growth factors. J Androl. 2001;22:104-10.

12. Hirano Y, Shibahara H, Obara H, Suzuki T, Takamizawa S, Yamaguchi C, Tsunoda H, Sato I. Relationships between sperm motility characteristics assessed by the computer-aided sperm analysis (CASA) and fertilization rates in vitro. J Assist Reprod Gen. 2001;18:213-8.

13. Aitken RJ. Sperm function tests and fertility. Int J Androl. 2006;29:69-74.

14. Paston MJ, Sarkar S, Oates RP, Badawy SZA. Computer-Aided Semen Analysis Variables as Predictors of Male-Fertility Potential. Arch Andrology. 1994;33:93-9.

15. Jodar M, Kalko S, Castillo J, Ballesca JL, Oliva R. Differential RNAs in the sperm cells of asthenozoospermic patients. Hum Reprod. 2012;27:1431-8.

16. Pelloni M, Paoli D, Majoli M, Pallotti F, Carlini T, Lenzi A, Lombardo F. Molecular study of human sperm RNA: Ropporin and CABYR in asthenozoospermia. J Endocrinol Invest. 2018;41:781-7.

17. Parthipan S, Selvaraju S, Somashekar L, Arangasamy A, Sivaram M, Ravindra JP. Spermatozoal transcripts expression levels are predictive of semen quality and conception rate in bulls (Bos taurus). Theriogenology. 2017;98:41-9.

18. Salas-Huetos A, Blanco J, Vidal F, Godo A, Grossmann M, Pons MC, S FF, Garrido N, Anton E. Spermatozoa from patients with seminal alterations exhibit a differential micro-ribonucleic acid profile. Fertil Steril. 2015;104:591-601.

19. Capra E, Turri F, Lazzari B, Cremonesi P, Gliozzi TM, Fojadelli I, Stella A, Pizzi F. Small RNA sequencing of cryopreserved semen from single bull revealed altered miRNAs and piRNAs expression between High- and Low-motile sperm populations. BMC Genomics. 2017;18.

20. Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. Nat Biotechnol. 2014;32:453-61.

21. Abe N, Abe H, Ito Y. Dumbbell-shaped nanocircular RNAs for RNA interference. J Am Chem Soc. 2007;129:15108-9.

22. Cocquerelle C, Mascrez B, Hetuin D, Bailleul B. Mis-splicing yields circular RNA molecules. FASEB J. 1993;7:155-60.

23. Cortes-Lopez M, Miura P. Emerging Functions of Circular RNAs. Yale J Biol Med. 2016;89:527-37.

24. Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, Lai EC. Genome-wide Analysis of Drosophila Circular RNAs Reveals Their Structural and Sequence Properties and Age-Dependent Neural Accumulation. Cell Rep. 2014;9:1966-80.

25. Guo JU, Agarwal V, Guo HL, Bartel DP. Expanded identification and characterization of mammalian circular RNAs. Genome Biol. 2014;15.

26. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495:333-8.

27. Xia SY, Feng J, Lei LJ, Hu J, Xia LJ, Wang J, Xiang Y, Liu LJ, Zhong S, Han L, He CJ. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. Brief Bioinform. 2017;18:984-92.

28. Liang GM, Yang YL, Niu GL, Tang ZL, Li K. Genome-wide profiling of Sus scrofa circular RNAs across nine organs and three developmental stages. DNA Res. 2017;24:523-35.

29. Venø MT, Hansen TB, Venø ST, Clausen BH, Grebing M, Finsen B, Holm IE, Kjems J. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. Genome Biol. 2015;16:245.

30. Zhang Z, Yang T, Xiao J. Circular RNAs: Promising Biomarkers for Human Diseases. EBioMedicine. 2018;34:267-74.

31. Quan GB, Li JL. Circular RNAs: biogenesis, expression and their potential roles in reproduction. J Ovarian Res. 2018;11.

32. Cheng J, Huang J, Yuan S, Zhou S, Yan W, Shen W, Chen Y, Xia X, Luo A, Zhu D, Wang S. Circular RNA expression profiling of human granulosa cells during maternal aging reveals novel transcripts associated with assisted reproductive technology outcomes. Plos One. 2017;12:e0177888.

33. Qian YT, Lu YQ, Rui C, Qian YJ, Cai MH, Jia RZ. Potential Significance of Circular RNA in Human Placental Tissue for Patients with Preeclampsia. Cell Physiol Biochem. 2016;39:1380-90.

34. Gòdia M, Estill M, Castelló A, Balasch S, Rodríguez-Gil JE, Krawetz SA, Sánchez A, Clop A. A RNA-seq analysis to describe the boar sperm transcriptome and its seasonal changes. Front Genet. 2018;10.

35. Gòdia M, Swanson G, Krawetz SA. A history of why fathers' RNA matters. Biol Reprod. 2018;99:147-59.

36. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014;20:1666-70.

37. Rybak-Wolf A, Stottmeister C, Glazar P, Jens M, Pino N, Giusti S, Hanan M, Behm M, Bartok O, Ashwal-Fluss R, et al. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. Mol Cell. 2015;58:870-85.

38. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA. 2013;19:141-57.

39. Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. Cell-type specific features of circular RNA expression. PLoS Genet. 2013;9:e1003777.

40. Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, Zhu S, Yang L, Chen LL. Circular intronic long noncoding RNAs. Mol Cell. 2013;51:792-806.

41. Zhou T, Xie X, Li M, Shi J, Zhou JJ, Knox KS, Wang T, Chen Q, Gu W. Rat BodyMap transcriptomes reveal unique circular RNA features across tissue types and developmental stages. RNA. 2018;24:1443-56.

42. Gruner H, Cortes-Lopez M, Cooper DA, Bauer M, Miura P. CircRNA accumulation in the aging mouse brain. Sci Rep. 2016;6.

43. Banaszewska D, Kondracki S. An Assessment of the Breeding Maturity of Insemination Boars Based on Ejaculate Quality Changes. Folia Biol. 2012;60:151-62.

44. Dong WW, Li HM, Qing XR, Huang DH, Li HG. Identification and characterization of human testis derived circular RNAs and their existence in seminal plasma. Sci Rep. 2016;6.

45. Marjanovic M, Sanchez-Huertas C, Terre B, Gomez R, Scheel JF, Pacheco S, Knobel PA, Martinez-Marchal A, Aivio S, Palenzuela L, et al. CEP63 deficiency promotes p53-dependent microcephaly and reveals a role for the centrosome in meiotic recombination. Nat Commun. 2015;6:7676.

46. Ota K, Jaiswal MK, Ramu S, Jeyendran R, Kwak-Kim J, Gilman-Sachs A, Beaman KD. Expression of a2 Vacuolar ATPase in Spermatozoa is Associated with Semen Quality and Chemokine-Cytokine Profiles in Infertile Men. Plos One. 2013;8:e70470.

47. Asghari A, Marashi SA, Ansari-Pour N. A sperm-specific proteome-scale metabolic network model identifies non-glycolytic genes for energy deficiency in asthenozoospermia. Syst Biol Reprod Med. 2017;63:100-12.

48. Shivaji S, Kota V, Siva AB. The role of mitochondrial proteins in sperm capacitation. J Reprod Immunol. 2009;83:14-8.

49. Aitken RJ, Baker MA, Nixon B. Are sperm capacitation and apoptosis the opposite ends of a continuum driven by oxidative stress? Asian J Androl. 2015;17:633-9.

50. Yanagiya A, Delbes G, Svitkin YV, Robaire B, Sonenberg N. The poly(A)-binding protein partner Paip2a controls translation during late spermiogenesis in mice. J Clin Invest. 2010;120:3389-400.

51. Schwab KR, Smith GD, Dressler GR. Arrested spermatogenesis and evidence for DNA damage in PTIP mutant testes. Dev Biol. 2013;373:64-71.

52. Rosok O, Pedeutour F, Ree AH, Aasheim HC. Identification and characterization of TESK2, a novel member of the LIMK/TESK family of protein kinases, predominantly expressed in testis. Genomics. 1999;61:44-54.

53. Luo JP, McGinnis LK, Carlton C, Beggs HE, Kinsey WH. PTK2b function during fertilization of the mouse oocyte. Biochem Bioph Res Co. 2014;450:1212-7.

54. Mi YJ, Shi Z, Li J. Spata19 Is Critical for Sperm Mitochondrial Function and Male Fertility. Mol Reprod Dev. 2015;82:907-13.

55. Rigau T, Rivera M, Palomo MJ, Fernandez-Novell JM, Mogas T, Ballester J, Pena A, Otaegui PJ, Guinovart JJ, Rodríguez-Gil JE. Differential effects of glucose and fructose on hexose metabolism in dog spermatozoa. Reproduction. 2002;123:579-91.

56. Sendler E, Johnson GD, Mao SH, Goodrich RJ, Diamond MP, Hauser R, Krawetz SA. Stability, delivery and functions of human sperm RNAs at fertilization. Nucleic Acids Res. 2013;41:4104-17.

57. Bachmayr-Heyda A, Reiner AT, Auer K, Sukhbaatar N, Aust S, Bachleitner-Hofmann T, Mesteri I, Grunt TW, Zeillinger R, Pils D. Correlation of circular RNA abundance with proliferation - exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. Sci Rep. 2015;5:8057.

58. Oatley JM, Brinster RL. Regulation of spermatogonial stem cell self-renewal in mammals. Annu Rev Cell Dev Biol. 2008;24:263-86.

59. Franca LR, Silva VA, Jr., Chiarini-Garcia H, Garcia SK, Debeljuk L. Cell proliferation and hormonal changes during postnatal development of the testis in the pig. Biol Reprod. 2000;63:1629-36.

60. Jumeau F, Chalmel F, Fernandez-Gomez FJ, Carpentier C, Obriot H, Tardivel M, Caillet-Boudin ML, Rigot JM, Rives N, Buee L, et al. Defining the human sperm microtubulome: an integrated genomics approach. Biol Reprod. 2017;96:93-106.

61. Abu-Halima M, Hammadeh M, Schmitt J, Leidinger P, Keller A, Meese E, Backes C. Altered microRNA expression profiles of human spermatozoa in patients with different spermatogenic impairments. Fertil Steril. 2013;99:1249-55.e16.

62. Muñoz X, Mata A, Bassas L, Larriba S. Altered miRNA Signature of Developing Germ-cells in Infertile Patients Relates to the Severity of Spermatogenic Failure and Persists in Spermatozoa. Sci Rep. 2015;5.

63. Bolusani S, Young BA, Cole NA, Tibbetts AS, Momb J, Bryant JD, Solmonson A, Appling DR. Mammalian MTHFD2L encodes a mitochondrial methylenetetrahydrofolate dehydrogenase isozyme expressed in adult tissues. J Biol Chem. 2011;286:5166-74.

64. Wallock L, Jacob R, Woodall A, Ames B. Nutritional status and positive relation of plasma folate to fertility indices in nonsmoking men. FASEB J. 1997;11:1068-.

65. Li FQ, Chen X, Fisher C, Siller SS, Zelikman K, Kuriyama R, Takemaru KI. BAR Domain-Containing FAM92 Proteins Interact with Chibby1 To Facilitate Ciliogenesis. Mol Cell Biol. 2016;36:2668-80.

66. Suchocki T, Szyda J. Genome-wide association study for semen production traits in Holstein-Friesian bulls. J Dairy Sci. 2015;98:5774-80.

67. Ma JD, Fan Y, Zhang JW, Feng SY, Hu ZH, Qiu WL, Long KR, Jin L, Tang QZ, Wang X, et al. Testosterone-Dependent miR-26a-5p and let-7g-5p Act as Signaling Mediators to Regulate Sperm Apoptosis via Targeting PTEN and PMAIP1. Int J Mol Sci. 2018;19.

68. Bao X, Tang J, Lopez-Pajares V, Tao S, Qu K, Crabtree GR, Khavari PA. ACTL6a enforces the epidermal progenitor state by suppressing SWI/SNF-dependent induction of KLF4. Cell Stem Cell. 2013;12:193-203.

69. Alsheimer M, Drewes T, Schutz W, Benavente R. The cancer/testis antigen CAGE-1 is a component of the acrosome of spermatids and spermatozoa. Eur J Cell Biol. 2005;84:445-52.

70. Zhou R, Wang R, Qin Y, Ji J, Xu M, Wu W, Chen M, Wu D, Song L, Shen H, et al. Mitochondria-related miR-151a-5p reduces cellular ATP production by targeting CYTB in asthenozoospermia. Sci Rep. 2015;5:17743.

71. Marin-Briggiler CI, Jha KN, Chertihin O, Buffone MG, Herr JC, Vazquez-Levin MH, Visconti PE. Evidence of the presence of calcium/calmodulin-dependent protein kinase IV in human sperm and its involvement in motility regulation. J Cell Sci. 2005;118:2013-22.

72. Liu Y, DeBoer K, de Kretser DM, O'Donnell L, O'Connor AE, Merriner DJ, Okuda H, Whittle B, Jans DA, Efthymiadis A, et al. LRGUK-1 Is Required for Basal Body and Manchette Function during Spermatogenesis and Male Fertility. PLoS Genet. 2015;11.

73. Ehrmann I, Dalgliesh C, Liu Y, Danilenko M, Crosier M, Overman L, Arthur HM, Lindsay S, Clowry GJ, Venables JP, et al. The tissue-specific RNA binding protein T-STAR controls regional splicing patterns of neurexin pre-mRNAs in the brain. PLoS Genet. 2013;9:e1003474.

74. Ferlin A, Raicu F, Gatta V, Zuccarello D, Palka G, Foresta C. Male infertility: role of genetic background. Reprod Biomed Online. 2007;14:734-45.

75. Gadea J. Sperm factors related to in vitro and in vivo porcine fertility. Theriogenology. 2005;63:431-44.

76. R Developmental Core Team. R: A language and environment for statistical computing. 2010.

77. Gòdia M, Mayer FQ, Nafissi J, Castelló A, Rodríguez-Gil JE, Sánchez A, Clop A. A technical assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-seq analysis. Syst Biol Reprod Med. 2018;64:291-303.

78. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-20.

79. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841-2.

80. Kuhn RM HD, Kent WJ. . The UCSC genome browser and associated tools Brief Bioinform. 2013;14:144-61.

81. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10-2.

82. Rueda A, Barturen G, Lebron R, Gomez-Martin C, Alganza A, Oliver JL, Hackenberg M. sRNAtoolbox: an integrated collection of small RNA research tools. Nucleic Acids Res. 2015;43:W467-73.

83. Kozomara A G-JS. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39:D152-D7.

84. Reverter A, Chan EK. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics. 2008;24:2491-7.

85. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. MicroRNA targets in Drosophila. Genome Biol. 2003;5:R1.

86. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498-504.

87. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res. 2010;38:D204-10.

88. Panda AC, Gorospe M. Detection and Analysis of Circular RNAs by RT-PCR. Bio Protoc. 2018;8.

Paper IV

An integrative systems biology approach to identify the molecular basis of sperm quality in swine

Marta Gòdia[1], Antonio Reverter[2], Rayner González-Prendes[3], Yuliaxis Ramayo-Caldas[4], Anna Castelló[1,5], Joan-Enric Rodríguez-Gil[6], Armand Sánchez[5] and Alex Clop[1,7*]

[1]Animal Genomics Group, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, 08193, Cerdanyola del Vallès (Barcelona), Spain.

[2]CSIRO Agriculture and Food, Queensland Bioscience Precinct, 306 Carmody Rd., St. Lucia, Brisbane, QLD, 4067, Australia.

[3]Animal Breeding and Genomics, Wageningen University & Research, 6708PB, Wageningen, the Netherlands.

[4]Animal Breeding and Genetics Program, Institute for Research and Technology in Food and Agriculture (IRTA), Torre Marimon, 08140, Caldes de Montbui, Catalonia, Spain.

[5]Unit of Animal Science, Department of Animal and Food Science, Autonomous University of Barcelona, 08193, Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[6]Unit of Animal Reproduction, Department of Animal Medicine and Surgery, Autonomous University of Barcelona, 08193, Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[7]Consejo Superior de Investigaciones Científicas (CSIC), 08003, Barcelona, Catalonia, Spain.

*Corresponding author:

Manuscript in preparation

Abstract

Background: The molecular processes affecting sperm quality remain largely unexplored. Genomic pressure in animal breeding industries has turned the interest to select for boars with high sperm quality to maximize ejaculate doses and fertility rates. In this study, we sought to identify key candidate genes, pathways and DNA variants associated to sperm quality traits in swine by analyzing 25 sperm-related phenotypes with a systems biology approach combining GWAS and RNA-seq (total and small).

Results: In the GWAS, we identified 12 QTL regions associated to head and neck abnormalities, abnormal acrosomes and motility. Some candidate genes were *CHD2, KATNAL2, SLC14A2* or *ABCA1.* For the RNA-seq analysis, we detected 6,128 significant correlations between the sperm traits and gene abundances. Gene-gene interactions shared between the GWAS and RNA-seq approach were used to build a network, which also included information on gene's interactions with miRNAs, and included genes as *LARP4, SPATC1* or *THADA*. The final network contained genes involved in game generation and development, meiotic cell cycle, DNA repair or embryo implantation. A selection of SNPs from the network's genes and lead GWAS and eGWAS tagged SNPs were used to build a SNP array with 74 polymorphisms that could explain between 5 to 36% of the phenotypic variance of the studied sperm traits.

Conclusions: Sperm quality is a complex trait influenced by several confounded factors and genes. By using a systems biology approach we could identify key genes that might hold a potential in the molecular processes and pathways of the different sperm quality characters studied. Furthermore, we have developed an SNP array that might able to explain a substantial part of the genetic variance in Pietrain boars and thus be used for animal breeding and selection.

Keywords: sperm quality, sytems biology, sperm RNA, GWAS, swine

Background

Sperm carries the paternal genome and a wide repertoire of molecules including RNAs, which are essential for fertilization and the development of a new organism. Spermatogenesis, the process in which germ cells proliferate and develop into mature spermatozoa, is highly orchestrated and controlled by multiple factors. Both DNA polymorphisms and gene expression have been linked to sperm quality and/or fertility in several mammalian species including swine (reviewed in: Gòdia et al., 2018a; Krausz et al., 2015). High-quality sperm is decisive to maximize the propagation of the best genetic material in livestock and the sustainability of the pig breeding sector. For this reason, ejaculated sperm is subjected to strict quality filters in boar artificial insemination (AI) studs. AI farms regularly evaluate the quality of ejaculates measuring traits such as concentration, morphology, viability and motility kinetics, as a way to predict their fertilizing ability (Gadea, 2005). Although these traits have been found to be low to moderately heritable (Smital et al., 2005; Wolf, 2009), the molecular processes and genetic mechanisms controlling sperm quality are far from being fully understood and boar replacement due to insufficient sperm quality remains an economic hurdle for the sector (Robinson and Buhr, 2005).

Currently, there are few studies employing high-throughput techniques to investigate the genetic basis of sperm quality in swine. Genome-wide association studies (GWAS) have been carried by 3 independent research groups. Diniz et al. (Diniz et al., 2014) identified a single quantitative trait loci (QTL) region associated to sperm motility in Large White. Two years later, Zhao and collaborators (Zhao et al., 2016) reported 3 multi-SNP QTL regions associated with epididymal weight, sperm concentration and total sperm per ejaculate, respectively and 7 singleton QTLs related to sperm motility, semen temperature, seminiferous tubule diameter and number of ejaculates in a White Duroc x Erhualian F_2 pedigree. More recently, Marques et al. (Marques et al.,

2018) detected 16 and 6 QTL regions in Large White and Landrace, respectively, associated to sperm motility, number of cells per ejaculate and morphological abnormalities. None of these regions were shared by both populations (Marques et al., 2018), nor across studies.

The presence of RNA molecules in the boar sperm is well documented (Gòdia et al., 2019a; Gòdia et al., 2018b), but their relation to sperm quality has been just thinly explored. Porcine sperm RNAs are highly fragmented and their gene abundances are mostly associated to prior events linked to spermatogenesis, fertility and embryo development (Gòdia et al., 2019a). A complex suite of RNAs are comprised in sperm, including coding (mRNA), long noncoding RNAs (e.g. circular RNA –circRNA-) and short noncoding RNAs (e.g. microRNA –miRNA- or Piwi interacting RNA –piRNA) (Gòdia et al., 2019a). In the temperate climate zones, sperm quality typically drops in the warm summer months (Li et al., 2019). Seasonal differences on mRNA and miRNA levels have also been detected and may be linked to this drop in semen quality (Gòdia et al., 2019a; Yang et al., 2010). Several studies have reported a relation between RNA abundances and semen quality in mammals (Capra et al., 2017; Wang et al., 2019)(Jodar et al., 2015). In swine, Curry et al. performed quantitative RT-PCR (RT-qPCR) targeting 10 miRNAs and identified 5 and 2 miRNAs associated to sperm morphology and motility, respectively (Curry et al., 2011). To the best of our knowledge, no association between mRNA and sperm quality is swine has been reported thus far. Other RNA classes, with still unreported links with sperm quality, may be also related to these traits. Our group has recently characterized the catalog of circRNAs of the porcine sperm using RNA-seq and has identified several circRNAs associated to sperm motility parameters (Gòdia et al., 2019b).

The genetic complexity of sperm quality involves several molecular mechanisms and pathways that are highly interconnected. In this context, a

systems biology approach to assess gene connections and functional interactions using genomics and transcriptomics is an attractive alternative to the classical "one-gene one-trait" analysis of a stand-alone GWAS or a differential gene expression analysis. Evaluating modules of interacting genes rather than single genes can provide a wider and more holistic picture to predict their functions and the regulation of complex traits (Cho et al., 2012). Furthermore, it can be used to design knowledge-based technologies and tools for their application to animal breeding.

The aim of this study was to indentify the strongest candidate genes, pathways and DNA variants associated to sperm quality in pigs combining in a systems biology approach, GWAS and RNA-seq.

Methods

Sample collection and phenotype measurement

Three hundred fresh sperm ejaculates each from a different Pietrain boar from commercial farms were collected by specialized professionals between September 2014 and January 2017. Sperm was obtained using the hand glove method, immediately diluted (1:2) in commercial extender and kept at 16°C for up to 2 hours until phenotype assessment. Blood samples were collected from specialists during their routine sample collection and gDNA was extracted using a phenol-choloform based method. The ejaculates were purified to remove somatic cells as described in (Gòdia et al., 2018b). Purified spermatozoa were stored with Trizol® at -80°C until further use.

Phenotypic records from fresh sperm were measured as previously described (Gòdia et al., 2018b) and included: sperm concentration (CON), cell viability (VIAB), morphologically abnormal acrosomes (ACRO), osmotic resistance test (ORT), sperm abnormalities (of the head –HABN-, neck –NABN- and tail –TABN) and cytoplasmatic droplets (proximal –PDROP- and distal –DDROP).

Sperm motility traits were also assessed using the computer-assisted semen analysis (CASA) system (Integrated Sperm Analysis System V1.0; Proiser) and included the percentage of motile cells (MT), Curvilinear Velocity (VCL) (μm/s), Straight Line Velocity (VSL) (μm/s) and Average Path Velocity (VAP) (μm/s). All phenotypes, except sperm concentration, ORT, sperm abnormalities and cytoplasmatic droplets, were assessed at 5 and 90 min after incubation of the samples at 37°C.

Sperm phenotypes were then corrected for the fixed effects farm, season and year of collection and boar age with the "lm" function of R (R Developmental Core Team, 2010) using a multiple linear regression model. Phenotypes were then normalized by z-score. The 90 min / 5 min incubation ratios were also calculated. In total, 25 phenotypic measures per sample were recorded. Correlations across traits were assessed with the R package "corrplot" (Taiyun and Viliam, 2017).

The different analyses are described below and the complete outline is summarized in Additional Figure 1.

Genome Wide Association Study (GWAS)

Two hundred and eighty-eight boars were genotyped using the high-density (660K markers) Axiom™ Porcine Genotyping Array (Thermo Fisher Scientific). The resulting genotype dataset was stringently filtered by excluding these samples with a genotype call rate below 96%. SNP locations were converted from Sscofa10.2 to Sscrofa11.1 coordinates using plink v1.9 (Purcell et al., 2007). We then excluded SNPs which (i) had a minimum allele frequency below 0.05, (ii) did not conform to Hardy-Weinberg expectations (P-value < 0.001) and (iii) showed above 5% of missing genotypes. Single-SNP association analysis was carried with the GCTA v.1.91.5 software (Yang et al., 2011) with the following model:

$$Y_{ijkl} = \mu + \mathrm{SNP}_i + \mathrm{Farm_j} + \mathrm{SeasonYear_k} + \mathrm{Age_l} + \mathrm{e}_{ijkl}$$

where (Y_{ijkl}) is the phenotype modeled as a function of the population mean (μ), fixed effect of each SNP (SNP_i), fixed effect of farm ($\mathrm{Farm_j}$), season and year ($\mathrm{SeasonYear_k}$), age ($\mathrm{Age_l}$) and a random residual effect (e_{ijkl}).

We adopted a SNP significance threshold of corrected P-values with FDR. Significantly associated SNPs with consecutive distance below 2 Mbp were considered to belong to the same GWAS interval. A new interval was called if the consecutive SNPs were > 2 Mbp apart. SNPs located in scaffolds were not considered for the analyses. Genetic heritability was assessed with GCTA v.1.91.5 (Yang et al., 2011). Manhattan plots were performed with the "qqman" R package (Turner, 2014).

RNA isolation, sequencing and gene annotation

RNA isolation from 40 sperm samples was performed as previously described (Gòdia et al., 2018b). 35 of these samples corresponded to boars included in the GWAS. The other 5 boars did not pass the genotyping quality control and were thus not included in the GWAS. Extracted RNA was subjected to quality control assays including quantification with the Qubit™ RNA HS Assay kit (Invitrogen), assessment of RNA integrity with the 2100 Bioanalyzer using the Agilent RNA 6000 Pico kit (Agilent Technologies) and evaluation by RT-qPCR of the sperm-specific *PRM1,* the somatic *PTPRC* mRNA and genomic DNA to confirm that the samples were free from somatic cell RNA and gDNA contaminations.

The ribosomal RNA (rRNA) from the 40 RNA samples was depleted with the Ribosomal RNA depletion Kit (Illumina) and libraries were prepared with the SMARTer Low Input Library Prep kit (Clontech) and sequenced to generate 75 bp pair-end reads in an Illumina's HiSeq2500. Undepleted total RNA was also subjected to short noncoding RNA (sncRNA) library preparation (34 of the

previous 40 samples) using the NEBNext library prep kit (New England Biolabs) and sequenced at 50 bp single-end in a Hiseq2000 (Illumina).

Total RNA-seq reads were evaluated for quality control with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Low quality reads and sequencing adaptors were trimmed with Trimmomatic v.0.36 (Bolger et al., 2014). Filtered reads were mapped to the porcine genome (Sscorfa 11.1) using HISAT2 v.2.1.0 (Kim et al., 2015). Duplicate reads were removed with Picard Tools v.2.18.29 (http://picard.sourceforge.net) Markduplicates. RNA levels of the genes annotated in the porcine genome (Ensembl v.91) were then quantified with StringTie v.1.3.4 (Pertea et al., 2015) . Only genes with average RNA abundances ≥ 10 Fragments Per Kilobase of exon per million reads mapped (FPKM) were kept for further analysis with the aim to discard low abundant genes and spuriously mapped reads.

The influence of fixed effects on gene abundance was estimated using the Variance Component Estimation (VCE) (Groeneveld, 1994) with the mixed model effect:

$$_{ijklmn} = \mu + G_i + \mathrm{A}_j + \mathrm{GR}_K + \mathrm{G}Y_l + \mathrm{GF}_m + \mathrm{G}A_n + \mathrm{e}_{ijklmn}$$

where FPKM (Y_{ijklmn}) was modeled as a function of the fixed effects of a gene (Gi) and animal (Aj) plus the random effects of gene by run (GR_k), gene by year and season (GYS_l), gene by farm (GF_m) and gene by age (GA_n). Random residual (e_{ijklmn}) was assumed to be independent and identically distributed.

The RNA abundance of each gene was stabilized with the log2 transformation. Corrected abundances were used to calculate the Pearson correlation of each gene with each of the 25 measured phenotypes.

For the sncRNA-seq data, trimming of adaptors and low quality bases was performed with Cutadapt v1.0 (Martin, 2011). Reads were mapped to the *Sus scrofa* genome (Sscrofa11.1) with the sRNAtoolbox v.6.17 (Rueda et al., 2015)

using default settings and with the porcine miRBase (Kozomara and Griffiths-Jones, 2011) release 21 database. Multi-adjusted read counts were normalized by library size as counts per million (CPM). Only miRNAs with average abundance > 1 CPM in all the samples were considered. miRNA abundance was stabilized with the log2 transformation.

SNP calling from RNA-seq data and Linkage Disequilibrium with GWAS lead SNP

Mapped RNA-seq reads of the 35 samples with RNA-seq and genotype data were subjected to SNP calling. Variant calling was performed with SAMtools mpileup and BCFtools v.1.9 (Li et al., 2009). Only SNP variants found in at least 10 samples with minimum Phred quality of 25 and minimum read depth of 10 were kept. The effect of the SNP on protein sequence was predicted with SnpEff v.4.3T (Cingolani et al., 2012). The new SNP genotypes were merged to the Axiom genotypes and Linkage Disequilibrium (LD) R^2 between GWAS lead SNPs and RNA-seq SNPs was assessed with PLINK v1.9 (Purcell et al., 2007).

Expression GWAS analysis

Expression GWAS (eGWAS) analysis included the 35 samples with RNA-seq and genotype data. The RNA abundances of the detected genes were taken as quantitative traits and tested for association with the genotypes that passed quality control using a linear model. Single-SNP association analysis was carried with the GCTA v.1.91.5 software (Yang et al., 2011).

Only eGWAS significant associations (FDR < 0.05) were considered if: (i) the expressed SNP (eSNP) was also a significant hit (FDR < 0.05) on the GWAS for sperm quality phenotypes and (b) the gene abundance correlated to the same phenotype as the GWAS SNP hit.

SNP co-association and gene co-abundance analyses

For the SNP co-association analysis, GWAS results were used to build an Associated Weight Matrix (AWM) (Fortes et al., 2010; Reverter and Fortes, 2013). Our study included viability at 5 min (VIAB_0) as key phenotype, as live cells with intact plasma membrane are essential for fertilization (Berger et al., 1996; Quintero-Moreno et al., 2004). The SNPs associated to this phenotype with P-value ≤ 0.01 were included in the AWM. The dependency among phenotypes was estimated based on the average number of non-key phenotypes associated (Ap) with these SNPs (P-value ≤ 0.01) (Ap ≥ 2). Then, SNPs located less than 2,500 bp or more than 1 Mbp from the nearest annotated gene (Ensembl v.91) were kept. The most significant SNP from each annotated gene was kept to build the AWM. The standardized SNP effects across phenotypes were computed in a hierarchical cluster analysis using Euclidean distance and the single method with the R package "dendextend" (Galili, 2015). Then, significant gene-gene interactions to build the SNP network were assessed with the Partial Correlation coefficient with Information Theory (PCIT) algorithm (Reverter and Chan, 2008). PCIT applies first-order partial correlation coefficients together with an information theory approach to identify meaningful gene–gene associations (Reverter and Chan, 2008). Only significant gene co-associations were kept in the SNP network.

For the RNA co-abundance analysis, significant gene-gene interactions to build the RNA network were also predicted with PCIT using the normalized RNA abundances. Interactions between genes and miRNAs were also assessed with PCIT (Reverter and Chan, 2008), and only negative significant correlations were kept.

Integration of SNP and RNA network and network visualization

To obtain a robust gene interaction network, only the pair-wise interactions present in both the SNP and the RNA networks were kept. The resulting

network was named the "Shared Network". In addition, these genes not present in the Shared Network but that presented abundance correlation with > 3 phenotypes were merged with the shared network to create the so-called "Final Network". This final network also included the interactions between miRNA and mRNA genes. Network visualization was performed with Cytoscape v3.6 (Shannon et al., 2003), and included information on: (i) the number of phenotypes associated to a gene, (ii) the highest correlated phenotype for each gene, (iii) whether the genes was annotated as a Transcription Factor (TF) or TF co-factor, and (iv) whether the gene was present in the Shared Network or not. TF and TF co-factors were extracted from the AnimalTFDB3.0 database (Hu et al., 2019a).

Development of a RNA and SNP model for the phenotypic prediction of sperm quality

The RNA abundance of a subset of genes of the network was used to identify which combination of thesewas a better predictor of the sperm quality phenotypes. For this, we first extracted 20 genes of the network. These genes were (i) correlated with at least 4 phenotypes, (ii) did not present interactions (edges) between them, (iii) all samples presented RNA abundance levels > 0 FPKM and (iv) were potentially relevant according to the existing literature . The RSQUARE method from the SAS software was used as an exploratory model building to evaluate all possible subsets of linear regressions using gene abundances and sperm phenotypes and extract the R^2 magnitude from each prediction. Then, we selected the subset of 10 genes that were most commonly present in all the phenotype models. This subset of common genes was then used for the STEPWISE method from the SAS software, which performs a linear regression analysis for each of the phenotypes to develop a model to predict the phenotype based on gene RNA levels.

We also developed a genome-wide SNP marker model to identify the polymorphisms that could better predict the phenotypic variance of sperm-related traits. The model included the lead SNPs from the GWAS and from the eGWAS hits and the most significant SNP for each of the genes included in the network that were: (i) correlated with at least 4 phenotypes and (ii) found in the shared network. The proportion of the variance explained by these polymorphisms was assessed with GCTA v.1.91.5 (Yang et al., 2011).

Results

Phenotype statistics

Three hundred ejaculates were phenotyped for 25 sperm quality traits (Table 1). Phenotype correlations (Additional Figure 2) were in agreement with their physiological similarities. In general, SNP-based heritabilities (Table 1) were low to moderate with motility related traits displaying higher values. MT_90 was the most heritable trait (h^2: 0.39). On the other side, motility ratios, NABN and VIAB_0 showed nearly null heritability (Table 1). The sperm phenotypes correlated with farm, boar age and Season per Year (Additional File 1). These parameters were thus included as fixed effects in the GWAS model and they were also used for correction in the correlation analysis.

Table 1. Descriptive statistics, genomic heritability (h^2) and number of significant SNPs for sperm quality parameters (n=300).

Trait	Acronym	Mean	SD	h^2	# SNPs FDR < 0.05
Concentration (sperm/ml)	CON	141.3	65.5	0.13	0
Viability 5 min	VIAB_0	90.1	6.3	$1x10^{-6}$	0
Viability 90 min	VIAB_90	77.4	17.3	0.14	0
Osmotic Resistance Test	ORT	79.8	12.5	0.13	0
Head abnormalities	HABN	2.1	5.9	0.16	44
Neck abnormalities	NABN	3.0	4.9	$1x10^{-6}$	21
Tail abnormalities	TABN	2.7	3.4	0.09	0
Proximal droplets	PDROP	3.5	5.1	0.12	1
Distal droplets	DDROP	4.5	4.5	0.06	0
Motility 5 min	MT_0	75.4	18.1	0.21	17
Average Path Velocity 5 min	VAP_0	34.0	10.2	0.17	0
Curvilinear Velocity 5 min	VCL_0	46.2	12.5	0.11	0
Straight Line Velocity 5 min	VSL_0	27.0	8.3	0.23	2
Motility 90 min	MT_90	64.1	22.0	0.39	17
Average Path Velocity 90 min	VAP_90	30.8	9.5	0.35	0
Curvilinear Velocity 90 min	VCL_90	39.7	10.2	0.35	0
Straight Line Velocity 90 min	VSL_90	25.9	8.3	0.34	0
Abnormal Acrosomes 5 min	ACRO_0	7.0	5.6	0.08	4
Abnormal Acrosomes 90 min	ACRO_90	16.4	12.6	0.06	0
Ratio Motility	R_MT	0.9	0.2	$1x10^{-6}$	0
Ratio Average Path Velocity	R_VAP	0.9	0.3	$1x10^{-6}$	0
Ratio Curvilinear Velocity	R_VCL	0.9	0.3	$1x10^{-6}$	0
Ratio Straight Line Velocity	R_VSL	1.0	0.3	0.06	0
Ratio Viability	R_VIAB	0.9	0.3	0.08	0
Ratio Acrosomes	R_ACRO	3.4	3.5	0.08	1

SD=Standard Deviation; h^2=heritability; All traits excepting concentration are presented as a percentage. The values shown are raw excepting the ratios which are corrected and stabilized.

