CLEAN COPY

Article Type: Research Communication

Title: A transcriptomic insight into the human sperm microbiome through next-generation sequencing

Author details: C. Corral-Vazquez [a] (ORCID: 0000-0002-0621-4590), J. Blanco [a] (ORCID: 0000-0003-0647-3856), R. Aiese Cigliano [b] (ORCID: 0000-0002-9058-6994), Z. Sarrate [a] (ORCID: 0000-0001-9677-1376), F. Vidal [a] (ORCID: 0000-0002-0985-7348), and E. Anton [a*] (ORCID: 0000-0002-8914-2667)

[a] *Genetics of Male Fertility Group, Unitat de Biologia Cel·lular (Facultat de Biociències), Universitat Autònoma de Barcelona, Cerdanyola del Vallès, 08193, Spain;* [b] *Sequentia Biotech SL, Barcelona, 08005, Spain;* * *Corresponding author, ORCID: 0000-0002-8914-2667, Tel: +34 93 581 3733; Fax: +34 93 581 2295; Email: ester.anton@uab.cat*

Data sharing statement: All sequence data mentioned in this article have been deposited into a database (NCBI BioProject accession number: PRJNA573604)

Short running head: Human sperm microbiome insight through RNA-seq

Length of the manuscript in number of words: 3,208

Number of figures: Two

Number of tables: Two (plus one supplemental)

Abstract

The purpose of this study is to provide novel information through Next Generation Sequencing (NGS) for the characterization of viral and bacterial RNA cargo of human sperm cells from healthy fertile donors. For this, RNA-seq raw data of poly(A) RNA from 12 sperm samples from fertile donors were aligned to microbiome databases using the GAIA software. Species of viruses and bacteria were quantified in Operational Taxonomic Units (OTU) and filtered by minimal expression level (>1% OTU in at least one sample). Mean expression values (and their standard deviation) of each species were estimated. A Hierarchical Cluster Analysis (HCA) and a Principal Component Analysis (PCA) were performed to detect common microbiome patterns among samples. Sixteen microbiome species, families, domains, and orders surpassed the established expression threshold. Of the 16 categories, nine corresponded to viruses (23.07% OTU) and seven to bacteria (2.77% OTU), among which the *Herperviriales* order and *Escherichia coli* were the most abundant, respectively. HCA and PCA displayed four clusters of samples with a differentiated microbiome fingerprint. This work represents a pilot study into the viruses and bacteria that make up the human sperm microbiome. Despite the high variability observed, some patterns of similarity among individuals were identified. Further NGS studies under standardized methodological procedures are necessary to achieve a deep knowledge of the semen microbiome and its implications in male fertility.

Keywords: bacteria, microbiome, Next-Generation Sequencing, sperm, viruses

List of abbreviations:

GAIA: Group for Artificial Intelligence Applications

HCA: Hierarchical Cluster Analysis

OTU: Operational Taxonomic Units

PC1: Principal Component 1

PC2: Principal Component 2

PCA: Principal Component Analysis

WGS: Whole Genome Sequencing

WTS: Whole Transcriptome Sequencing

Introduction

The microbiome is known as the collection of microorganisms that occupy a given niche. The presence of microbe communities has been detected in many organic hosts. In fact, the average number of bacteria in a human body ($3.8 \cdot 10^{13}$) is similar to the average number of human cells ($3.0 \cdot 10^{13}$) (Sender et al. 2016). Most human tissues host a microbiota composed of not only bacteria but also of other microorganisms such as viruses, protozoa, and fungi (Altmäe et al. 2019). Their existence is not trivial, since it is known that the microbiome has a vital influence on physiology, homeostasis and resistance to disease (D'Argenio and Salvatore 2015).

