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

Marta Godia
Centre for Research in Agricultural Genomics CRAG (CSIC-IRTA-UAB-UB)
Antonio Reverter
CSIRO Queensland Bioscience Precint Camrody
Rayner Gonzalez-Prendes
Centre for Research in Agricultural Genomics CRAG (CSIC-IRTA-UAB-UB)
Yuliaxis Ramayo-Caldas
Institut de Recerca i Tecnologia Agroalimentaries
Anna Castello
Centre for Research in Agricultural Genomics CRAG (CSIC-IRTA-UAB-UB)
Joan Enric Rodriguez-Gil
Universitat Autonoma de Barcelona
Armand Sanchez
Universitat Autonoma de Barcelona
Alex Clop (alex.clop@cragenomica.es)
CRAG https://orcid.org/0000-0001-9238-2728

Research

Keywords: sperm quality, systems biology, sperm RNA, GWAS, swine, small RNA-seq, miRNA

Posted Date: March 30th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-19366/v1

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

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

Affiliations:

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

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

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

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

*Corresponding author: alex.clop@cragenomica.es

Abstract

Background:

Genetic pressure in animal breeding is sparking the interest to select for elite boars with higher sperm quality to maximize ejaculate doses and fertility rates. However, the molecular basis of sperm quality remains largely unexplored. In this study, we sought to identify candidate genes, pathways and DNA variants associated to sperm quality in swine by analyzing 25 sperm-related phenotypes using a systems biology approach that integrates GWAS and RNA-seq.

Results:

By GWAS, we identified 12 QTL regions associated to the percentage of head and neck abnormalities, abnormal acrosomes and motile spermatozoa. Candidate genes included *CHD2, KATNAL2, SLC14A2* or *ABCA1.* By RNA-seq, we detected 6,128 significant correlations between sperm traits and gene RNA abundances. We built a gene interaction network with the GWAS and the RNA-seq data. To build a robust gene interaction network, only the pair-wise interactions present in both the genetic co-association and the RNA co-abundance network were kept. Moreover, we also included to the Final Network both the genes which RNA abundances correlated with more than 4 semen traits as well as the miRNAs interacting with the genes on the network. The Final Network was enriched for genes involved in gamete generation and development, meiotic cell cycle, DNA repair or embryo implantation. We finally designed a panel of 73 SNPs provided from the GWAS, eGWAS and the Final Network, that explains between 5 to 36% of the phenotypic variance of the sperm traits.

Conclusions:

By means of a systems biology approach, we identified potential key genes affecting sperm quality. Furthermore, we propose a SNP panel that might explain a substantial part of the genetic variance for semen quality in swine and may thus be of interest for the pig breeding sector.

Keywords: sperm quality, systems biology, sperm RNA, GWAS, swine, small RNA-seq, miRNA

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 whereby germ cells proliferate and develop into mature spermatozoa, is controlled by multiple factors. Both DNA polymorphisms and gene expression have been linked to sperm quality and/or fertility in several mammalian species including cattle (Boe-Hansen et al., 2018) and 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 (Diniz et al., 2014; Marques et al., 2018; 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. To date, 5 genome-wide association studies (GWAS) have been carried. 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 population. Marques et al. (Marques et al., 2018) detected 16 and 6 QTL regions in Large White and Landrace, respectively, associated with sperm motility, number of cells per ejaculate and morphological abnormalities. More recently, several QTL regions have been identified in a Duroc population associated to the number of sperm cells, sperm motility, sperm progressive motility, total morphological abnormalities, coiled tail, bent tail, proximal droplets, distal droplets and distal midpiece reflex (Gao et al., 2019; Zhao et al., 2020).

The presence of RNA molecules in the boar sperm is well documented (Gòdia et al., 2018b; Gòdia et al., 2019a), but their relation to sperm quality has been just tinely explored. Porcine sperm RNAs are highly fragmented and their gene abundances are mostly associated to prior transcriptional 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). Several studies have reported a relation between RNA abundances and semen quality in mammals (Capra et al., 2017; Jodar et al., 2015; Wang et al., 2019). 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). Moreover, our group has also identified a correlation between the abundance of some circRNAs (Gòdia et al., 2019b) and piRNAs (Ablondi et al., 2020) with semen quality parameters in the porcine species.

Resulting from this recent work, it is now apparent that the genetic complexity of sperm quality involves several molecular mechanisms and pathways that are highly

interconnected. For this reason, 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 identify candidate genes, pathways and DNA variants associated to sperm quality in pigs integrating 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-chloroform based method. The ejaculates were purified to remove somatic cells as described previously (Gòdia et al., 2018b) and 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), the percentage of viable

cells (VIAB), the percentage of morphologically abnormal acrosomes (ACRO), osmotic resistance test (ORT), the percentage of morphologically abnormal sperm cells (of the head –HABN-, neck –NABN- and tail –TABN-) and of cells with 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 spermatozoa cells (MT) (with Average Path Velocity -VAP- > 10 µm/s), average Curvilinear Velocity (VCL) (µm/s), average Straight Line Velocity (VSL) (µm/s) and average VAP (µm/s). All phenotypes were assessed after 5 and 90 min of incubation of the samples at 37°C, except for sperm concentration, ORT, sperm abnormalities and cytoplasmatic droplets that were measured only after 5 min of incubation of the samples at 37°C.

Sperm phenotypes were then corrected for the fixed effects of 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. 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 QC call rate below 96%. SNP locations were converted from Sscrofa10.2

to Sscrofa11.1 coordinates using plink v1.9 (Purcell et al., 2007). We then excluded SNPs which (i) had a minor allele frequency below 0.05, (ii) deviated from Hardy-Weinberg equilibrium (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) considering the genomic relatedness matrix (GRM) as a random effect to correct for the population structure with the following model:

$$Y_{ijkl} = \mu + SNP_i + Farm_j + SeasonYear_k + Age_l + e_{ijkl}$$

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 ($Farm_j$), season and year ($SeasonYear_k$), age (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 mapping to sexual chromosomes or to unmapped scaffolds were not taken into account. 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) and included 35 samples from boars analyzed 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 HiSeq2000/2500. 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 (https://www.bioinformatics.babraham.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 (Sscrofa 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 effect of external variables in gene expression was assessed using the following mixed effect model as in Reverter et al. (Reverter et al., 2005):

$$Y_{ijklmn} = \mu + L_i + G_j + GF_{jk} + GYS_{jl} + GA_{jm} + GR_{jn} + e_{ijklmn}$$

where Y_{ijklmn} represents the log2-transformed FPKM value from the i-th library (40 levels), j-th gene (4,120 levels), k-th farm (3 levels), l-th year-season (6 levels), m-th age (3 levels) and n-th assay run (4 levels). Accordingly, Y_{ijklmn} was modeled as a

function of the fixed effect of library (L_i) and the random effects of gene (G_j), gene by farm (GF_{jk}), gene by year-season (GYS_{jl}), gene by age (GA_{jm}) and gene by assay run (GR_{jn}). Random residuals in e_{ijklmn} were assumed to be independent and identically distributed. Variance component estimates and solutions to the model were obtained using VCE6 (Groeneveld, 1994; ftp://ftp.tzv.fal.de/pub/vce6/doc/vce6-manual-3.1-A4.pdf).

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.

The relationship between the 25 phenotypes and each of the log2-stabilized mRNA's and miRNA's abundances were calculated using the Pearson correlation coefficient.

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

Expression GWAS (eGWAS) 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), with the following model:

$$Y_i = \mu + SNP_i + e_i$$

where (Y_i) is the log2-transformed gene abundance modeled as a function of the population mean (μ), fixed effect of each SNP (SNP_i), and a random residual effect (e_i).

eGWAS significant associations (FDR ≤ 0.05) were considered only if: (i) the eGWAS associated SNP was also a significant hit (FDR ≤ 0.05) in the GWAS for sperm quality phenotypes and (ii) the gene's RNA abundance correlated to the same phenotype as the corresponding 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). The AWM was constructed from two matrices that contained row-wise SNPs and column-wise phenotypes. The first matrix included the P-values of the association between each SNP and the phenotype, and the second matrix corresponded to the SNP z-score standardized additive effect. As live cells with intact plasma membrane are essential for fertilization (Berger et al., 1996; Quintero-Moreno et al., 2004), the percentage of viable spermatozoa at 5 min (VIAB_5) was selected as key phenotype

and the associated SNPs (P-value ≤ 0.01) were included in the AWM. In the next step, 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 and represented using the hierarchical cluster analysis based on Euclidean distance with R package "dendextend" (Galili, 2015). Then, significant gene-gene interactions were assessed to build the SNP Network 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 stabilized 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 as 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 or miRNA, (ii) the phenotype with highest correlation for each gene, (iii) whether the gene was annotated as a Transcription Factor (TF) or TF co-factor, and (iv) whether the gene was present in the Shared Network or was only found in the Final Network. TF and TF co-factors were extracted from the AnimalTFDB3.0 database (Hu et al., 2019a).

Development of a RNA model and SNP panel 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 these was 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. The model is:

$$Y_i = intercept_i + GPE_{ij} + e_{ij}$$

where (Y_{ij}) represents the predicted phenotype value from i-th phenotypes (25 levels), j-th genes (10 levels). Y_{ij} was modeled as a function of the intercept value for the phenotype ($intercept_i$), the gene abundance by parameter estimate

304 (GPE_{ij}) and a residual term (e_{ij}). We also developed a genome-wide SNP marker
305 panel to identify the polymorphisms that could better predict the phenotypic variance
306 of sperm-related traits. The panel included the lead SNPs from the GWAS and from
307 the eGWAS hits and the GWAS most significant SNP for each of the genes included
308 in the network that also: (i) correlated with at least 4 phenotypes and (ii) were
309 identified in the Shared Network. The proportion of the variance explained by these
310 polymorphisms was assessed with GCTA v.1.91.5 (Yang et al., 2011).
311

312 **Results**

313 **Phenotype statistics**

314 Three hundred ejaculates were phenotyped for 25 sperm quality traits (Table 1).
315 Phenotype correlations (Additional Figure 2) were consistent with their physiological
316 similarities. In general, SNP-based heritabilities (Table 1) were low to moderate with
317 motility related traits displaying higher values. MT_90 was the most heritable trait (h^2:
318 0.39). On the other side, motility ratios, NABN and VIAB_5 showed nearly null
319 heritability (Table 1). The sperm phenotypes correlated with farm, boar age and
320 Season per Year (Additional file 1) and were thus included as fixed effects in the
321 GWAS model and phenotypes were also corrected for these effects to carry the
322 correlation analysis.