GWAS analysis

After quality control, 466,592 SNPs and 276 samples remained for the GWAS. The number of SNPs displaying significant associations (FDR < 0.05) for each trait is summarized in Table 2. A total of 19 genomic regions tagged by 71 significant SNPs were identified in chromosomes SSC1, 3, 4, 6, 7, 9, 13, 16 and X. Thirty-six additional significant SNPs had unknown positions in the genome. Seven sperm quality traits exhibited significant association signals (Figure 1),

and only one SNP signal was associated to more than 1 trait (Table 2). HABN and NABN presented the largest number of SNP signals with 42 and 18 associated SNPs, respectively (Figure 1. A and C). Seven of the 19 QTLs were represented by only 1 associated SNP and were discarded from further analyses (Table 2; Figure 1). The most significant SNPs (rs318575212 and rs332927981) of the study were associated to ACRO_0, (both with FDR = 0.006 and Additive effect = 4.11) (Table 2).

Table 2. Summary of the results of the genome wide association analysis for sperm quality traits.

SSC	Interval	#SNP	Interval, Mbp	Top SNP	Top SNP location, bp	Top SNP P-value	Top SNP FDR	Top SNP MAF	Trait
1	I1	1	-	rs339761632	13,501,755	$4.64x10^{-8}$	0.02	0.06	PDROP
1	I2	8	82.90-83.49	rs81354986	82,895,619	$1.69x10^{-6}$	0.03	0.07	HABN
1	I3	8	94.88-98.74	rs327733412	94,880,167	$1.61x10^{-7}$	0.02	0.07	HABN
1	I4	1	-	rs337166779	126,397,198	$2.05x10^{-6}$	0.03	0.06	HABN
1	I5	11	243.86-246.44	rs343194423	246,224,386	$1.72x10^{-7}$	0.01	0.07	NABN
1	I6	2	258.54-258.55	rs332256425	258,548,786	$1.76x10^{-6}$	0.04	0.06	NABN
3	I1	1	-	rs332055717	2,911,413	$6.35x10^{-8}$	0.01	0.09	HABN
3	I2	3	113.75-113.84	rs328292697	113,750,595	$1.09x10^{-7}$	0.01	0.07	NABN
4	I1	2	2.41-2.42	rs318575212, rs332927981	2,412,006, 2,415,239	$2.88x10^{-8}$	0.01	0.08	ACRO_0
6	I1	2	65.60-66.66	rs335394654	65,597,553	$1.86x10^{-7}$	0.03	0.14	ACRO_0
7	I1	2	6.20-6.38	rs326239534	6,377,172	$9.87x10^{-6}$	0.02	0.17	MT_0
7	I2	2	85.73-86.88	rs336588919	86,884,279	$4.13x10^{-8}$	0.01	0.06	NABN
9	I1	2	5.76-5.78	rs697275015	5,776,597	$1.55x10^{-7}$	0.02	0.07	HABN
9	I2	1	-	rs342738178	28,463,580	$1.53x10^{-5}$	0.03	0.14	MT_0, MT_90
9	I3	1	-	rs328217450	137,959,590	$4.77x10^{-8}$	0.02	0.18	R_ACRO
13	I1	18	25.36-28.47	rs690794887	25,535,100	$3.06x10^{-7}$	0.02	0.14	HABN
13	I2	3	33.82-37.65	rs327865244	33,819,549	$3.79x10^{-8}$	0.01	0.15	HABN
16	I1	1	-	rs324239602	6,476,358	$6.08x10^{-6}$	0.01	0.46	MT_90
X	I1	1	-	rs324336668	1,125,922	$2.04x10^{-7}$	0.02	0.07	HABN

SSC = S.scrofa chromosome; #SNP = number of single nucleotide polymorphisms significantly associated to the phenotype; Interval: region of the chromosome that spans the significant SNPs; FDR = False Discovery Rate; MAF = Minor Allele Frequency; ACRO_0 = Abnormal Acrosomes 5 min; HABN = Head abnormalities; NABN = Neck abnormalities; PDROP = Proximal droplets; R_ACRO = Ratio Acrosomes; MT_0 = Motility 5 min; MT_90 = Motility 90 min.

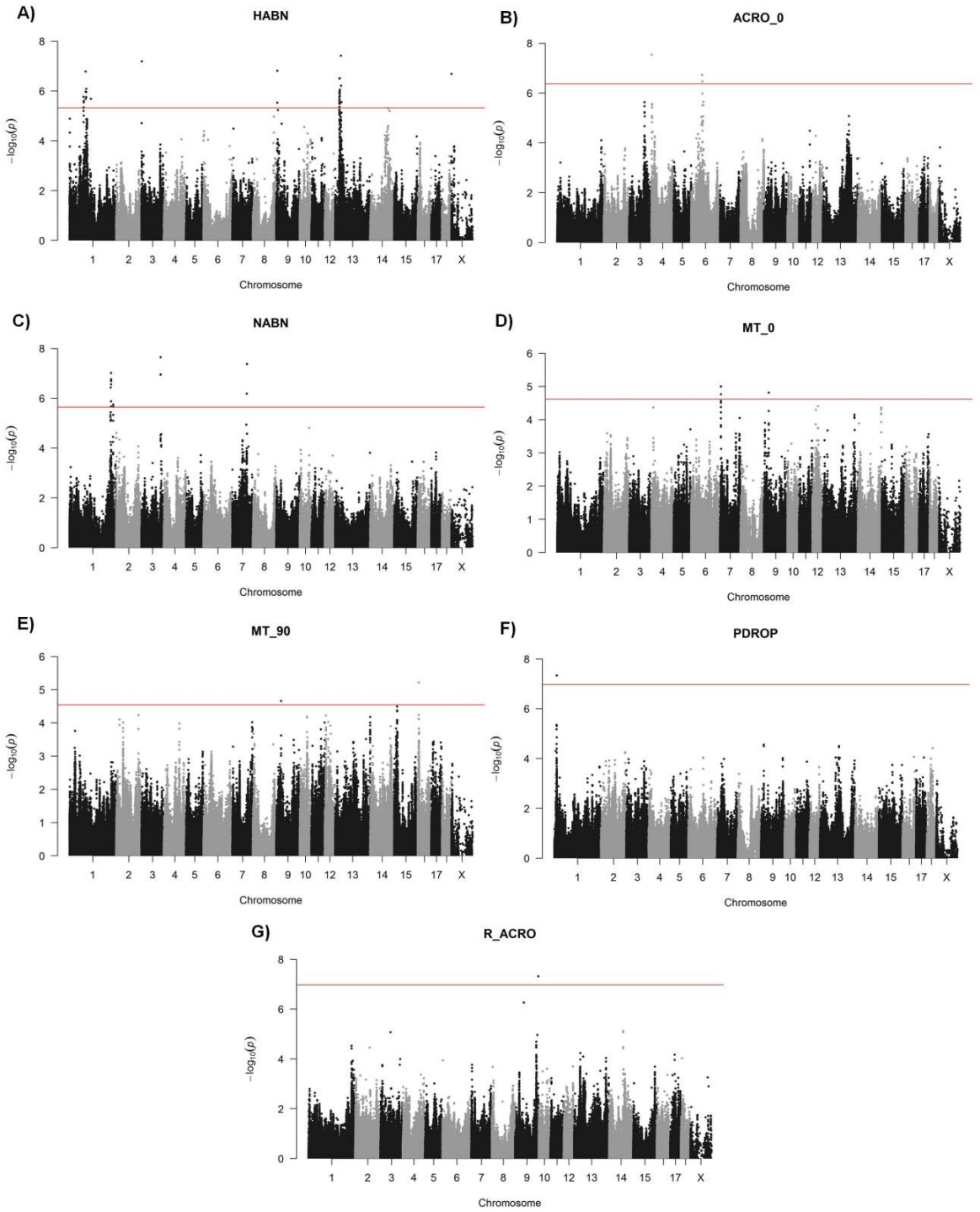


Figure 1. Manhattan plots depicting the genome-wide significant associations between SNP markers and sperm quality traits. Significant associations have been found in: **A)** Head abnormalities (HABN); **B)** Acrosomes at 5 min (ACRO_0); **C)** Neck abnormalities (NABN); **D)** Motility at 5 min (MT_0); **E)** Motility at 90 min (MT_90); **F)** Proximal Droplets (PRDOP); **G)** Acrosome's Ratio (R_ACRO). The *x*-axis represents chromosome length (Mb), and the *y*-axis shows the negative $\log_{10}$ *P*-values of the associations found. The horizontal red line represents the significance threshold (FDR ≤ 0.05).

Sperm RNA isolation, RNA-seq and bioinformatics analysis

Isolated RNA from mature spermatozoa was free from somatic cell RNA. Total RNA-seq resulted in an average of 40.7 M reads per sample and 98.2% of the reads passed the quality control filters (Additional File 2). An average of 83% of the reads mapped to the porcine genome and after duplicate removal and RNA abundance filters, we identified 4,120 genes. RNA levels were minimally affected by external factors with cumulative variance (< 2%). For short RNA-seq, we obtained an average of 7.3 M of reads per sample. Of these, 99.2% passed quality control and 81.5% mapped to the porcine genome (Additional File 2). We identified 95 miRNAs out of the 306 that are annotated in swine (Additional File 3).

SNP calling from RNA-seq and Linkage Disequilibrium with GWAS hits

In order to help predicting whether a GWAS hit could be tagging a causal variant altering protein sequence and function and to identify additional SNPs with the potential to be better markers in a GWAS , we sought to identify variants in annotated genes using the RNA-seq. As a requisite, these variants had to be in LD with the cognate GWAS hit. After filtering, we identified 7,719 expressed variants, 37 of which mapped within the genomic intervals identified by the GWAS (Table 2). Twenty-three SNPs were predicted to have low impact effect on protein sequence (synonymous variants and 5' UTR premature start codon), 13 SNPs showed moderate effect (missense variants) and 1 SNP was predicted to have a high impact on protein sequence (splice donor variant).

SSC13 I1 associated to HABN, harbored 21 expressed SNPs (7 and 14 with moderate and low effect, respectively). The polymorphism rs331304027 (a missense variant with moderate effect on the *ULK4* gene) was in moderate LD (LD=0.40) with the strongest GWAS SNP hit of the interval (rs690794887) (Table 3). SSC13 I2, also associated to HABN, presented 11 SNPs (1 high, 5 moderate and 5 with low effect). Of these, the variant with highest LD (LD=0.2) with the

GWAS hit (rs327865244) was a 5′ UTR premature start codon gain (low effect) SNP (rs323872641) in the *ABHD14A* gene(Table 3). This interval was the only one that presented a high effect SNP (rs323144103), a splice donor variant, but this SNP was in low LD (LD=0.02) with the GWAS hit. Chromosome SSC7 I2, associated to NABN, encompassed 2 expressed SNPs (both with low effect). rs330912302 (a synonymous SNP in the *CHD2* gene) presented a moderate LD (LD=0.4) with the strongest hit of the interval (rs336588919) (Table 3). The chromosome SSC1 I3 associated to HABN harbored 3 expressed SNPs (1 with moderate and 2 with low effect) (Table 3).

Table 3. SNPs identified trough SNP calling in the GWAS regions.

SSC	Interval	Top SNP of the GWAS interval	# SNP called	Highest LD	SNP with highest LD	SNP effect	Gene	Trait
1	I3	rs327733412	3	0.07	rs710447566	Low	*KATNAL2*	HABN
7	I2	rs336588919	2	0.40	rs330912302	Low	*CHD2*	NABN
13	I1	rs69079487	21	0.40	rs331304027	Moderate	*ULK4*	HABN
13	I2	rs32786544	11	0.20	rs323872641	Low	*ABHD14A*	HABN

SSC=S.scrofa chromosome; #SNP called=number of single nucleotide polymorphisms identified in the SNP calling analysis. The column SNP effect and gene refer to the information of the highest LD SNP of the region. LD=linkage disequilibrium; HABN=Head Abnormalities; NABN=Neck Abnormalities.

Expression GWAS analysis

In order to predict whether the GWAS hits were tagging a causal variant altering gene expression we carried an eGWAS. eGWAS was performed with the genotypes of 464,020 SNPs that passed the quality control and normalized RNA abundances. Correlation analysis of the 4,120 genes and the 25 phenotypes resulted in 6,128 significant correlations (P-value < 0.05) involving 3,007 genes and the 25 traits. These genes, presented between 1 and 9 significant

correlations with the different semen quality traits (Additional File 4). 344 genes were significantly correlated with ≥ 4 traits. We then focused only on the associations between GWAS SNP hits (with FDR < 0.05) and transcripts which abundances correlated with the same phenotype. We identified 45 eSNPs (FDR < 0.05) located in 3 genomic regions related to 2 ACRO_0 and HABN (Table 4). Six eSNPs had unknown positions in the genome after liftover from Sscrofa10.2 to Sscrofa11.1. The remaining eGWAS hits were in chromosomes SSC4, 6 and 13 (Table 4; Additional file 5). All the eSNPs had a trans effect, related to genes located in different chromosomes. The eQTL identified in SSC4, was related to ACRO_0 and was associated to 3 genes, *NCLN, ASCC1* and *AATF*. Also involving ACRO_0, the eQTL in SSC6 was associated to the *IQCJ* gene. Finally, the eQTL in SSC13 for HABN, included SNPs associated to *HARS, ACTR2, EPB41L3* and *RAB1B*.

Table 4. Summary of the results of the within-trait expression genome wide association analysis.

SSC	Interval	# eSNP /transcipts	Top eGWAS	Top eGWAS location, bp	Top eGWAS P-value	Top eGWAS FDR	Top eGWAS MAF	Trait	Correlation	Associated Gene
4	I1	2	rs318575212, rs332927981	2,412,006, 2,415,239	7.36×10^{-3}	0.03	0.09	ACRO_0	-0.33	*NCLN*
		2	rs318575212, rs332927981	2,412,006, 2,415,239	1.83×10^{-4}	0.03	0.09	ACRO_0	-0.46	*ASCC1*
		2	rs318575212, rs332927981	2,412,006, 2,415,239	2.87×10^{-4}	4.83×10^{-2}	0.09	ACRO_0	-0.4	*AATF*
6	I1	2	rs335394654	65,597,553	5.63×10^{-5}	0.02	0.11	ACRO_0	-0.35	*IQCJ*
13	I1	31	rs328397029	25,684,259	1.84×10^{-5}	2.95×10^{-3}	0.09	HABN	-0.38	*HARS, ACTR2, EPB41L3, RAB1B*

SSC = S.scrofa chromosome; # eSNP/trancsipts = number of single nucleotide polymorphisms significantly associated to a transcript; MAF= Minor Allele Frequency; ACRO_0 = Abnormal Acrosomes 5 min; HABN = Head abnormalities.

Gene network analysis

To carry the SNP co-association analysis, 2,648 of the 466,592 SNPs that passed the quality control filters were retained in the AWM. Hierarchical cluster distributions were in agreement with the biological similarities and phenotypic correlations identified (Additional Figure 2 and 3). A clear separation between (i) morphological abnormalities and motility parameters and (ii) cell viability and ORT was observed based on the additive effects of the SNPs calculated in the association analysis. In keeping with previous studies (Ramayo-Caldas et al., 2016; Snelling et al., 2013), the SNPs detected with the AWM explained 74.1% of the key phenotypic variance (VIAB_0). The SNP network predicted with PCIT (Reverter and Chan, 2008) resulted in significant correlations involving 2,648 nodes (all the genes) connected by 2,984,616 edges.

For the RNA network analysis, the RNA levels of the 4,120 detected genes were used to identify potential connections with PCIT (Reverter and Chan, 2008). The RNA network included 4,120 nodes (all the genes) connected by 1,173,995 edges. PCIT also built 4,539 significant interactions between 95 miRNAs and 630 genes.

To obtain the Shared Network, common SNP and RNA network edges were extracted thus focusing in the shared set of interacting genes from both approaches. This comparison resulted in 613 nodes connected by 16,591 edges. The Final Network included a set of 700 additional genes (as they correlated with > 3 phenotypes) and their interactions. Moreover, the Final Network also involved 1,564 edges connecting 202 genes and 94 miRNAs. Of the 1,313 genes included in the Final Network, the abundance of 1,135 correlated with at least one phenotype, 68 have been reported as TFs and 89 as TFcos (Figure 2.A). Nearly one quarter of the genes (282 out of the 1,313) presented at least 200 edges. The genes that presented more interactions were *PLCH2* (579 edges, present in the Final but not in Shared Network and correlated with 3

phenotypes), *CEP152* (399 edges, in the Shared Network and correlated with 4 traits) and *SLC41A2* (382 edges, in the Shared Network).

Gene ontology analysis of the genes included in the Final Network presented enrichment for DNA repair (e.g. *RAD51, SETX, SOD1*), meiotic cell cycle (e.g. *BAG6, HSPA2, RAD51*), gamete generation (e.g. *TSSK3, PRDM14, PRKAR1A*) and spermatogenesis (e.g. *BAG6, CAPZA3, HSPA2*) (Additional File 6).

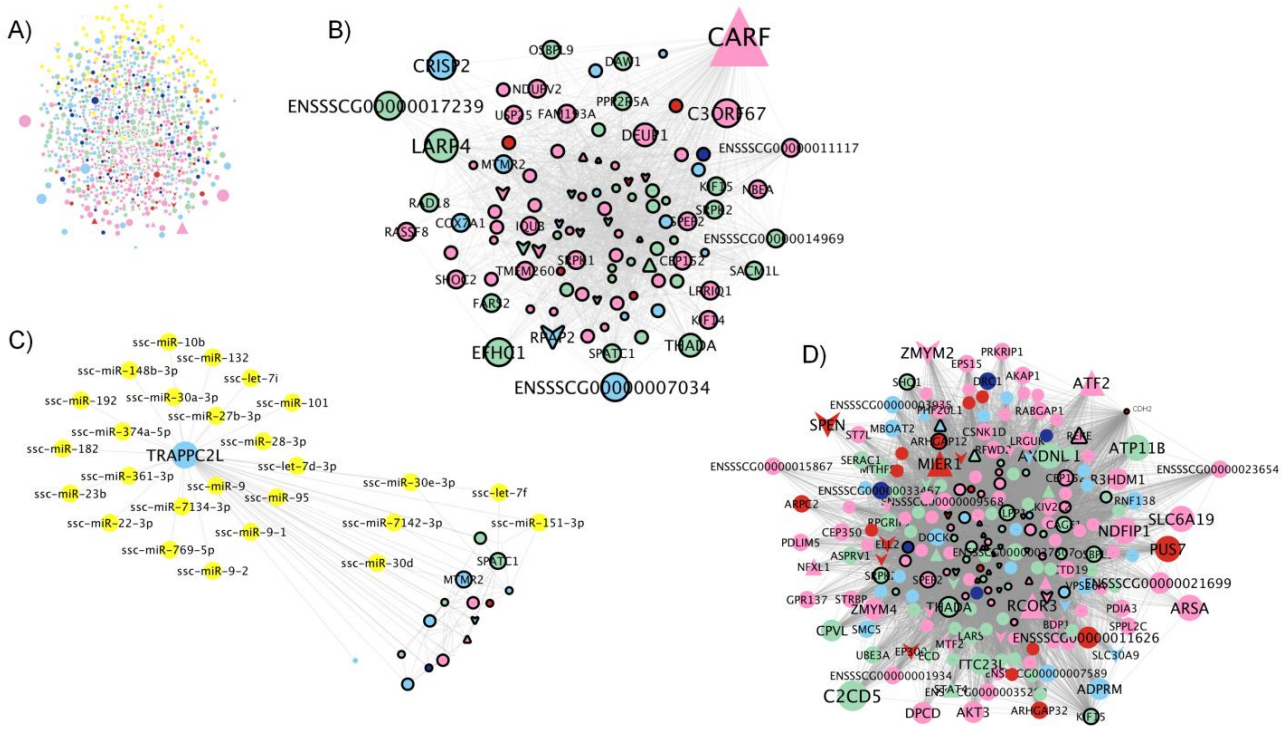


Figure 2. Co-association network based on the AWM approach and trancriptomics data. **A)** Full network with 1,313 genes and 94 miRNAs. **B)** Subset of the network showing the transcription factor *CARF* and all its predicted interactions. **C)** Subset of the network with the *TRAPPC2L* interactions, which included several miRNAs. **D)** Subset of the network with the *CDH2* gene interactions. The node color corresponds to the phenotype group with the highest correlation value, as follows: concentration (red), acrosomes (green), abnormalities and droplets (pink), osmotic resistance test (orange), motility (light blue) and viability (dark blue). miRNAs are depicted in yellow. Node size and text correspond to the number of significant phenotypes correlated with that gene. Nodes with a black line border correspond to genes identified in the shared network. Node shape indicates classification as: triangle (TF), V (TF co-factor) and ellipse (other genes and miRNAs).

Development of a RNA and a SNP models

The R^2 model predicted that the RNA levels of 20 genes could explain between 55 to 78% of the phenotypic variation across traits. The selection of 10 genes that were most commonly present in all the phenotype models explained the vast

majority (93 to 99%) of the phenotypic variation that was predicted by the model. The final set of 10 genes included in the linear regression model was: *MICAL3*, *EFHC1*, *TRAPPC2L*, *ATP9A*, *THADA*, *MOBKL3*, *BLVRB*, *LARP4*, *CARS2* and *NDUFV2*. The analysis resulted in significant models for 10 of the 25 phenotypes (Table 5). The most significant model was for PDROP and could predict the phenotype with an efficiency of 68% (Table 4). The estimated parameters of the significant models can be found in Additional File 7.

The SNP-based model was built with 74 polymorphisms (19 lead SNPs from GWAS hits, 2 lead SNPs from the eGWAS hits, 53 SNPs from the Shared Network and ≥ 4 phenotypes). These polymorphisms could explain between 5 to 36% of the phenotypic variance across the 25 traits (Table 5). A moderate proportion (>20%) of the phenotypic variance could be explained for 17 of the 25 traits. The best predictions were for sperm abnormalities (NABN, HABN, TABN) and sperm motility related traits (e.g. MT_0, VAP_90 and VCL_90) (Table 5).

Table 5. Results from the RNA and SNP models.

	RNA Model		SNP Model	
Acronym	**R^2**	**P-value**	**Phenotipic variance**	**SE**
CON	0.17	0.82	0.05	0.05
VIAB_0	0.43	0.06	0.27	0.07
VIAB_90	0.23	0.61	0.28	0.07
ORT	0.22	0.62	0.24	0.07
HABN	0.16	0.84	0.29	0.06
NABN	0.22	0.64	0.36	0.07
TABN	0.26	0.49	0.26	0.07
PDROP	0.68	<.0001	0.17	0.07
DDROP	0.42	0.07	0.06	0.05
MT_0	0.46	0.03	0.31	0.07
VAP_0	0.58	0.002	0.34	0.07
VCL_0	0.61	0.001	0.33	0.07
VSL_0	0.36	0.16	0.31	0.07
MT_90	0.34	0.22	0.30	0.07
VAP_90	0.55	0.005	0.34	0.07
VCL_90	0.55	0.01	0.34	0.07
VSL_90	0.61	0.001	0.33	0.07
ACRO_0	0.50	0.02	0.21	0.06
ACRO_90	0.21	0.68	0.23	0.07
R_MT	0.30	0.35	0.13	0.06
R_VAP	0.18	0.79	0.18	0.07
R_VCL	0.28	0.42	0.14	0.07
R_VSL	0.21	0.68	0.21	0.07
R_VIAB	0.44	0.05	0.06	0.04
R_ACRO	0.57	0.003	0.19	0.07

SE: Standard Error

Discussion

GWAS analysis

Investigating the genomic regions and molecular processes controlling boar sperm quality has become a focus of interest (Diniz et al., 2014; Marques et al., 2018; Zhao et al., 2016) for its relevance on the sustainability of pig breeding and production. In fact, our results as well as studies from other groups (Diniz et al., 2014; Marques et al., 2018; Smital et al., 2005; Wolf, 2009), have shown that boar sperm quality has a genetic basis and that it can thus be selected for breeding strategies.

In our study, sperm motility traits presented the largest heritabilities, followed by HABN and VIAB_90 (Table 1). In addition, nearly all sperm traits measured after 90 min of incubation experienced a significant drop of quality when compared to their 5 min counterparts (Table 1). VIAB, ACRO, MT and VCL (Wilcoxon test, P-value < $2.2x\ 20^{-16}$), VAP (P-value: $4.0x\ 20^{-9}$) and VSL (P-value: 0.03) showed significant differences.

The GWAS revealed 12 QTL regions represented by 2 or more significant SNPs and several positional candidate genes for HABN, NABN, ACRO_0 and MT_0. The highest signals were on SSC4 for ACRO_0 (~2.41-2.42 Mbp), ~69 kb upstream of the Solute Carrier Family 45 Member 4 (*SLC45A4*) gene. *SLC45A4* encodes a proton-coupled sugar transporter implicated in the nutrition of spermatozoa during their maturation in epididymis (Vitavska and Wieczorek, 2017) where acrosome assembly continues its posttesticular sperm maturation (Olson et al., 2003). On SSC7, we identified the Solute Carrier Family 35 Member B3 (*SLC35B3)* as a potential candidate for the MT_0 QTL. *SLC35B3* located less than 1 Mbp away from this QTL. Although a role in sperm has not been reported thus far, the SLC35 gene family has been postulated to play a role as nucleotide sugar transporter (Song, 2013) and we propose that it may also play a role in the nutritional support of spermatozoa.

We detected several significant regions for HABN. The QTL on SSC1 I2 (~94.9-98.8 Mbp) included interesting candidate genes such as the Katanin Catalytic Subunit A1 Like 2 (*KATNAL2*). Dunleavy et al. (Dunleavy et al., 2017) reported that *Katnal2* is a critical regulator of male germ cell development affecting sperm head shaping, acrosome attachment and sperm tail growth. Other candidate genes in that region were the solute carrier *SLC14A2,* encoding the urea transporter A, suggested to participate in sperm head formation by reducing its volume though excreting urea (Li et al., 2012), or the SMAD Family Member 2 (*SMAD2*) involved in spermatogonial differentiation (Wu et al.,

2017). On SSC13 I1, we identified two candidate genes: the Testis and Ovary-specific PAZ domain gene 1 (*TOPAZ1*) and the IQ Motif Containing F1 (*IQCF1*). Luangpraseuth-Prosper et al. (Luangpraseuth-Prosper et al., 2015) demonstrated that *Topaz1* knockout mice presented meiotic arrest and caused male infertility. As for *IQCF1,* Fang et al. (Fang et al., 2015) reported that this gene localizes in the acrosome and that it is involved in sperm capacitation in mice. *Iqcf1*$^{-/-}$ mice were significantly less fertile than wild type mice (Fang et al., 2015). The QTL region on SSC13 I2 included the candidate Protein Kinase C Delta (*PRKCD*) gene. *PRKCD* has been involved in spermatogenesis and embryonic development (Suh et al., 2003) and was found associated in a GWAS for semen volume in Holstein-Friesian bulls (Hering et al., 2014).

Four QTL regions were identified for NABN. The QTL on SSC1 I5 included as a candidate gene the transporter ATP Binding Cassette Subfamily A Member 1 (*ABCA1*). In humans, ABCA1 localizes in the dorsal side of the sperm head and in the middle piece of the tail (Morales et al., 2008). It has been suggested to contribute to cholesterol transport and fertilization capacity (Morales et al., 2008). The QTL in SSC7 I2 included two genes of interest, the Chromodomain Helicase DNA-binding protein 2 (*CHD2*) and the Sialyltransferase 2 (*ST8SIA2*). *CHD2* may be playing an important role in DNA damage response and genome stability maintenance (Nagarajan et al., 2009). In humans, *CHD2* has been associated with non-obstructive azoospermia (Qin et al., 2014). Simon et al. (Simon et al., 2013) demonstrated that the protein encoded by *ST8SIA2* is located in the post-acrosomal region of human sperm. It generates polysialic acid, which is suggested to act as a cytoprotective element to increase the number of vital/live sperm (Simon et al., 2013).

Our study allowed the identification of a moderate number of QTLs associated to sperm quality. Nevertheless, there were no overlapping regions with any of the previously published GWAS studies (Diniz et al., 2014; Marques et al., 2018;

Zhao et al., 2016). Only 1 of our QTLs (SSC1 I6), associated to NABN, mapped 335 kbp downstream to another QTL for boar sperm abnormalities and motility (Marques et al., 2018). The discrepancies across studies could arise due to different technical (e.g., sample size), environmental or genetic causes..

SNP calling from RNA-seq data

Calling genomic variants from RNA-seq data can be a complementary method to detect previously unknown or ungenotyped polymorphisms in transcribed genes that might carry important functional implications or may be better genetic markers for that given trait. Should these genes be involved in related phenotypes and these variants be: (i) in LD with the GWAS lead SNP and (ii) have a predicted effect on protein sequence, these polymorphisms could be suggested as potential causal candidates. For that purpose, we sought to identify transcribed variants in the QTL regions and assessed their LD with the lead SNP hit of the QTL.

For HABN we found new genetic variants in genes of physiological interest (Table 3). On SSC13 I1, we discovered several variants in the Unc-51 Like Kinase 4 (*ULK4*) gene in moderate LD with the lead SNP of this GWAS hit (Table 3). Although *ULK4* has not been associated to sperm defects, Liu et al. (Liu et al., 2016) showed that *Ulk4* has an essential role in ciliogenesis, the process of formation of cilium or flagellum, a microtubular structure located in the center of all motile cilia and flagella, also in sperm. In fact, the disruption of another ciliogenesis-related gene (*IFT25*) has resulted in infertile males with round sperm heads and abnormal tails (Liu et al., 2017). On SSC13 I2 we identified one variant with predicted high effect in the *IQCF4* gene. The IQ Motif family of proteins have been reported in myosins and promote calcium regulation (Bahler and Rhoads, 2002). Myosins are actin-based motors that translocate along actin filaments in an ATP-depending manner and have been implicated in various aspects of spermatogenesis (Hu et al., 2019b). In sperm,

actin filaments are located in the acrosomal region (Breitbart et al., 2005). Interestingly, the previously discussed GWAS positional and physiological candidate genes *CHD2* and *KATNAL2*, also presented genetic variants in LD with the lead SNPs at SSC7 I2 (low effects: rs330912302 LD = 0.4 and rs339719658 LD = 0.37) and SSC1 I3 (low effects: rs700749617 LD = 0.01, rs710447566 LD = 0.07 or moderate effect: rs690151450 LD = 6.9×10^{-3}), respectively.

eGWAS

In this study, we also performed a within-trait eGWAS linking for each phenotype, GWAS lead SNPs with genes which RNA abundance correlated with the same trait. We identified 3 eQTLs all with a *trans*-effect. *trans*-eQTL hotspots are of particular interest as their SNPs could harbor important regulatory roles and variations influencing gene expression and thus are more likely to contribute to the phenotype. The *trans*-eQTL on SSC6 was correlated with the abundance of the IQ Motif Containing J (*IQCJ*) gene, both SNP and mRNA were associated to ACRO_0. *IQCJ* is a member of the previously discussed IQ Motif family proteins. Although it has not been studied in sperm, Martin et al. (Martin et al., 2008) reported the presence of the *IQCJ-SCHIP-1* isoform in mammalian neurons and its role in calcium mediated responses. We hypothesize that *IQCJ* may also mediate calcium response in sperm. In fact, calcium has been involved in the regulation of motility, hyperactivation, capacitation and acrosome reaction (reviewed in: Sun et al., 2017).

The *trans*-eQTL on SSC13 for HABN was associated to several genes including the Actin Related Protein 2 (*ACTR2*) and Histidyl-TRNA Synthetase (*HARS*). Heid et al. (Heid et al., 2002) identified ACTR2 in bull sperm head and suggested that it serves for sperm capacitation and acrosome reaction. On the other side, HARS has been involved in attaching histidines to its corresponding tRNA molecules, a fundamental cellular process for the translation of mRNA

into protein (Ibba and Söll, 2000). Waldron et al. (Waldron et al., 2019) showed that *HARS* zebrafish knockout presented severe defects in high proliferative cells. Although its role in sperm remains to be resolved, HARS protein has been found overexpressed in sperm of low-fertility bulls (Aslam et al., 2019) and we do not rule out a potential involvement of this gene in spermatogenesis.

Gene network analysis

Despite the considerable number of candidate genes identified in our GWAS, many genes might have been missed by this traditional single-trait approach due to the lack of an acceptable significant association (FDR > 0.05). After all, sperm quality is a complex phenotype influenced by many factors, such as genetics, husbandry, environment, or testicular pathologies that influence an intricate network of genes and molecular processes. An alternative strategy to exploit GWAS information is to perform an AWM analysis that extracts SNPs that while having strong yet below the significance threshold of genetic association, are also associated to a certain number of traits (Fortes et al., 2010). The association of 1 SNP to more than 1 trait provides additional robustness to the potential relevance of that SNP – to semen quality in our case. This, followed by a PCIT analysis to study gene-gene interactions can provide information on the relevant genes and pathways for certain phenotypes and then search for SNPs in or affecting them. Obviously, transcriptomics data can contribute additional valuable information in the description of these genes and pathways. For this reason, we have addressed the genetics behind the boar's sperm quality through a systems biology integrative approach. The gene co-association and co-abundance interactions revealed a number of appealing features such as new candidate genes, TFs, TF-cos and miRNAs that belong to biological processes and relevant functions related to sperm.

The TF with the highest number of predicted interactions (129) was encoded by the Calcium Responsive Transcription Factor (*CARF*) gene, which RNA

abundance was in turn, correlated with 9 phenotypes (Figure 2.B; Study IV: Additional File 4). CARF acts as a transcriptional activator promoted by calcium influx (Tao et al., 2002). Since calcium ions are essential in spermatogenesis (reviewed in: Sun et al., 2017), we infer that this TF could be involved in pathways related to sperm maintenance and functioning. Some of the *CARF* predicted target genes from our analysis include interesting candidates such as La Ribonucleoprotein Domain Family Member 4 (*LARP4*), THADA Armadillo Repeat Containing (*THADA*) and EF-Hand Domain Containing 1 (*EFHC1*) gene. LARP4, has been proposed to regulate mRNA stability and translation of mRNAs (Blagden et al., 2009). Blagden et al. (Blagden et al., 2009) reported *Drosophila larp* knockout mutants resulted in a considerable proportion of spermatocytes with meiotic defects. Although the role of *THADA* remains uncertain in sperm, Moraru et al. (Moraru et al., 2017) showed that in *Drosophila,* THADA modulates though calcium signalling energy storage and thermogenesis balance. *EFHC1* encodes for a myoclonin1 protein and has been detected in sperm flagella in mice testis (Suzuki et al., 2008). Although *Efhc1*-deficient mice were fertile, mutants presented a reduced ciliary (flagellar) beating frequency (Suzuki et al., 2009).