Concerning the reproductive system, an association between their microbiome composition and human fertility has been stablished by some authors. Studies in women have revealed certain consistencies in the vaginal microbiome, being *Lactobacillus* the predominant bacterial genus (Mändar et al. 2015). Imbalances in this vaginal and uterine microbiome are associated with adverse pregnancy outcomes, infertility treatment failure, and irregular endometrial function (Koedooder et al. 2019). In men, specific microbiota has been detected in the genital tract (Koedooder et al. 2019). Further, studies focused on the seminal microbiome have revealed a relation between the microbiome composition and seminal quality, acrosome reaction and sperm DNA fragmentation. While the abundance of bacterial genus like *Lactobacillus* or *Gardnerella* show an association with fertility and good quality semen, others like *Anaerococcus*, *Prevotella*, *Neisseria*, *Klebsiella*, and *Pseudomonas* have been associated with infertility and seminal impairments (Hou et al. 2013; Liu et al. 2014; Weng et al. 2014; Mändar et al. 2017; Chen et al. 2018; Monteiro et al. 2018; Baud et al. 2019; Koedooder et al. 2019). Intriguingly, some authors have revealed that couples exhibit some kind of complementation between the semen and vagina microbiota (Mändar et al. 2015; Mändar et al. 2018).

Several RNA-based techniques have been developed to characterize the human microbiota. Although amplification of the bacterial 16S rRNA has been the most recurrent strategy, the use of next-generation techniques has opened doors to more powerful and complete studies. In particular, the application of RNA-seq to sperm microbiome characterization is still incipient, both in human (Monteiro et al. 2018; Swanson et al. 2020), and in other mammal models (Gòdia et al. 2020). Therefore, there is a need for additional studies to determine the main characteristics of the sperm microbiome and its possible influence on male fertility. In fact, the coexistence of bacterial species and other microorganisms such as viruses in semen samples of fertile men is a matter of concern that awaits exploration.

In this context, the present work represents a pilot study that provides a characterization of the bacterial and viral cargo of sperm samples detected in a population of healthy fertile donors through the analysis of non-human sperm poly(A) RNA-seq reads.

Results

All the analyzed sperm RNA samples fulfilled the requirements evaluated through the quality control analyses, indicating the absence of contamination of sperm DNA or DNA/RNA from non-sperm cells.

From the total reads analyzed, 6.27% were aligned by the GAIA software (n=1,560,701.17) (**Supplemental Table 1**). Of them, a mean of 94.32% sequences aligned with the human database, 3.26% belonged to bacteria, 2.40% sequences corresponded to viruses, and 0.01% sequences matched with archaea.

According to the RNA abundance level of each microbiome (expressed in percentage of Operational Taxonomic Units; OTU), a total of 16 elements (species, genera, families,

orders, or domains) surpassed the threshold (>1% OTU in at least one sample; **Table 1**). Of these 16 categories, nine corresponded to viruses and seven to bacteria.

The most prevalent groups of viruses, according to the amount of aligned RNA (highest mean percentage of OTU) referred to the *Herpesviridae* family and order (12.25% and 5.04% respectively), the species *Gallid alphaherpesvirus 2* (0.85%), the genus *Roseolovirus* (0.64%), *Proteus phage VB PmiS-Isfahan* (0.54%), *Guanarito mammarenavirus* (0.38%), *Hepacivirus C* (0.32%) and *Bacillus phage Stitch* (0.30%). Regarding bacteria, the species with the highest RNA quantification was *Escherichia coli* (0.46%), followed by a set of unknown species from the *Enterobacteriaceae* family (0.36%), *Bacillum megaterium* (0.36%), *Cutibacterium acnes* (0.28%), *Staphylococcus simulans* (0.23%) and *Gardnerella vaginalis* (0.13%).

The hierarchical clustering of the samples (**Figure 1A**) revealed four main group distributions: Cluster 1 included samples 1-2, 4 and 10; Cluster 2 included samples 6-9; Cluster 3 included samples 5, 11 and 12 and finally, sample 3 was clustered apart and was categorized independently as Cluster 4. The Principal Component Analysis (PCA) of the data (**Figure 1B**) revealed a similar sample distribution when Principal Component 1 (PC1) and Principal Component 2 (PC2) were represented.