323 [Table 1 appears here]

324 **GWAS analysis**

325 After quality control, 466,592 SNPs and 276 samples remained for the GWAS. A
326 total of 324 SNPs across autosomal chromosomes and unplaced scaffolds displayed
327 genetic associations (FDR ≤ 0.05) with 1 or more sperm quality phenotype (Table 1;

328 Additional file 2). Of these, 255 SNPs mapped in unplaced scaffolds and were not
329 considered for further data analysis (Additional file 2). A total of 19 chromosomal
330 regions tagged by 69 significant SNPs were identified in *Sus scrofa* chromosomes
331 (SSC) 1, 3, 4, 6, 7, 9, 13 and 16. The number of SNPs displaying significant
332 associations (FDR ≤ 0.05) for each trait is summarized in Table 2.

333 Seven sperm quality traits exhibited significant association signals (Figure 1. A – G;
334 Additional file 2), and only one SNP was associated to more than 1 trait (Table 2;
335 Figure 1. D – E; Additional file 2). HABN and NABN presented the largest number of
336 SNP signals with 41 and 18 associated SNPs, respectively (Figure 1. A and C;
337 Additional file 2). Six of the 19 QTLs were represented by only 1 associated SNP
338 and were discarded from further analyses (Table 2; Figure 1). The most significant
339 SNPs (rs318575212 and rs332927981) of the study were associated to ACRO_5
340 (both with FDR = 0.006 and Additive effect = 4.11) (Table 2).

341 [Table 2 appears here]

342 [Figure 1 appears here]

343 **Sperm RNA isolation, RNA-seq and bioinformatics analysis**

344 Isolated RNA from mature spermatozoa was free from somatic cell RNA. Total RNA-
345 seq resulted in an average of 40.7 M reads per sample and 98.2% of the reads
346 passed the quality control filters (Additional file 3). An average of 83% of the reads
347 mapped to the porcine genome and after duplicate removal and RNA abundance
348 filters, we identified 4,120 genes (Additional file 4). The Variance Component
349 Estimate mixed model explained 84% (80% due to the main effect of gene) of the
350 variation in gene abundance. Consequently, RNA abundances were not corrected
351 for external effects. 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 3). We identified 95 miRNAs out of the 306 that are annotated in swine (Additional file 4).

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

Under the hypothesis that some of the GWAS hits may be tagging a causal variant altering protein sequence and function, and to identify additional SNPs with the potential to be better genetic markers than these identified in the GWAS, we sought to identify variants in annotated genes using the RNA-seq data. 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 in the GWAS (Table 2; Additional file 5). 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 as a splice donor variant and to have, thus, a high impact on protein sequence (Additional file 5).

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 with high, 5 with moderate and 5 with low effect on protein sequence). 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; Additional file 5). This interval was the only one that presented a SNP with high effect (novel), a splice donor variant in the *IQCF5* gene, but the SNP was in low LD (LD=0.02) with the GWAS hit

(Additional file 5). The interval 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 SSC1 I3 region associated to HABN harbored 3 expressed SNPs (1 with moderate and 2 with low effect) (Table 3; Additional file 5).

[Table 3 appears here]

Correlation of gene's and miRNA's abundances with sperm quality traits

The 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 (Additional file 6). These genes presented between 1 and 9 significant correlations with the different semen quality traits (Additional file 6). 344 genes were significantly correlated with ≥ 4 traits. For the miRNAs, the abundance of the 95 miRNAs and the studied phenotypes resulted in 306 significant correlations (P-value ≤ 0.05) which involved 87 miRNAs and 17 semen traits (Additional file 7). The miRNAs presented between 1 and 9 significant correlations with the semen quality traits studied (Additional file 7).

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 the normalized RNA abundances. We then only focused on the associations between GWAS SNP hits (with FDR ≤ 0.05) and transcripts which abundances correlated with the same phenotype. We identified 45 SNPs (FDR ≤ 0.05) located in 3 genomic regions related to ACRO_5 and HABN (Table 4). Six SNPs had unknown positions in the

genome after liftover from Sscrofa10.2 to Sscrofa11.1. The remaining eGWAS hits were in SSC4, 6 and 13 (Table 4; Additional file 8). All the SNPs had a *trans* effect, related to genes located in different chromosomes. The eQTL identified in SSC4, was related to ACRO_5 and was associated to 3 genes, *NCLN, ASCC1* and *AATF*. Also involving ACRO_5, 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 appears here]

Gene network analysis

After SNP selection, 2,648 of the 466,592 SNPs were retained to build the AWM. Trait hierarchical cluster distributions were in agreement with the biological similarities and phenotypic correlations (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 phenotypic variance of the key phenotype (VIAB_5). 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 (Table 5).

425 **Table 5.** Number of nodes (genes) and edges (interactions).

Network	Nodes	Edges	Observations
SNP Network	2,648	2,984,616	
RNA Network	4,120	1,173,995	
Shared Network	613	16,591	
Final Network			
miRNAs	94	1,564	
Protein coding genes	1,313	81,733	
	1,135		Correlated with > 1 phenotype
	68		TF
	89		TF co-factor

426 TF = Transcription Factor

427

428 For the RNA network analysis, the RNA levels of the 4,120 detected genes were
429 used to identify potential connections with PCIT (Reverter and Chan, 2008). The
430 RNA network included 4,120 nodes (all the genes) connected by 1,173,995 edges
431 (Table 5). PCIT also built 4,539 significant interactions between 95 miRNAs and 630
432 genes.

433 To obtain the Shared Network, common SNP and RNA network edges were
434 extracted, thus, focusing only in the shared set of interacting genes from both
435 approaches. This comparison resulted in 613 nodes connected by 16,591 edges
436 (Table 5). The Final Network included a set of 700 additional genes (as they
437 correlated with > 3 phenotypes) and their interactions. Moreover, the Final Network
438 also involved 1,564 edges connecting 202 genes and 94 miRNAs (Table 5). Of the
439 1,313 genes included in the Final Network, the abundance of 1,135 correlated with at

least 1 phenotype, 68 have been reported as TFs and 89 as TFcos (Figure 2.A; Table 5). Nearly a 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 9).

Development of a RNA model and a SNP panel

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 6). The most significant model was for PDROP, which could predict the phenotype with an efficiency of 68% (Table 6). The estimated parameters of the significant models can be found in Additional file 10.

464 **Table 6.** R2 and phenotypic variance for each trait from the RNA model and SNP
465 panel.

Acronym	RNA model		SNP panel
	R_2	P-value	Phenotipic variance explained (SE)
CON	0.17	0.82	0.05 (0.05)
VIAB_5	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	<0.0001	0.17 (0.07)
DDROP	0.42	0.07	0.06 (0.05)
MT_5	0.46	0.03	0.31 (0.07)
MT_90	0.34	0.22	0.30 (0.07)
VAP_5	0.58	0.002	0.34 (0.07)
VAP_90	0.55	0.005	0.34 (0.07)
VCL_5	0.61	0.001	0.33 (0.07)
VCL_90	0.55	0.01	0.34 (0.07)
VSL_5	0.36	0.16	0.31 (0.07)
VSL_90	0.61	0.001	0.33 (0.07)
ACRO_5	0.5	0.02	0.21 (0.06)
ACRO_90	0.21	0.68	0.23 (0.07)
R_MT	0.3	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	0.23 (0.07)
R_ACRO	0.57	0.003	0.19 (0.07)

466 Acronym descriptions can be found in Table 1. SE: Standard Error

The SNP-based panel was built with 73 polymorphisms (18 lead SNPs from GWAS hits, 2 lead SNPs from the eGWAS hits, 53 SNPs from the Shared Network and correlated ≥ 4 phenotypes) (Additional file 11). These polymorphisms could explain between 5 to 36% of the phenotypic variance across the 25 traits (Table 6). A moderate proportion (>20%) of the phenotypic variance could be explained for 18 of the 25 traits. The best predictions were for sperm abnormalities (NABN, HABN, TABN) and sperm motility related traits (e.g. MT_5, VAP_90 and VCL_90) (Table 6).

Discussion

GWAS analysis

Investigating the genomic regions and molecular processes controlling sperm quality has become a focus of interest in human and in livestock including swine, in this case for its relevance on the sustainability of pig breeding and production (Diniz et al., 2014; Gao et al., 2019; Marques et al., 2018; Zhao et al., 2020; Zhao et al., 2016). In fact, our results as well as data obtained by 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.

The GWAS revealed 12 QTL regions represented by 2 or more significant SNPs and several positional candidate genes for HABN, NABN, ACRO_5 and MT_5 (Table 2). The highest signals were on SSC4 for ACRO_5 (~2.41-2.42 Mbp) (Table 2; Additional file 2), ~69 kb upstream of the Solute Carrier Family 45 Member 4 (*SLC45A4*) gene. *SLC45A4* encodes a proton-coupled sugar transporter implicated

491 in the nutrition of spermatozoa during their maturation in epididymis (Vitavska and
492 Wieczorek, 2017) where acrosome assembly continues its posttesticular sperm
493 maturation (Olson et al., 2003). The Solute Carrier Family 35 Member B3 (*SLC35B3)*
494 was selected as a potential candidate for the MT_5 QTL on SSC7 (Table 2;
495 Additional file 2). *SLC35B3* maps 0.6 Mbp away from this QTL. Although a role in
496 sperm has not been reported thus far, the SLC35 gene family has been postulated to
497 play a role as nucleotide sugar transporter (Song, 2013) and we propose that it may
498 also be relevant for the nutritional support of spermatozoa.