Other putative TFs were the SMAD Family Member 4 (*SMAD4*) gene (interacting with 32 genes) and the Lysine Demethylase 3A (*KDM3A*) gene (281 gene interactions), both potentially targeting a set of genes enriched for cellular macromolecular complex assembly processes (Additional File 6). TFs involved in DNA repair, such as the Bromodomain Adjacent To Zinc Finger Domain 1B (*BAZ1B*), were also identified. Its closest paralog, *BAZ1A* encodes a member of the chromatin remodeling complex (Racki et al., 2009). Dowdle et al. (Dowdle et al., 2013) showed that *Baz1a*$^{-/-}$ mice were infertile because of spermatogenesis defects tied to changes in chromatin composition. Another TF of interest was the *ESR1* gene, detected through the shared network. *ESR1* has been already

associated with pig sperm motility and cytoplasmatic droplets (Gunawan et al., 2011). Polymorphisms in *ESR1* have been suggested to influence estrogen levels which in turn, affect sperm motility (Carreau et al., 2002).

The network comprised new candidate genes for sperm quality. The Trafficking Protein Particle Complex 2 Like (*TRAPPC2L*) gene, correlated with 27 miRNAs including miR-30d, a miRNA that was dysregulated in oligozoospermic infertile individuals (Salas-Huetos et al., 2015) (Figure 2.C). *TRAPPC2L* belongs to the TRAPPC family, with a reported role in ciliogeneisis (Westlake et al., 2011). Interestingly, we identified *TRAPPC2L* was found associated in the final network with the Spermatogenesis And Centriole Associated 1 (*SPATC1*) gene, a gene that has been localized in the neck region of the mouse and human sperm (Goto et al., 2010). Disruption of its homolog *Spatc1l* in mice led to male sterility due to separation of sperm heads from tails, thereby advocating a role in sperm head-tail integrity (Kim et al., 2018). The network also included *DNAI2*, correlated with 4 phenotypes. Mutations in *DNAI2* have been associated with ciliary defects and with males showing reduced fertility due to impaired sperm tail function (Loges et al., 2008). *DNAI2* has been related to boar sperm motility in a previous GWAS (Marques et al., 2018). *CHD2* is another interesting gene in the network as it was also identified as a candidate gene in our GWAS analysis. This gene also presented new DNA variants in LD with GWAS lead SNPs which would be worth testing in a genetic association study (Figure 2.D; Table 3). *CHD2* was hidroxymethylated in human sperm after exposure to bisphenol A, an epigenetic modifier that causes spermatogenesis defects and alters sperm motility (Zheng et al., 2017).

Of the 94 miRNAs identified in sperm and included in the final network, 30 were found interacting with at least 20 genes. Some of them have been previously associated to sperm quality and fertility. Noteworthy, miR-16, a miRNA that was down-regulated in the semen of infertile males with sperm

abnormalities (Liu et al., 2012), correlated with 67 genes in our study (e.g. *ATP9A*, found in the shared network and included in the RNA model). Similarly, miR-10b, previously associated with human infertile semen samples (Tian et al., 2017), correlated with 32 genes (including *TRAPPC2L* included in the Final Network).

Development of a RNA and a SNP models

In this study, we provide a novel and innovative approach to develop a RNA model to estimate the phenotypes based on gene abundances. The model, including 10 genes, was predicted to be significant for 10 phenotypes. The model performed best for PDROP and some of the motility related traits (Table 5). The model of PDROP reported a highly significant role of the *THADA* gene (Additional File 7), which at the same time has been found in the shared network positively correlated with PDROP. *THADA* regulates the metabolism via calcium signaling by binding the sarco/ER Ca^{2+} ATPase transporter mechanism (Harper et al., 2005). The *CARS2* gene was also a strong contributor in the model for PDROP. *CARS2* was also identified in the shared network and has a critical role in protein synthesis.

Although SNPs have become the marker of choice for the genetic improvement of livestock species, the development of a SNP array for the prediction of the boar's sperm quality remains to be done. Here, we have developed a SNP model with 74 SNPs including the polymorphisms identified through the GWAS, eGWAS and the shared network. The model holds promising potential for its application in animal breeding programs. This panel of 74 SNPs could estimate between 5 to 36% of the phenotypic variance across the 25 traits that were evaluated. These SNPs were better predictors for the phenotypes related to sperm abnormalities and motility (Table 5). Remarkably, when only considering the 20 GWAS SNP hits, the panel would explain between 0.04 to 0.26% of the phenotypic variance and only for 3 traits (HABN, NABN and

TABN), the model would be able to predict above 20% of the phenotypic variance. Thus, a clear drop on the predictive potential is observed.

In a previous study for sperm motility and morphological abnormalities using two porcine lines, Marques et al. identified several QTLs that cumulatively explained 10.8% of the genetic variance (Marques et al., 2018) including 412 and 271 SNPs for each line. Our approach is able to predict 30-31% and 26-36% of the variance of the same group of traits with only 74 SNPs for motility and morphological-related traits, respectively. However, we have employed an integrated and informed approach based not only on the GWAS and eGWAS FDR significant associations but also in a robust network built from co-associated SNPs (identified at suggestive levels but across several phenotypes) as well as RNA co-abundant genes. Moreover, our SNPs were chosen to minimize LD between them and thus maximize the informativity of the panel. This allowed the informed inclusion of a large number of SNPs with independent marker potential and thus the development of a powerful panel for the prediction of semen quality in pigs.

Although the results only hold in our population and the validation of the panel will require additional evaluations in other populations, the integrative approach proposed in this study to ultimately build a SNP array provides compelling results of its application to any type of complex trait with a genetic basis. This offers another avenue to improve traits confounded by several genes that are of interest for the animal breeding industry.

Conclusions

In summary, our results suggest that genetic variants identified in the 12 QTL regions mapped to - or near - *CHD2*, *KATNAL2*, *SLC14A2*, *IQCF1* and *ABCA1*, together with other candidate genes based on a systems biology approach including *LAPR4*, *THADA*, *EFHC1*, *SMADA4*, *SPATC1* or *TRAPPC2L*, among others, may modulate sperm quality in pigs. This network also includes TFs

such as *CARF*, with a large number of potential interactions with target genes are likely to be key players in shaping the complex inheritance of the sperm quality traits. We have developed a DNA marker panel based on a systems biology approach that may be able to explain higher phenotypic variance than what could have been found from a stand-alone GWAS. The model included GWAS lead SNPs, top eSNPs and SNPs from genes identified in the Shared Network and could potentially explain over 30% of the phenotypic variance of sperm quality traits such as motility and morphology. Although our results are considerably promising for the improvement of the sector, caution should be taken due to the sample size of our study. Further work should include the validation of the RNA and SNP model in a large number of pigs belonging to different breeds and populations. The implications of this research are broad, ranging from applications to animal breeding strategies to modeling the biology of infertility in mammals.

List of abbreviations

ACRO: abnormal acrosomes
AI: Artificial Insemination
AWM: Associated Weight Matrix
CASA: computer-assisted semen analysis
circRNA: circular RNA
CON: concentration
CPM: counts per million
DDROP: distal droplet
eGWAS: Expression GWAS
eSNP: expressed SNP
FPKM: Fragments Per Kilobase of exon per million reads mapped
GWAS: Genome Wide Association Study
HABN: head sperm abnormalities

LD: Linkage Disequilibrium

miRNA: micro RNA

MT: percentage of motile cells

NABN: neck sperm abnormalities

ORT: osmotic resistance test

PCIT: Partial Correlation coefficient with Information Theory

PDROP: proximal droplet

piRNA:_Piwi interacting RNA

QTL: Quantitative Trait Loci

rRNA: ribosomal RNA

RT-qPCR: quantitative real time PCR

sncRNA: short non-coding RNA

TABN: tail sperm abnormalities

TF: Transcription Factor

VAP: Average Path Velocity

VCL: Curvilinear Velocity

VIAB: cell viability

VSL: Straight Line Velocity

Additional Files:

Additional figure 1 (TIFF)

Summary outline of the different steps of the analysis.

Framework of the dataset, analyses and methodologies included in the study.

Additional figure 2 (TIFF)

Correlation across boar sperm quality traits

Heatmap plot of the correlations among the 25 sperm characters from 300 boars.

Additional figure 3 (TIFF)

Cluster dendogram

Dendogram of the standardized SNP effects across the 25 sperm characters.

Additional file 1 (XLS)

Effect of external factors in the sperm quality traits. Effect of farm, age and season per year across the sperm quality related phenotypes. *=p-value < 0.05; **=p-value < 0.001; ***=p-value < 0.0001; ns=Not Significant.

Additional file 2 (XLS)

RNA-seq extraction and mapping statistics. Average and Standard Deviation (SD) for the 40 samples processed, including: amount of RNA extracted and several bioinformatics statistics for total RNA-seq (40 samples) and short RNA-seq (34 samples).

Additional file 3 (XLS)

List of identified porcine sperm miRNAs. Average and Standard Deviation (SD) for the 34 samples processed. miRNA abundances are expressed in counts per million (CPM).

Additional file 4 (XLS)

Correlations between gene abundances and phenotypes. P-values are given when (P-value < 0.05). The parenthesis include the correlation value. ns=Not Significant.

Additional file 5 (XLS)

Sperm eGWAS eSNP/transcript associations. There were 39 eSNPs/transcript associations with FDR < 0.05. The eSNPs were also identified in the GWAS (FDR < 0.05) and the RNA levels of targeted gene, was significantly correlated with the GWAS phenotype. Chr: chromosome. FDR = False Discovery Rate; ACRO_0 = Abnormal Acrosomes 5 min; HABN = Head abnormalities.

Additional file 6 (XLS)

Gene Ontology analysis of the genes included in the final network. GO biological process terms with significant Bonferroni corrected p-values (FDR < 0.05) and their associated genes.

Additional file 7 (DOC)

Parameter estimates for the significant models. For each of the phenotypes, the model outputs the estimated values for the 10 genes obtained from the GRM regression analysis. The lower the value of Pr > |t|, the higher the involvement of the gene abundance on the total phenotypic variance.

References

Aslam MKM, Kumaresan A, Yadav S, Mohanty TK and Datta TK. Comparative proteomic analysis of high- and low-fertile buffalo bull spermatozoa for identification of fertility-associated proteins. Reprod Domest Anim. 2019;54**:**786-94.

Bahler M, and Rhoads A. Calmodulin signaling via the IQ motif. FEBS Lett. 2002;513**:**107-13.

Berger T, Anderson DL and Penedo MCT. Porcine sperm fertilizing potential in relationship to sperm functional capacities. Anim Reprod Sci. 1996;44**:**231-9.

Blagden SP, Gatt MK, Archambault V, Lada K, Ichihara K et al. Drosophila Larp associates with poly(A)-binding protein and is required for male fertility and syncytial embryo development. Dev Biol. 2009;334**:**186-97.

Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30**:**2114-20.

Breitbart H, Cohen G and Rubinstein S. Role of actin cytoskeleton in mammalian sperm capacitation and the acrosome reaction. Reproduction. 2005;129**:**263-8.

Capra E, Turri F, Lazzari B, Cremonesi P, Gliozzi TM et al. Small RNA sequencing of cryopreserved semen from single bull revealed altered miRNAs and piRNAs expression between High- and Low-motile sperm populations. Bmc Genomics. 2017;18**:**14.

Carreau S, Bourguiba S, Lambard S, Galeraud-Denis I, Genissel C et al. Reproductive system: aromatase and estrogens. Mol Cell Endocrinol. 2002;193**:**137-43.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6**:**80-92.

Curry E, Safranski TJ and Pratt SL. Differential expression of porcine sperm microRNAs and their association with sperm morphology and motility. Theriogenology. 2011;76**:**1532-9.

Cho DY, Kim YA and Przytycka TM. Chapter 5: Network Biology Approach to Complex Diseases. PLoS Comput Biol. 2012;8:e1002820.

Diniz DB, Lopes MS, Broekhuijse ML, Lopes PS, Harlizius B et al. A genome-wide association study reveals a novel candidate gene for sperm motility in pigs. Anim Reprod Sci. 2014;151:201-7.

Dowdle JA, Mehta M, Kass EM, Vuong BQ, Inagaki A et al. Mouse BAZ1A (ACF1) Is Dispensable for Double-Strand Break Repair but Is Essential for Averting Improper Gene Expression during Spermatogenesis. PLoS Genet. 2013;9.

Dunleavy JEM, Okuda H, O'Connor AE, Merriner DJ, O'Donnell L et al. Katanin-like 2 (KATNAL2) functions in multiple aspects of haploid male germ cell development in the mouse. PLoS Genet. 2017;13:e1007078.

Fang P, Xu W, Li D, Zhao X, Dai J et al. A novel acrosomal protein, IQCF1, involved in sperm capacitation and the acrosome reaction. Andrology. 2015;3:332-44.

Fortes MR, Reverter A, Zhang Y, Collis E, Nagaraj SH et al. Association weight matrix for the genetic dissection of puberty in beef cattle. Proc Natl Acad Sci U S A. 2010;107:13642-7.

Gadea J. Sperm factors related to in vitro and in vivo porcine fertility. Theriogenology. 2005;63:431-44.

Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015;31:3718-20.

Gòdia M, Estill M, Castelló A, Balasch S, Rodríguez-Gil JE et al. A RNA-Seq Analysis to Describe the Boar Sperm Transcriptome and Its Seasonal Changes. Front Genet. 2019a;10:299.

Gòdia M, Castelló A, Rocco M, Cabrera B, Rodríguez-Gil JE et al. Identification of circular RNAs in porcine sperm and their relation to sperm motility. bioRxiv. 2019b:608026.

Gòdia M, Swanson G and Krawetz SA. A history of why fathers' RNA matters. Biol Reprod. 2018a;99:147-59.

Gòdia M, Mayer FQ, Nafissi J, Castelló A, Rodríguez-Gil JE et al. A technical assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-seq analysis. Syst Biol Reprod Med. 2018b;64:291-303.

Goto M, O'Brien DA and Eddy EM. Speriolin is a novel human and mouse sperm centrosome protein. Hum Reprod. 2010;25:1884-94.

Groeneveld E. A reparameterization to improve numerical optimization in multivariate REML (co)variance component estimation. Genet Select Evol. 1994;26:537.

Gunawan A, Kaewmala K, Uddin MJ, Cinar MU, Tesfaye D et al. Association study and expression analysis of porcine ESR1 as a candidate gene for boar fertility and sperm quality. Anim Reprod Sci. 2011;128:11-21.

Harper C, Wootton L, Michelangeli F, Lefièvre L, Barratt C et al. Secretory pathway Ca2+-ATPase (SPCA1) Ca2+ pumps, not SERCAs, regulate complex [Ca2+]i signals in human spermatozoa. J Cell Sci. 2005;118:1673-85.

Heid HW, Figge U, Winter S, Kuhn C, Zimbelmann R et al. Novel actin-related proteins Arp-T1 and Arp-T2 as components of the cytoskeletal calyx of the mammalian sperm head. Exp Cell Res. 2002;279:177-87.

Hering DM, Olenski K, Rusc A and Kaminski S. Genome-wide association study for semen volume and total number of sperm in Holstein-Friesian bulls. Anim Reprod Sci. 2014;151:126-30.

Hu H, Miao YR, Jia LH, Yu QY, Zhang Q et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. 2019a;47:D33-D8.

Hu J, Cheng S, Wang H, Li X, Liu S et al. Distinct roles of two myosins in C. elegans spermatid differentiation. PLoS Biol. 2019b;17:e3000211.

Ibba M, and Söll D. Aminoacyl-tRNA Synthesis. Annu Rev Biochem. 2000;69:617-50.

Jodar M, Sendler E, Moskovtsev SI, Librach CL, Goodrich R et al. Absence of sperm RNA elements correlates with idiopathic male infertility. Sci Transl Med. 2015;7:295re6.

Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357-60.

Kim J, Kwon JT, Jeong J, Kim J, Hong SH et al. SPATC1L maintains the integrity of the sperm head-tail junction. EMBO Rep. 2018;19.

Kozomara A, and Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39:D152-7.

Krausz C, Escamilla AR and Chianese C. Genetics of male infertility: from research to clinic. Reproduction. 2015;150:R159-74.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-9.

Li X, Chen G and Yang B. Urea transporter physiology studied in knockout mice. Front Physiol. 2012;3:217.

Li X, Jiang B, Wang X, Liu X, Zhang Q et al. Estimation of genetic parameters and season effects for semen traits in three pig breeds of South China. J Anim Breed Genet. 2019;136:183-9.

Liu H, Li W, Zhang Y, Zhang ZG, Shang XJ et al. IFT25, an intraflagellar transporter protein dispensable for ciliogenesis in somatic cells, is essential for sperm flagella formation. Biol Reprod. 2017;96:993-1006.

Liu M, Guan ZL, Shen Q, Lalor P, Fitzgerald U et al. Ulk4 Is Essential for Ciliogenesis and CSF Flow. J Neurosci. 2016;36:7589-600.

Liu T, Cheng W, Gao Y, Wang H and Liu Z. Microarray analysis of microRNA expression patterns in the semen of infertile men with semen abnormalities. Mol Med Rep. 2012;6:535-42.

Loges NT, Olbrich H, Fenske L, Mussaffi H, Horvath J et al. DNAI2 mutations cause primary ciliary dyskinesia with defects in the outer dynein arm. Am J Hum Genet. 2008;83:547-58.

Luangpraseuth-Prosper A, Lesueur E, Jouneau L, Pailhoux E, Cotinot C et al. TOPAZ1, a germ cell specific factor, is essential for male meiotic progression. Dev Biol. 2015;406:158-71.

Marques DBD, Bastiaansen JWM, Broekhuijse M, Lopes MS, Knol EF et al. Weighted single-step GWAS and gene network analysis reveal new candidate genes for semen traits in pigs. Genet Select Evol. 2018;50:40.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10-2.

Martin PM, Carnaud M, del Cano GG, Irondelle M, Irinopoulou T et al. Schwannomin-interacting protein-1 isoform IQCJ-SCHIP-1 is a late component of nodes of Ranvier and axon initial segments. J Neurosci. 2008;28:6111-7.

Morales CR, Marat AL, Ni X, Yu Y, Oko R et al. ATP-binding cassette transporters ABCA1, ABCA7, and ABCG1 in mouse spermatozoa. Biochem Bioph Res Co. 2008;376:472-7.

Moraru A, Cakan-Akdogan G, Strassburger K, Males M, Mueller S et al. THADA Regulates the Organismal Balance between Energy Storage and Heat Production. Dev Cell. 2017;41:72-81.

Nagarajan P, Onami TM, Rajagopalan S, Kania S, Donnell R et al. Role of chromodomain helicase DNA-binding protein 2 in DNA damage response signaling and tumorigenesis. Oncogene. 2009;28:1053-62.

Olson GE, Winfrey VP and Nagdas SK. Structural modification of the hamster sperm acrosome during posttesticular development in the epididymis. Microsc Res Tech. 2003;61:46-55.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290-5.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.

Qin Y, Ji J, Du G, Wu W, Dai J et al. Comprehensive pathway-based analysis identifies associations of BCL2, GNAO1 and CHD2 with non-obstructive azoospermia risk. Hum Reprod. 2014;29:860-6.

Quintero-Moreno A, Rigau T and Rodriguez-Gil JE. Regression analyses and motile sperm subpopulation structure study as improving tools in boar semen quality analysis. Theriogenology. 2004;61:673-90.

R Developmental Core Team. R: A language and environment for statistical computing. 2010.

Racki LR, Yang JG, Naber N, Partensky PD, Acevedo A et al. The chromatin remodeller ACF acts as a dimeric motor to space nucleosomes. Nature. 2009;462:1016-21.

Ramayo-Caldas Y, Renand G, Ballester M, Saintilan R and Rocha D. Multi-breed and multi-trait co-association analysis of meat tenderness and other meat quality traits in three French beef cattle breeds. Genet Select Evol. 2016;48:37.

Reverter A, and Fortes MR. Association weight matrix: a network-based approach towards functional genome-wide association studies. Methods Mol Biol. 2013;1019:437-47.

Reverter A, and Chan EK. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics. 2008;24:2491-7.

Robinson JA, and Buhr MM. Impact of genetic selection on management of boar replacement. Theriogenology. 2005;63:668-78.

Rueda A, Barturen G, Lebron R, Gomez-Martin C, Alganza A et al. sRNAtoolbox: an integrated collection of small RNA research tools. Nucleic Acids Res. 2015;43:W467-73.

Salas-Huetos A, Blanco J, Vidal F, Godo A, Grossmann M et al. Spermatozoa from patients with seminal alterations exhibit a differential micro-ribonucleic acid profile. Fertil Steril. 2015;104:591-601.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498-504.

Simon P, Baumner S, Busch O, Rohrich R, Kaese M et al. Polysialic Acid Is Present in Mammalian Semen as a Post-translational Modification of the Neural Cell Adhesion Molecule NCAM and the Polysialyltransferase ST8SiaII. J Biol Chem. 2013;288:18825-33.

Smital J, Wolf J and De Sousa LL. Estimation of genetic parameters of semen characteristics and reproductive traits in AI boars. Anim Reprod Sci. 2005;86:119-30.

Snelling WM, Cushman RA, Keele JW, Maltecca C, Thomas MG et al. Breeding and Genetics Symposium: networks and pathways to guide genomic selection. J Anim Sci. 2013;91:537-52.

Song Z. Roles of the nucleotide sugar transporters (SLC35 family) in health and disease. Mol Aspects Med. 2013;34:590-600.

Suh KS, Tatunchak TT, Crutchley JM, Edwards LE, Marin KG et al. Genomic structure and promoter analysis of PKC-delta. Genomics. 2003;82:57-67.

Sun XH, Zhu YY, Wang L, Liu HL, Ling Y et al. The Catsper channel and its roles in male fertility: a systematic review. Reprod Biol Endocrinol. 2017;15:65.

Suzuki T, Inoue I, Yamagata T, Morita N, Furuichi T et al. Sequential expression of Efhc1/myoclonin1 in choroid plexus and ependymal cell cilia. Biochem Bioph Res Co. 2008;367:226-33.

Suzuki T, Miyamoto H, Nakahari T, Inoue I, Suemoto T et al. Efhc1 deficiency causes spontaneous myoclonus and increased seizure susceptibility. Hum Mol Genet. 2009;18:1099-109.

Taiyun W, and Viliam S, 2017 R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Tao X, West AE, Chen WG, Corfas G and Greenberg ME. A calcium-responsive transcription factor, CaRF, that regulates neuronal activity-dependent expression of BDNF. Neuron. 2002;33:383-95.

Tian H, Li ZL, Peng D, Bai XG and Liang WB. Expression difference of miR-10b and miR-135b between the fertile and infertile semen samples (p). Forens Sci Int-Gen S. 2017;6:E257-E9.

Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv. 2014:005165.

Vitavska O, and Wieczorek H. Putative role of an SLC45 H(+)/sugar cotransporter in mammalian spermatozoa. Pflug Arch Eur J Phy. 2017;469:1433-42.

Waldron A, Wilcox C, Francklyn C and Ebert A. Knock-Down of Histidyl-tRNA Synthetase Causes Cell Cycle Arrest and Apoptosis of Neuronal Progenitor Cells in vivo. Front Cell Dev Biol. 2019;7:67.

Wang X, Yang C, Guo F, Zhang Y, Ju Z et al. Integrated analysis of mRNAs and long noncoding RNAs in the semen from Holstein bulls with high and low sperm motility. Sci Rep. 2019;9:2092.

Westlake CJ, Baye LM, Nachury MV, Wright KJ, Ervin KE et al. Primary cilia membrane assembly is initiated by Rab11 and transport protein particle II (TRAPPII) complex-dependent trafficking of Rabin8 to the centrosome. Proc Natl Acad Sci U S A. 2011;108:2759-64.

Wolf J. Genetic Parameters for Semen Traits in AI Boars Estimated from Data on Individual Ejaculates. Reprod Domest Anim. 2009;44:338-44.

Wu FJ, Lin TY, Sung LY, Chang WF, Wu PC et al. BMP8A sustains spermatogenesis by activating both SMAD1/5/8 and SMAD2/3 in spermatogonia. Sci Signal. 2017;10.

Yang CC, Lin YS, Hsu CC, Tsai MH, Wu SC et al. Seasonal effect on sperm messenger RNA profile of domestic swine (Sus Scrofa). Anim Reprod Sci. 2010;119:76-84.

Yang J, Lee SH, Goddard ME and Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76-82.

Zhao X, Zhao K, Ren J, Zhang F, Jiang C et al. An imputation-based genome-wide association study on traits related to male reproduction in a White Duroc x Erhualian F2 population. Anim Sci J. 2016;87:646-54.

Zheng H, Zhou X, Li DK, Yang F, Pan H et al. Genome-wide alteration in DNA hydroxymethylation in the sperm from bisphenol A-exposed men. PLoS One. 2017;12:e0178535.

Whole genome sequencing of porcine sperm identifies Allelic Ratio Distortion in genes related to spermatogenesis

Marta Gòdia[1], Joaquim Casellas[2], Joan-Enric Rodríguez-Gil[3], Anna Castelló[1,2], Armand Sánchez[1,2] and Alex Clop[1,4*]

[1]Animal Genomics Group, Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, 08193, Cerdanyola del Vallès (Barcelona), Spain.

[2]Unit of Animal Science, Department of Animal and Food Science, Autonomous University of Barcelona, 08193, Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[3]Unit of Animal Reproduction, Department of Animal Medicine and Surgery, Autonomous University of Barcelona, 08193, Cerdanyola del Vallès (Barcelona), Catalonia, Spain

[4]Consejo Superior de Investigaciones Científicas (CSIC), 08003, Barcelona, Catalonia, Spain.

*Corresponding author:

Manuscript in preparation

Abstract

Transmission Ratio Distortion (TRD) can be defined as the uneven transmission of an allele from a heterozygous parent to its offspring and may be caused by allelic differences in gametogenesis, fertilization or embryo development. TRD has been vaguely studied at a genomic scale but one report in pigs identified 84 variants in sire to offspring TRD out of 30,000 SNPs. No study has evaluated the extent of allelic ratio distortion in sperm at a genomic scale. We sequenced the diploid and haploid genomes of 3 Pietrain boars from leukocytes and purified ejaculated spermatozoa at 50x sequencing depth to shed light into the genetic basis of Allelic Ratio Distortion (ARD) in spermatogenesis. We used a Bayesian method to identify ARD SNPs and regions with the statistical power offered by the joint analysis with the 3 boars. This led to the identification of 55 ARD SNPs, 14 of which were less than 2 Mbp apart from TRD SNPs from the porcine sire to offspring study. We also identified 2 genes (*TOP3A* and *UNC5B*) that harboured ARD SNPs in the 3 boars. We found 4 ARD regions (defined by containing at least 3 ARD SNPs less than 2 Mbp apart between each) overlapping in the 3 pigs. These regions contained several candidate genes with functions related to spermatogenesis including *AK7*, *VRK1*, *ARID4B*, *PALB2*, *NID1*, *BDKRB2*, and *HTR2A*. Finally, we also searched for variants with moderate to high potential to create damaging impact on protein sequence in at least one boar. We identified 378 genes fulfilling these criteria. Four of these genes contained a potential high impact variants, including, *JAM2*, which has been directly associated to spermatogenic disruption. Several of these genes have been already associated to meiosis, adhesion of Sertoli cells and spermatogenic cells and spermatogenesis.

Introduction

Allelic transmission ratio distortion (TRD) can be defined as the preferential transmission of one allele from a heterozygous parent to the offspring and consequently, the departure from the expected transmission ratio of 0.5:0.5 under the Mendelian law of inheritance. Few studies have explored TRD at a genomic level in mammals despite their potential implications in male fertility, both for human medicine and animal breeding. These studies are mostly based on the genotypes of heterozygous parents and offspring in mouse (Casellas et al., 2012), pig (Casellas et al., 2014) and cattle (Id-Lahoucine et al., 2019), and have lead to the identification of hundreds of loci displaying TRD. In swine, Casellas *et al.* scanned the swine genome with 29,373 SNPs in 5 boars and their 352 offspring using a Bayesian Factor tool (Casellas et al., 2014). The authors identified 84 SNPs that were heterozygous in at least one boar and displayed strong TRD (Casellas et al., 2014). TRD analysis could become an approach complementary to GWAS to help mapping genomic regions influencing spermatogenesis and fertility which could have remained undetectable in a standard GWAS. TRD can be caused by defects at spermatogenesis (germline selection, meiotic drive), the reduced ability of the sperm cell to fertilize the egg or by compromising embryo development (Huang et al., 2013). The exploration of the potential impact on TRD caused by allelic ratio distortion (ARD) in sperm due to defects in spermatogenesis has not been explored thus far.

A Whole Genome Sequencing (WGS) approach to study TRD in sire:offspring designs is currently near to unfeasible due to the large number of animals that would need to be sequenced individually and pool sequencing is neither a practical option because this would not allow controlling for the maternal allelic contribution. This limitation does not exist when studying ARD in sperm as the sequencing of one ejaculate allows calculating the allelic ratio in the population of haploid spermatozoa, and thus determine the existence of this ARD. In other

words, each spermatozoon can be considered a single individual, carrying a haploid genome.

The aim of this study was to identify variants under ARD in the ejaculate of 3 boars from an artificial insemination stud. We have sequenced the genomes of these boars from leukocytes (a class somatic diploid cells) and ejaculated spermatozoa (haploid cells) and used the number of reads carrying each allele at heterozygous sites as proxies of the allelic frequency to estimate ARD in sperm. We hypothesize that these SNPs displaying ARD are indicating the presence of loci influencing the efficiency of spermatogenesis and that these may have an impact on sire to offspring TRD.

Materials and Methods

gDNA from blood from 3 Pietrain boars was extracted with the Maxwell® RSC Whole Blood DNA Kit (Promega) and treated with RNAse DNAse-free (Roche). Ejaculated sperm from the same animals was purified as described at Gòdia et al. (Gòdia et al., 2018) and gDNA was extracted as in (Hammoud et al., 2009). The 6 WGS libraries were prepared with TruSeq DNA PCR-Free Kit (Illumina) and sequenced to generate 150 bp PE in an Illumina's HiSeq X Ten system.

Raw sequencing reads were filtered to remove adaptors and low quality reads with Trimmomatic v.0.36 (Bolger et al., 2014). Filtered reads were aligned to the porcine reference genome (Sscrofa11.1) with Burrows-Wheeler Aligner (BWA) "mem" v.0.7.12 (Li, 2013) and duplicate reads were removed using Picard v.2.18.7 (http://broadinstitute.github.io/picard/). Variant calling was carried with GATK v.3.8.1 (DePristo et al., 2011) with base quality score recalibration. SNPs were discovered and filtered with standard hard filtering parameters along with a cluster filter (maximum of 3 variants in a cluster of 50 bp). The resulting variants were then filtered for a minimum read depth of 20 and a maximum of 2 standard deviations from the average coverage. The predicted effect of the variants was assessed with SnpEff v.4.3T (Cingolani et al., 2012).

Assessment of allelic ratio distortion in sperm

We used 2 statistical approaches to analyse ARD.

To compare the ARD in boar sperm from our analysis with the TRD in swine from Casellas *et al.* (Casellas et al. 2014), we used a Bayesian procedure based on the TRD study by (Casellas et al., 2014) but adapted for WGS. It was applied to all the variants that were heterozygous in the blood samples of all 3 boars. Within each of the 6 sequenced samples, we used the number of reads carrying each allele to calculate the ratio based on number of reads for a given allele divided by the total number of reads in that site. ARD was calculated in sperm (haploid) after correcting its allelic ratio by the ratio in white blood cells (diploid). The rationale behind this is that the ratio in blood should be 0.5 and any deviation from this should be considered technical and also affect sperm. Moreover, all the heterozygous variants with a ratio below 0.4 or above 0.6 in blood were considered as prone to technical errors and removed from the analysis.

To evaluate ARD individually within each animal, we first identified the heterozygous site in each pig in blood (again within the allele ratio 0.4 to 0.6) and then used the Fisher Exact Test to compare the allelic ratio between blood and sperm within each animal. Only variants in ARD in sperm above >0.6 or <0.4 were considered. This was applied to:

(i) identify coding variants with functional potential as assessed with SnpEff v.4.3T (Cingolani et al., 2012) in common genes in the 3 boars. This was based under the hypothesis that ARD variants may not be shared in the 3 boars but may affect common genes with similar functional consequences. SNPs located in coding regions were extracted with BEDTools intersect v.2.17.0 (Quinlan and Hall, 2010). Coding regions were extracted from the Ensembl (v96) porcine annotation.

(ii) identify ARD regions shared in the 3 boars which could be indicative of a common affected regulatory element. ARD regions were determined by identifying these genomic segments containing at least 3 ARD SNPs with consecutive distances between SNPs below 1 Mbp within each pig. The ARD regional overlap between the 3 pigs was evaluated with BEDTools closest and intersect v.2.17.0 (Quinlan and Hall, 2010).

(iii) identify ARD variants with moderate or high functional potential in genes known to be related to spermatogenesis or sperm quality in each pig regardless of whether they are shared or in these pigs. The hypothesis here was that a large number of different genes and biological pathways may led to ARD and thus each pig may have its own set of functions altered which may not be shared in the 3 boars.

Results and Discussion

WGS, mapping and variant calling

In average, 458 M PE reads were obtained per sample (Supplementary Table 1). Up to 99.5% of the reads mapped to the porcine genome (Sscrofa11.1). In average, 14.2% of the reads were duplicates and were thus discarded for further analysis. A sequencing depth between 46 and 55x was obtained (Supplementary Table 1). The average number of SNPs per sample was 10 M and 6.3 M of these passed quality control filters. The reference allele ratio of all the heterozygous SNPs displayed very similar distribution in both blood and sperm in the 3 boars (Supplementary Figure 1). From these, an average of 2.8 M SNPs were heterozygous in the blood of each animal (Supplementary Table 1).

Bayesian analysis to detect ARD in the SNPs heterozygous in the 3 boars

In order to take statistical advantage of analysing the 3 boars simultaneously and hypothesizing that ARD variants could be common in a population, we applied a Bayesian method to identify ARD among in the sites that were

heterozygous in the 3 boars, regardless of whether the ARD would be in one or more pigs. A total of 302,384 SNPs were heterozygous in the 3 samples.

Fifty-five SNPs displayed statistically significant ARD using this Bayesian method (Figure 1). Excluding 2 SNPs that located to unplaced scaffolds, these variants grouped into 44 regions containing 1 or more SNPs with consecutive distances below 1 Mbp. Thirty-seven, 5 and 2 regions contained 1, 2 and 3 SNPs, respectively (Supplementary Table 2). The previous work from Casellas et al. (2014), identified 84 SNPs in TRD. Of these, 7 SNPs could not be liftover into the coordinates of the Sscrofa11.1 genome assembly or mapped into unplaced scaffolds. The remaining 77 TRD SNPs were arranged by proximity in 63 regions (Supplementary Table 2). To estimate whether the ARD and TRD regions tended to co-locate, we calculated how many ARD regions were less than 2 Mbp apart from a TRD segment (Figure 2). Ten of the 44 regions, containing 12 ARD SNPs fulfilled this criteria and 1 additional ARD region marked by 2 SNPs was just 2.085 Mbp from a TRD segment (Table 1; Supplementary Table 2). These results suggest a possible shared biological basis and that a proportion of the TRD may be originated during spermatogenesis. These 11 ARD regions contained 14 ARD SNPs, which were less than 100 kbp away from 9 coding genes (Table 1). Two of the ARD variants (rs1111577152 and rs1113494508), located 16 bp apart, mapped 54 kbp downstream from *INO80* (Table 1), a member of the chromatin-remodelling complex expressed in developing spermatocytes, which plays a key role in DNA damage repair as is essential for successful meiosis and spermatogenesis in mice (Serber et al., 2016). Other ARD variants map within introns of genes with no reported links with spermatogenesis (Table 1).