The mean percentage of OTU and standard deviation were estimated for each cluster. **Table 2** contains these descriptive parameters corresponding to the 16 microbiome elements. In order to study the homogeneity and variability of each cluster, the number of samples that presented more than 1% OTU of the described species were identified and used to define the characteristics of the groups of samples **(Figure 2).** Cluster 1 displayed a homogeneous presence of *Herpesviridae* (order and family) since they were detected in three out of the four samples. It also presented a set of unknown viruses and bacteria. Cluster 2 showed a very homogeneous profile as all samples presented

Herpesvirales (family and order), *Gallid alphaherpesvirus 2,* and *Roseolovirus*, besides a fraction of unknown viruses. Cluster 3 displayed a heterogeneous profile since *Proteus phage VB PmiS-Isfahan, Escherichia coli, Guanarito mammarenavirus, Enterobacteriaceae* and *Cutibacterium acnes* were only detected in single samples. Finally, Cluster 4 (formed by one sample) displayed the presence of *Bacillus megaterium, Bacillus phage Stitch, Cutibacterium acnes, Escherichia coli, Gallid alphaherpesvirus 2, Gardnerella vaginalis, Guanarito mammarenavirus, Hepacivirus C* and *Proteus phage VB PmiS-Isfahan.*

Discussion

The influence of microbiome in fertility has been highlighted by many authors through the last decades. Nevertheless, although the microbial population of the female genital tract has been widely characterized, new studies are still needed to delve into the seminal and sperm microbiome. Knowing the physiological microbiome of semen is an additional step towards fully understanding its functions. Therefore, the results described in this article provide new data to a field that has been poorly explored up to now.

The analysis of the whole transcriptome by RNA-seq has advantages and limitations over other common strategies such are those based on the identification of specific bacterial 16S rRNA (Liu et al. 2014; Weng et al. 2014; Monteiro et al. 2018; Baud et al. 2019). A complete RNA-seq profiling of the microbiome allows a more in-depth characterization of the abundance of any species present in a given sample. Therefore, the analysis of the resulting amplicons is not limited to specific 16S rRNA files but can be compared with complete microbiome databases (e.g. GAIA in the present study) and lead to a more species-sensitive detection. Lastly, this technique provides the possibility of integrating

the sequencing of bacterial and viral RNA from the same sample in a single assay, which avoids possible biases from the combination of different experimental procedures.

As a limitation of the study, it should be mentioned that the RNA-seq library was constructed through poly(A) captured sequences. RNA-seq studies can be achieved by using a previous rRNA depletion or through a poly(A)-tail selection. As it is well known, both options entail some pros and cons, and both processes selectively omit distinct sets of RNAs. As commented before, by using this strategy we had assumed that, since polyadenylated transcripts represent only a fraction of the whole set of bacterial mRNAs (Sarkar 1997), only a specific portion of the transcriptome was taken into account in the analysis.

In line with previous studies, we have identified high variability between samples regarding expression levels and species content, which suggests that variability could be considered as an inherent feature of the sperm microbiome. Many factors, such as diet, hygiene practices, geographic location, circumcision status (Tomaiuolo et al. 2020), abnormal seminal parameters (Chen et al. 2018; Monteiro et al. 2018) and even sexual intercourse (Hou et al. 2013; Mändar et al. 2018), have been associated with changes in the bacterial semen profile, leading to variability among samples. Nevertheless, we cannot rule out the possibility that the reported differences could instead be related to methodological differences. For instance, since the pH and the molecular composition of seminal fluid favor microorganismal growth, it is crucial to collect, store and manipulate semen samples under specific and uniform conditions. Otherwise, some differences could be attributable to bacterial contamination and growth rather than a real variation in the 'physiological' sample's microbiome. Moreover, in a comparison between studies, the sample treatments employed to isolate the analyzed fraction should also be considered. In this sense, while some articles analyzed the total semen fraction without a

centrifugation step (Hou et al. 2013; Chen et al. 2018; Baud et al. 2019). In other studies, samples were centrifuged to yield either the cell fraction (Weng et al. 2014; Mändar et al. 2017; Monteiro et al. 2018; Swanson et al. 2020) or the seminal fluid (Liu et al. 2015). Nevertheless, in these cases, neither the centrifugal force nor the time of centrifugation were comparable between studies, which represents another possible source of variation in the obtained outcomes.