499 We detected several significant regions for HABN (Table 2; Additional file 2). The
500 QTL on SSC1 I2 (~94.9-98.8 Mbp) included interesting candidate genes such as the
501 Katanin Catalytic Subunit A1 Like 2 (*KATNAL2*). Dunleavy et al. (Dunleavy et al.,
502 2017) reported that *Katnal2* is a critical regulator of male germ cell development
503 affecting sperm head shaping, acrosome attachment and sperm tail growth. Other
504 candidate genes in that region were the solute carrier *SLC14A2,* encoding the urea
505 transporter A, suggested to participate in sperm head formation by reducing its
506 volume though excreting urea (Li et al., 2012), or the SMAD Family Member 2
507 (*SMAD2*) involved in spermatogonial differentiation (Wu et al., 2017). On SSC13 I1,
508 we identified two candidate genes: the Testis and Ovary-specific PAZ domain gene 1
509 (*TOPAZ1*) and the IQ Motif Containing F1 (*IQCF1*). Luangpraseuth-Prosper et al.
510 (Luangpraseuth-Prosper et al., 2015) demonstrated that *Topaz1* knockout mice
511 presented meiotic arrest and caused male infertility. As for *IQCF1,* Fang et al. (Fang
512 et al., 2015) reported that this gene localizes in the acrosome and that it is involved
513 in sperm capacitation in mice. *Iqcf1*$^{-/-}$ mice were significantly less fertile than wild
514 type mice (Fang et al., 2015). The QTL region on SSC13 I2 included the candidate
515 Protein Kinase C Delta (*PRKCD*) gene. *PRKCD* has been involved in

516 spermatogenesis and embryonic development (Suh et al., 2003) and was highlighted
517 in a GWAS for semen volume in Holstein-Friesian bulls (Hering et al., 2014).

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

532 Four of our GWAS hits map near previously reported QTLs for semen quality traits.
533 This is the case for the SSC1 I6 QTL, associated to NABN, which mapped 335 kbp
534 downstream from a QTL associated to sperm abnormalities and motility in boars
535 (Marques et al., 2018). The QTL SSC3 I2, associated to NABN lies 350 kpb
536 upstream from a PDROP QTL (Zhao et al., 2020). The SSC4 I1 QTL, associated to
537 ACRO_5, resides 655 kpb upstream from a QTL for the Distal Midpiece Reflex (Zhao
538 et al., 2020) and the SSC7 I1 QTL, associated to MT_5 maps 123 kpb upstream
539 from a PDROP QTL (Zhao et al., 2020). The discrepancies across studies could
540 arise due to different technical (e.g. sample size, SNP arrays, QTL or phenotyping

accuracy), environmental (e.g. temperature, animal husbandry or sperm processing) or biological factors (e.g. genetic heterogeneity).

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; Additional file 5). 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; Additional file 5). Although *ULK4* has not been associated to sperm defects, Liu et al. (Liu et al., 2016) showed that this gene 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 *IQCF5* gene (Additional file 5). 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

565 filaments in an ATP-depending manner and have been implicated in various aspects
566 of spermatogenesis (Hu et al., 2019b). In sperm, actin filaments are located in the
567 acrosomal region (Breitbart et al., 2005). Interestingly, the previously discussed
568 GWAS positional and physiological candidate genes *CHD2* and *KATNAL2*, also
569 presented genetic variants in LD with the lead SNPs at SSC7 I2 (low effects:
570 rs330912302 LD = 0.4 and rs339719658 LD = 0.37) and SSC1 I3 (low effects:
571 rs700749617 LD = 0.01, rs710447566 LD = 0.07 or moderate effect: rs690151450
572 LD = 6.9 x 10^{-3}), respectively (Additional file 5).

573 **Correlation between genes and miRNAs with semen traits**

574 For mRNA transcripts, the strongest correlation was for *TTC28* and HABN (-0.71)
575 (Additional file 6). *TTC28* is required for the condensation of spindle microtubules
576 during mitosis and meiosis (Izumiyama et al., 2012). Other genes of interest included
577 *ABCA3*, which RNA levels correlated with 9 phenotypes (Additional file 6). This gene
578 is an ABC transporter that plays a role in flipin-cholesterol complexes as a
579 mechanism to remove cholesterol from the sperm membrane (Mengerink and
580 Vacquier, 2002). Although the molecular basis induced by cholesterol efflux from
581 sperm is not well understood, it has been reported to be required for sperm
582 capacitation (Visconti et al., 2002). Another example is *EFHC1,* which RNA levels
583 correlated with 6 phenotypes (Additional file 6). E*fhc1*$^{-/-}$ knockout mice presented
584 reduced flagellar beating frequency (Suzuki et al., 2009).

585 Several miRNAs of interest including miR-23a, miR-27a and miR-122 correlated with
586 7, 8 and 8 semen quality traits, respectively (Additional file 7). miR-23a, has been
587 found to be dysregulated in men's subfertility (Abu-Halima et al., 2019). miR-27a
588 abundance of spermatozoa has been related to lower progressive motility and
589 normal morphology (Zhou et al., 2017). miR-122 expression was associated with

abnormal sperm development (Liu et al., 2013) and dysregulated in subfertile men (Abu-Halima et al., 2013).

eGWAS

We also performed a within-trait eGWAS linking for each phenotype, GWAS lead SNPs with genes which RNA abundance correlated with the same trait. These GWAS regions could be tagging causal variants with regulatory functions on gene expression. We identified 3 eQTLs all with a *trans*-effect (Table 4; Additional file 8). 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_5 (Table 4). *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*) (Table 4; Additional file 8). 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

614 role in sperm remains to be resolved, HARS protein has been found overexpressed
615 in sperm of low-fertility bulls (Aslam et al., 2019) and we do not rule out a potential
616 involvement of this gene in spermatogenesis. *trans*-eQTL hotspots (these trans-
617 eQTLS involving several genes) are of particular interest as their SNPs could harbor
618 important regulatory roles and variations influencing gene expression and thus are
619 more likely to contribute to the phenotype.

620 **Gene network analysis**

621 Despite the considerable number of candidate genes identified in our GWAS, many
622 genes might have been missed by this traditional single-trait approach due to the
623 lack of an acceptable significant association (FDR > 0.05). After all, sperm quality is
624 a complex phenotype influenced by many factors, such as genetics, husbandry,
625 environment, or testicular pathologies that contribute to an intricate network of genes
626 and molecular processes. Moreover, many of the SNPs included in the GWAS may
627 not be at sufficiently high allelic frequency or be in strong LD with the causal variants
628 in the studied populations. An alternative strategy to exploit GWAS information is to
629 perform an AWM analysis that extracts SNPs that while having strong yet below the
630 significance threshold of genetic association, are also associated to a certain number
631 of traits (Fortes et al., 2010). The association of 1 SNP to more than 1 trait provides
632 additional robustness to the potential relevance of that SNP to semen quality in our
633 case. This, followed by a PCIT analysis to study gene-gene interactions can provide
634 information on the relevant genes and pathways for certain phenotypes and then
635 search for SNPs in or affecting them. Obviously, transcriptomics data can contribute
636 additional valuable information in the description of these genes and pathways. The
637 integration of both sources of information can also be used to improve the accuracy
638 of genomic predictions (Ramayo-Caldas et al., 2019). For this reason, we have

639 addressed the genetics behind boar's sperm quality through an integrative systems
640 biology approach. The genetic co-association and RNA co-abundance interactions
641 revealed a number of appealing features such as new candidate genes, TFs, TF-cos
642 and miRNAs that belong to biological processes and relevant functions related to
643 sperm.

644 The TF with the highest number of predicted interactions (129) was encoded by the
645 Calcium Responsive Transcription Factor (*CARF*) gene, which RNA abundance was
646 in turn, correlated with 9 phenotypes (Figure 2.B; Additional file 6). CARF acts as a
647 transcriptional activator promoted by calcium influx (Tao et al., 2002). Since calcium
648 ions are essential in sperm function (Publicover et al., 2007), we cannot discard the
649 possibility that this TF could be involved in pathways related to sperm maintenance
650 and functioning. Some of the *CARF* predicted target genes from our analysis include
651 interesting candidates such as La Ribonucleoprotein Domain Family Member 4
652 (*LARP4*), THADA Armadillo Repeat Containing (*THADA*) and EF-Hand Domain
653 Containing 1 (*EFHC1*) gene. LARP4, has been proposed to regulate mRNA stability
654 and translation of mRNAs (Blagden et al., 2009). Blagden et al. (Blagden et al.,
655 2009) reported *Drosophila larp* knockout mutants resulted in a considerable
656 proportion of spermatocytes with meiotic defects. Although the role of *THADA*
657 remains uncertain in sperm, Moraru et al. (Moraru et al., 2017) showed that in
658 *Drosophila,* THADA modulates calcium signaling, energy storage and thermogenesis
659 balance. *EFHC1* encodes for a myoclonin1 protein and has been detected in sperm
660 flagella in mice testis (Suzuki et al., 2008). Although *Efhc1*-deficient mice were
661 fertile, mutants presented a reduced ciliary (flagellar) beating frequency (Suzuki et
662 al., 2009).