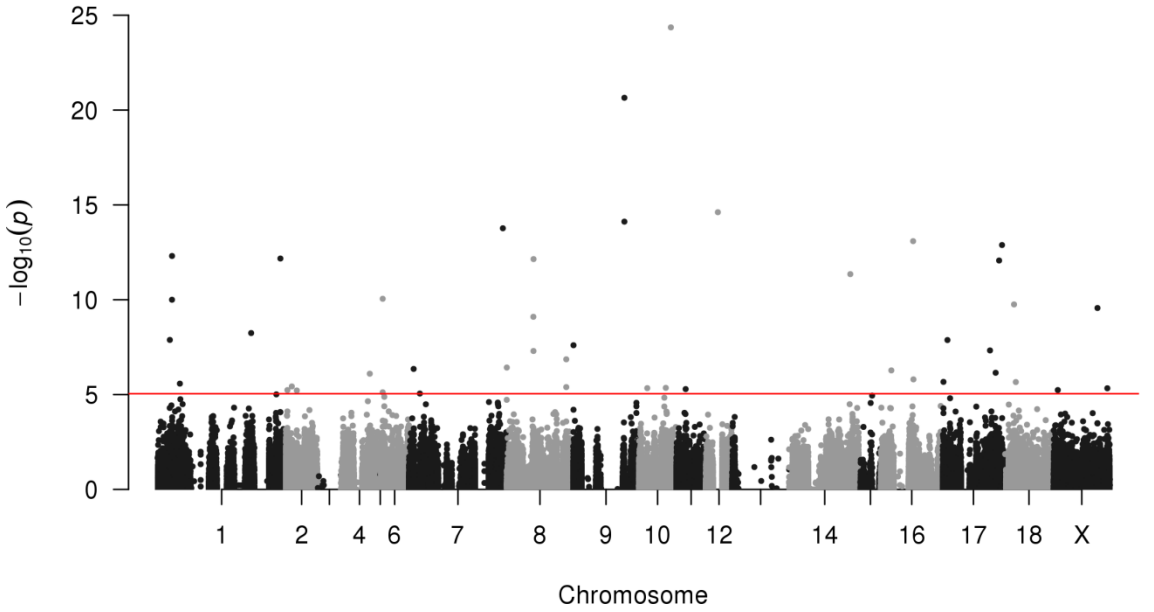


Figure 1. Manhattan plot of the allelic ratio distortion across the porcine chromosomes. The Bayesian approach identified 55 significant SNPs in allelic ratio distortion.

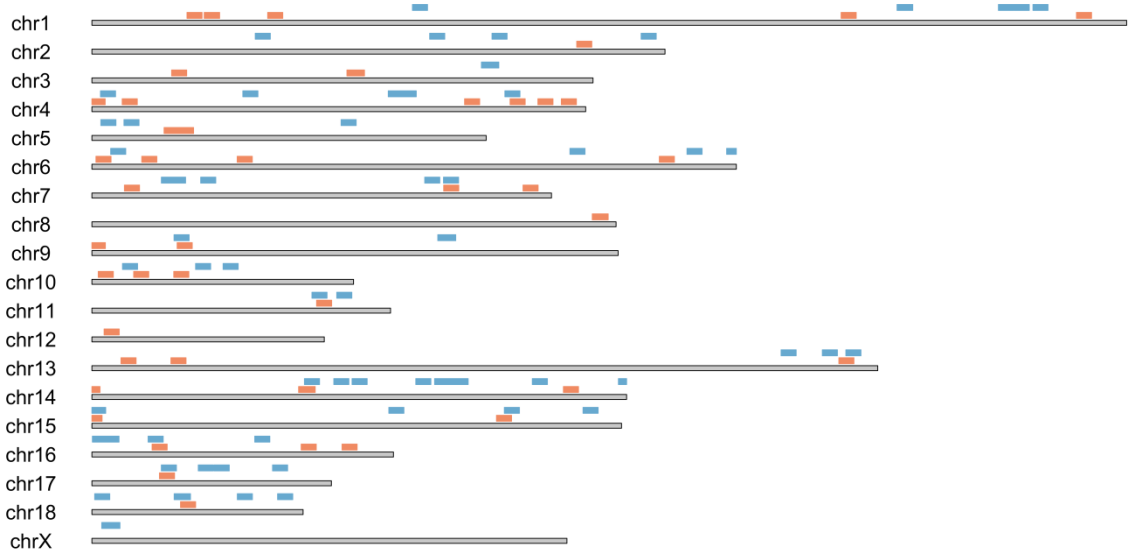


Figure 2. Distribution of the ARD and TRD regions in the pig genome. TRD intervals (Casellas *et al.* 2014), in blue and ARD regions (our study) in orange. The regions included the ARD and TRD SNP with 2 Mbp upstream and downstream extended limits.

Table 1. List of ARD regions in close proximity or overlapping to TRD segments.

ARD or TRD	rsID	Chr	Position	Distance between ARD and TRD regions	Closest gene distance between gene and ARD SNP)
TRD		*4*	*111,484,345*	1.38 Mbp	
ARD	rs327579254	4	112,863,646		None
TRD		*7*	*95,211,457*	52 kbp	
ARD	rs342810440	7	95,263,998		*SIPA1L1* (90 kbp)
TRD		*9*	*23,775,901*	0.82 Mbp	
ARD	rs342042877	9	24,599,014		None
TRD		*11*	*60,336,131*	1.22 Mbp	
ARD	rs339426473	11	61,564,228		None
ARD	novel	13	200,024,203	1.86 Mbp	*MORC3* (intronic)
TRD		13	201,881,431		
ARD	rs325570178	14	56,797,388	1.14 Mbp	*SLC35F3* (intronic)
ARD	rs340156423	14	57,037,751		
ARD	rs337352239	14	57,200,494		*KCNK1* (60 kbp)
TRD		*14*	*58,345,290*		
ARD	rs339246273	15	691,771	0.97 Mbp	*NEB* (intronic)
TRD		*15*	*1,657,293*		
ARD	rs1111577152	15	109,256,963	2.08 Mbp	*INO80* (54 kbp)
ARD	rs1113494508	15	109,256,979		
TRD		*15*	*111,342,012*		
TRD		*16*	*16,901,264*	1.05 Mbp	
ARD	rs325913039	16	17,955,129		*GOLPH3* ortholog
ARD	rs694882285	17	19,909,268	0.50 Mbp	None
TRD		*17*	*20,406,228*		
TRD		*18*	*23,809,561*	1.36 Mbp	
TRD		*18*	*24,134,625*		
ARD	rs788330877	18	25,496,780		*FAM3C* (70 kbp)

In italics, the SNPs identified in the TRD study by Casellas and co-authors (Casellas et al. 2014). Chr: chromosome. The rsID variant is only provided for the ARD variants identified in our study.

We also evaluated ARD within each animal by comparing the allelic events of blood and of sperm using the Fisher Exact Test. Most of the 55 variants identified with the Bayesian approach presented ARD with the Fisher Exact Test in only 1 pig. Furthermore, 3 of the 55 variants presented ARD in the same allelic direction (e.g. homozygous for the reference allele) in the 3 boars. These results suggest that most of the 55 variants are not the ARD causal variants and that ARD could be animal-specific. As a matter of fact, this approach included only those variants that were heterozygous in the 3 pigs, thereby discarding a large proportion of potential candidates. One of these 3 SNPs in ARD in all the 3 pigs mapped 260 kbp away from *GPAT3*. The *GPAT3* gene is involved in the synthesis of fatty acids which is required for spermatogenesis, acrosomal reaction and fertility (reviewed in: Martinez-Soto et al., 2013).

ARD coding variants in common genes

In parallel, we also hypothesized that the ARD variants may be rare, or at least not common, and thus not shared between the 3 pigs. Huang et al. (Huang et al., 2011) suggested that TRD variants tend to be rare because they are wiped out from the population as one allele is preferentially transmitted to the offspring over the other. We therefore sought to identify ARD variants, which although different in the 3 pigs, would affect a common gene or regulatory element.

The 3 pigs presented ARD variants in the genes *TOP3A* and *UNC5B* (Table 2). *TOP3A* was affected by 3 ARD variants. A synonymous ARD variant in *TOP3A* was shared by 2 boars and 1 of these boars also presented a missense ARD variant (Table 2). All the variants in *TOP3A* were novel whilst the variants in *UNC5B* were already annotated in dbSNP, although we do not know their allelic frequency in any population and they could be thus rare or uncommon. *TOP3A* controls chromatin's topologic states during transcription and it has been related to meiotic cell cycle (Hanai et al., 1996). *UNC5B* is a netrin receptor

required for axon guidance during neural development, (Bhat et al., 2019). Although its role in sperm remains to be elucidated, as axon guidance has been linked with ciliogenesis (Lancaster et al., 2011; Poretti et al., 2007), we hypothesise that mutations in this gene could also affect the formation of the flagellum during spermatogenesis. Thus, these two genes seem to have relevant functions in spermatogenesis.

Table 2. List of ARD variants affecting a common gene in the 3 boars.

Sample	Chr	Start	rsID	Closest gene	P-value	Ratio in blood	Ratio in sperm	snpEff
S2	12	60,452,676	novel	*TOP3A*	0.03	0.60	0.35	synonymous
S1	12	60,465,223	novel	*TOP3A*	0.04	0.58	0.35	missense
S1	12	60,466,709	novel	*TOP3A*	0.05	0.43	0.65	synonymous
S3	12	60,466,709	novel	*TOP3A*	0.03	0.54	0.28	synonymous
S3	14	74,186,268	rs324649834	*UNC5B*	0.04	0.60	0.38	synonymous
S2	14	74,199,368	rs339908015	*UNC5B*	0.02	0.59	0.32	synonymous
S1	14	74,204,519	rs337527282	*UNC5B*	0.04	0.60	0.38	synonymous

The ratios were calculated based on the reference allele. Chr: Chromosome. S: sample.

Shared ARD regions in the 3 boars

We also considered the possibility that ARD variants could be rare and only present in one of the 3 pigs but affect common regulatory regions of relevance in spermatogenesis or sperm quality. We extracted the regions in each pig that contained at least 3 SNPs with consecutive SNP distance below 1 Mbp and then selected these that overlapped or were less than 1 Mbp apart in the 3 pigs. We identified 4 of such regions (Table 3), which in total contained 55 genes (Table 3; Supplementary Figure 2). These candidate genes involved were interesting for their role on spermatogenesis (Table 3). These were *AK7*, related to spermatogenic failure and male infertility (Lorès et al., 2018), *VRK1*, reported to cause loss of spermatogonia and infertility in mice (Wiebe et al., 2010), *ARID4B* (Chen and Liu, 2015), *PALB2* (Yan et al., 2016) and *NID1* (Jeffreys and

Neumann, 2005), related to meiosis, *BDKRB2*, regulating the AQP9 water channel in the epydydymis, where sperm maturations takes place (Belleannée et al., 2009), and *HTR2A* which has been genetically associated to sperm count (Cortés-Rodriguez et al., 2018).

Table 3. List of ARD regions with overlap in the 3 samples

Chr	S1	S2	S3	Genes in the region
3	19,346,924-21,139,827 & 22,281,850-24,847,672 (14)	19,698,060-20,410,463 & 24,847,578-24,911,195 (7)	20,623,640-21,407,898 (5)	*GTF3C1, NSMCE1, ERN2, PALB2, NDUFAB1*, EARS2, GGA2, COG7, ENSSSCG00000031197, USP31, IGSF6, CDR2*, PDZD9, CRYM, ZP2*
7	116,030,235-117,205,191 (4)	117,052,038-117,122,913 (5)	116439959-118289816 (7)	*RF00322, GLRX5, RF02192, RF02193, TCL1B, C14orf132, BDKRB2, BDKRB1, GSK3B*, AK7, PAPOLA*, VRK1*
11	20,080,154-20,761,357 (5)	20,283,123-20,824,460 (4)	21,085,554-21,441,712 (3)	*HTR2A, ESD, RUBCNL, LCP1, ENSSSCG00000034648, CPB2, ZC3H13*
14	55,926,799-57,746,832 (9)	54,287,808-55,764,609 (6)	54,474,132-57,200,642 (11)	*RF00001, RF00019, HEATR1, ERO1B, NID1, LYST, GNG4, RF00026, B3GALNT2, ARID4B, RF00425, TOMM20*, RF00397, IRF2BP2, TARBP1, RF00026, PCNX2*

Columns 2, 3 and 4 indicate the ARD genomic intervals for each sample (S1, S2 and S3, respectively). The number of ARD SNPs representing these intervals in each sample is indicated between brackets. Chr: Chromosome. *: gene name from orthologous genes.

ARD in genes related to spermatogenesis within each boar

Finally, we also considered the possibility that ARD may originate from a large number of genes and processes throughout the post-meiotic phases of spermatogenesis, and thus, ARD variants may affect non-shared genes or regulatory regions. For each boar, we extracted the ARD variants with a predicted moderate or high damaging effect on protein sequence thereby

potentially altering the protein function. We identified 409 (132, 129 and 148 for Sample 1, 2 and 3, respectively) ARD variants with moderate or high protein damaging effect, none of them was shared between animals and they mapped to 378 genes (Supplementary Table 3). This catalogue was enriched for genes related to replication fork processing (FDR: $4.4x10^{-2}$), damage DNA checkpoint (FDR: $4.8x10^{-2}$) and filament cytoskeleton organization (FDR: $2.2x10^{-2}$). Of these variants, 4 had a predicted high impact on *TDRD15*, *PCDHGA9*, *JAM2* and *AOX4*. The TDRD family is associated to piwi RNA biology which is essential to keep genome stability during spermatogenesis and *TDRD15* has been shown to be upregulated in mature versus immature horse testes (Li et al., 2019). Little is known about the *PCDHGA9* protocadherin, but protocadherins have been linked to cell adhesion and also this gene is mainly expressed in human testes (Schmidt et al., 2018). *JAM2* has been directly linked to cell adhesion of Sertoli cells to form the blood-testis barrier and to spermatogenesis disruption (Paul and Robaire, 2013). The physiological function of *AOX4* is still unclear with no direct link to spermatogenesis, semen quality or fertility and suggested roles on adipogenesis, locomotion and diurnal rythms (Terao et al., 2016).

Remarkably, the *CDCP1* gene that presented an ARD variant (rs325749569) with moderate protein damaging effect, was located within a GWAS region associated to the percentage of boar sperm head abnormalities (unpublished results from our group). *CDCP1* has been reported to play a role in cell adhesion and motility in cancer cells (Orchard-Webb et al., 2014).

ARD requires that the genes harbouring these mutations function during meiosis or in the haploid phases of spermatogenesis. Several of the genes harbouring moderate or high impact variants are related to meiosis (Supplementary Table 3) and others are important for cell adhesion in Sertoli cells. Sertoli cells are adjacent to the maturing germ cells with junctions that need to be timely dissembled and reassembled to allow migration of the

spermatogenic lineage from the basement membrane to the luminal compartment of the seminiferous tubules (reviewed in: Ni et al., 2019). Moreover, Sertoli cells are responsible of removing, by endocytosis, the cytoplasm from elongated spermatids (reviewed in: Ni et al., 2019). Thus, our data suggests that the junction between Sertoli cells and the haploid cells (secondary spermatocytes, round and elongated spermatids and immature spermatozoa), may play an important role in ARD and TRD (Supplementary Table 3).

To the best of our knowledge, this pioneer study is the first to evaluate the potential forces of spermatogenesis that could drive TRD by evaluating ARD at the sperm level using WGS. WGS has an advantage over genotyping platforms as it allows the interrogation of practically the whole genome and has thus the potential to identify the causal variants. Moreover, WGS allows interrogating ARD at the sperm level, which would be impossible with genotyping arrays.

This is a preliminary study with 3 boars and WGS at 50x depth. Forthcoming studies should include additional boars and larger sequencing depth to drastically increase the power to identify ARD variants and clarify the biological basis of the spermatogenesis driven TRD. Moreover, the variants that we identified in this study should be confirmed with genotyping technology in the sequenced animals and ideally in a larger sire:offspring pedigree to assess their allelic frequency and confirm the TRD effect. Also, if frequent, these variants should be included in genetic association studies for sperm quality and male fertility to assess their potential implication on the male's reproductive ability.

Supplementary material:

Supplementary Figure 1. Density plot of the reference allele ratio distribution in heterozygous SNPs in blood and sperm for each boar.

Supplementary Figure 2. Overview of the genomic overlap in the 4 ARD regions shared in the 3 pigs. Each ARD region is formed by at least 3 ARD SNPs with consecutive SNP distance below 1 Mbp. We identified 4 of such regions in chromosome 3, 7, 11 and 14. ARD regions of sample S1 are depicted in blue, sample S2 in yellow and S3 in green.

Supplementary Table 1. Sequencing and mapping statistics for the 3 boars.

Supplementary Table 2. List of variants and regions in ARD (this study) and in TRD (Casellas *et al*., 2014).

Supplementary Table 3. SNPs in ARD in the 3 boars. For each SNP we provide information of the rsID, host gene, ARD P-value, reference allele ratio and SNP effect. Chr: chromosome.

References:

Belleannée C, Da Silva N, Shum WW, Marsolais M, Laprade R et al. (2009). Segmental expression of the bradykinin type 2 receptor in rat efferent ducts and epididymis and its role in the regulation of aquaporin 9. *Biol Reprod* 80**:**134-143.

Bhat SA, Gurtoo S, Deolankar SC, Fazili KM, Advani J et al. (2019). A network map of netrin receptor UNC5B-mediated signaling. *J Cell Commun Signal* 13**:**121-127.

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

Casellas J, Manunza A, Mercader A, Quintanilla R and Amills M (2014). A Flexible Bayesian Model for Testing for Transmission Ratio Distortion. *Genetics* 198**:**1357-1367.

Casellas J, Gularte RJ, Farber CR, Varona L, Mehrabian M et al. (2012). Genome Scans for Transmission Ratio Distortion Regions in Mice. *Genetics* 191**:**247-259.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T et al. (2012). A program for annotating and predicting the effects of single nucleotide

polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 6**:**80-92.

Cortés-Rodriguez M, Royo JL, Reyes-Palomares A, Lendinez AM, Ruiz-Galdon M et al. (2018). Sperm count and motility are quantitatively affected by functional polymorphisms of HTR2A, MAOA and SLC18A. *Reprod Biomed Online* 36**:**560-567.

Chen SR and Liu YX (2015). Regulation of spermatogonial stem cell self-renewal and spermatocyte meiosis by Sertoli cell signaling. *Reproduction* 149**:**R159-R167.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43**:**491-+.

Gòdia M, Mayer FQ, Nafissi J, Castelló A, Rodríguez-Gil JE et al. (2018). A technical assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-seq analysis. *Syst Biol Reprod Med* 64**:**291-303.

Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT et al. (2009). Distinctive chromatin in human sperm packages genes for embryo development. *Nature* 460**:**473-478.

Hanai R, Caron PR and Wang JC (1996). Human TOP3: a single-copy gene encoding DNA topoisomerase III. *Proc Natl Acad Sci U S A* 93**:**3653-3657.

Huang L, Labbe A and Infante-Rivard C (2011). Impact of transmission ratio distortion on the interpretation of genetic association studies and evolution of population parameters. 6th Annual Genetic Epidemiology and Statistical Genetic Meeting

Huang LO, Labbe A and Infante-Rivard C (2013). Transmission ratio distortion: review of concept and implications for genetic association studies. *Hum Genet* 132**:**245-263.

Id-Lahoucine S, Canovas A, Jaton C, Miglior F, Fonseca PAS et al. (2019). Implementation of Bayesian methods to identify SNP and haplotype regions with transmission ratio distortion across the whole genome: TRDscan v.1.0. *J Dairy Sci* 102**:**3175-3188.

Jeffreys AJ and Neumann R (2005). Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum Mol Genet* 14**:**2277-2287.

Lancaster MA, Schroth J and Gleeson JG (2011). Subcellular spatial regulation of canonical Wnt signalling at the primary cilium. *Nat Cell Biol* 13**:**700-707.

Li B, He X, Zhao Y, Bai D, Bou G et al. (2019). Identification of piRNAs and piRNA clusters in the testes of the Mongolian horse. *Sci Rep* 9**:**5022.

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997.

Lorès P, Coutton C, El Khouri E, Stouvenel L, Givelet M et al. (2018). Homozygous missense mutation L673P in adenylate kinase 7 (AK7) leads to primary male infertility and multiple morphological anomalies of the flagella but not to primary ciliary dyskinesia. *Hum Mol Genet* 27**:**1196-1211.

Martinez-Soto JC, Landeras J and Gadea J (2013). Spermatozoa and seminal plasma fatty acids as predictors of cryopreservation success. *Andrology-Us* 1**:**365-375.

Ni FD, Hao SL and Yang WX (2019). Multiple signaling pathways in Sertoli cells: recent findings in spermatogenesis. *Cell Death Dis* 10**:**541.

Orchard-Webb DJ, Lee TC, Cook GP and Blair GE (2014). CUB domain containing protein 1 (CDCP1) modulates adhesion and motility in colon cancer cells. *BMC Cancer* 14**:**754.

Paul C and Robaire B (2013). Impaired Function of the Blood-Testis Barrier during Aging Is Preceded by a Decline in Cell Adhesion Proteins and GTPases. *Plos One* 8.

Poretti A, Boltshauser E, Loenneker T, Valente EM, Brancati F et al. (2007). Diffusion tensor imaging in Joubert syndrome. *AJNR Am J Neuroradiol* 28**:**1929-1933.

Quinlan AR and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26**:**841-842.

Schmidt T, Samaras P, Frejno M, Gessulat S, Barnert M et al. (2018). ProteomicsDB. *Nucleic Acids Res* 46**:**D1271-D1281.

Serber DW, Runge JS, Menon DU and Magnuson T (2016). The Mouse INO80 Chromatin-Remodeling Complex Is an Essential Meiotic Factor for Spermatogenesis. *Biol Reprod* 94.

Terao M, Barzago MM, Kurosaki M, Fratelli M, Bolis M et al. (2016). Mouse aldehyde-oxidase-4 controls diurnal rhythms, fat deposition and locomotor activity. *Sci Rep* 6**:**30343.

Wiebe MS, Nichols RJ, Molitor TP, Lindgren JK and Traktman P (2010). Mice Deficient in the Serine/Threonine Protein Kinase VRK1 Are Infertile Due to a Progressive Loss of Spermatogonia. *Biol Reprod* 82**:**182-193.

Yan ZC, Fan DD, Meng QJ, Yang JJ, Zhao W et al. (2016). Transcription factor ZFP38 is essential for meiosis prophase I in male mice. *Reproduction* 152**:**431-437.

General discussion

Chapter 4

Traditional breeding programs in swine have focused on production (carcass and meat quality) and feed efficiency traits in males and reproduction traits in females. However, targeting only these traits in boars, have resulted in a decrease of sperm quality due to the negative correlations between them and semen quality (Arsenakis et al., 2015; Oh et al., 2006a)

A fast growing population worldwide together with the climate crisis poses food scarcity as a major global threat as recognized by the EU and most national research, development and innovation programs (e.g. The European Comission's Horizon 2020 research program). Improving the efficiency of food production systems, reducing their environmental and health impact and ultimately securing the sustainability of these activities are some of the top priorities of these programs. With this aim, the animal breeding sector is now starting to target novel traits in the genetic selection schemes including, among other traits, semen quality and male fertility. However, these traits are complex and influenced by a large number of genetic and non-genetic factors (Lopez Rodriguez et al., 2017) which remain to be resolved. Deciphering the genetic basis of sperm quality is necessary for the development of new genetic selection strategies involving these traits. Selecting the boars with the highest genetic merit for production traits that are at the same time able to efficiently transmit this genetic material (with good semen quality and fertility) will not only benefit the AI companies but will also help accelerating the genetic progress and benefit the pig breeding sector overall.

Several studies have estimated the heritability of different semen quality parameters, evaluated their correlations and assessed some potentially influencing external factors (Oh et al., 2006a; Oh et al., 2006b; Smital et al., 2005; Wolf, 2009; Wolf and Smital, 2009). Small differences in sperm quality traits have been observed between breeds (Schulze et al., 2014), and in fact, the individual variation may actually exceed breed variation. Several groups have

evaluated sperm phenotypic values in Pietrain (Schulze et al., 2014; Wolf, 2009; Wolf and Smital, 2009) as this is a commonly used breed in the porcine industry for its remarkable performance on growth and carcass quality. All the ejaculates included in this thesis were from Pietrain boars provided by 3 commercial farms. The AI studs provided the records for volume of ejaculate, sperm concentration and the number of doses per ejaculate. We assessed all the other phenotypic parameters within a few hours of collection in collaboration with experts on porcine sperm sciences at the UAB's Veterinary College. I processed the ejaculates from May 2015 to January 2017.

The sperm traits included motility, morphology and membrane integrity parameters. Most of these parameters were assessed twice, after 5 and 90 minutes of incubation at 37ºC to evaluate the effect of heat stress. Independent studies have reported that sperm motility may be a proxy of fertility in cattle (Farrell et al., 1998), horse (Love, 2011) and swine (Broekhuijse et al., 2012; Holt et al., 1997). The motility parameters recorded in this thesis were measured by the CASA system and included, among others, the percentage of motile cells, curvilinear velocity (VCL; μm/s; time-average velocity of a sperm head along its actual path), average path velocity (VAP; μm/s; time-average velocity of a sperm head projected along its special trajectory), straight-line velocity (VSL; μm/s; time-average velocity of a sperm head projected along straight line) and the amplitude of lateral head displacement (ALH**;** μm; mean head displacement along its curvilinear displacement). Sperm morphology has also been significantly associated to fertility in pigs (Alm et al., 2006; Lee et al., 2014; Tsakmakidis et al., 2010). Abnormal sperm may fail to reach the fertilization site or are unable to fertilize the ovum (Lee et al., 2014). In our work, sperm morphology parameters were estimated with Eosin-Nigrosin staining with no less than 200 cells examined per sample and it included the percentage of head, neck and tail sperm abnormalities, proximal and distal droplets and abnormal

acrosomes. Plasma membrane integrity was also evaluated with Eosin-Nigrosin staining by assessing the osmotic resistance with ORT and cell viability. In the ORT, spermatozoa is subjected to a osmotic challenge, and as a response, the cells with intact plasma membrane will swell (bending, coiling or shortening of the tail) (Revell and Mrode, 1994). Cell viability is determined by assessing the percentage of living cells with intact cytoplasmic membrane, identified by their resistance to absorb the staining dye (Moskovtsev and Librach, 2013). Correlations between all the phenotypes have been estimated and can be found in Additional Figure 2 of the Study IV of this thesis. Briefly, all the motility-related traits (e.g. percentage of motile cells, VCL, VSL, VAP) presented moderate to high correlations (>0.4) across them. The percentage of abnormal acrosomes was negatively correlated with the percentage of viable cells (-0.7) and ORT (-0.5) and moderately correlated with the percentage of neck abnormal morphologies (0.2). While neck, tail and cytoplasmatic droplets presented moderate correlations (~0.2), head abnormalities did not correlate with any trait.

In addition to the benefit for the animal breeding sector, research onto the genetic basis of sperm quality in pigs can also provide meaningful insights into male infertility. Unlike in livestock, were research is mostly oriented to improve production efficiency, studies in humans aim to understand the perturbations of infertile patients in order to provide a more preventive, predictive, and personalized approach. In humans, infertility affects approximately 15% of couples worldwide (Boivin et al., 2007), and male-related factors account for 15-30% of the cases (Ferlin et al., 2007). Several compelling independent studies have identified correlations between infertility and genetic variants, DNA fragmentation, chromatin modifications, protein or RNA levels (Cho and Agarwal, 2018; Jenkins and Carrell, 2012; Jodar et al., 2015; Jodar et al., 2017; Krausz et al., 2015; Salas-Huetos et al., 2014). Indeed, sperm RNAs have become

a good resource to study spermatogenesis, sperm quality, fertility and transgenerational epigenetics (Review: Gòdia et al., 2018a). Mature spermatozoa contain a complex suite of RNAs including long (>200 nt) coding and long noncoding (long noncoding RNA and circular RNA –circRNA-) and short noncoding RNAs (micro RNA –miRNA-, small interfering RNA –siRNA-, piwi-interacting RNA –piRNA- and transfer RNA –tRNA-). Several research efforts have partially resolved their role in spermatogenesis, fertilization, embryo development and offspring phenotype (Review: Gòdia et al., 2018a) but this is just the tip of the iceberg and multiple new studies will spur in the near future due to the relevance of these traits in breeding and medicine.

In this thesis, we have studied sperm RNAs to elucidate the molecular mechanisms behind sperm quality in swine. For this, we have carried high-throughput sequencing of the total and the short RNA fractions from ejaculates with semen quality phenotypic records and contrasted the existence of relationships between the RNA abundances and the traits. We also sought to identify potential genomic variants and regions associated to sperm quality. For this, we have employed high-throughput genotyping arrays on boars with sperm phenotype evaluations. Finally, we sought to map genetic variants in sperm with allelic deviations from the expected Mendelian ratio of 0.5 (allelic ratio distortion –ARD-), indicative of a genetic influence on spermatogenesis and sperm viability. This was achieved by WGS of matched blood (diploid and used to identify heterozygous sites) and sperm (haploid and used to evaluate ARD, potentially linked to spermatogenesis) samples. In the next sections, we will discuss the main results obtained in the studies presented in this thesis and their integration.

4.1 Selection of a protocol for the purification and RNA extraction from porcine sperm

In the Study I of the thesis we evaluated a series of protocols to obtain porcine sperm RNAs of sufficient quality to carry the subsequent experiments. The spermatozoon is a highly complex cell type with very tiny amounts of highly degraded RNA. Ejaculates contain somatic cells with high amounts of high quality RNA and if not removed, these could easily mask any RNA particularity in sperm. Thus, the study of sperm RNA requires specific steps for cell purification, RNA extraction, quality validation and analysis (Review: Gòdia et al., 2018a).

We have reported for the first time, that the sperm recovery rate (SRR) in boars is ~22% (Study I: Gòdia et al., 2018b). The SRR was estimated using 285 porcine ejaculates with the gradient centrifugation purification method (Study I: Gòdia et al., 2018b). This SRR number can be useful when designing future studies to ensure sufficient sperm cells are available. Remarkably, the purification method that we used was not influenced by sperm quality (e.g. low motility or high number of cell abnormalities). This indicates that the sperm quality of the purified and unpurified samples is similar and thus, the first can be used to infer the molecular basis of semen quality on raw ejaculates.

Sperm RNA extraction using commercial kits has been proven successful in previous studies (Goodrich et al., 2013; Goodrich et al., 2007). However, we obtained better performance with a modified Trizol-based protocol (Study I: Gòdia et al., 2018b). Despite this improvement, the average RNA yield per sperm cell was still low (1.6 fg), a value that is considerably lower than in human and mice (Goodrich et al., 2013; Pessot et al., 1989). Our study also pointed out that the RNA yield obtained per sperm cell is independent of the sperm quality of that sample (Study I: Gòdia et al., 2018b). Hence, the RNA yield cannot be used as a proxy of semen quality.

Although validation of the purification method through quantitative real time PCR (RT-qPCR) revealed that 57% of the purified samples (49/70) still presented traces (quantification cycles –Cq < 40) of the somatic cell marker *PTPRC*, only 2 samples presented $\Delta Cq_{PTPRC\text{-}PRM1} > 16$ (Study I: Gòdia et al., 2018b). This normalization method ($\Delta Cq_{PTPRC\text{-}PRM1}$) has been proved to be successful in our study as 4 of the samples subjected to RNA-seq that presented traces of *PTPRC* RNA by RT-qPCR (Cq between 33 to 36 and $\Delta Cq_{PTPRC\text{-}PRM1}$ between 17.4 to 19.1), did not show levels of *PTPRC* when screened by RNA-seq (Study I: Gòdia et al., 2018b).

4.2 The porcine sperm transcriptome

The porcine sperm transcriptome has been investigated in the Studies II, III and IV of this thesis. Even though these studies used the same RNA-seq technology, each addressed different questions of relevance for cell biology and sperm quality.

In Study II, we described a thorough characterization of the swine sperm transcriptome and its seasonal changes (Study II: Gòdia et al., 2019a). We used 10 ejaculates, each from a different boar. Half of the ejaculates were collected in summer and the others in winter. We performed total and small RNA-seq. In the first section of the study, we provided a detailed description of the transcriptome using a RNA Element discovery algorithm (Estill et al., 2019). Briefly, the algorithm detects expressed regions (covered by reads that are mapped in the genome). Then these regions are annotated based on their genomic positions and quantified based on the number of reads in the given position. We identified 185,037 sperm RNA elements (SREs), from which the 10% most abundant SREs (top decile) accounted for 65% of the read counts, thereby indicating that a small proportion of the genes' RNA are highly abundant while the majority of these are scarce (< 10 RPKM) in mature sperm. The top decile included 4,436 annotated genes (Study II: Gòdia et al., 2019a).

This number might seem low when compared to other porcine tissues such as muscle (Chen et al., 2011), fat (Corominas et al., 2013; Chen et al., 2011) or duodenum (Mach et al., 2014) each with over 15,000 expressed genes, but it is similar to data described in human sperm (4,765 genes) (Estill et al., 2019). Many of these low abundant SREs are likely to be mere remnants of spermatogenesis. Our analysis also suggested that transcript fragmentation follows a programmatic pattern (Study II: Gòdia et al., 2019a) and was found moderately conserved with humans (Sendler et al., 2013). Moreover, the high fragmentation level of most transcripts in the boar sperm has technical implications. First, RNA-seq to profile the sperm transcriptome cannot rely on poly-dT primers for library prep. Second, the design of primers for RT-qPCR assays has to take into account the gene's SRE structure.

Although it is generally assumed that sperm cells are transcriptionally and translationally silent, few independent studies have reported active mitochondrial gene expression in sperm cells of boars (Zhu et al., 2019) and mice (Alcivar et al., 1989). We detected that a large portion of the most abundant SREs corresponded to mitochondrial genes (Study II: Gòdia et al., 2019a). These results are not surprising, as spermatozoa contain a large number of active mitochondria to provide energy. In keeping, many of these genes are related to glycolysis and mitochondrial oxidative phosphorylation (OXPHOS) pathways (Ferramosca and Zara, 2014; Piomboni et al., 2012). Glycolysis takes place in the fibrous sheath of the flagellum, whereas OXPHOS occurs in the level of sperm mid-piece, where the mitochondria are mostly located (Ferramosca and Zara, 2014; Piomboni et al., 2012). The sperm mitochondrial organelles participate in crucial processes, noticeably including the production of adenosine triphosphate (ATP) through the mechanism of OXPHOS, that provides the energy necessary for sustaining sperm motility (Piomboni et al., 2012).

This initial characterization of the boar sperm transcriptome (Study II) also provided other interesting findings. Although some SREs and miRNAs presented stable abundances across samples (measured by the coefficient of variation –CV-), others were very variable (Study II: Gòdia et al., 2019a); Supplementary File 13). These differences were somehow related to the most likely source of these RNAs as has been reported based on their preferential RNA abundances (seminal fluid, testis, or sperm-enriched) (Jodar et al., 2016).SREs enriched in the seminal fluid showed the highest CV. This variability could be partly explained by the changes in the secretion of seminal exosomes by the male's accessory sex glands and by a varying efficiency on the exosome uptake by spermatozoa (Vojtech et al., 2014). Large CV was also observed in a fraction of the small nonocoding RNA load (miRNAs and tRNAs). This large CV could be partially given again by the large variability of seminal (Barceló et al., 2018; Vojtech et al., 2014) or epididymal (Reilly et al., 2016; Sharma et al., 2018) exosome secretion and sperm uptake. These coefficients together with the SRE characterization have been proven to be very useful to identify housekeeping genes for RT-qPCR evaluation. For example, while the most commonly used housekeeping *GAPDH* or *ACTB* genes did not present stable abundances (unpublished results), the *ISYNA1* and *GPR137* genes were very stable across samples (Study II: Gòdia et al., 2019a; Study III: Gòdia et al., 2019b).