In our series of results, even though all samples were collected from donors who satisfied a pre-stablished and well-defined list of inclusion criteria (see Materials and methods), and despite the fact that samples were collected, stored, and manipulated following the exact same procedure, the sperm microbiome profiles displayed certain differences. Nevertheless, it is important to highlight that all the identified bacteria have been described in other seminal microbiome characterization studies: *Escherichia coli* (Weidner et al. 1991; Kiessling et al. 2008; Hou et al. 2013; Weng et al. 2014; Swanson et al. 2020), *Bacillus* (Swanson et al. 2020), *Cutibacterium* (Swanson et al. 2020), *Staphylococcus* (Weidner et al. 1991; Kiessling et al. 2008; Hou et al. 2013; Weng et al. 2014; Swanson et al. 2020) and *Gardnerella* (Weng et al. 2014; Mändar et al. 2015; Mändar et al. 2017). The high prevalence of *Lactobacillus* in other semen studies (Kiessling et al. 2008; Hou et al. 2013; Weng et al. 2014), which has been previously associated with a good sperm quality (Farahani et al. 2020), stands out its absence in the present results. As indicated in the previous paragraph, controversial findings between studies may be influenced by differences in sample obtainments, treatments, and analyses. Regarding the viruses, although *Herpesviridae* and *Hepacivirus* have also been identified in seminal fluid before (Salam and Horby 2017), other viruses such as *Guanarito mammarenavirus*, and the phages *Proteus phage VB PmiS-Isfahan* and *Bacillus phage Stitch* have been detected in this study for the first time. When we

analyzed if the corresponding host bacteria described for these two phages (*Proteus mirabilis* and *Bacillus thuringiensis*, respectively) were present in the sperm bacterial cargo, we realized that they were not. We believe that these results can be explained by two not mutually exclusive facts. Firstly, a poly(A) transcript capture has been used when constructing our libraries. Poly(A) is associated with only 2-60% of the molecules of a given mRNA species (Sarkar 1997). In this sense, the absence of a certain transcript could either be attributed to a lack of expression (real absence) or a missing detection. Secondly, despite current models define most phages as host-specific, the use of more sophisticated analysis in this area has started emerging doubts regarding such a paradigm (Ross et al. 2016; de Jonge et al. 2019). That is, the use of less direct methods of analysis has allowed considering that the currently stablished "host range" could be wider for some phages.

Finally, in five individuals we have also identified the presence of the *Gallid alphaherpesvirus 2*, a virus that affects poultry health (also known as Marek's disease virus). Although such finding might seem surprising at first, the presence of this chicken virus in human sera has been already described in previously published studies (Laurent et al. 2001). In fact, zoonosis events appear to be especially common among herpesvirus (Tischer and Osterrieder 2010). This order has been observed to potentially infect almost all animal species (insects, fish, mollusks, reptiles, birds, and mammals including humans), and its components share several properties that potentially make them capable of crossing species barriers (Woźniakowski and Samorek-Salamonowicz 2015). Nevertheless, there are also controversial data about this topic in the literature (Hennig et al. 2003). We believe that more studies in this area will eventually help to clarify the potential threat that represent these zoonotic infections.

Our study has also allowed identifying some specific profiles through hierarchical cluster analysis. The presence of microbiome clusters in sperm samples has been described

before (Hou et al. 2013; Weng et al. 2014) suggesting that, despite the existence of individual differences, some sperm samples tend to share a common microbiome profile. In this case, Cluster 1 displayed a predominant presence of *Herpesvirus,* a sexually transmitted agent that can be found in the male genital tract and in seminal samples, mainly during herpes recurrences (Dejucq and Jégou 2001). Cluster 2, despite also containing a strong *Herpesvirus* presence (including *Gallid alphaherpesvirus 2)*, showed a generally low bacterial composition. In its turn, Cluster 3 contained a higher variability of both bacteria and viruses. *Cutibacterium acnes* is a bacillus typically found in human skin or the digestive tract (Dréno et al. 2018). The phage *VB PmiS-Isfahan* was characterized by Yazdi et al. as an active agent against the bacterium *Proteus mirabilis*, one of the most common causes of complicated urinary tract infections (Yazdi et al. 2019). *Guanarito mammarenavirus* is the source of the Venezuelan hemorrhagic fever, a zoonotic human illness (Tesh et al. 1994). Lastly, Cluster 4 contained a wider range of bacterial and viral species. Among them, the presence of *Gardnerella vaginalis* may be indicative of a bidirectional microbiome transmission occasioned by sexual intercourse (Mändar et al. 2015). Besides, the presence of *Gardnerella* in semen samples has been associated to normal sperm parameters (Weng et al. 2014). Nevertheless, it is necessary to keep into account that the abundance of species in Cluster 4 could be due to an overrepresentation given by the lack of more samples in this group.