663 Other TFs with a large number of interactions were the SMAD Family Member 4
664 (*SMAD4*) gene (interacting with 32 genes) and the Lysine Demethylase 3A (*KDM3A*)
665 gene (281 gene interactions), both potentially targeting a set of genes enriched for
666 cellular macromolecular complex assembly processes (Additional file 9). TFs
667 involved in DNA repair, such as the Bromodomain Adjacent To Zinc Finger Domain
668 1B (*BAZ1B*), were also identified. Its closest paralog, *BAZ1A* encodes a member of
669 the chromatin remodeling complex (Racki et al., 2009). Dowdle et al. (Dowdle et al.,
670 2013) showed that *Baz1a*$^{-/-}$ mice were infertile because of spermatogenesis defects
671 tied to changes in chromatin composition. Another TF of interest was the *Estrogen*
672 *Receptor 1* (*ESR1*), which was present in the shared network. *ESR1* has been
673 already associated with pig sperm motility and cytoplasmatic droplets (Gunawan et
674 al., 2011). Moreover, polymorphisms in *ESR1* have been suggested to influence
675 estrogen levels which in turn, affect sperm motility (Carreau et al., 2002).

676 The network comprised new candidate genes for sperm quality. The Trafficking
677 Protein Particle Complex 2 Like (*TRAPPC2L*) gene, correlated with 27 miRNAs
678 including miR-30d, a miRNA that was dysregulated in oligozoospermic infertile
679 individuals (Salas-Huetos et al., 2015) (Figure 2.C). *TRAPPC2L* belongs to the
680 TRAPPC family, with a reported role in ciliogeneisis (Westlake et al., 2011).
681 Interestingly, *TRAPPC2L* was found associated in the Final Network with the
682 Spermatogenesis And Centriole Associated 1 (*SPATC1*) gene, a gene that has been
683 localized in the neck region of the mouse and human sperm (Goto et al., 2010).
684 Disruption of its homolog *Spatc1l* in mice led to male sterility due to separation of
685 sperm heads from tails, thereby advocating for a role in sperm head-tail integrity
686 (Kim et al., 2018). The network also included *DNAI2*, which correlated with 4
687 phenotypes (Additional file 6). 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 interacted with at least 20 genes. Some of these 30 miRNAs correlated with sperm traits and have also been previously linked to sperm quality and fertility in other studies. Noteworthy, miR-16, a miRNA that was down-regulated in the semen of infertile males with sperm abnormalities (Liu et al., 2012), correlated with 4 sperm phenotypes (Additional file 7) and potentially interacted with 67 genes (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 a motility-related parameter (VCL) and interacted with 32 genes (including the previously discussed *TRAPPC2L* that is present 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 and performed best for PDROP and some of the motility related traits (Table 6). The model for PDROP reported a highly significant role of the *THADA* gene (Additional file 10), which at the

same time was present in the Shared Network and its RNA levels are 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 and was also identified in the Shared Network (Additional file 10). This gene has a critical role in protein synthesis but no direct link to spermatogenesis or sperm function has been reported.

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 propose a SNP model with 73 SNPs including the polymorphisms identified through the GWAS, eGWAS and gene : gene interaction and phenotypic correlation analysis (Additional file 11). The model could hold promising potential for its application in animal breeding programs. This panel of 73 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 6). Remarkably, when only considering the GWAS lead SNPs, the panel explained between 4 to 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, this systems biology approach allowed including an additional set of SNPs that increased the predictive potential of the panel.

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. Gao et al. (Gao et al., 2019) identified 20 and 16 QTLs that could explain

738 35.3% and 20.6% of sperm motility and morphological abnormalities traits in Duroc
739 boars, respectively. Our approach was able to predict 30-31% and 26-36% of the
740 variance of the same group of traits with only 73 SNPs for motility and
741 morphological-related traits, respectively (Table 6). However, we have employed an
742 integrated and informed approach based not only on the GWAS and eGWAS FDR
743 significant associations but also in a robust network built from co-associated SNPs
744 (identified at suggestive levels but across several phenotypes) as well as gene RNA
745 co-abundance. Moreover, our SNPs were chosen to minimize LD between them and
746 thus maximize the informativity of the panel. This allowed the informed inclusion of a
747 large number of SNPs with independent marker potential and thus the development
748 of a more powerful panel for the prediction of semen quality in pigs.

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

755

756 **Conclusions**

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

762 a large number of potential interactions with target genes that are likely to be key
763 players in shaping the complex inheritance of the sperm quality traits. We have
764 developed a DNA marker panel based on a systems biology approach that may be
765 able to explain higher phenotypic variance than what could have been found from a
766 stand-alone GWAS. The model included GWAS lead SNPs, top eGWAS SNPs and
767 SNPs from genes identified in the Shared Network and could potentially explain over
768 30% of the phenotypic variance of sperm quality traits such as motility and
769 morphology. Although our results are considerably promising for the improvement of
770 the sector, caution should be taken due to the sample size of our study. Further work
771 should include the validation of the RNA and SNP model in a large number of pigs
772 belonging to different breeds and populations. The implications of this research are
773 broad, ranging from applications to animal breeding strategies to modeling the
774 biology of infertility in mammals.

775

776 **List of abbreviations**

777 ACRO: abnormal acrosomes

778 AI: Artificial Insemination

779 Ap: Associated Phenotype

780 AWM: Associated Weight Matrix

781 CASA: computer-assisted semen analysis

782 circRNA: circular RNA

783 CON: concentration

784 CPM: counts per million

DDROP: distal droplet

eGWAS: Expression GWAS

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

Declarations

Ethics approval and consent to participate

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 and blood following standard routine monitoring procedures and relevant guidelines.

Consent for publication

Not applicable

Availability of data and material

The datasets generated and/or analysed during the current study are available at NCBI's BioProject PRJNA520978. The phenotypic and genotypic datasets used in the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

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

828 Development Fund (ERDF). We thank the Agency for Management of University and
829 Research Grants (AGAUR) of the Generalitat de Catalunya (Grant Numbers 2014
830 SGR 1528 and 2017 SGR 1060). We also acknowledge the support of the Spanish
831 Ministry of Economy and Competitivity for the Center of Excellence Severo Ochoa
832 2016–2019 (Grant Number SEV-2015-0533) grant awarded to the Centre for
833 Research in Agricultural Genomics (CRAG). MG acknowledges a Ph.D. studentship
834 from MINECO (Grant Number BES-2014-070560) and a Short-Stay fellowship from
835 MINECO (EEBB-I-18-12860) at AR's group.

836 **Authors' contributions**

837 MG, AS, and AlC conceived and designed the experiments. JR-G carried the
838 phenotypic analysis. MG performed sperm purifications and RNA extractions. AnC
839 designed and carried the RT-qPCR and their analyses. MG analyzed the data with
840 support from AR, RGP and YRC. MG and AlC wrote the manuscript. All authors
841 discussed the data and read and approved the contents of the manuscript.

842 **Acknowledgements**

843 We thank Betlem Cabrera (CRAG), Dr. Fabiana Quoos Mayer (Instituto de
844 Pesquisas Veterinárias Desidério Finamor) and Dr. Martina Rocco (CRAG) for their
845 laboratory support. We gratefully acknowledge Craig Lewis from Genus PIC and
846 Sam Balasch from Gepork for contributing the sperm samples.

847 **References**

848 Ablondi M, Gòdia M, Rodríguez-Gil JE, Sánchez A, Clop A. Characterisation of
849 sperm piRNAs and their correlation with semen quality traits in swine. bioRxiv.
850 2020.03.16.994178

851 Abu-Halima M, Hammadeh M, Schmitt J, Leidinger P, Keller A et al. Altered
852 microRNA expression profiles of human spermatozoa in patients with different
853 spermatogenic impairments. Fertil Steril. 2013;99:1249-55 e16.

854 Abu-Halima M, Ayesh BM, Hart M, Alles J, Fischer U et al. Differential expression of
855 miR-23a/b-3p and its target genes in male patients with subfertility. Fertil
856 Steril. 2019;112:323-35 e2.

857 Aslam MKM, Kumaresan A, Yadav S, Mohanty TK and Datta TK. Comparative
858 proteomic analysis of high- and low-fertile buffalo bull spermatozoa for
859 identification of fertility-associated proteins. Reprod Domest Anim.
860 2019;54:786-94.

861 Bahler M, and Rhoads A. Calmodulin signaling via the IQ motif. FEBS Lett.
862 2002;513:107-13.

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

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

868 Boe-Hansen GB, Fortes MRS and Satake N. Morphological defects, sperm DNA
869 integrity, and protamination of bovine spermatozoa. Andrology. 2018;6:627-
870 33.

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

873 Breitbart H, Cohen G and Rubinstein S. Role of actin cytoskeleton in mammalian
874 sperm capacitation and the acrosome reaction. Reproduction. 2005;129:263-
875 8.

876 Capra E, Turri F, Lazzari B, Cremonesi P, Gliozzi TM et al. Small RNA sequencing
877 of cryopreserved semen from single bull revealed altered miRNAs and
878 piRNAs expression between High- and Low-motile sperm populations. Bmc
879 Genomics. 2017;18:14.
880 Carreau S, Bourguiba S, Lambard S, Galeraud-Denis I, Genissel C et al.
881 Reproductive system: aromatase and estrogens. Mol Cell Endocrinol.
882 2002;193:137-43.
883 Cingolani P, Platts A, Wang le L, Coon M, Nguyen T et al. A program for annotating
884 and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs
885 in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly
886 (Austin). 2012;6:80-92.
887 Curry E, Safranski TJ and Pratt SL. Differential expression of porcine sperm
888 microRNAs and their association with sperm morphology and motility.
889 Theriogenology. 2011;76:1532-9.
890 Cho DY, Kim YA and Przytycka TM. Chapter 5: Network Biology Approach to
891 Complex Diseases. PLoS Comput Biol. 2012;8:e1002820.
892 Diniz DB, Lopes MS, Broekhuijse ML, Lopes PS, Harlizius B et al. A genome-wide
893 association study reveals a novel candidate gene for sperm motility in pigs.
894 Anim Reprod Sci. 2014;151:201-7.
895 Dowdle JA, Mehta M, Kass EM, Vuong BQ, Inagaki A et al. Mouse BAZ1A (ACF1) Is
896 Dispensable for Double-Strand Break Repair but Is Essential for Averting
897 Improper Gene Expression during Spermatogenesis. PLoS Genet.
898 2013;9:e1003945.