Sperm parameters follow a seasonal pattern in many mammalian species including swine, with a decrease of its quality in the summer months (Li et al., 2019b; Wolf and Smital, 2009). This seasonality has been associated to changes in the photoperiod, temperature and/or humidity in pig (Li et al., 2019b). This was tackled in the second section of Study II. Our data showed suggestive (P-value between 0.08 and 0.06), but not significant phenotype differences between the summer and winter sperm groups (Study II: Gòdia et al., 2019a). This lack of

significance was due to the low sample size (N = 10). In fact, re-analysis of the seasonal effect on semen quality using all the available dataset (300 samples), which included 29 summer and 117 winter ejaculates, showed strong significant differences between groups for sperm cell viability (Wilcoxon test, P-value = 3.856×10^{-10}), the percentage of abnormal acrosomes (P-value = 3.049×10^{-7}) and neck (P-value = 7.175×10^{-8}) and tail (P-value = 6.563×10^{-9}) abnormalities (unpublished). Despite the fact that pig farms have been incorporating novel technological systems (e.g. heating or positive pressure ventilation systems) to minimize the pernicious summer effect, the quality of the ejaculates is still subjected to seasonal fluctuations (Li et al., 2019b). We identified 36 differentially abundant transcripts and 7 dysregulated miRNAs (Study II: Gòdia et al., 2019a). These included *OSGIN1* and miR-106a, both associated to oxidative stress, a noxious process for sperm quality that is generally counteracted by antioxidant factors that present in spermatozoa and the seminal plasma. Interestingly, the RNA-seq SNP calling task that we carried within the frame of Study IV, allowed us to detect an heterozygous novel nonsense (premature stop codon) mutation in *OSGIN1* (c.1005C>A) in 32 of the 35 sequenced samples. However, the average read depth (DP) of this variant was low (DP = 10) and thus, this variant could be a potential sequencing error. Worth to mention, the differential expression analysis was carried on transcripts as units of measurement instead of SREs. This mostly aimed to reduce the statistical power needed as the number of SREs (185,037) was way larger than the number of transcripts (26,215).

4.3 Association of sperm RNA abundances with quality traits

In Studies III and IV of this thesis, we used the 10 samples from Study II but also additional samples for total and short RNA-seq. Altogether, we generated 40 and 34 datasets for total and short RNA-seq, respectively. The samples were selected from the total set of 300 ejaculates. Initially, we first planned to

compare phenotypic groups in a case (bad sperm) : control (good sperm) design, including at least 5 samples in each extreme group per trait for 4 selected traits (e.g. motility, viability, ORT and total morphologies). However, we decided to evaluate the relationship between RNA abundances and phenotypes using a linear regression approach, to include and exploit the information provided by all the samples.

Study III of this thesis aimed to characterize the population of circRNAs in porcine sperm and their potential as biomarkers of sperm motility (Study III: Gòdia et al., 2019b). To the best of our knowledge, this is the first study that reports a genome-wide characterization of the circRNAome in mature sperm although it has been followed by a more recent work in human sperm using microarray hybridization (Chioccarelli et al., 2019). To identify putative circRNAs, we used the *find_circ* software (Memczak et al., 2013), which does not rely in gene annotations but performs *de novo* predictions.

We detected circa 1,600 circRNAs (Study III: Gòdia et al., 2019b). Remarkably, this number is lower than in human sperm with 10,726 reported circRNAs (Chioccarelli et al., 2019). These discrepancies may be due to technical differences as Chioccarellie and co-authors kept all the circRNAs identified in at least one of the 3 samples on a microarray hybridization approach whilst we were stricter and selected only these circRNAs that were present in at least 30 of the 40 samples.

Most of these circRNAs were exonic but we also identified a small proportion (~20%) that included intergenic or intronic segments. Although circRNA biogenesis and function is not well understood, their cellular localization may shed light into their function. Whereas exonic circRNAs are mostly located in the cytoplasm and have been proposed to act as post-transcriptional gene regulators (reviewed in: Cortes-Lopez and Miura, 2016), exonic circRNAs with retained introns or intronic circRNAs have been predominantly detected in the

nucleus and have been suggested to function as transcriptional regulators (Li et al., 2015). The role of circRNAs as miRNA sponges is well described in the literature (reviewed in: Cortes-Lopez and Miura, 2016). For this reason, we built a co-expression network based on significant abundance interactions and *in silico* prediction of miRNA targets in the circRNA sequences (Study III. Figure 2). Some of these circRNAs host genes have been associated to spermatogenesis and fertility as is the case for *DCDC2C*, *FAM92A*, *WDR7* and *ACTL6A.*

Scientific evidence showing the potential of circRNAs as biomarkers is emerging. Here, we studied 4 sperm motility-related traits and found 179 circRNAs as potential biomarkers. A similar number (148) of differentially abundant circRNAs were detected in 3 human samples within-donor, studying the highly motile with good morphology versus the low motility with bad morphology sperm fractions (Chioccarelli et al., 2019). Remarkably, the pathways in these extreme quality groups (e.g. fatty acid metabolism, oocyte meiosis or focal adhesion) were not identified in our study. We identified biological processes as spermatogenesis, cilium assembly or chromatin organization.

We experimentally validated the presence of 9 of the 10 circRNAs by RT-PCR. Only the abundance of 2 circRNAs (ssc_circ_1458 from *LRBA* and ssc_circ_1321 from *PAPOLA*) still correlated with sperm motility parameters by RT-qPCR. The differences between the RNA : phenotype correlations calculated with the RNA-seq data and with the RT-qPCR could have several causes. First, the samples used for RNA-seq and for RT-qPCR were different. We chose different samples due to (i) the low amount of RNA identified in each sample and our interest to keep these RNAs for the validation of mRNAs and miRNAs for other studies and (ii) that the RT-qPCR samples were chosen due to their extreme values for the phenotypes of interest in Study III. Second, there are technical differences between both methods. In fact, a study comparing mRNA

abundances calculated by RNA-seq and by RT-qPCR found moderate correlations (0.56) between the two methods (Trost et al., 2015).Although our data provides new promising insights into the presence and role of circRNAs in the boar sperm, the study has few limitations. The diversity of some overlapping circRNAs with different alternative splicing, made it difficult to design specific primers for their validation. Furthermore, the discrepancy in abundances between RT-qPCR and RNA-seq algorithm suggests that the RNA-seq read count in junctions might be a poor predictor of the real quantifications. Moreover, the differences observed between species and tissues might partly be given by biological reasons but also by the circRNA prediction algorithm and filtering parameters, which were different in each study and play a crucial role in circRNA annotation (Szabo and Salzman, 2016). Of note, the circRNAs validated in this study were not subjected to the Ribonuclease R (RNase R) treatment. RNase R is a 3′ to 5′ exoribunuclease able to degrade linear but not circularized RNAs and it is widely used for circRNA validation. Nonetheless, there is a rising number of experiments reporting that RNase R can also degrade circRNAs and thus this treatment may not be indispensable (reviewed in: Szabo and Salzman, 2016). Moreover, the degradation of the linear transcript is typically achieved by mixing 1U of RNase R with 1μg of RNA (according to protocol). Consequently, and as the RNA yield obtained in our mature sperm samples was below 200 ng and we needed to save these amounts for other studies, we decided to skip the RNase R treatment. Bearing this is mind, although we acknowledge this limitation in our study, we believe that a considerable proportion of potential circRNAs detected are indeed formed through back splicing and are thus not false positives.

In Study IV of this thesis, we assessed the correlation between transcript abundances and 25 sperm quality-related phenotypes. Together, our data revealed a surprisingly high number of correlated transcripts across the

different phenotypes. 3,007 out of the 4,120 genes presented at least 1 significant correlation ($P < 0.05$) with a phenotype. In particular, 344 of these genes showed correlation with at least 4 traits (Table 4.1). The 3,007 genes showed in total 6,128 correlations. Considering the number of tests that were run (4,120 genes * 25 phenotypes = 103,000 tests), only 5.9% of the examined pairs showed significant correlation.

Table 4.1. Correlation between gene abundances and phenotypes

Number of phenotypes (P-value < 0.05)	Number of genes	Number of interactions
9	3	27
8	2	16
7	9	63
6	33	198
5	68	340
4	229	916
3	487	1461
2	931	1862
1	1245	1245
Total	3,007	6,128

Although not mentioned in Study IV, these correlations involved several genes of interest. The strongest correlation was for the RNA levels of *TTC28* and head abnormalities (-0.71). So far, *TTC28* has not been linked to sperm or fertility but the gene is required for the condensation of spindle microtubules during mitosis and meiosis (Izumiyama et al., 2012), which is of obvious relevance during spermatogenesis. Sperm quality comprises various criteria (e.g. motility, viability, tail abnormalities, etc). Therefore, these genes showing significant correlation with more than 1 trait are more likely to be relevant in the overall description of semen quality. This is the case of for example of *ABCA3*, which RNA levels correlated with 9 phenotypes (correlated with proximal and distal droplets, neck and tail abnormalities, ratio acrosomes and viability, VAP, VSL and VCL at 5 min). This gene is an ABC transporter that plays a role in flipin-cholesterol complexes as a mechanism to remove cholesterol from the sperm membrane (Mengerink and Vacquier, 2002). Although the molecular basis induced by cholesterol efflux from sperm is not well understood, it has been

reported to be required for sperm capacitation (Visconti et al., 2002). Another example is *EFHC1,* which RNA levels correlated with 6 phenotypes (e.g. proximal droplets, VAP, VSL). E*fhc1*$^{-/-}$ knockout mice presented reduced flagellar beating frequency (Suzuki et al., 2009).

Multiple lines of research in different mammals have encountered the complexity to study the genetic basis of sperm quality. Complex phenotypes are determined by several or a large number of genes (Sun, 2012). In fact, the holistic effect of a gene network on a phenotype is expected to be larger than the sum of small individual effect of each of its genes. For this reason, in Study IV we build a network combining GWAS and RNA-seq data. For the GWAS, we employed the AWM (Fortes et al., 2010), a systems biology approach that combines GWAS results from different traits to identify key SNPs with effects in several phenotypes. Then, these SNPs are used to build a network based on a SNP-to-SNP co-association evidence using the PCIT algorithm (Reverter and Chan, 2008). These interactions were then corroborated by the gene co-expression analysis, which was also carried by the PCIT algorithm. The final network included genes of important relevance in spermatogenesis, DNA repair, gamete generation or metabolic functions of the cell. Some of the genes included in the network were very interesting and previously identified in our Studies II and/or III.

For example, the gene *DENND1A*, also termed as connecdenn 1, with high abundances in our samples (Study II: Gòdia et al., 2019a) was predicted as a circRNA hotspot hosting 7 circRNAs (Study III: Gòdia et al., 2019b). DENND1A functions as a guanine nucleotide exchange that interacts with members of the Rab family (Marat et al., 2011). This family plays a role in the initial stages of the exoctytosis mechanism and have been identified in the acrosomal region of human sperm with a role in calcium-triggered exocytosis (Bustos et al., 2012). Noteworthy, *DENND1A* was a member of our RNA-seq and GWAS shared

network and it also showed significant negative correlations with the percentage of abnormal acrosomes at 5 min (-0.45) and the percentage of head abnormalities (-0.32).

One of the most intact and abundant transcript in our datasets was the Heat Shock Protein B9 (*HSPB9*) gene. *HSPB9*, which is responsible for mediating heat stress response, is highly expressed in testis and in the different sperm maturation cell stages (Kappe et al., 2001; Xun et al., 2015). Interestingly, mRNA levels of *HSPB9* were positively associated with membrane integrity parameters such as the ratio of viability (0.38) and ORT (0.42).

The sperm flagellum is a complex structure with several molecular compartments involved in its composition, assembly and function (Figure 4.1 A). Through our analyses, we identified 2 genes involved in ciliogenesis. These were *ARMC9* (Van De Weghe et al., 2017) and *ULK4* (Liu et al., 2016). The role of *ARMC9* in ciliogenesis has been validated by CRISPR/Cas9 in zebrafish (Van De Weghe et al., 2017). In our samples, *ARMC9* was present at moderate to high abundance in the RNA-seq and GWAS shared network and as a circRNA hotspot hosting 5 potential circRNAs. Its abundance did not correlate to any sperm trait. *ULK4* was found highly abundant in the shared network and presented 4 circRNAs. Interestingly, its mRNA levels were positively correlated with proximal droplets (0.39). This correlation could be potentially explained as *ULK4* is an autophagy-related (ATG) gene and a member of the autophagy regulatory network (Türei et al., 2015). Although its role in sperm remains to be resolved, a study that generated *Atg7* knockout mice identified inefficient cytoplasm removal in their spermatozoa (Shang et al., 2016), this together with other studies (Zheng et al., 2007) suggest that a genetic control may play a role in the cytoplasmatic removal. *ULK4* also presented several polymorphisms in LD with the lead SNP from the GWAS hit for head abnormalities in *Sus Scrofa* Chromosome (SSC) 13; Interval 3. Although in our study no correlation has

been found between proximal droplets and head abnormalities (0.06), others have identified that cytoplasmic droplets were significantly correlated with different head morphometry measurements (Gaggini et al., 2017).

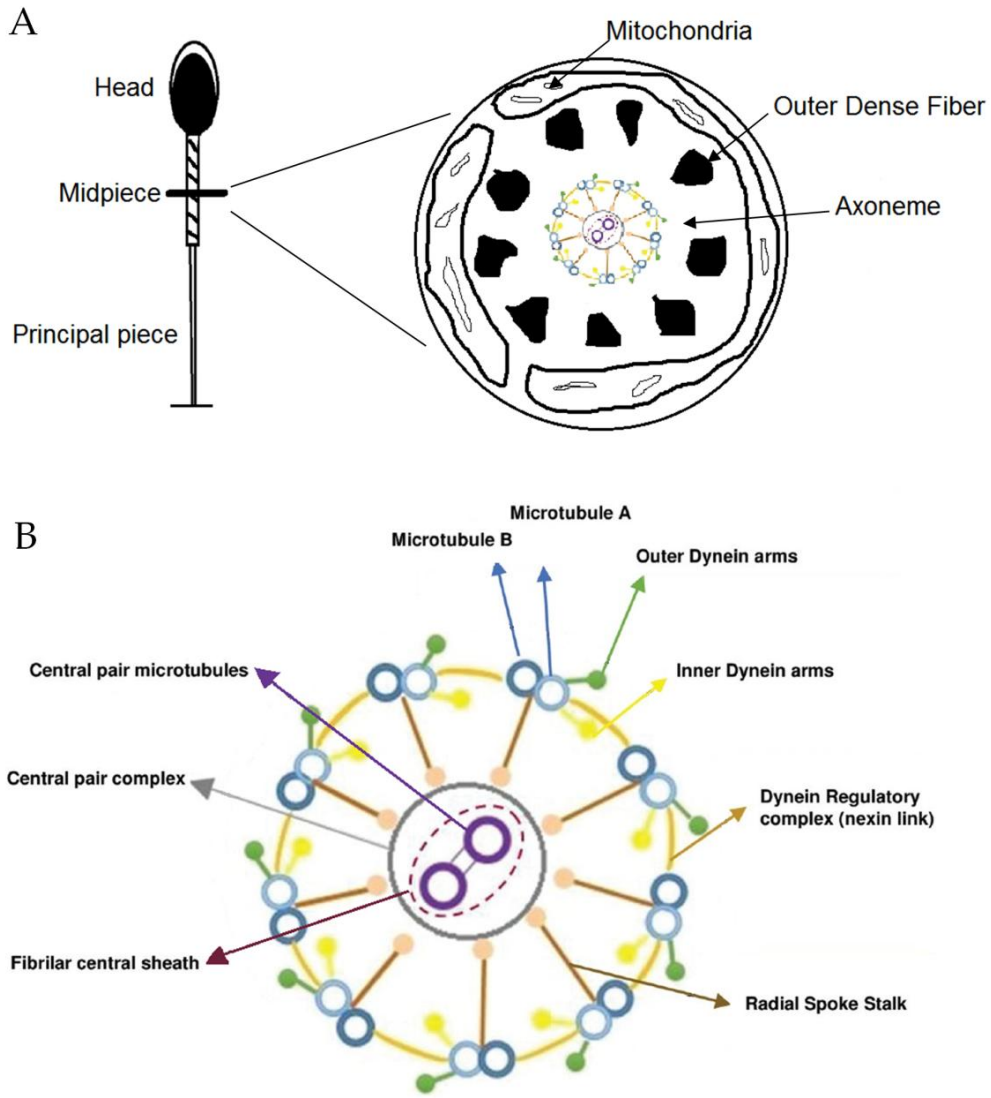


Figure 4.1. Schematic internal structure of the sperm flagella. A) drawing of the porcine sperm head and part of the tail. Cross section of the sperm midpice shows the axoneme in the center surrounded by outer dense fibers, mitochondria and the plasma membrane. B) Substructures comprising the flagellar axoneme. Modified from (Pereira et al., 2017).

The sperm mid piece contains the mitochondria and 9 Outer Dense Fibers (ODFs) surrounding the axonome (Figure 4.1 A). Our results revealed that, even though sperm mitochondrial genes are highly abundant in the transcriptome for their crucial role to generate ATP and maintaining sperm motility, only 3 genes showed significant abundance correlations with sperm motility. *ATP6*

correlated with the motility ratio (0.37), *ATP8* with VSL at 5 (0.34) and 90 min (0.37) and VAP at 90 min (0.32) and *ND1* with VSL (0.39) and VAP at 90 min (0.36). ODFs play a role in axoneme stabilization and in maintaining sperm motility (Zhao et al., 2018). In fact, *ODF2* was present in the shared network at high abundance and it was also positively correlated with sperm motility traits (motile cells : 0.41 and VCL: 0.33).

The axoneme is the core structure of the sperm flagellum and it is 1 of the most studied for its role in regulating sperm motion (Pereira et al., 2017). The axoneme is a microtubule-based structure with 9 outer doublet microtubules and 2 central doublets (9 + 2) associated with dynein arms and radial spokes (Figure 4.1 B). Dynein arms are motor proteins that convert the ATP energy into mechanical energy of movement (reviewed in: Lehti and Sironen, 2017). We identified several dynein genes in the porcine sperm transcriptome (*DRC1*, *DRC7*, *DNAH2*, *DNAH3*, *DNAH8*, *DNAH9*, *DNAH10* and *DNAI2*). Only *DNAH10* was present in the GWAS and RNA-seq shared network. *DNAI2* hosted 1 circRNA and its transcript mRNA levels were negatively correlated to the motility parameter VCL (-0.31). Interestingly, *DNAI2* has also been identified in a QTL region associated to motility in a previous GWAS study in swine (Marques et al., 2018). SNP calling showed 3 missense (moderate effect) variants in *DNAI2* (rs340229250, rs328778597 and rs703188460).

The regulation of the sperm function is dependent on post-translational processes and Calcium (Ca^{2+}) is pivotal in this regulation. Ca^{2+} in mammalian sperm is involved in the regulation of motility, hyperactivation, chemotaxis, capacitation and acrosome reaction (reviewed in: Sun et al., 2017). Ca^{2+} is also known to control sperm cell volume. There are 2 main sources of Ca^{2+} in sperm. It can be stored in the sperm head and in mitochondria or it can access the cell trough Ca^{2+} permeable-specific channels such as the CatSper (cation channels of sperm) channels. Several mutations in this channel have been associated to

infertility (Avenarius et al., 2009; Qi et al., 2007). We detected a subunit of the CatSper channel, the *CATSPERG*, which was highly abundant in the shared network. Albeit no significant correlation with phenotypes were found, it may be playing a relevant role as it was interacting with 120 genes in the final network.

Variant calling from RNA-seq data yielded the identification of several variants with high effects in genes of interest due to their known role in relevant spermatogenesis, sperm function or sperm quality. This included a heterozygous sample for a splice donor variant in *ATG16L1* (c.1146+2T>A), a gene involved in autophagy and also highlighted as a candidate for the sperm quality seasonality in Study II. We also identified a heterozygous sample for a stop gain in *SOD3* (c.315G>A), a member of the superoxide dismutase family which is involved in protecting the extracellular space from the toxic effect of ROS. These variants were only present in 1 individual and they may thus be false genotype calls.

4.4 GWAS reveals candidate genes associated to sperm quality traits

The GWAS described in Study IV was based on 285 samples and used the highest density marker chip of the market (Axiom_PigHDv1) (Groenen, 2015). Sperm quality is a complex phenotype and a low number of QTLs have been identified so far. Our study revealed 71 significant SNPs scattered in 19 genomic intervals, 12 of which presented at least 2 significant SNPs with consecutive distance below 2 Mbp.

Even though our study provided a moderate number of QTLs and positional candidate SNPs for semen quality, none matched the previously published GWAS intervals (Diniz et al., 2014; Marques et al., 2018; Zhao et al., 2016) or the candidate genes from the Pig QTL database (release 38) (Hu et al., 2019). Yet, we identified 1 of our QTL regions associated to neck abnormalities 335 kb apart from a QTL associated to motility and sperm abnormalities in another

study (Marques et al., 2018). There are several technical, biological and/or environmental factors that could be causing these discrepancies. Some technical factors could include the phenotyping accuracy, sample size, number of ejaculates evaluated per boar, the use of different genome assemblies, different statistical approaches and genotyping platforms (the Axiom_PigHDv1 contains 10 time more SNPs than the previous and widely used PorcineSNP60v2 BeadChip, and consequently, the multiple testing correction requires higher statistical genetic association for a marker to remain significant). The biological factors include the potential for genetic heterogeneity. The animals screened in each study belong to different breeds or populations with diverse genetic background. While we analysed a Pietrain population, the other studies interrogated Large White, Landrace and White Duroc × Erhualian F_2 animals. The environmental effects include factors that we and others could or did not control for, such as temperature, humidity, photoperiod, age, days between sperm collection, housing conditions and diet. We corrected our data for age, farm and season, this last being in a way is related to temperature and photoperiod.

One of the most relevant factors affecting the sensitivity and accuracy for detecting QTL associations are the sample size and the number of ejaculates evaluated per boar. Apart of 1 study (Zhao et al., 2016), all the genetic association analyses published to date used the average of a variable number of ejaculates per sample (Diniz et al., 2014; Marques et al., 2018). In fact, large within-boar variability in the ejaculates has been observed for some phenotypic records of semen quality. Wolf et al., (Wolf, 2009) estimated the repeatability for several sperm traits in 215,830 ejaculates from 3,675 boars of different breeds. In particular for Pietrain, 156 boars with an average of 48 ejaculates each showed moderately-low repeatability: semen volume (0.46), sperm concentration (0.38), motility (0.29), percentage of abnormal sperm (0.46) and total number of sperm

cells in an ejaculate (0.30). One of the major limitations of our study is that we evaluated only 1 ejaculate per boar. The reason for this is that we did not have access to any database with years of entries of phenotypic records from thousands of ejaculates and boars like the other studies from different countries. We had to process and measure the traits for all the ejaculates and within the time frame work of this thesis. Moreover, as our systems biology approach uses RNA-seq data from these samples, it requires available fresh sperm ejaculates from living animals. This did not allow us to analyse a large number of samples nor to have replicates from each boar. Nevertheless, for 28 boars we had access to 2 (24 boars) or 3 (4 boars) ejaculates each. As noted before (Wolf, 2009), we observed high variability between the ejaculates of the same pig. This is easily visualized in a Principal Component Analysis (PCA) plot (Figure 4.2), a statistical procedure that allows visualizing the variation present in a dataset with many variables (20 sperm traits) through a linear transformation in the dataset.

Our GWAS together with previous studies (Diniz et al., 2014; Marques et al., 2018; Zhao et al., 2016), suggest that sperm quality is a polygenic character with small individual effects from probably a large number of genes. Using a systems biology approach we designed a SNP panel with 74 SNPs that could predict between 5 to 36% of the phenotypic variance (Study IV). This SNP panel is very useful as (i) no major locus or gene was found shared in the different populations; (ii) sperm quality is a trait that is expressed late in life; (iii) it is measured only in 1 sex; (iv) it is difficult and expensive to be recorded. The panel included genes that have been associated to cell motility (*EFHC1, IQUB*), electron transport (*NDUFV2, COX7A1*), adherens junction maintenance (*RASSF8, RDX*), sperm development (*SPEF2*) or reproduction (*MTMR2*). These SNPs hold promising potential to be applied in genomic selection and perhaps included in the current panels for animal breeding selection strategies.

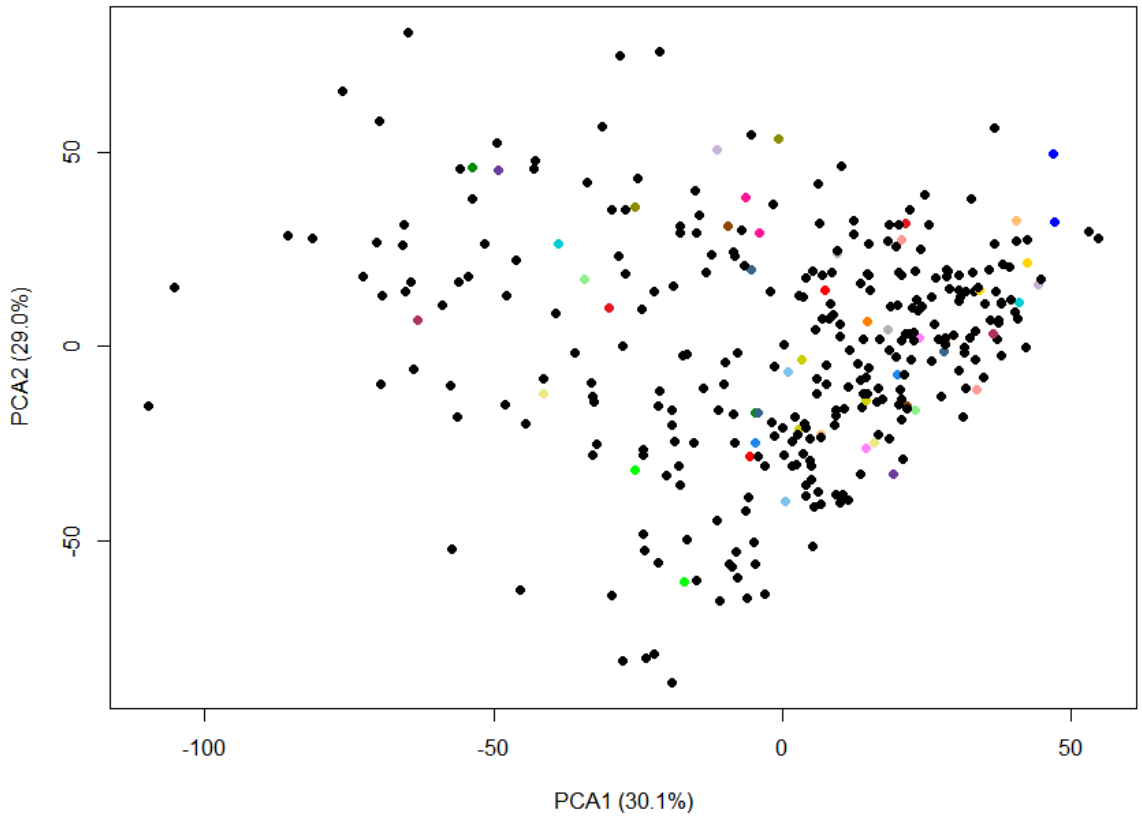


Figure 4.2. Principal Component Analysis for 20 sperm quality traits in 300 different boars. The PCA analysis includes 20 sperm quality traits with data from 1 ejaculate for 276 samples (black) and 24 samples with at least 2 ejaculates (ejaculates from the same individual present the same color).

For the eGWAS, we sought to identify whether the previous GWAS SNP hits were showing this association via marking a regulatory variant altering gene expression or via a coding variant changing protein sequence and function. For this, we interrogated the association of these variants with the RNA abundance of genes which transcript levels correlated with the same associated phenotype. This lead to the identification of 45 expressed SNPs (eSNPs) (FDR < 0.05) related to abnormal acrosome and head abnormalities and involved genes as *ACTR2, HARS, NCLN* or *IQCJ* (Study IV; Table 4 and Supplementary Table 5). This analysis was initially carried using all the SNPs although only these SNPs with FDR significant associations in the GWAS were considered thereafter. However, the correction for multiple testing included the initial set of SNPs and was thus very strict.

4.5 Allelic Ratio Distortion in sperm

ARD in sperm is expected to contribute a proportion of the TRD existing in the offspring of a sire. ARD may occur as a consequence of unequal spermatogenic efficiency or cell viability between 2 alleles in the haploid stages of spermatogenesis. ARD can provide further information on the genetic and molecular basis of spermatogenesis and semen quality. As variants under ARD may result in Hardy-Weinberg disequilibrium (reviewed in: Huang et al., 2013), these variants and their flagging markers in LD are likely to be removed from GWAS (reviewed in: Huang et al., 2013). Thus, ARD analysis may complement GWAS to obtain a more comprehensive picture on the key genes and genetic variants influencing semen quality.

We sequenced the matched diploid (from blood buffy coat) and haploid (from purified ejaculated spermatozoa) genomes of 3 boars included in the GWAS with the aim to identify heterozygous sites in blood and evaluate the existence of ARD in all these variants in sperm. We carried 2 different statistical analyses. First, we used a Bayesian method to exploit the data from the 3 samples simultaneously to identify ARD in variants that were heterozygous in the 3 boars and compare it to a previous TRD study in swine that used a similar Bayesian approach (Casellas et al., 2014). We identified that 25% of the ARD regions were in close proximity to TRD genomic segments. This high percentage suggests that these location concordances are not random and indicate a common genetic basis. However, we realized that the 3 pigs shared very few ARD variants. Under the hypothesis that TRD may led to rare variants (reviewed in: Huang et al., 2013) we searched for ARD variants within each boar using a Fisher exact test to compare allelic ratios in blood (diploid) and sperm (haploid). We classified them by these that are affecting a common gene in the 3 boars, a common genomic region, or just a – non-necessarily shared - candidate gene with reported relevance in spermatogenesis or semen quality.

Some of the identified ARD variants were predicted to have a moderate or high impact effect on the protein sequence of genes of interest for spermatogenesis or embryo development. For example, the alternative allele was found in a lower proportion of sperm cells than expected by chance for the variants identified in *TDRD6* and *TDRD15* genes. The Tudor domain-containing (TDRD) proteins are involved in piRNA biogenesis and have a crucial role in spermatogenesis. They inhibit the activity and movements of transposons during spermatogenesis by forming piRNAs and Piwi protein complexes, to maintain the germline integrity (Li et al., 2019a). A moderate effect missense variant was identified in *TDRD6* (rs336951504; c.3740A>G). TDRD6 is a testis-specific protein that has been associated to male infertility and embryo loss in humans (Sha et al., 2018). *Tdrd6*$^{-/-}$ mice showed post-meiotic germ cell arrest in round spermatids and several dysregulated miRNAs (Sha et al., 2018). A paralog of this gene is *TDRD15,* in which we identified a high effect stop gained variant (rs321799011; c.3728C), yet no relation with sperm has been found. Another interesting gene is the *Thyroid Adenoma Associated* (*THADA*). *THADA* has been connected to metabolism (type 2 diabetes) and adaptation to climate, and the cellular and molecular basis of its function is beginning to be elucidated. THADA binds the sarco/ER Ca^{2+} ATPase (SERCA) and regulate the metabolism via calcium signaling. SERCA is a major transporter involved in refilling the CA^{2+} stores and whose protein has been mainly localized in the neck region of human sperm cell (Sepulveda et al., 2007). Our results suggest that sperm cells with the *THADA* moderate missense variant (c.1724C) could present some type of fitness disadvantage. Interestingly, *THADA* was found in the shared network and was significantly positively correlated with proximal and distal droplets –associated with sperm immaturity- and the ratio of viable cells.

We also sought to compare the genomic positions of the ARD SNPs found in our study (Study V) with the genomic regions associated with sperm quality

traits identified in the GWAS (Study IV). Only 1 of the 55 ARD SNPs identified with the Bayesian approach (SSC4:1,517,094, 887 bp upstream the *ARC* gene) was found less than 1 Mbp distance away from a GWAS hit associated to percentage of abnormal acrosomes (SSC4 I1; Study IV: Table 2). Remarkably, an ARD SNP (rs325749569) identified in 1 boar through the Fisher Test approach, causing a missense variant in *CDCP1* gene, fell within the GWAS region (SSC13 I1) associated to head abnormalities (Study IV: Table 2). *CDCP1* has been reported to play a role in cell adhesion and motility in cancer cells (Orchard-Webb et al., 2014).

This is a preliminary study to grasp the potential existence of ARD in sperm, which has never been explored before. A similar study with larger sample size and deeper sequencing depth should be more robust and sensitive to provide a better ARD map of the boar sperm. This dataset could be also utilized to evaluate the extent of *de novo* germline mutations that could be transmitted to the offspring. One of the key questions in the sustainability of animal breeding is maintaining genetic diversity. Should the frequency of *de novo* germline mutations be high, this would indicate that high genetic pressure in animal breeding would have less pernicious effects on genetic diversity. For this reason, this is an interesting topic of research in livestock sciences. Furthermore, the ARD SNPs could be genotyped in our samples and included for a GWAS to analyze a potential linked effect on the phenotype.

4.6 Additional studies: the boar sperm microbiome and epigenomic map

Semen quality and male fertility can be influenced by other biological layers. Among these additional layers, the sperm microbiome and epigenome are of particular relevance. Bacterial composition has been found to affect a plethora of phenotypes in animals, including sperm quality or fertility in humans (Weng et al., 2014). Likewise, the sperm epigenome has been also linked to male infertility (Hammoud et al., 2011).

After RNA-seq and genome mapping of the sequencing reads described in Studies I-IV, we were intrigued by the high percentage of unmapped reads (between 9.1 and 28.7%). We hypothesized that a proportion of these unmapped reads could correspond to microbes present in the samples.

We aligned the unmapped reads of each sample to microbial genomes databases using the Kraken software (Wood and Salzberg, 2014). Despite the fact that the ejaculates are stored mixed with boar extenders that contain antibiotics and that the samples were purified to remove microbes, we identified a rich population of bacteria. Early results identified 18 phyla with the most abundant being *Proteobacteria*, *Firmicutes, Actinobacteria* and *Bacterioidetes* (Figure 4.3). This composition was more similar to the microbiome described in human sperm (Weng et al., 2014) than to the one described in porcine gut (Ramayo-Caldas et al., 2016). The most abundant bacterial species in porcine sperm were from environmental sources suggesting that the bacteria present in pig sperm contaminate the samples upon or after ejaculation. These species included for example *Bacillus megaterium* and *Clostridium hominis*. These results are not surprising as boars are in contact with soil, feces and water, which are the main sources of these bacteria.

We also sought to explore the relationship between the microbe abundances and sperm quality traits. For this we used canonical correlation comparing bacterial abundances with each phenotypic measurement. With this approach, we identified 25 significant correlations (FDR < 0.05) including a negative correlation between *Cutibacterium acnes* and viability at 5 min (-0.57) or *Corynebacterium aurimucosum* with VCL at 90 min (-0.54). Our results suggest, as previously reported in human (Weng et al., 2014), that bacterial contamination can have a detrimental effect on semen quality and thus, caution should be taken to minimize its impact. This data is now being further analysed and an article is being drafted.