All in all, we cannot rule out the fact that the size of the population is a limitation of the study. Our results come from a homogeneous population that fulfilled all requirements established in the study design. These strict criteria were intended to guarantee the proven fertility of the whole cohort and thus the “normality” of the analyzed sperm microbiome. Therefore, our work must be considered as a pilot study that provides new information for the characterization of viral and bacterial RNA cargo of human sperm cells. This will

be a starting point for further studies to start focusing on how alterations in the composition of this community of microorganisms could eventually modify the reproductive capacity of individuals. We believe that further studies in groups of patients selected for presenting different reproductive handicaps of idiopathic etiology should be of great interest in order to assess the microbiome impact in their fertility hindrances.

From all these compiled data, it becomes evident that the influence of the semen microbiome on male fertility requires additional work. Further studies must be performed under standardized criteria of sample procurement, manipulation, and analysis with two main objectives: first, to define the physiological microbiome of semen samples; and second, to figure out the different fingerprints that can be discerned in different human populations and their implications in reproductive health.

Materials and methods

Sample collection and processing

Raw sperm RNA-seq data were selected from a previous publication of our group (Corral-Vazquez et al. 2021) (NCBI BioProject accession number of the raw data: PRJNA573604). These data were obtained from a cohort of 12 fertile normozoospermic donors with proven fertility, normal karyotypes, no previous exposure to any genotoxic agent, and no history of chemotherapy, radiotherapy, or chronic illness. Their average age was 24.08 years (range: 19–31 years).

Briefly, semen samples were collected from the donors by masturbation after 3-7 days of sexual abstinence, and were centrifuged at 2500 X g for 10 min. The cell fraction was incubated in a lysis buffer (0.1% Sodium Dodecyl Sulfate and 0.5% Triton X-100 in milliQ water) in order to lysate those possible somatic cells present in the ejaculate (Goodrich et al. 2007). This was further verified by optical microscopic examination.

Afterwards, sperm samples were subjected to RNA extraction, and several quality controls were performed: RNA purity was assessed (full-spectrum UV-Vis spectrophotometer NanoDrop© 2000, Thermo Fisher Scientific); the size distribution of the isolated RNA molecules was evaluated (RNA 6000 Pico chip and the Agilent DNA High Sensitivity Bioanalyzer, Thermo Fisher Scientific); and the absence of RNA from non-sperm cells was verified by checking that no peaks corresponding to 28S and 18S rRNAs were present. Finally, by performing RT-PCR for *PRM1, GAPDH*, and *CDH1* genes (High-Capacity cDNA Reverse Transcription kit, Thermo Fisher) we confirmed the lack of RNA from somatic or germs cells as well as the absence of DNA contamination.

RNA sequencing

The Universal Plus mRNAseq with the NuQuant kit (NuGEN) was used to construct cDNA libraries with an initial poly(A) RNA selection. Libraries were sequenced in paired-end 125 bp mode using the HiSeq 2500 platform (Illumina). The Bcl2Fastq 2.0.2 version of the Illumina pipeline was used to produce raw data and to perform de-multiplexing. The raw sequencing data were deposited in NCBI BioProject under accession number PRJNA573604.

Data analysis and statistics

The quality of the raw reads was assessed with FASTQC v.0.11.8, and then trimming and clipping were performed using BBDuk by setting a minimum base quality of 1 and a minimum read length of 35 bp. The obtained reads were aligned to the microbiome genome and transcriptome and analyzed using the GAIA software (www.metagenomics.cloud) (Paytuví et al. 2019). For this purpose, several databases were employed: Metatranscriptomics (16S, 18S and Internal Transcribed Spacer);

Prokaryotes (Whole Genome Sequencing [WGS] and Whole Transcriptome Sequencing [WTS], release 2020); and Viruses (WGS and WTS, release 2020). The Human WGS and WTS databases (release 2020) were used to filter out human-specific reads. Species were identified and classified in any of the non-eukaryote databases according to the alignment analysis. When this alignment did not allow identification of an organism to the taxonomic rank of species, the detected features were classified into less specific ranks: genus, families, orders, or domains. The RNA abundance levels of each microbiome element were expressed in percentage of Operational Taxonomic Unit (OTU). The percentages of OTU were calculated by dividing the quantified reads of the microbiome element by the total number of reads in its taxonomic rank. Therefore, the different identified species were comparable with each other, and also the different genus, families, etc.