899 Dunleavy JEM, Okuda H, O'Connor AE, Merriner DJ, O'Donnell L et al. Katanin-like
900 2 (KATNAL2) functions in multiple aspects of haploid male germ cell
901 development in the mouse. PLoS Genet. 2017;13:e1007078.
902 Fang P, Xu W, Li D, Zhao X, Dai J et al. A novel acrosomal protein, IQCF1, involved
903 in sperm capacitation and the acrosome reaction. Andrology. 2015;3:332-44.
904 Fortes MR, Reverter A, Zhang Y, Collis E, Nagaraj SH et al. Association weight
905 matrix for the genetic dissection of puberty in beef cattle. Proc Natl Acad Sci
906 U S A. 2010;107:13642-7.
907 Gadea J. Sperm factors related to in vitro and in vivo porcine fertility.
908 Theriogenology. 2005;63:431-44.
909 Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of
910 hierarchical clustering. Bioinformatics. 2015;31:3718-20.
911 Gao N, Chen Y, Liu X, Zhao Y, Zhu L et al. Weighted single-step GWAS identified
912 candidate genes associated with semen traits in a Duroc boar population.
913 Bmc Genomics. 2019;20:797.
914 Gòdia M, Estill M, Castelló A, Balasch S, Rodríguez-Gil JE et al. A RNA-Seq
915 Analysis to Describe the Boar Sperm Transcriptome and Its Seasonal
916 Changes. Front Genet. 2019a;10:299.
917 Gòdia M, Castelló A, Rocco M, Cabrera B, Rodríguez-Gil JE et al. Identification of
918 circular RNAs in porcine sperm and their relation to sperm motility. bioRxiv.
919 2019b:608026.
920 Gòdia M, Swanson G and Krawetz SA. A history of why fathers' RNA matters. Biol
921 Reprod. 2018a;99:147-59.

922 Gòdia M, Mayer FQ, Nafissi J, Castelló A, Rodríguez-Gil JE et al. A technical
923 assessment of the porcine ejaculated spermatozoa for a sperm-specific RNA-
924 seq analysis. Syst Biol Reprod Med. 2018b;64:291-303.
925 Goto M, O'Brien DA and Eddy EM. Speriolin is a novel human and mouse sperm
926 centrosome protein. Hum Reprod. 2010;25:1884-94.
927 Gunawan A, Kaewmala K, Uddin MJ, Cinar MU, Tesfaye D et al. Association study
928 and expression analysis of porcine ESR1 as a candidate gene for boar fertility
929 and sperm quality. Anim Reprod Sci. 2011;128:11-21.
930 Harper C, Wootton L, Michelangeli F, Lefièvre L, Barratt C et al. Secretory pathway
931 Ca2+-ATPase (SPCA1) Ca2+ pumps, not SERCAs, regulate complex [Ca2+]i
932 signals in human spermatozoa. J Cell Sci. 2005;118:1673-85.
933 Heid HW, Figge U, Winter S, Kuhn C, Zimbelmann R et al. Novel actin-related
934 proteins Arp-T1 and Arp-T2 as components of the cytoskeletal calyx of the
935 mammalian sperm head. Exp Cell Res. 2002;279:177-87.
936 Hering DM, Olenski K, Rusc A and Kaminski S. Genome-wide association study for
937 semen volume and total number of sperm in Holstein-Friesian bulls. Anim
938 Reprod Sci. 2014;151:126-30.
939 Hu H, Miao YR, Jia LH, Yu QY, Zhang Q et al. AnimalTFDB 3.0: a comprehensive
940 resource for annotation and prediction of animal transcription factors. Nucleic
941 Acids Res. 2019a;47:D33-D8.
942 Hu J, Cheng S, Wang H, Li X, Liu S et al. Distinct roles of two myosins in C. elegans
943 spermatid differentiation. PLoS Biol. 2019b;17:e3000211.
944 Ibba M, and Söll D. Aminoacyl-tRNA Synthesis. Annu Rev Biochem. 2000;69:617-
945 50.

946 Izumiyama T, Minoshima S, Yoshida T and Shimizu N. A novel big protein TPRBK
947 possessing 25 units of TPR motif is essential for the progress of mitosis and
948 cytokinesis. Gene. 2012;511:202-17.
949 Jodar M, Sendler E, Moskovtsev SI, Librach CL, Goodrich R et al. Absence of sperm
950 RNA elements correlates with idiopathic male infertility. Sci Transl Med.
951 2015;7:295re6.
952 Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory
953 requirements. Nat Methods. 2015;12:357-60.
954 Kim J, Kwon JT, Jeong J, Kim J, Hong SH et al. SPATC1L maintains the integrity of
955 the sperm head-tail junction. EMBO Rep. 2018;19.
956 Kozomara A, and Griffiths-Jones S. miRBase: integrating microRNA annotation and
957 deep-sequencing data. Nucleic Acids Res. 2011;39:D152-7.
958 Krausz C, Escamilla AR and Chianese C. Genetics of male infertility: from research
959 to clinic. Reproduction. 2015;150:R159-74.
960 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. The Sequence
961 Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-9.
962 Li X, Chen G and Yang B. Urea transporter physiology studied in knockout mice.
963 Front Physiol. 2012;3:217.
964 Liu H, Li W, Zhang Y, Zhang ZG, Shang XJ et al. IFT25, an intraflagellar transporter
965 protein dispensable for ciliogenesis in somatic cells, is essential for sperm
966 flagella formation. Biol Reprod. 2017;96:993-1006.
967 Liu M, Guan ZL, Shen Q, Lalor P, Fitzgerald U et al. Ulk4 Is Essential for
968 Ciliogenesis and CSF Flow. J Neurosci. 2016;36:7589-600.

969 Liu T, Huang Y, Liu J, Zhao Y, Jiang L et al. MicroRNA-122 influences the
970 development of sperm abnormalities from human induced pluripotent stem
971 cells by regulating TNP2 expression. Stem Cells Dev. 2013;22:1839-50.
972 Liu T, Cheng W, Gao Y, Wang H and Liu Z. Microarray analysis of microRNA
973 expression patterns in the semen of infertile men with semen abnormalities.
974 Mol Med Rep. 2012;6:535-42.
975 Loges NT, Olbrich H, Fenske L, Mussaffi H, Horvath J et al. DNAI2 mutations cause
976 primary ciliary dyskinesia with defects in the outer dynein arm. Am J Hum
977 Genet. 2008;83:547-58.
978 Luangpraseuth-Prosper A, Lesueur E, Jouneau L, Pailhoux E, Cotinot C et al.
979 TOPAZ1, a germ cell specific factor, is essential for male meiotic progression.
980 Dev Biol. 2015;406:158-71.
981 Marques DBD, Bastiaansen JWM, Broekhuijse M, Lopes MS, Knol EF et al.
982 Weighted single-step GWAS and gene network analysis reveal new candidate
983 genes for semen traits in pigs. Genet Select Evol. 2018;50:40.
984 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing
985 reads. EMBnet J. 2011;17:10-2.
986 Martin PM, Carnaud M, del Cano GG, Irondelle M, Irinopoulou T et al.
987 Schwannomin-interacting protein-1 isoform IQCJ-SCHIP-1 is a late
988 component of nodes of Ranvier and axon initial segments. J Neurosci.
989 2008;28:6111-7.
990 Mengerink KJ, and Vacquier VD. An ATP-binding cassette transporter is a major
991 glycoprotein of sea urchin sperm membranes. J Biol Chem. 2002;277:40729-
992 34.

993 Morales CR, Marat AL, Ni X, Yu Y, Oko R et al. ATP-binding cassette transporters
994 ABCA1, ABCA7, and ABCG1 in mouse spermatozoa. Biochem Bioph Res Co.
995 2008;376:472-7.
996 Moraru A, Cakan-Akdogan G, Strassburger K, Males M, Mueller S et al. THADA
997 Regulates the Organismal Balance between Energy Storage and Heat
998 Production. Dev Cell. 2017;41:72-81.
999 Nagarajan P, Onami TM, Rajagopalan S, Kania S, Donnell R et al. Role of
1000 chromodomain helicase DNA-binding protein 2 in DNA damage response
1001 signaling and tumorigenesis. Oncogene. 2009;28:1053-62.
1002 Olson GE, Winfrey VP and Nagdas SK. Structural modification of the hamster sperm
1003 acrosome during posttesticular development in the epididymis. Microsc Res
1004 Tech. 2003;61:46-55.
1005 Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT et al. StringTie
1006 enables improved reconstruction of a transcriptome from RNA-seq reads. Nat
1007 Biotechnol. 2015;33:290-5.
1008 Publicover S, Harper CV and Barratt C. [Ca2+]i signalling in sperm — making the
1009 most of what you've got. Nature Cell Biology. 2007;9:235-42.
1010 Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA et al. PLINK: a tool set for
1011 whole-genome association and population-based linkage analyses. Am J
1012 Hum Genet. 2007;81:559-75.
1013 Qin Y, Ji J, Du G, Wu W, Dai J et al. Comprehensive pathway-based analysis
1014 identifies associations of BCL2, GNAO1 and CHD2 with non-obstructive
1015 azoospermia risk. Hum Reprod. 2014;29:860-6.