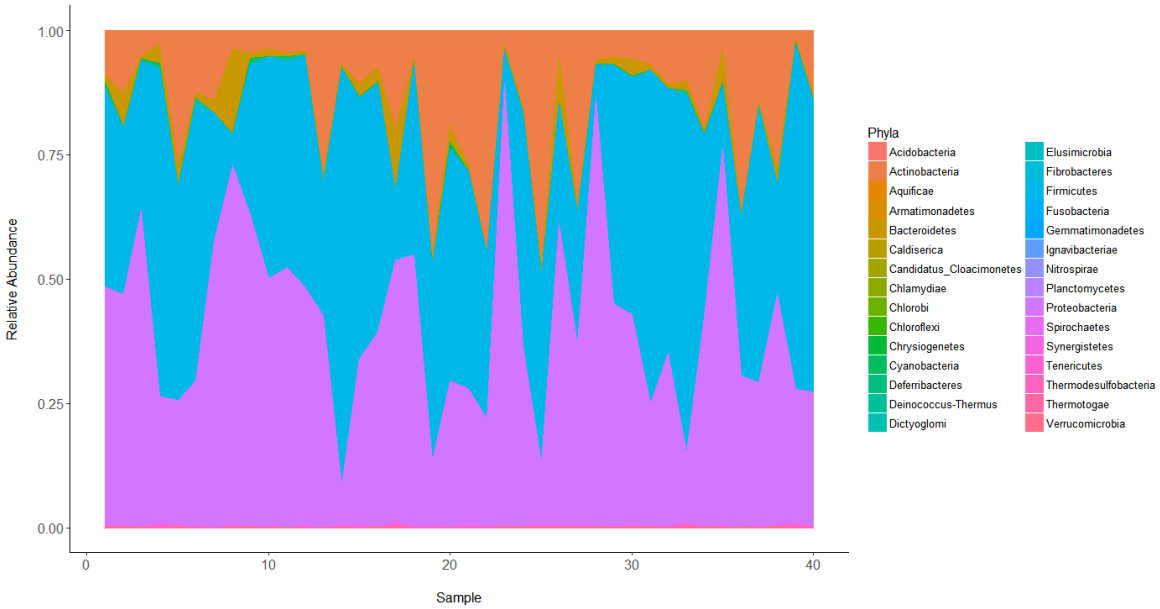


Figure 4.3. Stacked area plot of the porcine sperm microbiome phyla. The most abundant phyla across the 40 samples. *Proteobacteria, Firmicutes, Actinobacteria* and *Bacterioidetes* were the most abundant phyla.

We also characterized the localization of mononucleosomes (MN) (and thus of the retained histones) in the boar sperm. Initially, we wanted to map the genomic location of retained histone modifications (H3K4me1, H3K4me3 and H3K27ac) in sperm. As previously mentioned in the introduction, the sperm chromatin is extensively replaced by protamine forming a highly compact complex. The role of this condensation is to protect the paternal genome for fertilization (reviewed in: Ward, 2010). Abnormal protamine packaging is associated with reduced sperm concentration, motility, abnormalities, increase of DNA fragmentation and embryo development (reviewed in: Steger and Balhorn, 2018). Interestingly, our results showed that the porcine sperm chromatin appears to be more compacted than the human or mice counterparts (Hammoud et al., 2009) and several difficulties were found to detect chromatin specific histone marks when compared to human sperm (Figure 4.4 A-D).

The sperm chromatin from 4 samples was separated into protamine and histone-bound fractions through micrococcal nuclease (MNase) digestion followed by gel electrophoresis. In the histone-bound fraction, we observed a tiny presence of the genome as mononucleosome (MN; 150 bp) and subnucleosome (SN; ~90 bp) (Figure 4.4 E). As the DNA extracted from the MN

fraction was below the minimum required for studying histone modifications, both the MN and SN from 2 samples were subjected to MNase sequencing.

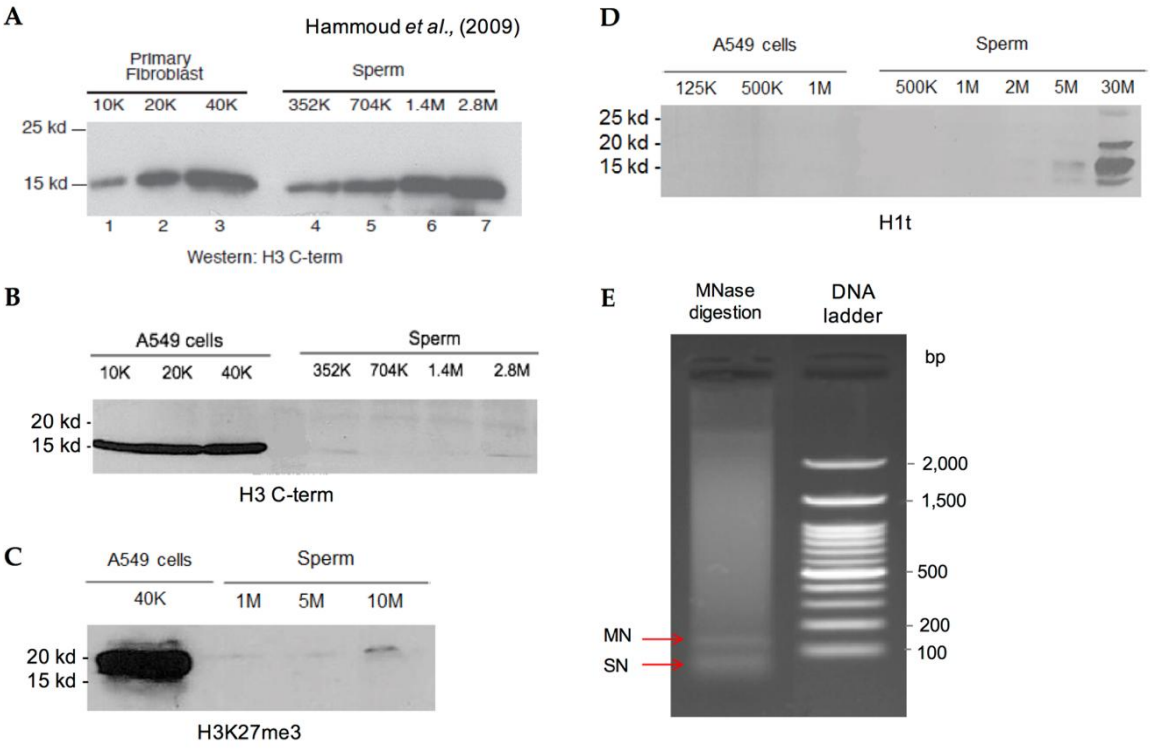


Figure 4.4. **Composition of boar sperm chromatin. A.** In human sperm cells, Hammoud et al., quantified histone content of primary fibroblast and human sperm cells by inmmunoblot analysis with H3C terminus antibody. The authors identified protein from the starting amounts of 352K cells. **B.** In porcine sperm cells, we did not quantify H3C terminus in sperm cells whereas the histone was present in A549 adenocarcinomic cells (15 kd). **C.** Western analysis for H3K27me3 (17 kd) in A549 and mature porcine sperm cells. A small quantification was obtained when processing 10 M of sperm cells. **D.** Western analysis of the Histone 1 testis-specific (H1t) (22 kd). No quantification was obtained for the control A549 cells. **E.** Digestion of porcine sperm chromatin with micrococal nuclease (MNase) resulted in 2 bands: mononucelosomes (MN) with an average size of 150 bp and subnucleosomes (SN) with an estimated size of ~90 bp.

Bioinformatics analysis showed that retained histones via MN or SN mapping presented a consistent genome-wide distribution with apparently non-random retention in gene promoters (Figure 4.5.A) and enriched in the promoters of genes related to embryo development such as the HOX family and to genes related to embryo development such as the HOX family and to spermatogenesis such as the *SPATA16* gene (Figure 4.5.B). Mono-nucleosome genomic locations in the pig sperm tended to overlap to their locations in human sperm (Hammoud et al., 2009) (adjusted P-value < 1.0×10^{-275}), which suggest a non-

random functional implication of the location of these mono-nucleosomes in sperm.

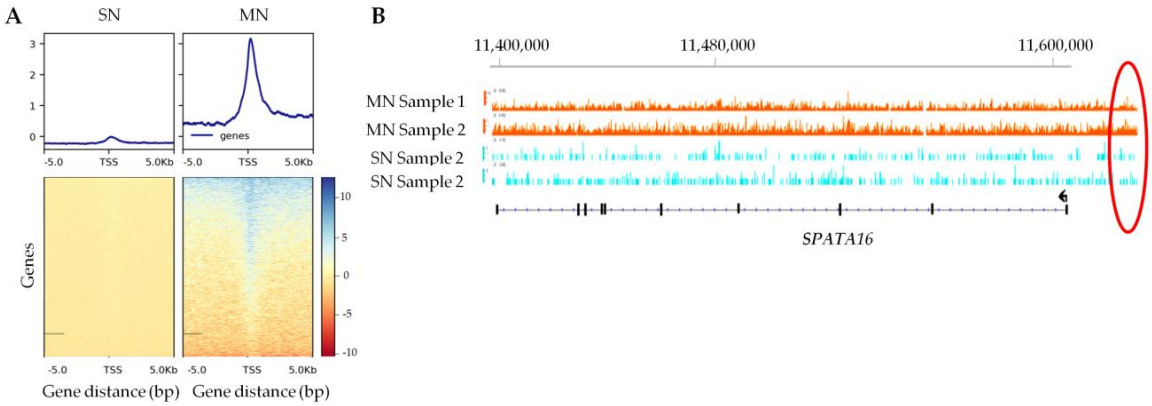


Figure 4.5 Porcine sperm chromatin nuclease-sensitivity visualization. **A.** Histone retention was mostly associated in promoter regions for both SN and MN regions. **B.** Read coverage profiling of MN and SN fractions after peak finding of the *SPATA16* (*Spermatogenesis associated 16*) gene, with a clear peak in the promoter region for the MN fractions. The y-axis is the normalized difference score for sequencing. TSS: Transcription Start Site. MN: mono-nucleosome; SN: sub-nucleosome.

4.7 Challenges and future directions

The studies carried within the frame of this thesis have yielded substantial advances in understanding the genetic and molecular basis of sperm quality. Nonetheless, we have just started to grasp the biology underlying semen traits in pigs and given the complexity of spermatogenesis and the heterogeneity of ejaculates influenced by both genetic and environmental factors, this will continue to be a challenging field of research with promising positive outcomes for pig breeding and will thus continue to be an active area of research for many years to come.

The future elucidation of the molecular mechanisms controlling sperm quality and male fertility will be strongly dependent on experimental design and methodological improvements. First, experimental designs will have to include large sample sizes with multiple ejaculates per boar and control for a large number of environmental factors such as age, days between ejaculates, season, and other factors that we might not be aware of to reduce the non-genetic phenotypic variability observed between the ejaculates of a boar. This will

require the monitoring and recording of these factors, which could include information such as diet, drug administration, health status, age, frequency of sperm collection, dilutions, storage, sow's information, etc. Second, different breeds and populations will have to be tested to identify the shared and the population specific genetic factors. Third, male fertility traits such as conception rate and litter size are now starting to be recorded by the largest breeding companies and having access to this data will be a significant advance in the field. Epigenetics, metagenomics and metatranscriptomics are also emerging as promising fields of research with remarkable potential in male reproduction and animal breeding.

Single cell analysis in testis to understand the key transcriptional and regulatory events evolving during spermatogenesis has been successfully carried in mice (Green et al., 2018). A similar approach in swine could be used to identify molecular and functional sperm subpopulations that may hold different tasks with consequences on fertility.

Genome editing tools can be useful to validate target genes or DNA variants related to spermatogenesis or fertility. In a foreseeable future, these techniques will be also implemented in animal breeding to quickly introgress beneficial genetic variants to the commercial populations. Nonetheless, this will still require some time before consumers, ethical boards and policy makers accept these strategies in the livestock production industry.

The goal for us and in the future is to identify genetic variants with a predictive value for semen quality and fertility. Moreover, genomic selection schemes may want to include these SNPs to have a more informed and efficient use of genetic information. These schemes will also aim to include other layers of information such as epigenetics and metagenomics.

Conclusions

Chapter 5

1. The purification of sperm cells from swine ejaculates results in a relatively low recovery (22%) but it is not influenced by the measured sperm quality parameters or by the RNA yield obtained upon extraction.

2. A global view of the porcine sperm transcriptome has been obtained by total and small RNA-seq analysis of 10 samples. We identified 4,436 coding genes with varying RNA abundances and transcript integrity. Moreover, we identified novel isoforms as well as protein coding genes. On the other hand, piRNAs were the most abundant class within the small noncoding RNA payload, followed by miRNAs and tRNAs.

3. The abundance of 37 mRNA transcripts and 7 miRNAs followed a seasonal pattern and could thus be related to the decrease of semen quality in the warm summer months. This group of transcripts included several genes related to oxidative stress.

4. Nearly 1,600 circRNAs were identified in at least 30 of the 40 boar sperm RNA-seq datasets. The abundance of a proportion of these circRNAs (3.5%) correlated with the levels of 31 miRNAs and also contained potential miRNA target sequences thereby suggesting they may act as miRNA sponges.

5. The RNA levels of 179 circRNAs correlated with sperm motility traits. This was validated by RT-qPCR in 2 of the 6 circRNAs that were tested.

6. A GWAS carried for 25 sperm quality traits identified 12 QTL regions associated to the percentage of head and neck abnormalities, abnormal

acrosomes and motility, and included candidate genes as *CHD2, KATNAL2* and *IQCF1*.

7. 6,128 interactions between gene RNA abundances and the 25 sperm-related traits were identified.

8. A systems biology approach integrating RNA-seq and GWAS for sperm quality traits showed a highly interconnected network with 1,313 genes and 94 miRNAs. This set of genes was enriched for biological processes such as DNA repair, meiotic cell cycle and spermatogenesis and included core genes such as *TRAPPC2L ,CARF, EFHC1* or *LAPR4.*

9. A panel of 74 SNPs that explains between 5 to 36% of the genetic variance of sperm related traits is proposed. This panel should be tested in additional boar populations to check its predictive potential, its robustness and thus its value for the animal breeding sector.

10. Whole Genome Sequencing allowed the identification of genetic variants in Allellic Ratio Distortion in sperm. These variants and affected genes could impact on semen quality and boar fertility. Some of the genes under ARD included *ARID4, AK7, GEN1* and *RAD9B.*

References

Chapter 6

Aitken RJ (2006). Sperm function tests and fertility. *Int J Androl* 29**:**69-75.

Alcivar AA, Hake LE, Millette CF, Trasler JM and Hecht NB (1989). Mitochondrial gene expression in male germ cells of the mouse. *Dev Biol* 135**:**263-271.

Alm K, Peltoniemi OA, Koskinen E and Andersson M (2006a). Porcine field fertility with two different insemination doses and the effect of sperm morphology. *Reprod Domest Anim* 41**:**210-213.

Alm K, Peltoniemi OA, Koskinen E and Andersson M (2006b). Porcine field fertility with two different insemination doses and the effect of sperm morphology. *Reprod Domest Anim* 41**:**210-213.

Amills M, Clop A, Ramírez Ó and Pérez-Enciso M (2010). Origin and Genetic Diversity of Pig Breeds. In: Encyclopedia of Life Sciences (ELS). John Wiley and Sons, Ltd: Chichester. 1-10.

Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* 16**:**57.

Archibald AL, Bolund L, Churcher C, Fredholm M, Groenen MA et al. (2010). Pig genome sequence-analysis and publication strategy. *BMC Genomics* 11**:**438.

Arsenakis I, Appeltant R, Sarrazin S, Rijsselaere T, Soom AV et al. (2015). Relationship between carcass quality characteristics and semen quality in Pietrain boars. *Reprod Domest Anim* 50**:**121-121.

Avenarius MR, Hildebrand MS, Zhang Y, Meyer NC, Smith LL et al. (2009). Human male infertility caused by mutations in the CATSPER1 channel protein. *Am J Hum Genet* 84**:**505-510.

Ballester M, Ramayo-Caldas Y, Revilla M, Corominas J, Castelló A et al. (2017). Integration of liver gene co-expression networks and eGWAs analyses highlighted candidate regulators implicated in lipid metabolism in pigs. *Sci Rep* 7**:**46539.

Banaszewska D and Kondracki S (2012). An assessment of the breeding maturity of insemination boars based on ejaculate quality changes. *Folia Biol (Krakow)* 60**:**151-162.

Barceló M, Mata A, Bassas L and Larriba S (2018). Exosomal microRNAs in seminal plasma are markers of the origin of azoospermia and can predict the presence of sperm in testicular tissue. *Human Reprod* 33**:**1087-1098.

Bauer H, Veron N, Willert J and Herrmann BG (2007). The t-complex-encoded guanine nucleotide exchange factor Fgd2 reveals that two opposing

signaling pathways promote transmission ratio distortion in the mouse. *Genes Dev* 21**:**143-147.

Bauer H, Schindler S, Charron Y, Willert J, Kusecek B et al. (2012). The Nucleoside Diphosphate Kinase Gene Nme3 Acts as Quantitative Trait Locus Promoting Non-Mendelian Inheritance. *PLoS Genet* 8**:**e1002567.

Boivin J, Bunting L, Collins JA and Nygren KG (2007). International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care. *Hum Reprod* 22**:**1506-1512.

Bonet S, Garcia E and Sepúlveda L (2013). The Boar Reproductive System. In: Boar Reproduction. Springer, Berlin, Heidelberg. 65-107.

Botstein D, White RL, Skolnick M and Davis RW (1980). Construction of a Genetic-Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am J Hum Genet* 32**:**314-331.

Briz M (1994). Microscopical analysis of the ejaculated sperm and the sperm epididymal maturation of sus domesticus. Doctoral Thesis

Briz M and Fàbrega A (2013). The Boar Spermatozoon. In: Boar Reproduction. Springer, Berlin, Heidelberg. 3-47.

Broekhuijse ML, Sostaric E, Feitsma H and Gadella BM (2012a). Application of computer-assisted semen analysis to explain variations in pig fertility. *J Anim Sci* 90**:**779-789.

Broekhuijse MLWJ, Sostaric E, Feitsma H and Gadella BM (2012b). The value of microscopic semen motility assessment at collection for a commercial artificial insemination center, a retrospective study on factors explaining variation in pig fertility. *Theriogenology* 77**:**1466-1479.

Brykczynska U, Hisano M, Erkek S, Ramos L, Oakeley EJ et al. (2010). Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nat Struct Mol Biol* 17**:**679-687.

Bustos MA, Lucchesi O, Ruete MC, Mayorga LS and Tomes CN (2012). Rab27 and Rab3 sequentially regulate human sperm dense-core granule exocytosis. *Proc Natl Acad Sci U S A* 109**:**E2057-E2066.

Casellas J, Manunza A, Mercader A, Quintanilla R and Amills M (2014). A flexible bayesian model for testing for transmission ratio distortion. *Genetics* 198**:**1357-1367.

Cooke HJ and Saunders PT (2002). Mouse models of male infertility. *Nat Rev Genet* 3**:**790-801.

Corominas J, Ramayo-Caldas Y, Puig-Oliveras A, Estelle J, Castello A et al. (2013). Analysis of porcine adipose tissue transcriptome reveals

differences in de novo fatty acid synthesis in pigs with divergent muscle fatty acid composition. *BMC Genomics* 14**:**843.

Cortes-Lopez M and Miura P (2016). Emerging Functions of Circular RNAs. *Yale J Biol Med* 89**:**527-537.

Champroux A, Cocquet J, Henry-Berger J, Drevet JR and Kocer A (2018). A Decade of Exploring the Mammalian Sperm Epigenome: Paternal Epigenetic and Transgenerational Inheritance. *Front Cell Dev Biol* 6**:**50.

Chavatte-Palmer P, Velazquez MA, Jammes H and Duranthon V (2018). Review: Epigenetics, developmental programming and nutrition in herbivores. *Animal* 12**:**s363-s371.

Chen C, Ai H, Ren J, Li W, Li P et al. (2011). A global view of porcine transcriptome in three tissues from a full-sib pair with extreme phenotypes in growth and fat deposition by paired-end RNA sequencing. *BMC Genomics* 12**:**448.

Chen R, Yu SA, Ren F, Lv XY and Pan CY (2016). Detection of one large insertion/deletion (indel) and two novel SNPs within the SPEF2 gene and their associations with male piglet reproduction traits. *Arch Anim Breed* 59**:**275-283.

Chioccarelli T, Manfrevola F, Ferraro B, Sellitto C, Cobellis G et al. (2019). Expression Patterns of Circular RNAs in High Quality and Poor Quality Human Spermatozoa. *Front Endocrinol (Lausanne)* 10**:**435.

Cho CL and Agarwal A (2018). Role of sperm DNA fragmentation in male factor infertility: A systematic review. *Arab J Urol* 16**:**21-34.

Dekkers JC (2012). Application of genomics tools to animal breeding. *Curr Genomics* 13**:**207-212.

Diniz DB, Lopes MS, Broekhuijse ML, Lopes PS, Harlizius B et al. (2014). A genome-wide association study reveals a novel candidate gene for sperm motility in pigs. *Anim Reprod Sci* 151**:**201-207.

Diniz WJS, Mazzoni G, Coutinho LL, Banerjee P, Geistlinger L et al. (2019). Detection of Co-expressed Pathway Modules Associated With Mineral Concentration and Meat Quality in Nelore Cattle. *Front Genet* 10**:**210.

Estill MS, Hauser R and Krawetz SA (2019). RNA element discovery from germ cell to blastocyst. *Nucleic Acids Res* 47**:**2263-2275.

FAO (1994). A manual for the primary animal health care worker. 85-107.

FAO (2017). http://www.fao.org/home/ [accessed: 1 May 2019]

Farrell PB, Presicce GA, Brockett CC and Foote RH (1998). Quantification of bull sperm characteristics measured by computer-assisted sperm analysis (CASA) and the relationship to fertility. *Theriogenology* 49**:**871-879.

Fawcett DW (1975). The mammalian spermatozoon. *Dev Biol* 44**:**394-436.

Ferlin A, Raicu F, Gatta V, Zuccarello D, Palka G et al. (2007). Male infertility: role of genetic background. *Reprod Biomed Online* 14**:**734-745.

Ferramosca A and Zara V (2014). Bioenergetics of mammalian sperm capacitation. *Biomed Res Int* 2014**:**902953.

Flowers WL (2009). Selection for boar fertility and semen quality--the way ahead. *Soc Reprod Fertil Suppl* 66**:**67-78.

Fortes MR, Reverter A, Zhang Y, Collis E, Nagaraj SH et al. (2010). Association weight matrix for the genetic dissection of puberty in beef cattle. *Proc Natl Acad Sci U S A* 107**:**13642-13647.

Gadea J (2003). Review: semen extenders used in the artificial inseminarion of swine. *Span J Agric Res* 1**:**11.

Gadea J and Matas C (2000). Sperm factors related to in vitro penetration of porcine oocytes. *Theriogenology* 54**:**1343-1357.

Gaggini TS, Rocha LO, Souza ET, de Rezende FM, Antunes RC et al. (2017). Head morphometry and chromatin instability in normal boar spermatozoa and in spermatozoa with cytoplasmic droplets. *Anim Reprod* 14**:**1253-1258.

Garcia-Gil N, Pinart E, Sancho S, Badia E, Bassols J et al. (2002). The cycle of the seminiferous epithelium in Landrace boars. *Anim Reprod Sci* 73**:**211-225.

Garner DL and Hafez ESE (2000). Spermatozoa and Seminal Plasma. In: Reproduction in Farm Animals. Lippincott Williams and Wilkins, Maryland. 96-109.

Geldermann H (1975). Investigations on Inheritance of Quantitative Characters in Animals by Gene Markers. 1. Methods. *Theor Appl Genet* 46**:**319-330.

Giuffra E, Tuggle CK and Consortium F (2019). Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu Rev Anim Biosci* 7**:**65-88.

Gòdia M, Swanson G and Krawetz SA (2018a). A history of why fathers' RNA matters. *Biol Reprod* 99**:**147-159.

Gòdia M, Mayer FQ, Nafissi J, Castelló A, Rodríguez-Gil JE et al. (2018b). A technical assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-seq analysis. *Syst Biol Reprod Med* 64**:**291-303.

Gòdia M, Estill M, Castelló A, Balasch S, Rodríguez-Gil JE et al. (2019a). A RNA-Seq Analysis to Describe the Boar Sperm Transcriptome and Its Seasonal Changes. *Front Genet* 10**:**299.

Gòdia M, Castelló A, Rocco M, Cabrera B, Rodríguez-Gil JE et al. (2019b). Identification of circular RNAs in porcine sperm and their relation to sperm motility. *bioRxiv*:608026.

Goodrich R, Anton E and Krawetz SA (2013). Isolating mRNA and small noncoding RNAs from human sperm. *Methods Mol Biol* 927:385-396.

Goodrich R, Johnson G and Krawetz SA (2007). The preparation of human spermatozoal RNA for clinical analysis. *Arch Androl* 53:161-167.

Green CD, Ma Q, Manske GL, Shami AN, Zheng X et al. (2018). A Comprehensive Roadmap of Murine Spermatogenesis Defined by Single-Cell RNA-Seq. *Dev Cell* 46:651-667 e610.

Groenen MAM, Schook LB and Archibald AL (2011). Pig genomics. In: The genetics of the pig. CABI, Wallingford, UK. 179-199.

Groenen MAMArchibald ALUenishi HTuggle CKTakeuchi Y et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491:393-398.

Groenen MAM (2015). Development of a high-density Axiom® porcine genotyping array to meet research and commercial needs. Animal Genome XXIII Conference, San Diego, CA.

Groenen MAM (2016). A decade of pig genome sequencing: a window on pig domestication and evolution. *Genet Sel Evol* 48:23.

Gunawan A, Cinar MU, Uddin MJ, Kaewmala K, Tesfaye D et al. (2012). Investigation on Association and Expression of ESR2 as a Candidate Gene for Boar Sperm Quality and Fertility. *Reprod Domest Anim* 47:782-790.

Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT et al. (2009). Distinctive chromatin in human sperm packages genes for embryo development. *Nature* 460:473-478.

Hammoud SS, Nix DA, Hammoud AO, Gibson M, Cairns BR et al. (2011). Genome-wide analysis identifies changes in histone retention and epigenetic modifications at developmental and imprinted gene loci in the sperm of infertile men. *Human Reprod* 26:2558-2569.

Hirai M, Boersma A, Hoeflich A, Wolf E, Foll J et al. (2001). Objectively measured sperm motility and sperm head morphometry in boars (Sus scrofa): Relation to fertility and seminal plasma growth factors. *J Androl* 22:104-110.

Hogarth CA and Griswold MD (2010). The key role of vitamin A in spermatogenesis. *J Clin Invest* 120:956-962.

Holstein A-F, Schulze W and Davidoff M (2003). Understanding spermatogenesis is a prerequisite for treatment. *Reprod Biol Endocrinol* 1**:**107.

Holt C, Holt WV, Moore HD, Reed HC and Curnock RM (1997). Objectively measured boar sperm motility parameters correlate with the outcomes of on-farm inseminations: results of two fertility trials. *J Androl* 18**:**312-323.

Hu ZL, Park CA and Reecy JM (2019). Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res* 47**:**D701-D710.

Huang LO, Labbe A and Infante-Rivard C (2013). Transmission ratio distortion: review of concept and implications for genetic association studies. *Hum Genet* 132**:**245-263.

Huang SY, Chen MY, Lin EC, Tsou HL, Kuo YH et al. (2002). Effects of single nucleotide polymorphisms in the 5'-flanking region of heat shock protein 70.2 gene on semen quality in boars. *Anim Reprod Sci* 70**:**99-109.

Humphray SJ, Scott CE, Clark R, Marron B, Bender C et al. (2007). A high utility integrated map of the pig genome. *Genome Biol* 8**:**R139.

IDESCAT (2017). https://www.idescat.cat/ [accessed: 1 May 2019]

Izumiyama T, Minoshima S, Yoshida T and Shimizu N (2012). A novel big protein TPRBK possessing 25 units of TPR motif is essential for the progress of mitosis and cytokinesis. *Gene* 511**:**202-217.

Jenkins TG and Carrell DT (2012). The sperm epigenome and potential implications for the developing embryo. *Reproduction* 143**:**727-734.

Jodar M, Sendler E and Krawetz SA (2016). The protein and transcript profiles of human semen. *Cell Tissue Res* 363**:**85-96.

Jodar M, Sendler E, Moskovtsev SI, Librach CL, Goodrich R et al. (2015). Absence of sperm RNA elements correlates with idiopathic male infertility. *Sci Transl Med* 7**:**295re296.

Jodar M, Soler-Ventura A, Oliva R, Molecular Biology of R and Development Research G (2017). Semen proteomics and male infertility. *J Proteomics* 162**:**125-134.

Jones RE and Lopez KH (2014). Chapter 4 - The Male Reproductive System. In: Human Reproductive Biology (Fourth Edition). Academic Press. 67-83.

Juonala T, Lintukangas S, Nurttila T and Andersson M (1998). Relationship between semen quality and fertility in 106 AI-boars. *Reprod Domest Anim* 33**:**155-158.

Juyena NS and Stelletta C (2012). Seminal plasma: an essential attribute to spermatozoa. *J Androl* 33**:**536-551.

Kaewmala K, Uddin MJ, Cinar MU, Grosse-Brinkhaus C, Jonas E et al. (2012). Investigation into Association and Expression of PLCz and COX-2 as Candidate Genes for Boar Sperm Quality and Fertility. *Reprod Domest Anim* 47**:**213-223.

Kaewmala K, Uddin MJ, Cinar MU, Grosse-Brinkhaus C, Jonas E et al. (2011). Association study and expression analysis of CD9 as candidate gene for boar sperm quality and fertility traits. *Anim Reprod Sci* 125**:**170-179.

Kappe G, Verschuure P, Philipsen RL, Staalduinen AA, Van de Boogaart P et al. (2001). Characterization of two novel human small heat shock proteins: protein kinase-related HspB8 and testis-specific HspB9. *Biochim Biophys Acta* 1520**:**1-6.

Kmiec M, Ziemak J, Dybus A and Matusiak S (2002). Analysis of relations between polymorphism in steroid 21-hydroxylase gene (CYP21) and quantitative and qualitative characters of boar semen. *Czech J Anim Sci* 47**:**194-199.

Knox RV (2016). Artificial insemination in pigs today. *Theriogenology* 85**:**83-93.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK and Mardis ER (2013). The next-generation sequencing revolution and its impact on genomics. *Cell* 155**:**27-38.

Kondracki S (2003). Breed differences in semen characteristics of boars used in artificial insemination in Poland. *Pig News Inf* 24**:**119N-122N.

Krausz C, Escamilla AR and Chianese C (2015). Genetics of male infertility: from research to clinic. *Reproduction* 150**:**R159-174.

Krawetz SA (2005). Paternal contribution: new insights and future challenges. *Nat Rev Genet* 6**:**633-642.

Langendijk P, Bouwman E, Kidson A, Kirkwood R, Soede N et al. (2002). Role of myometrial activity in sperm transport through the genital tract and in fertilization in sows. *Reproduction* 123**:**683-690.

Larson G, Dobney K, Albarella U, Fang MY, Matisoo-Smith E et al. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* 307**:**1618-1621.

Lee WY, Lee R, Kim HC, Lee KH, Cui XS et al. (2014). Pig Spermatozoa Defect in Acrosome Formation Caused Poor Motion Parameters and Fertilization Failure through Artificial Insemination and In vitro Fertilization. *Asian Austral J Anim* 27**:**1417-1425.

Leenhouwers JI, Bergsma R, Knol EF and Feitsma H (2008). Genetic Parameters for Fertility of Boars and Impact of Selection against Boar Taint. *Reprod Domest Anim* 43**:**100-100.

Lehti MS and Sironen A (2017). Formation and function of sperm tail structures in association with sperm motility defects. *Biol Reprod* 97**:**522-536.

Li B, He X, Zhao Y, Bai D, Bou G et al. (2019a). Identification of piRNAs and piRNA clusters in the testes of the Mongolian horse. *Sci Rep* 9**:**5022.

Li X, Jiang B, Wang X, Liu X, Zhang Q et al. (2019b). Estimation of genetic parameters and season effects for semen traits in three pig breeds of South China. *J Anim Breed Genet* 136**:**183-189.

Li Z, Huang C, Bao C, Chen L, Lin M et al. (2015). Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* 22**:**256-264.

Lin C, Tholen E, Jennen D, Ponsuksili S, Schellander K et al. (2006a). Evidence for effects of testis and epididymis expressed genes on sperm quality and boar fertility traits. *Reprod Domest Anim* 41**:**538-543.

Lin CL, Ponsuksili S, Tholen E, Jennen DG, Schellander K et al. (2006b). Candidate gene markers for sperm quality and fertility of boar. *Anim Reprod Sci* 92**:**349-363.

Liu M, Guan ZL, Shen Q, Lalor P, Fitzgerald U et al. (2016). Ulk4 Is Essential for Ciliogenesis and CSF Flow. *J Neurosci* 36**:**7589-7600.

Lopez Rodriguez A, Van Soom A, Arsenakis I and Maes D (2017). Boar management and semen handling factors affect the quality of boar extended semen. *Porcine Health Manag* 3**:**15.

Love CC (2011). Relationship between sperm motility, morphology and the fertility of stallions. *Theriogenology* 76**:**547-557.

Mackay TF, Stone EA and Ayroles JF (2009). The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* 10**:**565-577.

Mach N, Berri M, Esquerre D, Chevaleyre C, Lemonnier G et al. (2014). Extensive expression differences along porcine small intestine evidenced by transcriptome sequencing. *PLoS One* 9**:**e88515.

Maes D, López Rodríguez A, Rijsselaere T, Vyt P and Van Soom A (2011). Artificial Insemination in Pigs. In: Artificial Insemination in Farm Animals. InTech. 79-94.

Maes D, Nauwynck H, Rijsselaere T, Mateusen B, Vyt P et al. (2008). Diseases in swine transmitted by artificial insemination: An overview. *Theriogenology* 70**:**1337-1345.

Marat AL, Dokainish H and McPherson PS (2011). DENN Domain Proteins: Regulators of Rab GTPases. *J Biol Chem* 286**:**13791-13800.

Marete A, Lund MS, Boichard D and Ramayo-Caldas Y (2018). A system-based analysis of the genetic determinism of udder conformation and health

phenotypes across three French dairy cattle breeds. *PLoS One* 13:e0199931.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.

Marques DBD, Bastiaansen JWM, Broekhuijse MLWJ, Lopes MS, Knol EF et al. (2018). Weighted single-step GWAS and gene network analysis reveal new candidate genes for semen traits in pigs. *Genet Sel Evol* 50:40.

Marques DBD, Lopes MS, Broekhuijse M, Guimaraes SEF, Knol EF et al. (2017). Genetic parameters for semen quality and quantity traits in five pig lines. *J Anim Sci* 95:4251-4259.

Matsumoto AM and Bremner WJ (2016). Chapter 19 - Testicular Disorders. In: Williams Textbook of Endocrinology 694-784.

Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495:333-338.

Mengerink KJ and Vacquier VD (2002). An ATP-binding cassette transporter is a major glycoprotein of sea urchin sperm membranes. *J Biol Chem* 277:40729-40734.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet* 11:31-46.

Meyerholz DK (2016). Lessons learned from the cystic fibrosis pig. *Theriogenology* 86:427-432.