Subsequent statistical analyses were performed using the R statistical program in RStudio (RStudio Team 2015). Based on previous empirical testing and professional advice, the identified microbial species were filtered to remove possible background amplifications, so a threshold of 1% OTU in at least one sample was established. To identify possible differentiated microbiome patterns among the samples, a Hierarchical Cluster Analysis (HCA) (Ward method) and a PCA were performed. For the total population and for each cluster individually, the mean percentage of OTU, the standard deviation, and the number of samples with more than 1% OTU were estimated.

Ethics approval

Written informed consent was obtained from all donors. The study was approved by the Ethics Committee on Animal and Human Experimentation of the Universitat Autònoma de Barcelona (ref. CEEAH1884).

Funding

This work was supported by Agència de Gestió d'Ajuts Universitaris i de Recerca, Generalitat de Catalunya, under Grant 2017/SGR-503; Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, Gobierno de España, under Grant PI21/00564; and Universitat Autònoma de Barcelona under Grant UAB CF-180034, UAB/PIF2015.

Disclosure of interests

The authors report there are no competing interests to declare.

Author contributions

Conceived and designed the experiments: CC-V, EA, FV, JB, RAC, ZS; Performed the experiments: CC-V; Analyzed the data: CC-V, RAC; Contributed reagents/ materials/ analysis tools: RAC; Wrote the manuscript: CC-V; Final edit of paper: EA, FV, JB, RAC, ZS.

References

Altmäe S, Franasiak JM, Mändar R. 2019. The seminal microbiome in health and disease. Nat Rev Urol. 16(12):703–721.

Baud D, Pattaroni C, Vulliemoz N, Castella V, Marsland BJ, Stojanov M. 2019. Sperm microbiota and its impact on semen parameters. Front Microbiol. 10(FEB):1–9.

Chen H, Luo T, Chen T, Wang G. 2018. Seminal bacterial composition in patients with obstructive and non-obstructive azoospermia. Exp Ther Med. 15(3):2884–2890.

Corral-Vazquez C, Blanco J, Aiese Cigliano R, Sarrate Z, Rivera-Egea R, Vidal F, Garrido N, Daub C, Anton E. 2021. The RNA content of human sperm reflects prior events in spermatogenesis and potential post-fertilization effects. Mol Hum Reprod. 27(6):gaab035.

D'Argenio V, Salvatore F. 2015. The role of the gut microbiome in the healthy adult status. Clin Chim Acta. 451:97–102.

Dejucq N, Jégou B. 2001. Viruses in the Mammalian Male Genital Tract and Their Effects on the Reproductive System. Microbiol Mol Biol Rev. 65(2):208–231.

Dréno B, Pécastaings S, Corvec S, Veraldi S, Khammari A, Roques C. 2018. Cutibacterium acnes (Propionibacterium acnes) and acne vulgaris: a brief look at the latest updates. J Eur Acad Dermatology Venereol. 32:5–14.

Farahani L, Tharakan T, Yap T, Ramsay JW, Jayasena CN, Minhas S. 2020. The semen microbiome and its impact on sperm function and male fertility: A systematic review and meta-analysis. Andrology.(April 2020):115–144.

Gòdia M, Ramayo-Caldas Y, Zingaretti LM, Darwich L, López S, Rodríguez-Gil JE, Yeste M, Sánchez A, Clop A. 2020. A pilot RNA-seq study in 40 pietrain ejaculates

to characterize the porcine sperm microbiome. Theriogenology. 157:525–533.

Goodrich RJ, Johnson GD, Krawetz S. 2007. The preparation of human spermatozoal RNA for clinical analysis. Arch Androl. 53(3):161–167.