1016 Quintero-Moreno A, Rigau T and Rodriguez-Gil JE. Regression analyses and motile
1017 sperm subpopulation structure study as improving tools in boar semen quality
1018 analysis. Theriogenology. 2004;61:673-90.
1019 R Developmental Core Team. R: A language and environment for statistical
1020 computing. 2010.
1021 Racki LR, Yang JG, Naber N, Partensky PD, Acevedo A et al. The chromatin
1022 remodeller ACF acts as a dimeric motor to space nucleosomes. Nature.
1023 2009;462:1016-21.
1024 Ramayo-Caldas Y, Renand G, Ballester M, Saintilan R and Rocha D. Multi-breed
1025 and multi-trait co-association analysis of meat tenderness and other meat
1026 quality traits in three French beef cattle breeds. Genet Select Evol.
1027 2016;48:37.
1028 Ramayo-Caldas Y, Marmol-Sanchez E, Ballester M, Sanchez JP, Gonzalez-Prendes
1029 R et al. Integrating genome-wide co-association and gene expression to
1030 identify putative regulators and predictors of feed efficiency in pigs. Genet Sel
1031 Evol. 2019;51:48.
1032 Reverter A, and Fortes MR. Association weight matrix: a network-based approach
1033 towards functional genome-wide association studies. Methods Mol Biol.
1034 2013;1019:437-47.
1035 Reverter A, Barris W, McWilliam S, Byrne KA, Wang YH et al. Validation of
1036 alternative methods of data normalization in gene co-expression studies.
1037 Bioinformatics. 2005;21:1112-20.
1038 Reverter A, and Chan EK. Combining partial correlation and an information theory
1039 approach to the reversed engineering of gene co-expression networks.
1040 Bioinformatics. 2008;24:2491-7.

1041 Robinson JA, and Buhr MM. Impact of genetic selection on management of boar
1042 replacement. Theriogenology. 2005;63:668-78.
1043 Rueda A, Barturen G, Lebron R, Gomez-Martin C, Alganza A et al. sRNAtoolbox: an
1044 integrated collection of small RNA research tools. Nucleic Acids Res.
1045 2015;43:W467-73.
1046 Salas-Huetos A, Blanco J, Vidal F, Godo A, Grossmann M et al. Spermatozoa from
1047 patients with seminal alterations exhibit a differential micro-ribonucleic acid
1048 profile. Fertil Steril. 2015;104:591-601.
1049 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. Cytoscape: a software
1050 environment for integrated models of biomolecular interaction networks.
1051 Genome Res. 2003;13:2498-504.
1052 Simon P, Baumner S, Busch O, Rohrich R, Kaese M et al. Polysialic Acid Is Present
1053 in Mammalian Semen as a Post-translational Modification of the Neural Cell
1054 Adhesion Molecule NCAM and the Polysialyltransferase ST8SiaII. J Biol
1055 Chem. 2013;288:18825-33.
1056 Smital J, Wolf J and De Sousa LL. Estimation of genetic parameters of semen
1057 characteristics and reproductive traits in AI boars. Anim Reprod Sci.
1058 2005;86:119-30.
1059 Snelling WM, Cushman RA, Keele JW, Maltecca C, Thomas MG et al. Breeding and
1060 Genetics Symposium: networks and pathways to guide genomic selection. J
1061 Anim Sci. 2013;91:537-52.
1062 Song Z. Roles of the nucleotide sugar transporters (SLC35 family) in health and
1063 disease. Mol Aspects Med. 2013;34:590-600.
1064 Suh KS, Tatunchak TT, Crutchley JM, Edwards LE, Marin KG et al. Genomic
1065 structure and promoter analysis of PKC-delta. Genomics. 2003;82:57-67.

1066 Sun XH, Zhu YY, Wang L, Liu HL, Ling Y et al. The Catsper channel and its roles in
1067 male fertility: a systematic review. Reprod Biol Endocrinol. 2017;15:65.
1068 Suzuki T, Inoue I, Yamagata T, Morita N, Furuichi T et al. Sequential expression of
1069 Efhc1/myoclonin1 in choroid plexus and ependymal cell cilia. Biochem Bioph
1070 Res Co. 2008;367:226-33.
1071 Suzuki T, Miyamoto H, Nakahari T, Inoue I, Suemoto T et al. Efhc1 deficiency
1072 causes spontaneous myoclonus and increased seizure susceptibility. Hum
1073 Mol Genet. 2009;18:1099-109.
1074 Taiyun W, and Viliam S, 2017 R package "corrplot": Visualization of a Correlation
1075 Matrix (Version 0.84). Available at: https://github.com/taiyun/corrplot
1076 Tao X, West AE, Chen WG, Corfas G and Greenberg ME. A calcium-responsive
1077 transcription factor, CaRF, that regulates neuronal activity-dependent
1078 expression of BDNF. Neuron. 2002;33:383-95.
1079 Tian H, Li ZL, Peng D, Bai XG and Liang WB. Expression difference of miR-10b and
1080 miR-135b between the fertile and infertile semen samples (p). Forens Sci Int-
1081 Gen S. 2017;6:E257-E9.
1082 Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and
1083 manhattan plots. bioRxiv. 2014:005165.
1084 Visconti PE, Westbrook VA, Chertihin O, Demarco I, Sleight S et al. Novel signaling
1085 pathways involved in sperm acquisition of fertilizing capacity. J Reprod
1086 Immunol. 2002;53:133-50.
1087 Vitavska O, and Wieczorek H. Putative role of an SLC45 H(+)/sugar cotransporter in
1088 mammalian spermatozoa. Pflug Arch Eur J Phy. 2017;469:1433-42.

1089 Waldron A, Wilcox C, Francklyn C and Ebert A. Knock-Down of Histidyl-tRNA
1090 Synthetase Causes Cell Cycle Arrest and Apoptosis of Neuronal Progenitor
1091 Cells in vivo. Front Cell Dev Biol. 2019;7:67.
1092 Wang X, Yang C, Guo F, Zhang Y, Ju Z et al. Integrated analysis of mRNAs and
1093 long noncoding RNAs in the semen from Holstein bulls with high and low
1094 sperm motility. Sci Rep. 2019;9:2092.
1095 Westlake CJ, Baye LM, Nachury MV, Wright KJ, Ervin KE et al. Primary cilia
1096 membrane assembly is initiated by Rab11 and transport protein particle II
1097 (TRAPPII) complex-dependent trafficking of Rabin8 to the centrosome. Proc
1098 Natl Acad Sci U S A. 2011;108:2759-64.
1099 Wolf J. Genetic Parameters for Semen Traits in AI Boars Estimated from Data on
1100 Individual Ejaculates. Reprod Domest Anim. 2009;44:338-44.
1101 Wu FJ, Lin TY, Sung LY, Chang WF, Wu PC et al. BMP8A sustains
1102 spermatogenesis by activating both SMAD1/5/8 and SMAD2/3 in
1103 spermatogonia. Sci Signal. 2017;10:eaal1910.
1104 Yang J, Lee SH, Goddard ME and Visscher PM. GCTA: a tool for genome-wide
1105 complex trait analysis. Am J Hum Genet. 2011;88:76-82.
1106 Zhao X, Zhao K, Ren J, Zhang F, Jiang C et al. An imputation-based genome-wide
1107 association study on traits related to male reproduction in a White Duroc x
1108 Erhualian F2 population. Anim Sci J. 2016;87:646-54.
1109 Zhao Y, Gao N, Li X, El-Ashram S, Wang Z et al. Identifying candidate genes
1110 associated with sperm morphology abnormalities using weighted single-step
1111 GWAS in a Duroc boar population. Theriogenology. 2020;141:9-15.

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.

Zhou JH, Zhou QZ, Yang JK, Lyu XM, Bian J et al. MicroRNA-27a-mediated repression of cysteine-rich secretory protein 2 translation in asthenoteratozoospermic patients. Asian J Androl. 2017;19**:**591-5.

Figure legends:

Figure 1 (TIFF)

Manhattan plots depicting the genetic associations between SNPs and the sperm quality traits that showed genome-wide significant values. Significant associations have been found with the percentage of: **A)** Percentage of cells with head abnormalities (HABN); **B)** Percentage of cells with abnormal acrosomes after 5 min incubation at 37ºC (ACRO_5); **C)** Percentage of cells with of neck abnormalities (NABN); **D)** Percentage of motile spermatozoa after 5 min incubation at 37ºC (MT_5); **E)** Percentage of motile spermatozoa after 90 min incubation at 37ºC (MT_90); **F)** Percentage of cells with proximal droplets (PDROP); **G)** Ratio of the percentage of abnormal acrosomes at 5 min versus 90 min incubation times (R_ACRO). The *x*-axis represents chromosome length (Mb), and the *y*-axis shows the negative $\log_{10}$ P-values of the genetic associations. The horizontal red line represents the significance threshold (FDR ≤ 0.05).

Figure 2 (TIFF)

Co-association network based on the AWM 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 *CHD2* gene interactions. The node color corresponds to the phenotype group with the highest correlation value, as follows: concentration (red), abnormal 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 or miRNA. 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).

Additional files:

Additional figure 1 (TIFF)

Outline of the analysis pipeline.

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

Additional figure 2 (PNG)

Correlation across boar sperm quality traits.

Heatmap plot of the correlations among the 25 sperm traits measured on 300 boars. CON=Concentration; VIAB_5= Viability 5 min; VIAB_90= Viability 90 min;ORT= Osmotic Resistance Test; HABN= Head abnormalities; NABN= Neck abnormalities; TABN= Tail abnormalities; PDROP= Proximal droplets; DDROP= Distal droplets; MT_5= Motility 5 min; VAP_5= Average Path Velocity 5 min; VCL_5= Curvilinear Velocity 5 min;_5= Straight Line Velocity 5 min; MT_90= Motility 90 min; VAP_90= Average Path Velocity 90 min; VCL_90= Curvilinear Velocity 90 min; VSL_90=

Straight Line Velocity 90 min; ACRO_5= Abnormal Abnormal Acrosomes 5 min; ACRO_90= Abnormal Acrosomes 90 min; R_MT= Ratio Motility; R_VAP= Ratio Average Path Velocity; R_VCL= Ratio Curvilinear Velocity; R_VSL= Ratio Straight Line Velocity; R_VIAB= Ratio Viability; R_ACRO= Ratio Acrosomes.

Additional figure 3 (PNG)

SNP based dendrogram for the 25 semen parameters.

Dendrogram of the standardized SNP effects across the 25 sperm traits.