Moskovtsev SI and Librach CL (2013). Methods of sperm vitality assessment. *Methods Mol Biol* 927:13-19.

Neeteson-van Nieuwenhoven A-M, Knap P and Avendaño S (2013). The role of sustainable commercial pig and poultry breeding for food security. *Animal Frontiers* 3:52-57.

Ng P and Kirkness E (2010). Whole Genome Sequencing. In: Genetic Variation. Methods in Molecular Biology. Humana Press. 215-226.

Oh SH, See MT, Long TE and Galvin JM (2006a). Estimates of genetic correlations between production and semen traits in boar. *Asian-Australas J Anim Sci* 19:160-164.

Oh SH, See MT, Long TE and Galvin JM (2006b). Genetic parameters for various random regression models to describe total sperm cells per ejaculate over the reproductive lifetime of boars. *J Anim Sci* 84:538-545.

Oliva R (2006). Protamines and male infertility. *Hum Reprod Update* 12:417-435.

Orchard-Webb DJ, Lee TC, Cook GP and Blair GE (2014). CUB domain containing protein 1 (CDCP1) modulates adhesion and motility in colon cancer cells. *BMC Cancer* 14**:**754.

Pareek CS, Smoczynski R and Tretyn A (2011). Sequencing technologies and genome sequencing. *J Appl Genet* 52**:**413-435.

Paston MJ, Sarkar S, Oates RP and Badawy SZA (1994). Computer-Aided Semen Analysis Variables as Predictors of Male-Fertility Potential. *Arch Andrology* 33**:**93-99.

Pegolo S, Mach N, Ramayo-Caldas Y, Schiavon S, Bittante G et al. (2018). Integration of GWAS, pathway and network analyses reveals novel mechanistic insights into the synthesis of milk proteins in dairy cows. *Sci Rep* 8**:**566.

Pereira R, Sa R, Barros A and Sousa M (2017). Major regulatory mechanisms involved in sperm motility. *Asian J Androl* 19**:**5-14.

Pessot CA, Brito M, Figueroa J, Concha, II, Yanez A et al. (1989). Presence of RNA in the sperm nucleus. *Biochem Biophys Res Commun* 158**:**272-278.

Piomboni P, Focarelli R, Stendardi A, Ferramosca A and Zara V (2012). The role of mitochondria in energy production for human sperm motility. *Int J Androl* 35**:**109-124.

Qi HY, Moran MM, Navarro B, Chong JA, Krapivinsky G et al. (2007). All four CatSper ion channel proteins are required for male fertility and sperm cell hyperactivated motility. *Proc Natl Acad Sci U S A* 104**:**1219-1223.

Quintero-Moreno A, Rigau T and Rodríguez-Gil J (2004). Regression analyses and motile sperm subpopulation structure study as improving tools in boar semen quality analysis. *Theriogenology* 61**:**673-690.

Ramayo-Caldas Y, Renand G, Ballester M, Saintilan R and Rocha D (2016a). Multi-breed and multi-trait co-association analysis of meat tenderness and other meat quality traits in three French beef cattle breeds. *Genet Sel Evol* 48**:**37.

Ramayo-Caldas Y, Mach N, Lepage P, Levenez F, Denis C et al. (2016b). Phylogenetic network analysis applied to pig gut microbiota identifies an ecosystem structure linked with growth traits. *ISME J* 10**:**2973-2977.

Ramayo-Caldas Y, Ballester M, Fortes MRS, Esteve-Codina A, Castello A et al. (2014). From SNP co-association to RNA co-expression: Novel insights into gene networks for intramuscular fatty acid composition in porcine. *BMC Genomics* 15**:**232.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL et al. (2009). Design of a High Density SNP Genotyping Assay in the Pig Using

SNPs Identified and Characterized by Next Generation Sequencing Technology. *PLoS One* 4:e6524.

Raudsepp T and Chowdhary B (2011). Cytogenetics and chromosome maps. In: The genetics of the pig. CABI, Wallingford, UK. 134-178.

Reilly JN, McLaughlin EA, Stanger SJ, Anderson AL, Hutcheon K et al. (2016). Characterisation of mouse epididymosomes reveals a complex profile of microRNAs and a potential mechanism for modification of the sperm epigenome. *Sci Rep* 6:31794.

Revell SG and Mrode RA (1994). An Osmotic Resistance Test for Bovine Semen. *Anim Reprod Sci* 36:77-86.

Reverter A and Chan EK (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24:2491-2497.

Ritchie MD, Holzinger ER, Li R, Pendergrass SA and Kim D (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* 16:85-97.

Robinson JA and Buhr MM (2005). Impact of genetic selection on management of boar replacement. *Theriogenology* 63:668-678.

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

Rothschild MF (1996). Genetics and reproduction in the pig. *Anim Reprod Sci* 42:143-151.

Salas-Huetos A, Blanco J, Vidal F, Mercader JM, Garrido N et al. (2014). New insights into the expression profile and function of micro-ribonucleic acid in human spermatozoa. *Fertil Steril* 102:213-222.

Sancho S and Vilagran I (2013). The Boar Ejaculate: Sperm Function and Seminal Plasma Analyses. In: Boar Reproduction. Springer, Berlin, Heidelberg. 471-516.

Sanger F (1975). The Croonian Lecture, 1975. Nucleotide sequences in DNA. *Proc R Soc Lond B Biol Sci* 191:317-333.

Schadt EE, Zhang B and Zhu J (2009). Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments. *Genetica* 136:259-269.

Schagdarsurengin U and Steger K (2016). Epigenetics in male reproduction: effect of paternal diet on sperm quality and offspring health. *Nat Rev Urol* 13:584-595.

Schill WB, Topfer-Petersen E and Heissler E (1988). The sperm acrosome: functional and clinical aspects. *Hum Reprod* 3:139-145.

Schulze M, Buder S, Rudiger K, Beyerbach M and Waberski D (2014). Influences on semen traits used for selection of young AI boars. *Anim Reprod Sci* 148**:**164-170.

Sendler E, Johnson GD, Mao SH, Goodrich RJ, Diamond MP et al. (2013). Stability, delivery and functions of human sperm RNAs at fertilization. *Nucleic Acids Res* 41**:**4104-4117.

Sepulveda MR, Berrocal M, Marcos D, Wuytack F and Mata AM (2007). Functional and immunocytochemical evidence for the expression and localization of the secretory pathway Ca2+-ATPase isoform 1 (SPCA1) in cerebellum relative to other Ca2+ pumps. *J Neurochem* 103**:**1009-1018.

Sha YW, Wang X, Su ZY, Wang CR, Ji ZY et al. (2018). TDRD6 is associated with oligoasthenoteratozoospermia by sequencing the patient from a consanguineous family. *Gene* 659**:**84-88.

Shang YL, Wang HN, Jia PF, Zhao HC, Liu C et al. (2016). Autophagy regulates spermatid differentiation via degradation of PDLIM1. *Autophagy* 12**:**1575-1592.

Sharma U, Sun F, Conine CC, Reichholf B, Kukreja S et al. (2018). Small RNAs Are Trafficked from the Epididymis to Developing Mammalian Sperm. *Dev Cell* 46**:**481-494.

Shipley C (1999). Breeding soundness examination in the boar. *Swine Health Prod* 7**:**117-120.

Sironen A, Uimari P, Nagy S, Paku S, Andersson M et al. (2010). Knobbed acrosome defect is associated with a region containing the genes STK17b and HECW2 on porcine chromosome 15. *BMC Genomics* 11**:**699.

Sironen A, Thomsen B, Andersson M, Ahola V and Vilkki J (2006). An intronic insertion in KPL2 results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. *Proc Natl Acad Sci U S A* 103**:**5006-5011.

Smital J, Wolf J and De Sousa LL (2005). Estimation of genetic parameters of semen characteristics and reproductive traits in AI boars. *Anim Reprod Sci* 86**:**119-130.

Steger K and Balhorn R (2018). Sperm nuclear protamines: A checkpoint to control sperm chromatin quality. *Anat Histol Embryol* 47**:**273-279.

Sun XH, Zhu YY, Wang L, Liu HL, Ling Y et al. (2017). The Catsper channel and its roles in male fertility: a systematic review. *Reprod Biol Endocrinol* 15**:**65.

Sun YV (2012). Integration of biological networks and pathways with genetic association studies. *Hum Genet* 131**:**1677-1686.

Suzuki T, Miyamoto H, Nakahari T, Inoue I, Suemoto T et al. (2009). Efhc1 deficiency causes spontaneous myoclonus and increased seizure susceptibility. *Hum Mol Genet* 18**:**1099-1109.

Szabo L and Salzman J (2016). Detecting circular RNAs: bioinformatic and experimental challenges. *Nat Rev Genet* 17**:**679-692.

Trost B, Moir CA, Gillespie ZE, Kusalik A, Mitchell JA et al. (2015). Concordance between RNA-sequencing data and DNA microarray data in transcriptome analysis of proliferative and quiescent fibroblasts. *Roy Soc Open Sci* 2.

Tsakmakidis IA, Lymberopoulos AG and Khalifa TAA (2010). Relationship between sperm quality traits and field-fertility of porcine semen. *J Vet Sci* 11**:**151-154.

Türei D, Földvári-Nagy L, Fazekas D, Módos D, Kubisch J et al. (2015). Autophagy Regulatory Network-A systems-level bioinformatics resource for studying the mechanism and regulation of autophagy. *Autophagy* 11**:**155-165.

Urban T and Kuciel J (2001). The effect of point mutation in RYR1 gene on the semen quality traits in boars of Large White and Landrace breeds. *Czech J Anim Sci* 46**:**460-464.

Van De Weghe JC, Rusterholz TDS, Latour B, Grout ME, Aldinger KA et al. (2017). Mutations in ARMC9, which Encodes a Basal Body Protein, Cause Joubert Syndrome in Humans and Ciliopathy Phenotypes in Zebrafish. *Am J Hum Genet* 101**:**23-36.

Véron N, Bauer H, Weiße AY, Lüder G, Werber M et al. (2009). Retention of gene products in syncytial spermatids promotes non-Mendelian inheritance as revealed by the t complex responder. *Genes Dev* 23**:**2705-2710.

Verstegen J, Iguer-Ouada M and Onclin K (2002). Computer assisted semen analyzers in andrology research and veterinary practice. *Theriogenology* 57**:**149-179.

Visconti PE, Westbrook VA, Chertihin O, Demarco I, Sleight S et al. (2002). Novel signaling pathways involved in sperm acquisition of fertilizing capacity. *J Reprod Immunol* 53**:**133-150.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 101**:**5-22.

Vojtech L, Woo S, Hughes S, Levy C, Ballweber L et al. (2014). Exosomes in human semen carry a distinctive repertoire of small non-coding RNAs with potential regulatory functions. *Nucleic Acids Res* 42**:**7290-7304.

Vyt P, Maes D, Rijsselaere T, Dewulf J, de Kruif A et al. (2007). Semen handling in porcine artificial insemination centres: the Belgian situation. *Vlaams Diergeneeskundig Tijdschrift*:195-200.

Waddington CH (1942). The epigenotype. *Endeavour* 1:18-20.

Wang X and Kadarmideen HN (2019). An Epigenome-Wide DNA Methylation Map of Testis in Pigs for Study of Complex Traits. *Front Genet* 10:405.

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

Ward WS (2010). Function of sperm chromatin structural elements in fertilization and development. *Mol Hum Reprod* 16:30-36.

Weng SL, Chiu CM, Lin FM, Huang WC, Liang C et al. (2014). Bacterial communities in semen from men of infertile couples: metagenomic sequencing reveals relationships of seminal microbiota to semen quality. *PLoS One* 9:e110152.

Widmann P, Reverter A, Fortes MRS, Weikard R, Suhre K et al. (2013). A systems biology approach using metabolomic data reveals genes and pathways interacting to modulate divergent growth in cattle. *BMC Genomics* 14:798.

Wimmers K, Lin CL, Tholen E, Jennen DGJ, Schellander K et al. (2005). Polymorphisms in candidate genes as markers for sperm quality and boar fertility. *Anim Genet* 36:152-155.

Woelders H, Te Pas MF, Bannink A, Veerkamp RF and Smits MA (2011). Systems biology in animal sciences. *Animal* 5:1036-1047.

Wolf J (2009). Genetic parameters for semen traits in AI boars estimated from data on individual ejaculates. *Reprod Domest Anim* 44:338-344.

Wolf J and Smital J (2009). Quantification of factors affecting semen traits in artificial insemination boars from animal model analyses. *J Anim Sci* 87:1620-1627.

Wolf J (2010). Heritabilities and genetic correlations for litter size and semen traits in Czech Large White and Landrace pigs. *J Anim Sci* 88:2893-2903.

Wood DE and Salzberg SL (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15.

Xun W, Shi L, Cao T, Zhao C, Yu P et al. (2015). Dual functions in response to heat stress and spermatogenesis: characterization of expression profile of small heat shock proteins 9 and 10 in goat testis. *Biomed Res Int* 2015:686239.

Yeh CM, Liu ZJ and Tsai WC (2018). Advanced Applications of Next-Generation Sequencing Technologies to Orchid Biology. *Curr Issues Mol Biol* 27**:**51-70.

Yeste M, Sancho S, Briz M, Pinart E, Bussalleu E et al. (2010). A diet supplemented with L-carnitine improves the sperm quality of Pietrain but not of Duroc and Large White boars when photoperiod and temperature increase. *Theriogenology* 73**:**577-586.

Zak LJ, Gaustad AH, Bolarin A, Broekhuijse M, Walling GA et al. (2017). Genetic control of complex traits, with a focus on reproduction in pigs. *Mol Reprod Dev* 84**:**1004-1011.

Zhang B and Horvath S (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 4**:**Article17.

Zhao W, Li Z, Ping P, Wang G, Yuan X et al. (2018). Outer dense fibers stabilize the axoneme to maintain sperm motility. *J Cell Mol Med* 22**:**1755-1768.

Zhao X, Zhao K, Ren J, Zhang F, Jiang C et al. (2016). An imputation-based genome-wide association study on traits related to male reproduction in a White Duroc x Erhualian F2 population. *Anim Sci J* 87**:**646-654.

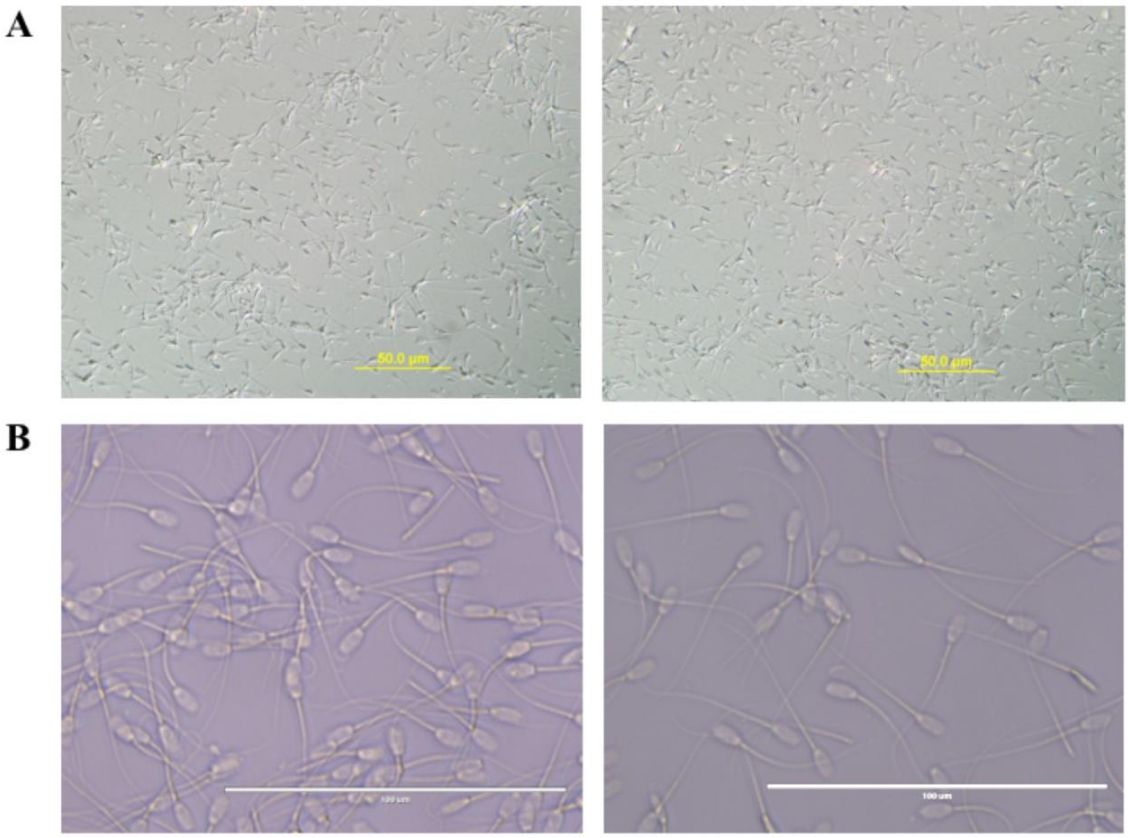
Zheng HL, Stratton CJ, Morozumi K, Jin JL, Yanagimachi R et al. (2007). Lack of Spem1 causes aberrant cytoplasm removal, sperm deformation, and male infertility. *Proc Natl Acad Sci U S A* 104**:**6852-6857.

Zhu Z, Umehara T, Okazaki T, Goto M, Fujita Y et al. (2019). Gene Expression and Protein Synthesis in Mitochondria Enhance the Duration of High-Speed Linear Motility in Boar Sperm. *Front Physiol* 10**:**252.

Annexes

Chapter 7

Supplementary material for Study I: "A technical assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-seq analysis"



Paper I: Figure S1: Optical microscopy inspection to determine the success of somatic and non-mature spermatozoa cell removal. The sperm purification protocol that we implemented is sufficient to remove non-spermatozoa and immature sperm cells. 20µl of purified sperm were smeared onto a microscope slide (Linealab) and analyzed with AixoPhot microscopy. (A) 2 samples with 20X magnification; scale bars: 50µm.(B) 2 samples with 40X magnification; scale bars: 100µm.

Paper I: Table S1: RNA levels of 14 tissue specific genes in the purified spermatozoa transcriptome. This data is based on the total

				Abundance (FPKM)						
Specific tissue	**Gene symbol**	**Gene name**	**Ensembl Gene ID**	**Sample_1**	**Sample_2**	**Sample_3**	**Sample_4**	**Sample_5**	**Sample_6**	**Average**
Spermatozoa	*OAZ3*	Ornithine Decarboxylase Antizyme 3	ENSSSCG00000027091	18.010,8	16.835,0	27.210,3	27.279,1	32.008,0	14.679,3	22.670,4
Spermatozoa	*PRM1*	Protamine 1	ENSSSCG00000021337	8.374,3	13.283,4	23.123,3	18.433,1	11.802,0	17.192,1	15.368,0
Spermatozoa	*PRM2*	Protamine 2	ENSSSCG00000024307	1.567,7	3.976,0	3.013,1	3.004,3	4.592,4	2.584,1	3.122,9
Spermatozoa	*TNP1*	Transition Protein 1	ENSSSCG00000016179	3.126,1	10.799,8	3.289,3	3.111,6	8.235,0	17.889,0	7.741,8
Spermatozoa	*ODF1*	Outer Dense Fiber Of Sperm Tails 1	ENSSSCG00000006056	323,5	883,9	493,9	522,9	379,6	2.071,1	779,2
Spermatozoa	*SMCP*	Sperm Mitochondria Associated Cysteine Rich Protein	ENSSSCG00000006594	814,6	1.027,8	1.734,1	1.397,2	1.479,4	962,1	1.235,9
Immature germ cells	*KIT*	KIT Proto-Oncogene Receptor Tyrosine Kinase	ENSSSCG00000008842	0,5	0,9	0,0	0,0	0,3	0,0	0,3
Epithelial cells	*CDH1*	Cadherin 1	ENSSSCG00000013481	3,1	1,6	3,5	1,9	3,2	6,2	3,3
Keratinocytes	*KRT10*	Keratin 10	ENSSSCG00000028522	0,0	0,0	0,0	0,2	0,0	0,0	0,0
Keratinocytes	*KRT1*	Keratin 1	ENSSSCG00000000251	0,0	0,0	0,0	0,0	0,1	0,0	0,0
Leukocytes	*PTPRC*	Protein Tyrosine Phosphatase, Receptor Type C	ENSSSCG00000010908	0,2	0,9	0,5	0,0	0,0	0,0	0,3
Leukocytes	*IL8*	Interleukin 8	ENSSSCG00000008953	3,5	24,9	5,8	0,0	5,2	16,5	9,3
Whole blood	*HBB*	Hemoglobin Subunit Beta	ENSSSCG00000014725	18,2	7,1	3,5	0,0	6,7	0,0	5,9
Prostate	*KLK3*	Kallikrein Related Peptidase 3	ENSSSCG00000015799	0,0	0,0	0,3	0,0	0,0	0,0	0,1

Supplementary material for Study II: "A RNA-Seq Analysis to Describe the Boar Sperm Transcriptome and Its Seasonal Changes"

Paper II: Table S1: RNA-seq quality and mapping statistics. Average and Standard Deviation (SD) for the 10 boar sperm samples processed, including: amount of RNA extracted and several RNA-seq bioinformatics statistics for both total and small RNA-seq.

	Average	**SD**
Starting amount of RNA (fg per cell)	2.1	0.6
Total RNA-seq reads	23631767	13945595
Reads passing quality control	23199420	187600
Proportion of reads passing quality control	0.98	0.01
Proportion of mapped reads in the Sscrofa11.1 genome	0.81	0.03
Unique reads (duplicates removed)	5606879	2988090
Proportion of non duplicated reads	0.19	0.07
Unmapped reads for *De Novo* assembly	5061315	2763218
sncRNA-seq reads	6609540	1123089
Reads passing quality control	6075152	1451518
Proportion of reads passing quality control	0.92	0.10
Proportion of mapped reads	0.82	0.02
Remaining unmapped reads	1116482	359301

Paper II: Table S2: Distribution of the top decile most abundant SREs (Sperm RNA Elements) into SRE types and gene biotypes. Number of SREs (within the top decile) for each SRE type (exonic, intronic, upstream/downstream 10 kb and orphan). Total non-redundant number of genes and their biotype for each SRE class.

SRE type	Number of SREs	Number of unique genes	Biotype
Exonic	12860	3470	
		3325	protein coding
		37	snoRNA
		30	snRNA
		22	Mt tRNA
		21	lincRNA
		13	misc RNA
		9	miRNA
		8	scaRNA
		2	Mt rRNA
		2	ribozyme
		1	pseudogene
Intronic	2129	1474	
		1440	protein coding
		32	lincRNA
		2	pseudogene
Upstream/downstream 10 kb	848	631	
		587	protein coding
		13	snoRNA
		9	lincRNA
		9	snRNA
		7	misc RNA
		2	miRNA
		1	Mt tRNA
		1	pseudogene
		1	rRNA
		1	sRNA
Orphan	2667		
Total	18504	4436	

Paper II: Table S3: List of human and bovine genes identified by syntenic alignment of the orphan SREs. Orphan SRE genome coordinates were liftover to human and bovine coordinates, and the genes mapped in these regions were extracted. A total of 45 genes shared in both species were found. From these genes, 44 were already annotated in the *Sscrofa* Ensembl v.91 annotation. 17 of these genes were also detected by exonic, intronic and/or upstream/downstream 10 kb SREs. This suggests that orphan SREs could correspond to unannotated isoforms or to paralogous genes.

Gene symbol	**Gene annotated in *Sscrofa* Ensembl v.91**	**Gene detected in the SRE pipeline**	**SRE biotype**
ANXA3	No	No	
CDYL	Yes	No	
RORA	Yes	No	
PPP1R14D	Yes	No	
KCNB2	Yes	No	
TACC3	Yes	No	
CCDC3	Yes	No	
XRN1	Yes	No	
ZNF300	Yes	No	
EDA	Yes	No	
ZNF385B	Yes	No	
EFCAB6	Yes	No	
RIMS2	Yes	No	
WDPCP	Yes	No	
MS4A13	Yes	No	
CCDC7	Yes	No	
SRGAP3	Yes	No	
FHIT	Yes	No	
DNHD1	Yes	No	
UNC79	Yes	No	
DNAI1	Yes	No	
EEA1	Yes	No	
FMR1NB	Yes	No	
LELP1	Yes	No	
SMIM23	Yes	No	
TSG101	Yes	No	
TTN	Yes	No	
UBE2N	Yes	No	
UBL7	Yes	Yes	Exon
ABCA3	Yes	Yes	Exon
CLIP4	Yes	Yes	Exon

Gene symbol	Gene annotated in *Sscrofa* Ensembl v.91	Gene detected in the SRE pipeline	SRE biotype
CYB5R4	Yes	Yes	Exon
CDRT4	Yes	Yes	Exon
RAB3GAP1	Yes	Yes	Exon / upstream-downstream 10 kb
BOLL	Yes	Yes	Exon / upstream-downstream 10 kb
CATSPERE	Yes	Yes	Exon / upstream-downstream 10 kb
PEBP4	Yes	Yes	Exon / intronic
TTC28	Yes	Yes	Exon / intronic
ZEB1	Yes	Yes	Exon / intronic
ANKRD26	Yes	Yes	Exon / intronic
ADARB2	Yes	Yes	Exon / intronic
PPP1R12A	Yes	Yes	Exon / intronic
ST5	Yes	Yes	Exon / intronic / upstream-downstream 10 kb
ADAM32	Yes	Yes	Exon / intronic / upstream-downstream 10 kb
ANHX	Yes	Yes	Exon / intronic / upstream-downstream 10 kb

Paper II: Table S4: Gene Ontology analysis of the genes including the top decile most abundant and the orphan SREs detected in the SRE pipeline. GO biological process terms with significant Bonferroni corrected *p*-values (*p*-val < 0.05) and their associated genes.

See table at:

https://www.frontiersin.org/articles/10.3389/fgene.2019.00299/full#supplementary-material

Paper II: Table S5: Gene Ontology analysis of the different SRE abundance variance groups. GO biological process terms with significant Bonferroni corrected *p*-values (*p*-val < 0.05) and their associated genes.

SRE variance	GO Term	P-value (Bonferroni corrected)	Associated Genes Found
< 25 %	diencephalon development	15.0E-3	*BAX, GATA2, NR4A2*
	regulation of histone modification	14.0E-3	*GATA2, OTUB1, PHF1, PYGO2*
	positive regulation of extrinsic apoptotic signaling pathway	17.0E-3	*BAX, HYAL2, STK3*
	histone acetylation	4.0E-3	*BRD8, GATA2, KAT2A, MSL1, PYGO2*
	histone H4 acetylation	16.0E-3	*BRD8, KAT2A, MSL1*
	DNA-templated transcription, elongation	13.0E-3	*GTF2F1, TCEB2, ZMYND11*
	transcription elongation from RNA polymerase II promoter	16.0E-3	*GTF2F1, TCEB2, ZMYND11*
	spermatid differentiation	1.5E-3	*BAX, LOC100513764, PYGO2, TSSK2, TSSK6*
	spermatid development	13.0E-3	*LOC100513764, PYGO2, TSSK2, TSSK6*
	spermatid nucleus differentiation	1.8E-3	*LOC100513764, PYGO2, TSSK6*
	regulation of carbohydrate metabolic process	17.0E-3	*AP2A1, ARFGEF1, KAT2A, PGAM2*
	monosaccharide biosynthetic process	14.0E-3	*GPI, KAT2A, PGAM2*
	hexose biosynthetic process	16.0E-3	*GPI, KAT2A, PGAM2*
	gluconeogenesis	16.0E-3	*GPI, KAT2A, PGAM2*
	glycolytic process	9.5E-3	*GPI, PFKM, PGAM2*
	ATP generation from ADP	9.5E-3	*GPI, PFKM, PGAM2*
	sequestering of metal ion	5.3E-3	*BAX, CASQ1, DHRS7C*
	regulation of sequestering of calcium ion	13.0E-3	*BAX, CASQ1, DHRS7C*
	sequestering of calcium ion	13.0E-3	*BAX, CASQ1, DHRS7C*
	negative regulation of sequestering of calcium ion	16.0E-3	*BAX, CASQ1, DHRS7C*
	regulation of calcium ion transport into cytosol	14.0E-3	*BAX, CASQ1, DHRS7C*
	release of sequestered calcium ion into cytosol	16.0E-3	*BAX, CASQ1, DHRS7C*
	regulation of release of sequestered calcium ion into cytosol	15.0E-3	*BAX, CASQ1, DHRS7C*
> 75 %	single fertilization	80.0E-6	*AQN-1, BSP1, PSP-I, SPMI*

Paper II: Table S6: Correlation between transcript integrity across samples, with transcript abundance and coding sequence length. Correlation of the TIN (Transcripts Integrity Number) between samples, with the transcript abundance and with the coding sequence length of the transcripts. This table shows the correlation of the TIN (Transcripts Integrity Number) between each pair of samples, the correlation of the TIN with the transcript average abundance in FPKM (Fragments per Kilobase per Million mapped reads) across the 10 samples, and the correlation of the TIN with the length of coding sequence of the transcripts.

	Sample_1	**Sample_2**	**Sample_3**	**Sample_4**	**Sample_5**	**Sample_6**	**Sample_7**	**Sample_8**	**Sample_9**	**Sample_10**	**Average FPKM**	**CDS length**
Sample_1	1	0.88	0.88	0.89	0.90	0.82	0.89	0.83	0.80	0.88	0.19	0.16
Sample_2		1	0.84	0.93	0.90	0.82	0.84	0.85	0.73	0.93	0.18	0.17
Sample_3			1	0.85	0.90	0.72	0.86	0.73	0.78	0.85	0.14	0.19
Sample_4				1	0.91	0.83	0.85	0.85	0.76	0.92	0.18	0.17
Sample_5					1	0.8	0.88	0.81	0.80	0.90	0.17	0.18
Sample_6						1	0.84	0.91	0.78	0.80	0.25	0.16
Sample_7							1	0.83	0.85	0.83	0.20	0.20
Sample_8								1	0.76	0.83	0.24	0.14
Sample_9									1	0.73	0.21	0.16
Sample_10										1	0.18	0.18

Paper II: Table S7: Summary statistics of the *de novo* transcriptome assembly. Summary statistics of the Trinity output based on the number of potential novel genes and transcripts, and size (in bp) of the contigs based on all transcripts isoforms or based only on the longest isoform for each potential gene.

		Mean	SD
Counts of transcripts, etc.	**Total trinity 'genes':**	5162	4949
	Total trinity transcripts:	8459	9718
	Percent GC:	51	2
Stats based on ALL transcript contigs:	**Contig N10:**	411	22
	Contig N20:	342	17
	Contig N30:	305	13
	Contig N40:	279	10
	Contig N50:	259	8
	Median contig length:	245	5
	Average contig:	269	7
	Total assembled bases:	2259991	2650673
Stats based on ONLY LONGEST ISOFORM per 'GENE':	**Contig N10:**	422	27
	Contig N20:	348	20
	Contig N30:	307	14
	Contig N40:	280	11
	Contig N50:	259	8
	Median contig length:	243	5
	Average contig:	270	7
	Total assembled bases:	1378675	1366476

Paper II: Table S8: List of proteins identified by *de novo* analysis, with the species in which they were detected and transcript abundance. *De novo* analysis of the unmapped reads resulted in 1,060 proteins which passed the quality control filters. For each protein, we include the cognate species, the predicted RNA mean abundance in the 10 samples (in FPKM), the Standard Deviation (SD) of their RNA abundance and the gene ID symbol retrieved from Uniprot (https://www.uniprot.org/). FPKM: Fragments per Kilobase per Million mapped reads.

See table at:

https://www.frontiersin.org/articles/10.3389/fgene.2019.00299/full#supplementary-material

Paper II: Table S9: Non-redundant list of genes identified by *de novo* analysis. 768 potentially novel genes were identified from the unmapped reads. The gene symbol IDs were retrieved with Uniprot from the Trinity output protein names. These genes were detected in at least one species (detailed in column 2 of **Supplementary File S8**). The majority of these genes were annotated in the porcine Ensembl v.91 but 29 were identified as novel genes. 40 of the genes annotated in the porcine genome were not detected with the SREs pipeline which indicates that none of their cognate reads mapped to the genome even though these genes are annotated.

See table at:

https://www.frontiersin.org/articles/10.3389/fgene.2019.00299/full#supplementary-material

Paper II: Table S10: List of long non-coding RNAs detected in porcine sperm. Ensembl IDs of the lncRNAs identified in this study, their genome coordinates, average RNA abundance across the 10 samples and length. Most of the lncRNAs presented, as an average across all samples, low RNA abundances.

Ensembl ID	Chr	Start	End	Mean (FPKM)	SD (FPKM)	Length
ENSSSCG00000035295	13	61955128	61969856	373.59	141.70	465
ENSSSCG00000034939	8	97432154	97450700	72.73	22.23	1703
ENSSSCG00000035981	18	27856577	28014521	64.29	23.87	341
ENSSSCG00000033972	1	233025365	233336379	61.85	26.26	1425
ENSSSCG00000032052	4	79322942	79328061	21.65	9.48	1291
ENSSSCG00000031291	3	24513403	24520855	12.94	5.16	729
ENSSSCG00000036690	11	69511284	69651155	12.25	6.07	833
ENSSSCG00000038695	17	8959711	8962539	11.87	6.52	307
ENSSSCG00000032301	10	24956426	24976482	7.68	2.97	3142
ENSSSCG00000032000	10	35448134	357612691	7.02	2.41	1037
ENSSSCG00000031683	10	15279669	15314943	5.71	2.47	6128
ENSSSCG00000034856	12	16373726	16418238	4.33	1.92	1143
ENSSSCG00000040839	17	53705212	545574861	3.74	0.91	3611

Ensembl ID	Chr	Start	End	Mean (FPKM)	SD (FPKM)	Length
ENSSSCG00000036379	3	21638532	216502811	3.60	2.85	5757
ENSSSCG00000038464	12	16418630	16442927	2.97	1.07	1481
ENSSSCG00000036144	14	96622430	97038967	2.19	1.49	1870
ENSSSCG00000033132	6	7661357	78467871	1.84	0.49	790
ENSSSCG00000031841	12	16445936	16483292	1.67	0.58	5763
ENSSSCG00000039306	13	166979297	167254436	1.63	0.88	3018
ENSSSCG00000035395	10	3699470	37137201	1.57	0.72	678
ENSSSCG00000037652	3	72251	140482	1.52	0.47	1548
ENSSSCG00000036070	8	9305256	94876731	1.47	0.71	1869
ENSSSCG00000036117	10	66741291	669717461	1.38	0.59	4286
ENSSSCG00000034578	5	81118160	811948591	1.24	0.77	1023
ENSSSCG00000040829	18	30192179	302339921	1.11	0.66	440
ENSSSCG00000037166	7	14264736	146333571	1.05	0.46	6265
ENSSSCG00000036344	2	145419449	145644372	0.95	0.53	708

Paper II: Table S11: Distribution of the short RNA-seq reads mapping to different RNA types. Proportion and standard deviation (SD) across the 10 samples.