Additional file 1 (XLS)

Effect of external factors on 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)

Details on the SNPs showing significant associations (FDR ≤ 0.05) in the GWAS across autosomal chromosomes and unplaced scaffolds. Chr: chromosome; BP: base pairs (location); Beta = additive effect; FDR = False Discovery Rate; HABN = Head abnormalities; MT_5 = Percentage of motile spermatozoa at 5 min; MT_90 = Percentage of motile spermatozoa at 90 min; NABN= Neck abnormalities; PDROP= Proximal droplets; R_ACRO= Ratio Abnormal Acrosomes.

Additional file 3 (XLS)

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

Additional file 4 (XLS)

List of protein coding genes and miRNAs identified in sperm. Average and Standard Deviation (SD) for the samples processed. Protein coding and miRNA abundances are expressed in Fragments Per Kilobase per Million mapped reads (FPKM) and counts per million (CPM), respectively.

Additional file 5 (XLS)

SNPs identified in the RNA-seq data mapping within the GWAS regions. Chr=chromosome. LD=linkage disequilibrium. Allelic frequency for each of the genotypes. # samples called=number of samples with reads in the given SNP position.

Additional file 6 (XLS)

Correlations between gene abundances and phenotypes. P-values are given when (P-value ≤ 0.05). The correlation value is indicated between brackets. ns=Not Significant.

Additional file 7 (XLS)

Correlations between miRNA abundances and phenotypes. P-values are given when (P-value < 0.05). The correlation value is indicated between brackets. ns=Not Significant.

Additional file 8 (XLS)

Associations identified in the within trait eGWAS. Thirty-nine SNPs showed significant associations (FDR ≤ 0.05) with semen phenotypes in the GWAS and also displayed significant association with the abundance of genes which abundance correlated with the same phenotype (P-value ≤ 0.05). Chr: chromosome. FDR =

False Discovery Rate; ACRO_5 = Abnormal Acrosomes 5 min; HABN = Head abnormalities.

Additional file 9 (XLSX)

Gene Ontology analysis of the genes included in the Final Network. GO biological process terms with significant Bonferroni corrected P-values and their associated genes.

Additional file 10 (PDF)

Parameter estimates for the significant RNA 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.

Additional file 11 (XLS)

Description of the SNPs included in the SNP panel. Chromosome, position, SNP ID and analysis from which the SNP was extracted.

Table 1. Descriptive statistics, genomic heritability (h2) and number of significant SNPs in the GWAS for sperm quality parameters (n=300).

Trait	**Acronym**	**Mean (SD)**	**h^2 (SE)**	**Number of SNPs in autosomal chromosomes**	**Number of SNPs in unplaced scaffolds**
Concentration (sperm/ml)	CON	141.3 (65.5)	0.13 (0.11)	0	0
Viability 5 min	VIAB_5	90.1 (6.3)	$1x10^{-6}$ (0.11)	0	0
Viability 90 min	VIAB_90	77.4 (17.3)	0.14 (0.13)	0	0
Osmotic Resistance Test	ORT	79.8 (12.5)	0.13 (0.12)	0	0
Head abnormalities	HABN	2.1 (5.9)	0.16 (0.11)	41	0
Neck abnormalities	NABN	3.0 (4.9)	$1x10^{-6}$ (0.13)	18	0
Tail abnormalities	TABN	2.7 (3.4)	0.09 (0.12)	0	0
Proximal droplets	PDROP	3.5 (5.1)	0.12 (0.15)	1	0

Distal droplets	DDROP	4.5 (4.5)	0.06 (0.11)	0	0
Motility 5 min	MT_5	75.4 (18.1)	0.21 (0.15)	3	217
Motility 90 min	MT_90	64.1 (22.0)	0.39 (0.14)	2	252
Average Path Velocity 5 min (µm/seg)	VAP_5	34.0 (10.2)	0.17 (0.11)	0	0
Average Path Velocity 90 min (µm/seg)	VAP_90	30.8 (9.5)	0.35 (0.13)	0	0
Curvilinear Velocity 5 min (µm/seg)	VCL_5	46.2 (12.5)	0.11 (0.10)	0	0
Curvilinear Velocity 90 min (µm/seg)	VCL_90	39.7 (10.2)	0.35 (0.13)	0	0
Straight Line Velocity 5 min (µm/seg)	VSL_5	27.0 (8.3)	0.23 (0.13)	0	38
Straight Line Velocity 90 min (µm/seg)	VSL_90	25.9 (8.3)	0.34 (0.13)	0	0
Abnormal Acrosomes 5 min	ACRO_5	7.0 (5.6)	0.08 (0.11)	4	0
Abnormal Acrosomes 90 min	ACRO_90	16.4 (12.6)	0.06 (0.10)	0	0
Ratio Motility	R_MT	0.9 (0.2)	$1x10^{-6}$ (0.11)	0	0

Ratio Average Path Velocity	R_VAP	0.9 (0.3)	$1x10^{-6}$ (0.08)	0	0
Ratio Curvilinear Velocity	R_VCL	0.9 (0.3)	$1x10^{-6}$ (0.09)	0	0
Ratio Straight Line Velocity	R_VSL	1.0 (0.3)	0.06 (0.10)	0	0
Ratio Viability	R_VIAB	0.9 (0.3)	0.08 (0.11)	0	0
Ratio Acrosomes	R_ACRO	3.4 (3.5)	0.08 (0.11)	1	0

All traits except stated are presented as a percentage. # SNPs = GWAS number of single nucleotide polymorphisms significantly associated (FDR) with the trait; The values shown are raw excepting the ratios which were previously corrected and stabilized. SD: Standard Deviation; SE: Standard Error

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	Beta	Trait
1	I1	1	-	rs339761632	13,501,755	4.64×10^{-8}	0.02	0.06	4.84	PDROP
1	I2	8	82.90-83.49	rs81354986	82,895,619	1.69×10^{-6}	0.03	0.07	5.05	HABN
1	I3	8	94.88-98.74	rs327733412	94,880,167	1.61×10^{-7}	0.02	0.07	5.65	HABN
1	I4	1	-	rs337166779	126,397,198	2.05×10^{-6}	0.03	0.06	5.02	HABN
1	I5	11	243.86-246.44	rs343194423	246,224,386	1.72×10^{-7}	0.01	0.07	3.17	NABN
1	I6	2	258.54-258.55	rs332256425	258,548,786	1.76×10^{-6}	0.04	0.06	3.44	NABN
3	I1	1	-	rs332055717	2,911,413	6.35×10^{-8}	0.01	0.09	5.07	HABN
3	I2	3	113.75-113.84	rs328292697	113,750,595	1.09×10^{-7}	0.01	0.07	3.41	NABN

4	I1	2	2.41-2.42	rs318575212, rs332927981	2,412,006, 2,415,239	2.88×10^{-8}	0.01	0.08	4.11	ACRO_5
6	I1	2	65.60-66.66	rs335394654	65,597,553	1.86×10^{-7}	0.03	0.14	3.04	ACRO_5
7	I1	2	6.20-6.38	rs326239534	6,377,172	9.87×10^{-6}	0.02	0.17	-9.15	MT_5
7	I2	2	85.73-86.88	rs336588919	86,884,279	4.13×10^{-8}	0.01	0.06	3.75	NABN
9	I1	2	5.76-5.78	rs1110111787	5,776,597	1.55×10^{-7}	0.02	0.07	5.43	HABN
9	I2	1	-	rs342738178	28,463,580	1.53×10^{-5}	0.03	0.14	-10.42	MT_5, MT_90
9	I3	1	-	rs328217450	137,959,590	4.77×10^{-8}	0.02	0.18	2.36	R_ACRO
13	I1	18	25.36-28.47	rs690794887	25,535,100	3.06×10^{-7}	0.02	0.14	3.78	HABN
13	I2	3	33.82-37.65	rs327865244	33,819,549	3.79×10^{-8}	0.01	0.15	4.28	HABN
16	I1	1	-	rs324239602	6,476,358	6.08×10^{-6}	0.01	0.46	9.07	MT_90

SSC = S. scrofa chromosome; #SNP = number of single nucleotide polymorphisms significantly associated (FDR) with the trait; Interval = region of the GWAS interval; Beta = additive effect; FDR = False Discovery Rate; MAF = Minor Allele Frequency; ACRO_5 = Abnormal Acrosomes 5 min; HABN = Head abnormalities; NABN = Neck abnormalities; PDROP = Proximal droplets; R_ACRO = Ratio Acrosomes; MT_5 = Motility 5 min; MT_90 = Motility 90 min.

Table 3. Summary of the SNPs identified from the RNA-seq datasets in genes mapping within the GWAS regions.

SSC	Interval	Top SNP of the GWAS interval	# SNPs called	Highest LD	SNP with highest LD	Gentoypic frequency (0/0; 0/1; 1/1)	# called samples	SNP effect	Gene	Trait
1	I3	rs327733412	3	0.07	rs710447566	0.34; 0.54; 0.11	35	Low	*KATNAL2*	HABN
7	I2	rs336588919	2	0.4	rs330912302	0.63; 0.12; 0.25	32	Low	*CHD2*	NABN
13	I1	rs690794887	21	0.4	rs331304027	0.06; 0.09; 0.85	33	Moderate	*ULK4*	HABN
13	I2	rs327865244	11	0.2	rs323872641	0.49; 0.37; 0.14	35	Low	*ABHD14A*	HABN

SSC = S. scrofa chromosome; #SNP called = number of single nucleotide polymorphisms identified in the SNP calling analysis. LD = linkage disequilibrium; Allelic frequency for each of the genotypes. # called samples = number of samples with reads in the given SNP position. The column SNP effect and gene refer to the SNP with highest LD in the region. HABN = Head abnormalities; NABN = Neck abnormalities.

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

SSC	Interval	# SNP : transcipts	Top eGWAS	Top eGWAS location, bp	Top eGWAS P-value	Top eGWAS FDR	Top eGWAS MAF	Beta	Trait	RNA abundance correlation	Associated Gene
		2	rs318575212, rs332927981	2,412,006, 2,415,239	7.36 x10^{-3}	0.03	0.09	-0.39	ACRO_5	-0.33	*NCLN*
4	I1	2	rs318575212, rs332927981	2,412,006, 2,415,239	1.83 x10^{-4}	0.03	0.09	-1.8	ACRO_5	-0.46	*ASCC1*
		2	rs318575212, rs332927981	2,412,006, 2,415,239	2.87 x10^{-4}	4.83 x10^{-2}	0.09	-1.1	ACRO_5	-0.4	*AATF*
6	I1	2	rs335394654	65,597,553	5.63 x10^{-5}	0.02	0.11	-1.65	ACRO_5	-0.35	*IQCJ*

13	I1	31	rs328397029	25,684,259	1.84×10^{-5}	2.95×10^{-3}	0.09	-1.03	HABN	-0.38	*HARS, ACTR2, EPB41L3, RAB1B*

SSC = S. scrofa chromosome; # SNP : transcripts = number of single nucleotide polymorphisms significantly associated to a transcript; Beta = additive effect; MAF = Minor Allele Frequency; ACRO_5 = Abnormal Acrosomes 5 min; HABN = Head abnormalities.

Figures

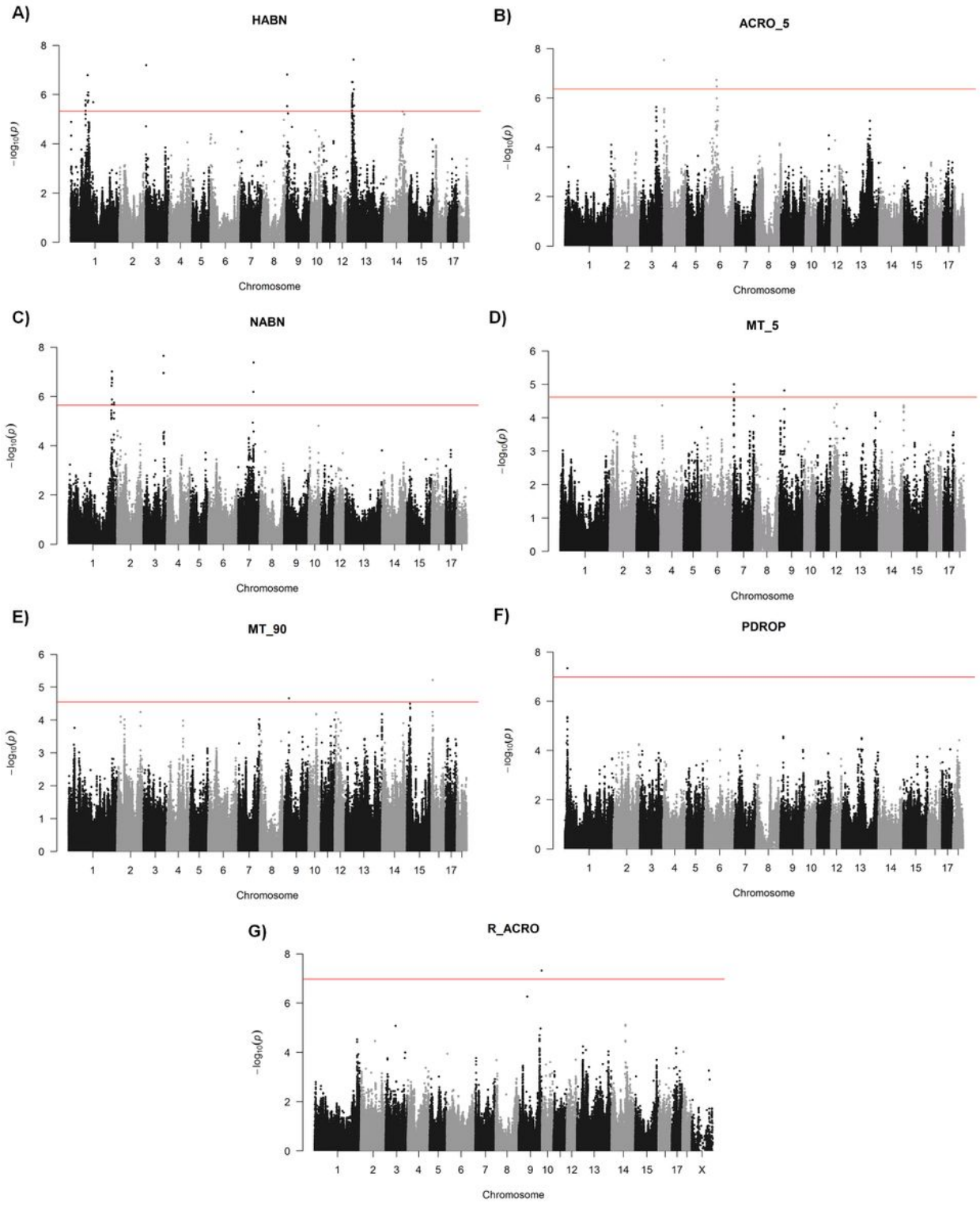

Figure 1

Manhattan plots depicting the genetic associations between SNPs and the sperm quality traits that showed genome-wide significant values. Significant associations have been found with the percentage of: A) Percentage of cells with head abnormalities (HABN); B) Percentage of cells with abnormal

acrosomes after 5 min incubation at 37ºC (ACRO_5); C) Percentage of cells with of neck abnormalities (NABN); D) Percentage of motile spermatozoa after 5 min incubation at 37ºC (MT_5); E) Percentage of motile spermatozoa after 90 min incubation at 37ºC (MT_90); F) Percentage of cells with proximal droplets (PDROP); G) Ratio of the percentage of abnormal acrosomes at 5 min versus 90 min incubation times (R_ACRO). The x-axis represents chromosome length (Mb), and the y-axis shows the negative log10 P-values of the genetic associations. The horizontal red line represents the significance threshold (FDR ≤ 0.05).

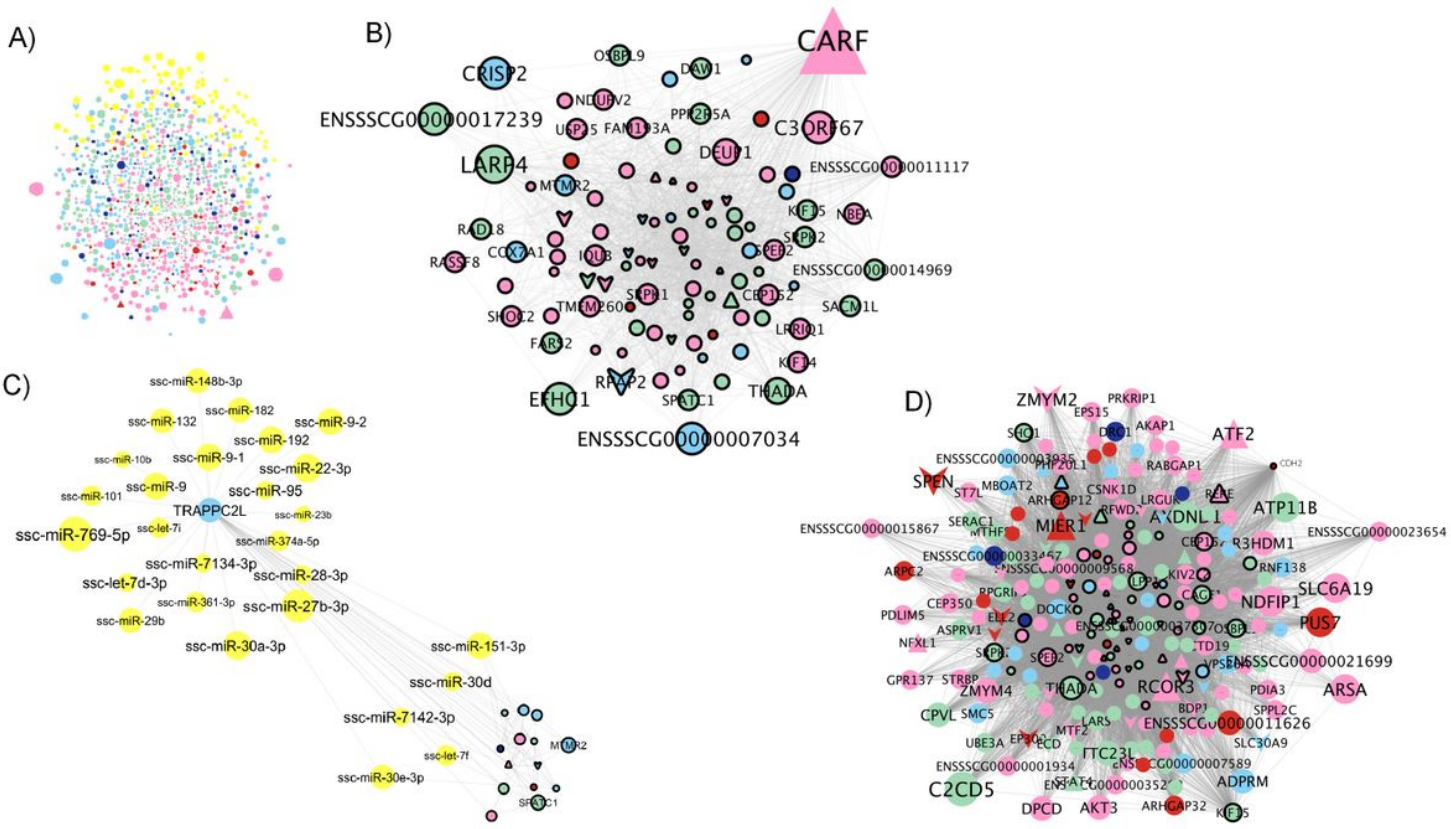

Figure 2

Co-association network based on the AWM 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 CHD2 gene interactions. The node color corresponds to the phenotype group with the highest correlation value, as follows: concentration (red), abnormal 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 or miRNA. 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).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplemental.zip