RNA biotype	Proportion of the total reads (Mean)	Proportion of the total reads (SD)		
snRNA	0.0409	0.0166		
snoRNA	0.0008	0.0002		
rRNA	0.0268	0.0195		
piRNA	0.1250	0.0588		
miRNA	0.0681	0.0399		
tRNA	0.0763	0.0465	Total proportion of sncRNAs	0.34
Mt tRNA	0.4412	0.1317		
Mt rRNA	0.0731	0.0221		
miscRNA	0.0686	0.0246		
Protein Coding	0.0789	0.0491		
LincRNA	0.0004	0.0002	Total proportion of other RNA classes	0.66

Paper II: Table S12: Concordance of miRNA identification between our dataset and other sperm RNA-seq studies. Comparison of the miRNAs identified in our study with other sperm RNA-seq experiments in pig, in human, and cattle.
See table at:

https://www.frontiersin.org/articles/10.3389/fgene.2019.00299/full#supplementary-material

Paper II: Table S13: RNA abundance levels and coefficient of variation of miRNAs, tRNAs, and piRNAs in the porcine sperm. RNA abundance is measured in CPM (Counts Per Million) across the 10 samples. We only considered the miRNAS with >0 CPMs in all the samples. The genomic coordinates of piRNAs refer to the Sscrofa10.2 built instead of Sscrofa11.1 as provided by the piRNAs cluster database [40].

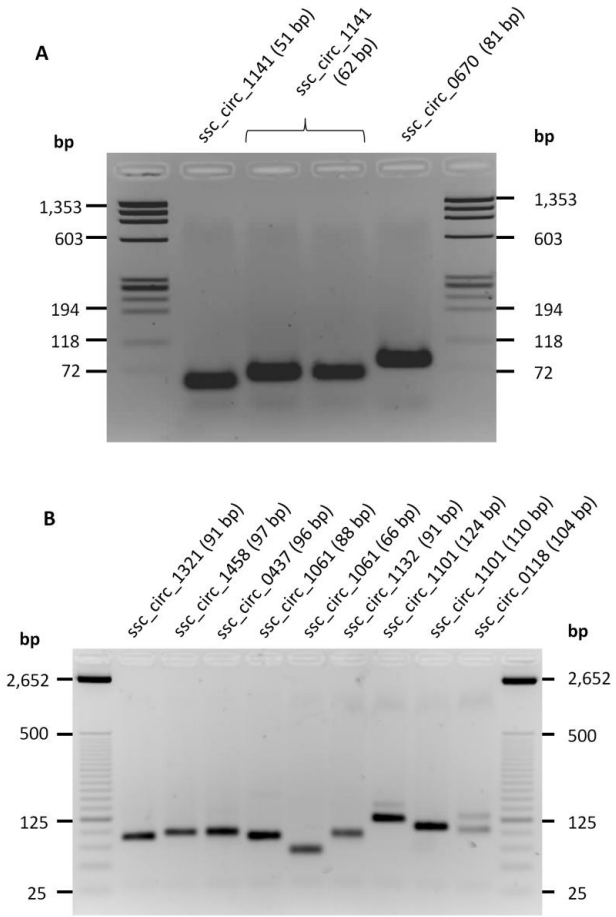
See table at:

https://www.frontiersin.org/articles/10.3389/fgene.2019.00299/full#supplementary-material

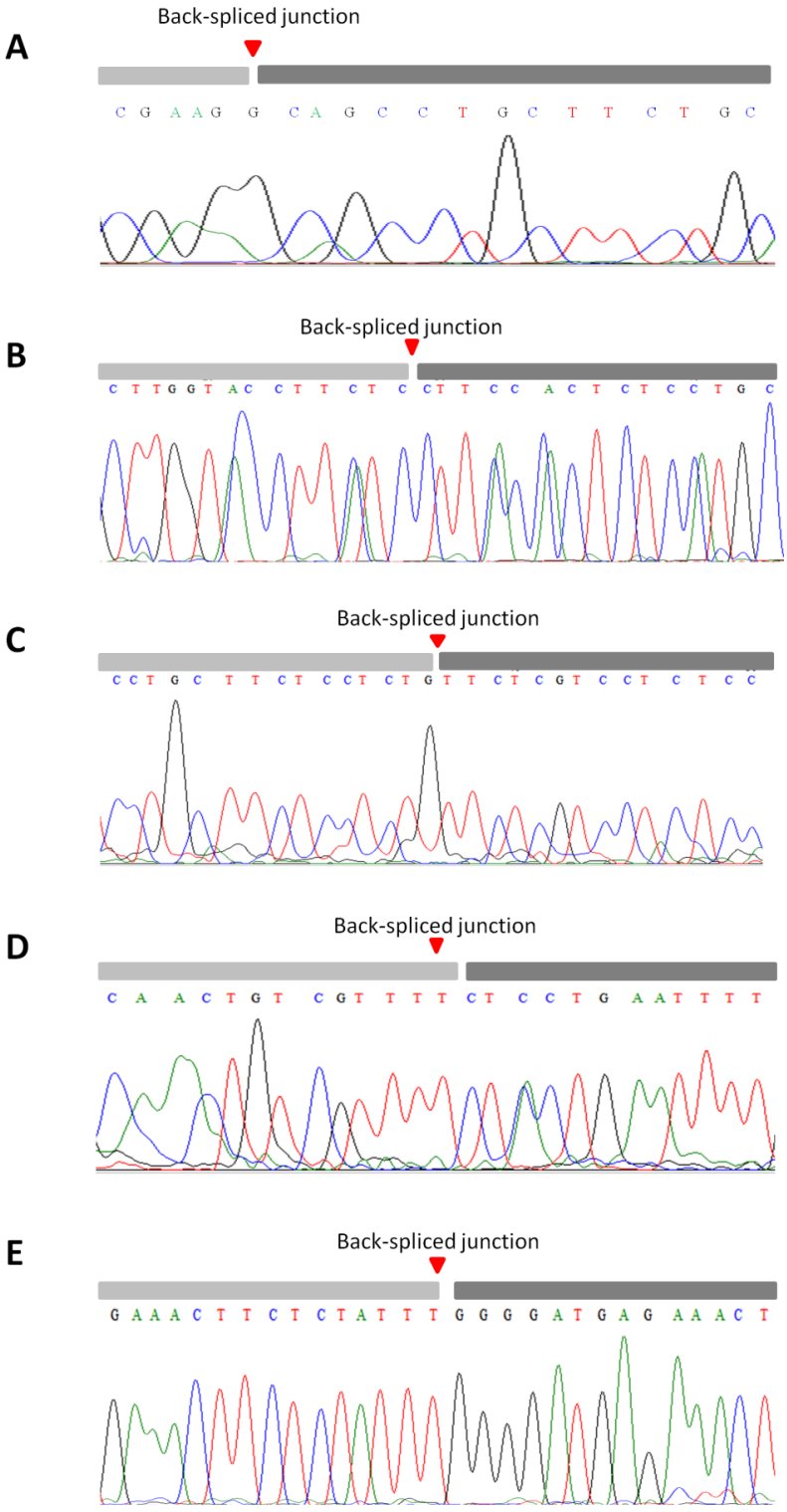
Paper II: Table S14: Novel piRNA clusters identified in the pig sperm RNA. We detected 17 potential clusters of piRNAs that were found in at least 3 of the 10 samples analyzed in this study. Mean and standard deviation (SD) in CPM (Counts Per Million).

Chromosome	Start	End	Mean (CPM)	SD (CPM)	Length (bp)	Number of samples in which this piRNA cluster was detected
13	119924726	119933132	585.0	188.1	8406	10
7	4775171	4783632	329.6	114.6	8461	10
14	67372347	67376436	284.1	95.9	4089	10
14	71966940	71971281	241.4	57.9	4341	10
13	201704570	201713054	229.0	372.2	8484	3
8	595049	602180	163.6	38.4	7131	10
AEMK02000694.1	162144	177203	105.0	28.5	15059	10
AEMK02000390.1	36273	38630	69.4	63.3	2357	7
11	7908960	7964989	40.4	44.0	56029	5
4	811190	817220	18.9	31.0	6030	3
2	4317724	4322885	17.2	30.4	5161	3
AEMK02000328.1	480375	486936	17.2	23.0	6561	4
18	10134892	10141526	16.5	21.7	6634	4
6	67117339	67120107	15.8	26.2	2768	3
6	93836060	93840355	15.7	27.5	4295	3
7	96887245	96895588	11.3	18.8	8343	3
7	94298196	94303762	10.3	16.6	5566	3

Supplementary material for Study III: "Identification of circular RNAs in porcine sperm and their relation to sperm motility"



Paper III. Figure S1. Figure displaying the validation of the amplified set of circRNAs by agarose-gel electrophoresis. A. Amplification of the 2 randomly selected circRNAs. Two different primer sets for ssc_circ_1141 from *PTGES3* were tested (primer pair a in lane 2 and b in lanes 3 and 4) and ssc_circ_0670 from *BAZ2B* (lane 5). **B.** Validation of the circRNAs correlated to sperm motility parameters: ssc_circ_1321 from *PAPOLA* (*ENSSSCG00000002505)*, ssc_circ_1458 from *LRBA*, ssc_circ_0437 from *ULK4*, two different primer sets for ssc_circ_1061 from *ZNHIT6* primer pair a and b, ssc_circ_1132 from *LIN7A*, two different primer sets for ssc_circ_1101 from *KHDRBS3*, primer pair a (which resulted in amplification of two splicing forms and was excluded) and b, and ssc_circ_0118 from *PDE10A* that resulted in amplification of two splicing forms (and excluded from further analysis).



Paper III. Figure S2. Figure showing the Sanger sequencing based validation of the set of circRNAs. A. ssc_circ_1141 from *PTGES3.* **B**. ssc_circ_0670 from *BAZ2B.* **C.** ssc_circ_0437 from *ULK4.* **D.** ssc_circ_1061 from *ZNHIT6.* **E.** ssc_circ_1101 from *KHDRBS3*.

Paper III: Table S1. List of the 1,598 circRNAs identified in sperm with their genomic coordinates, mean abundance (in CPM) and Standard Deviation (SD) in the 40 samples, and the host gene of the exonic circRNAs.

See table at:

https://drive.google.com/open?id=1AVVmrndQbvu56AHK17oqrZxqeCo4ZB3O

Paper III: Table S2. Gene Ontology analysis and FDR value of the circRNA host genes.

See table at:

https://drive.google.com/open?id=1RYb_ockhdojjGDVPMzA5H54fXPEWEJE1

Paper III: Table S3. Correlation between circRNA abundance and sperm motility parameters. The table includes information on the genomic coordinates of the circRNAs, p-values of the correlation with sperm motility parameters, host gene, whether it was tested for RT-qPCR validation, and the article reference for these host genes that have previously been associated to sperm biology or male fertility. MT: total percentage of motile cells; VCL: curvilinear velocity; VSL: straight line velocity; VAP: velocity of the sperm cells; ns: not significant.

See table at:

https://drive.google.com/open?id=1SbdY8lM2Vt0rpT-UfrfoCZTVdTkAQnUb

Paper III: Table S4. Concordance on the list of circRNAs present in 15 porcine tissues.

Liang GM, Yang YL, Niu GL, Tang ZL, Li K. Genome-wide profiling of Sus scrofa circular RNAs across nine organs and three developmental stages. DNA Res 2017; 24:523-535. doi: 10.1093/dnares/dsx022.

Venø MT, Hansen TB, Venø ST, Clausen BH, Grebing M, Finsen B, Holm IE, Kjems J. Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. Genome Biol 2015; 16:245. doi: 10.1186/s13059-015-0801-3.

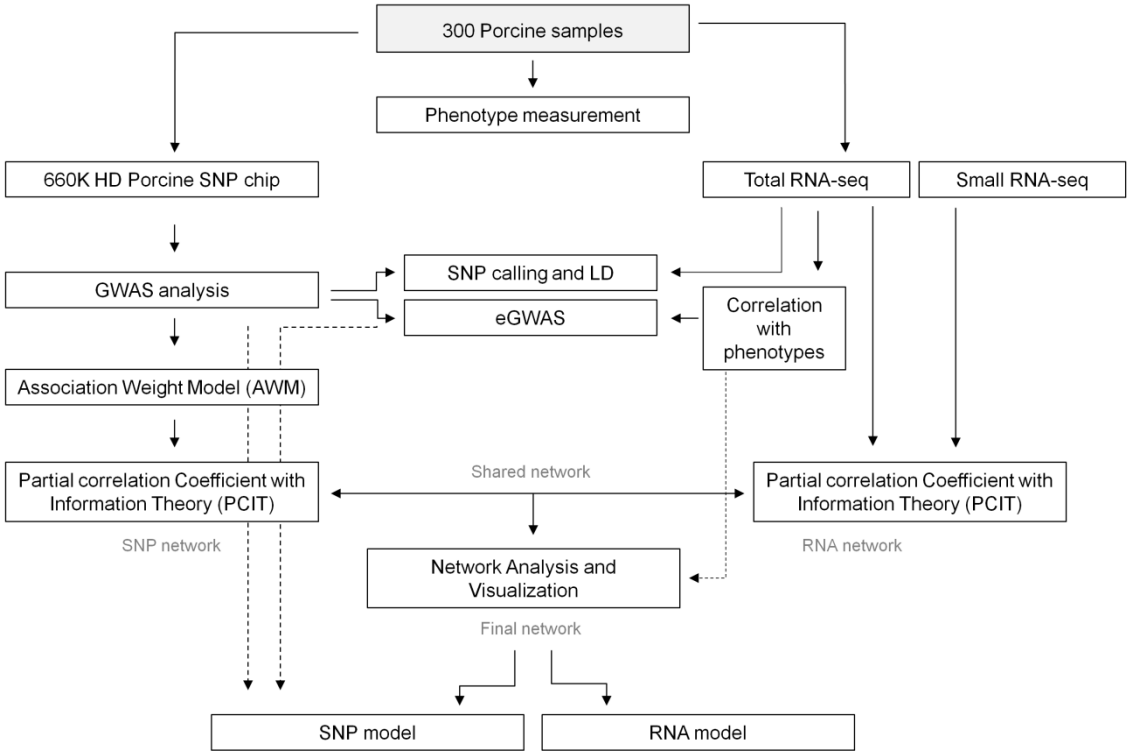
Number of circRNAs	456	820	1,061	2,163	549	241	683	469	353	494	2,685	652	539	532
Tissue	**Basal_ ganglia**	**Brain_ stem**	**Cerebellum**	**Cortex**	**Hippocampus**	**Spleen**	**Lung**	**Kidney**	**Liver**	**Fat**	**Testes**	**Ovarium**	**Heart**	**Muscle**
Basal_ganglia	100%	30%	26%	16%	36%	21%	11%	17%	15%	13%	6%	16%	15%	11%
Brain_stem	54%	100%	37%	26%	50%	23%	17%	20%	19%	18%	9%	22%	19%	16%
Cerebellum	60%	48%	100%	33%	57%	27%	20%	25%	24%	20%	13%	25%	25%	20%
Cortex	77%	70%	67%	100%	73%	35%	27%	31%	30%	26%	19%	34%	30%	27%
Hippocampus	43%	33%	30%	19%	100%	18%	14%	19%	16%	16%	8%	20%	17%	14%
Spleen	11%	7%	6%	4%	8%	100%	14%	16%	18%	15%	4%	13%	12%	11%
Lung	17%	14%	13%	8%	18%	41%	100%	33%	32%	28%	11%	28%	23%	21%
Kidney	18%	12%	11%	7%	16%	31%	22%	100%	25%	21%	8%	22%	18%	18%
Liver	12%	8%	8%	5%	10%	26%	17%	19%	100%	18%	6%	15%	15%	15%
Fat	14%	11%	10%	6%	15%	31%	20%	22%	25%	100%	7%	17%	18%	19%
Testis	38%	31%	32%	23%	37%	45%	42%	44%	44%	41%	100%	43%	34%	33%
Ovarium	23%	17%	16%	10%	23%	36%	26%	30%	28%	23%	11%	100%	22%	19%
Heart	18%	13%	13%	8%	17%	28%	18%	21%	23%	20%	7%	18%	100%	21%
Muscle	13%	10%	10%	7%	14%	23%	17%	20%	23%	20%	7%	15%	21%	100%
Sperm	9%	10%	8%	8%	9%	9%	6%	6%	9%	8%	7%	9%	6%	8%

Paper III: Table S5. List of primers designed and used for the RT-qPCR to assess the abundance of target circRNAs and reference genes.

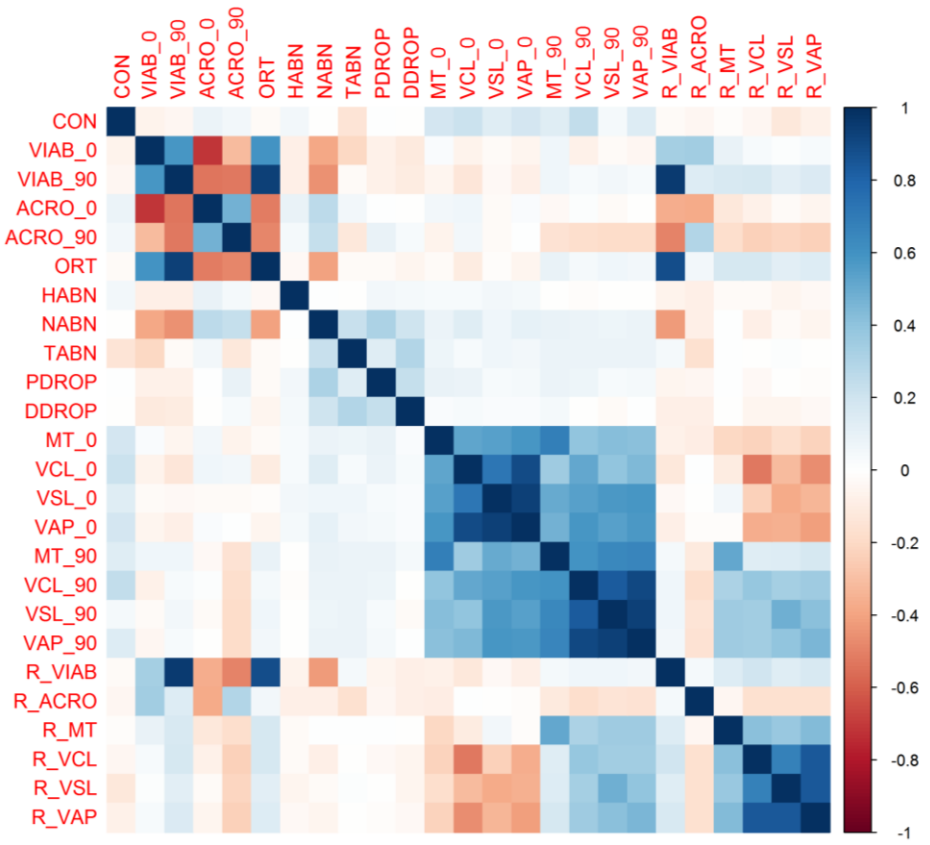
circRNA name	circRNA coordinates	Primer name	Sequence (5' to 3')
ssc_circ_0118	1:3302131..3302802	PDE10A_Fw	TTTCCTTGGCGAGTGCAATAA
		PDE10A_Rv	GGTTGTCTCCTCCTGTGTCCA
ssc_circ_1321	7:117733118..117735703	ENSSSCG00000002505_Fw	AACTCAATCAGAAACCATTCAGACAG
		ENSSSCG00000002505_Rv	TCCACTCAAAGCAAGACAGTTAGC
ssc_circ_1458	8:78370923..78381727	LRBA_Fw	CTTCAGGAAAATGGACCCCA
		LRBA_Rv	GGGTTACGCACAAATCGTCG
ssc_circ_1132	5:100579291..100627916	LIN7A_Fw	GGTGGCTGAAAGACACGGAG
		LIN7A_Rv	TTTCCAGTAATTCAATTGCTCTGG
ssc_circ_1061	4:130322062..130325218	ZNHIT6_Fw_a	TGATGAAGGTTGAACATATGCAGC
		ZNHIT6_Rv_a	GCTTGACTCTGAGGAAACTGCAG
ssc_circ_1061	4:130322062..130325218	ZNHIT6_Fw_b	CAGCAAAATTCAGTGAGAAAACGA
		ZNHIT6_Rv_b	TTGACTCTGAGGAAACTGCAGC
ssc_circ_1101	4:6367950..6392035	KHDRBS3_Fw_a	GCCCTGGAGGAAATCAAGAAGT
		KHDRBS3_Rv_a	ACTTTCTGTCCCAGCTTCATGTTC
ssc_circ_1101	4:6367950..6392035	KHDRBS3_Fw_b	CAAGAAGTTTCTCATCCCCAAATAG
		KHDRBS3_Rv_b	ACTTTCTGTCCCAGCTTCATGTTC
ssc_circ_0437	13:25706236..25715600	ULK4_Fw	TTTGCTATTTGTGTGTGGTGGC
		ULK4_Rv	TGCAGTGCTAGTTCTTGGAGTAGG
ssc_circ_0839	2:141254413..141254577	PAIP2_Fw	ATGTGGATGGAAAATGAAGAGGA
		PAIP2_Rv	GAAACGAATCCAACTAGGAACCA
ssc_circ_1141	5:22008783..22009750	PTGES3a_Fw	CGGTTAACAAAAGAAAGGGCG
		PTGES3a_Rv	TCGTACCACTTTGCAGAAGCAG
ssc_circ_1141	5:22008783..22009750	PTGES3b_Fw	AGTCATGGCCACGGTTAACAA
		PTGES3b_Rv	TCGTACCACTTTGCAGAAGCAG
ssc_circ_0670	15:66291709..66293017	BAZ2B_Fw	ATTTTTGTAGAAGGCATGGAGAGTG
		BAZ2B_Rv	TCAATCTGTTTTTCTAGCTCTTCAACAC
		ISYNA1_Fw*	CATCCGTGACTTCCGGTCC
		ISYNA1_Rv	CGCAGAAGCGCTCTGTGTT
		GPR137_Fw*	CGCCTTCGATTACGACTGGTA
		GPR137_Rv	CAGGTAGCCTTTGTTCCCCA

* Reference gene

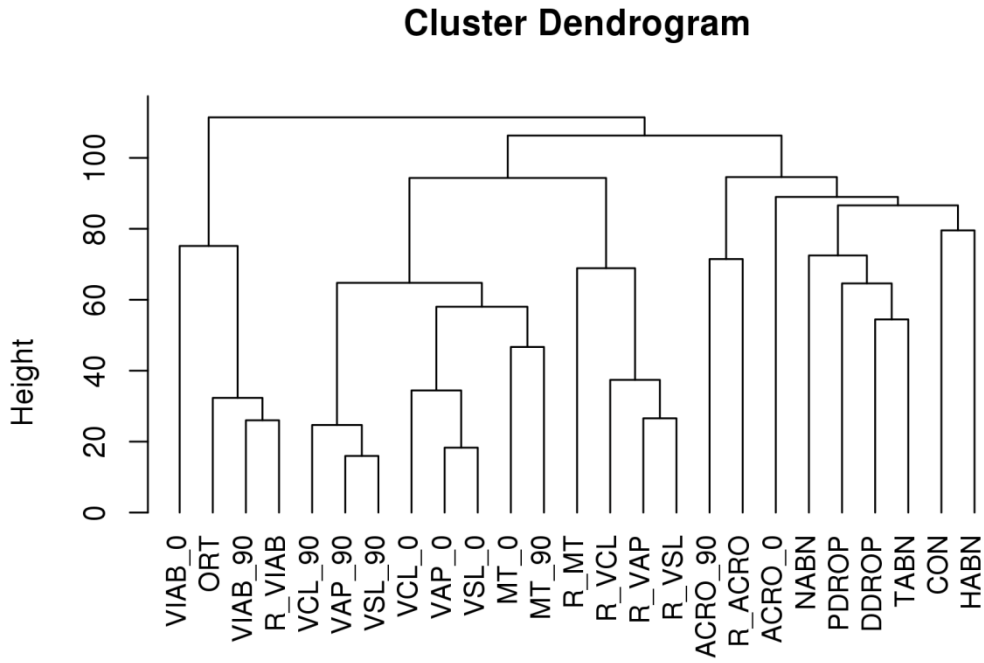
Supplementary material for Study IV: "An integrative systems biology approach to identify the molecular basis of sperm quality in swine"



Paper IV. Figure S1. Summary outline of the different steps of the analysis. Framework of the dataset, analyses and methodologies included in the study.



Paper IV. Figure S2. Correlation across boar sperm quality traits. Heatmap plot of the correlations among the 25 sperm characters from 300 boars.



Paper IV. Figure S3. Cluster dendogram. Dendogram of the standardized SNP effects across the 25 sperm characters.

Paper IV: Table S1. Effect of external factors in the sperm quality traits. Effect of farm, age and season per year across the sperm quality related phenotypes. *=p-value < 0.05; **=p-value < 0.001; ***=p-value < 0.0001; ns=Not Significant.

Trait	Farm	Age	Season per year
Concentration	***	ns	***
Viability 5 min	ns	ns	***
Viability 90 min	ns	ns	***
Osmotic Resistance Test	ns	ns	***
Head abnormalities	ns	ns	*
Neck abnormalities	ns	ns	***
Tail abnormalities	*	ns	***
Proximal droplets	ns	ns	.
Distal droplets	***	ns	***
Motility 5 min	ns	**	*
Average Path Velocity 5 min	ns	**	**
Curvilinear Velocity 5 min	ns	.	***
Straight Line Velocity 5 min	.	**	**
Motility 90 min	.	***	*
Average Path Velocity 90 min	ns	**	***
Curvilinear Velocity 90 min	ns	***	***
Straight Line Velocity 90 min	ns	***	***
Abnormal Acrosomes 90 min	ns	ns	***
Abnormal Acrosomes 5 min	ns	ns	***

Paper IV: Table S2. RNA-seq extraction and mapping statistics. Average and Standard Deviation (SD) for the 40 samples processed, including: amount of RNA extracted and several bioinformatics statistics for total RNA-seq (40 samples) and short RNA-seq (34 samples).

	Average	SD
Starting amount of RNA (fg per cell)	2.16	0.77
Total RNA-seq reads	40,707,821	20,085,879
Reads passing quality control	40,096,875	20,169,826
Proportion of reads passing quality control	0.982	0.013
Proportion of mapped reads in the Sscrofa11.1 genome	0.827	0.044
Unmapped reads	4,815,145	2,068,389
sncRNA-seq reads	7,321,663	2,586,723
Reads passing quality control	7,260,967	2,569,742
Proportion of reads passing quality control	0.992	0.005
Proportion of mapped reads	0.815	0.024
Remaining unmapped reads	1,343,003	62,629

Paper IV: Table S3. List of identified porcine sperm miRNAs. Average and Standard Deviation (SD) for the 34 samples processed. miRNA abundances are expressed in counts per million (CPM).

See table at:

https://drive.google.com/open?id=1xzemun1XF-Ne3kE6G-SylcrTG4SKs-ir

Paper IV: Table S4. Correlations between gene abundances and phenotypes. P-values are given when (P-value < 0.05). The parenthesis include the correlation value. ns=Not Significant.

See table at:

https://drive.google.com/open?id=137KuOJS5wDo8_3Zefrq2TmwXlXukd8zH

Paper IV: Table S5. Sperm eGWAS eSNP/transcript associations. There were 39 eSNPs/transcript associations with FDR < 0.05. The eSNPs were also identified in the GWAS (FDR < 0.05) and the RNA levels of targeted gene, was significantly correlated with the GWAS phenotype. Chr: chromosome. FDR = False Discovery Rate; ACRO_0 = Abnormal Acrosomes 5 min; HABN = Head abnormalities.

See table at:

https://drive.google.com/open?id=14RLNHOGfvlA1fMy1k2MtT9NtzMaUDSQH

Paper IV: Table S6. Gene Ontology analysis of the genes included in the final network. GO biological process terms with significant Bonferroni corrected p-values (FDR < 0.05) and their associated genes.

See table at:

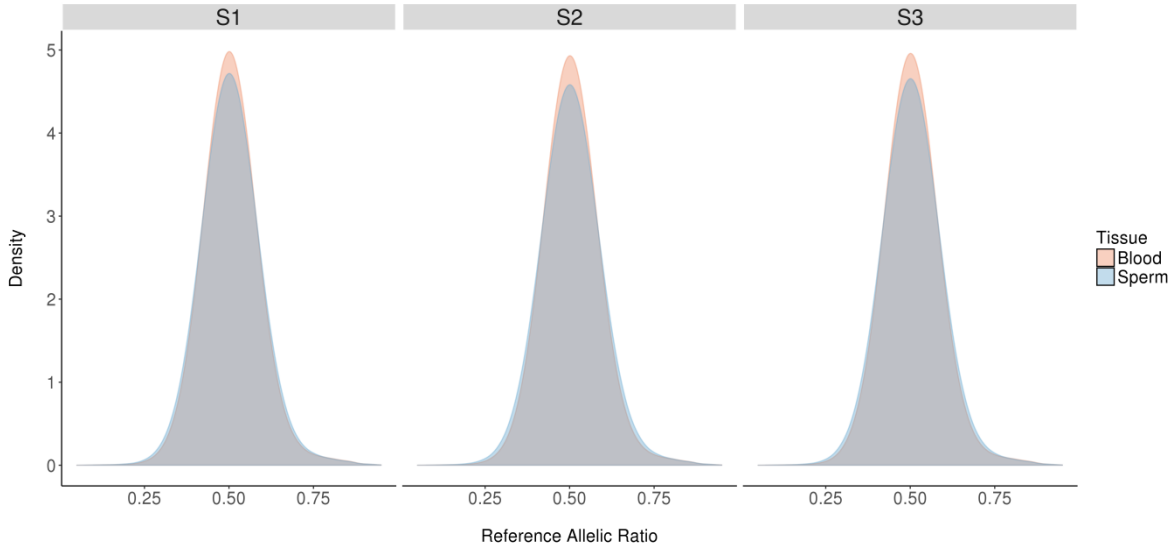
https://drive.google.com/open?id=1GkU3PSIUEX38jSaTjwIU_n8K45NXqq33

Paper IV: Table S7. Parameter estimates for the significant models. For each of the phenotypes, the model outputs the estimated values for the 10 genes obtained from the GRM regression analysis. The lower the value of Pr > |t|, the higher the involvement of the gene abundance on the total phenotypic variance.

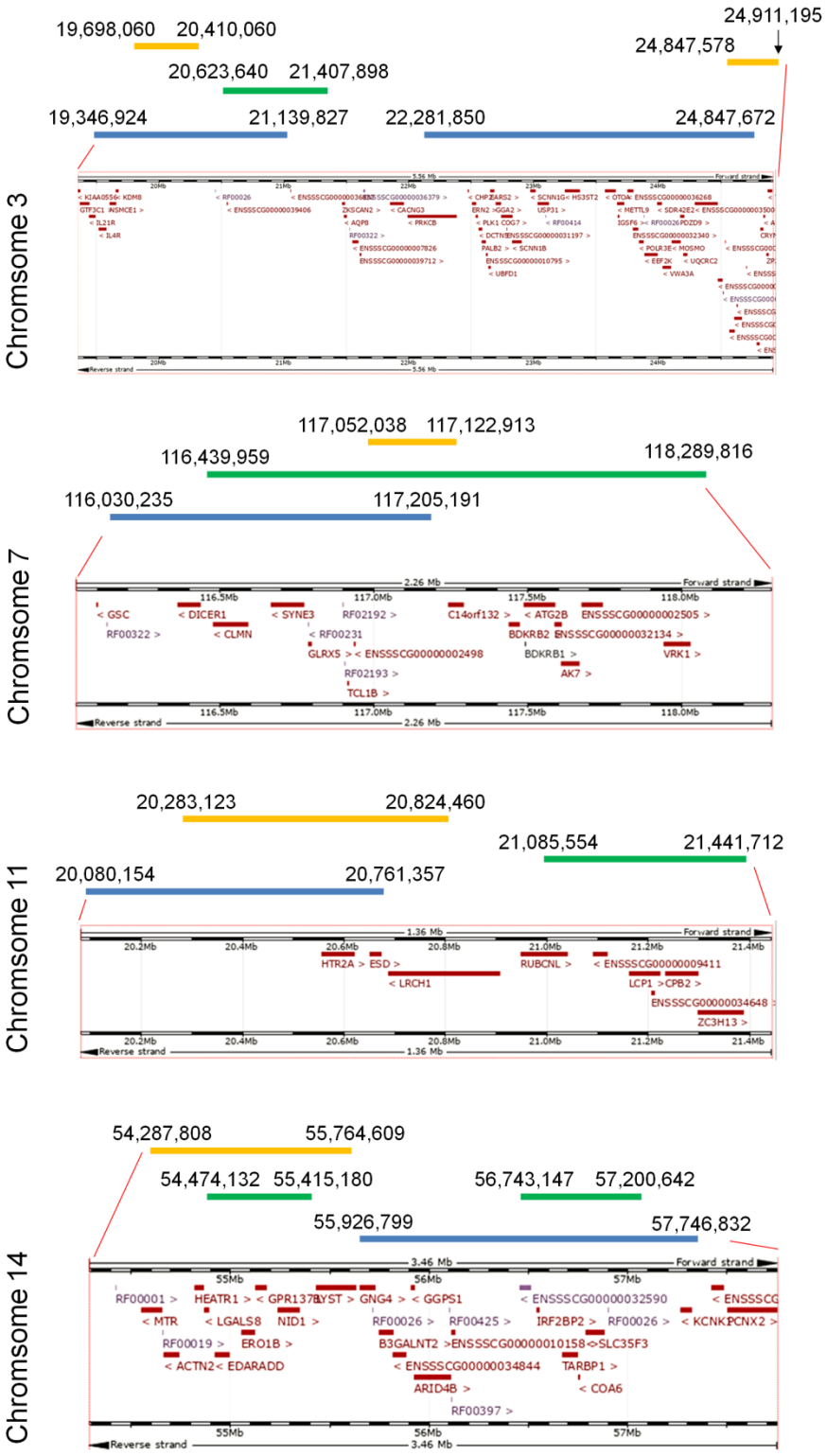
See tables at:

https://drive.google.com/open?id=1zmyHxsTLb5Enaw24JzO94o5IU3rceR4_

Supplementary material for Study V: "Whole genome sequencing of porcine sperm identifies Allelic Ratio Distortion in genes related to spermatogenesis"



Paper V: Figure S1. Density plot of the reference allele ratio distribution in heterozygous SNPs in blood and sperm for each boar.



Paper V: Figure S2. Overview of the genomic overlap in the 4 ARD regions shared in the 3 pigs. Each ARD region is formed by at least 3 ARD SNPs with consecutive SNP distance below 1 Mbp. We identified 4 of such regions in chromosome 3, 7, 11 and 14. ARD regions of sample S1 are depicted in blue, sample S2 in yellow and S3 in green.

Paper V: Table S1. Sequencing and mapping statistics for the 3 boars.

	S1		S2		S3		Average
	Sperm	Blood	Sperm	Blood	Sperm	Blood	
Raw reads (PE)	421,384,207	491,644,271	432,115,367	497,065,481	413,531,019	490,920,039	457,776,731
% Mapped reads	99.4	99.5	99.5	99.5	99.5	99.4	99.5
% of duplicate reads	11.4	15.5	13.4	17.5	11.4	16.2	14.2
Genome coverage (X)	47.1	55.0	48.3	55.6	46.3	54.9	51.2
# of SNPs	10,031,590		10,092,320		9,910,338		10,011,416
# of SNPs passed filter	6,264,314		6,274,958		6,247,364		6,262,212
# of SNPs hetozygous in blood	2,829,445		2,917,048		2,845,197		2,863,897

Paper V: Table S2. List of variants and regions in ARD (this study) and in TRD (Casellas et al., 2014).

See table at:

https://drive.google.com/open?id=1zTNZuB6vZWuFvHXUYDjUQeLhURdLYNA9

Paper V: Table S3. SNPs in ARD in the 3 boars. For each SNP we provide information of the rsID, host gene, ARD P-value, reference allele ratio and SNP effect. Chr: chromosome.

See table at:

https://drive.google.com/open?id=1xEog87y2MI53jq-qjjCYzEhjJawlUsXd