Hennig H, Osterrieder N, Müller-Steinhardt M, Teichert HM, Kirchner H, Wandinger KP. 2003. Detection of Marek's disease virus DNA in chicken but not in human plasma. J Clin Microbiol. 41(6):2428–2432.

Hou D, Zhou X, Zhong X, Settles ML, Herring J, Wang L, Abdo Z, Forney LJ, Xu C. 2013. Microbiota of the seminal fluid from healthy and infertile men. Fertil Steril. 100(5):1261–1269.

de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. 2019. Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends Microbiol. 27(1):51–63.

Kiessling AA, Desmarais BM, Yin HZ, Loverde J, Eyre RC. 2008. Detection and identification of bacterial DNA in semen. Fertil Steril. 90(5):1744–1756.

Koedooder R, Mackens S, Budding A, Fares D, Blockeel C, Laven J, Schoenmakers S. 2019. Identification and evaluation of the microbiome in the female and male reproductive tracts. Hum Reprod Update. 25(3):298–325.

Laurent S, Esnault E, Dambrine G, Goudeau A, Choudat D, Rasschaert D. 2001. Detection of avian oncogenic Marek's disease herpesvirus DNA in human sera. J Gen Virol. 82:233–240.

Liu CM, Osborne BJW, Hungate BA, Shahabi K, Huibner S, Lester R, Dwan MG, Kovacs C, Contente-Cuomo TL, Benko E, et al. 2014. The Semen Microbiome and Its Relationship with Local Immunology and Viral Load in HIV Infection. PLoS Pathog. 10(7):e1004262.

Liu Y, Niu M, Yao C, Hai Y, Yuan Q, Liu Y, Guo Y, Li Z, He Z. 2015. Fractionation of human spermatogenic cells using STA-PUT gravity sedimentation and their miRNA profiling. Sci Rep. 5:8084.

Mändar R, Punab M, Borovkova N, Lapp E, Kiiker R, Korrovits P, Metspalu A, Krjutškov K, Nlvak H, Preem JK, et al. 2015. Complementary seminovaginal microbiome in couples. Res Microbiol. 166(5):440–447.

Mändar R, Punab M, Korrovits P, Türk S, Ausmees K, Lapp E, Preem JK, Oopkaup K, Salumets A, Truu J. 2017. Seminal microbiome in men with and without prostatitis. Int J Urol. 24(3):211–216.

Mändar R, Türk S, Korrovits P, Ausmees K, Punab M. 2018. Impact of sexual debut on culturable human seminal microbiota. Andrology. 6(3):510–512.

Monteiro C, Marques PI, Cavadas B, Damião I, Almeida V, Barros N, Barros A, Carvalho F, Gomes S, Seixas S. 2018. Characterization of microbiota in male infertility cases uncovers differences in seminal hyperviscosity and oligoasthenoteratozoospermia possibly correlated with increased prevalence of infectious bacteria. Am J Reprod Immunol. 79(6):1–9.

Paytuví A, Battista E, Scippacercola F, Cigliano RA, Sanseverino W. 2019. GAIA: an integrated metagenomics suite. bioRxiv.

Ross A, Ward S, Hyman P. 2016. More is better: Selecting for broad host range bacteriophages. Front Microbiol. 7:1352.

RStudio Team. 2015. RStudio: Integrated Development for R. http://www.rstudio.com/

Salam AP, Horby PW. 2017. The Breadth of Viruses in Human Semen. Emerg Infect Dis. 23(11):1922–1924.

Sarkar N. 1997. Polyadenylation of mRNA in prokaryotes. Annu Rev Biochem. 66:173–197.

Sender R, Fuchs S, Milo R. 2016. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 14(8):1–14.

Swanson GM, Moskovtsev S, Librach C, Pilsner JR, Goodrich R, Krawetz SA. 2020. What human sperm RNA-Seq tells us about the microbiome. J Assist Reprod Genet. 37(2):359–368.

Tesh RB, Jahrling PB, Salas R, Shope RE. 1994. Description of Guanarito Virus (Arenaviridae: Arenavirus), the Etiologic Agent of Venezuelan Hemorrhagic Fever. Am J Trop Med Hyg. 50(4):452–459.

Tischer BK, Osterrieder N. 2010. Herpesviruses - a zoonotic threat? Mol Cell Biochem. 140(3–4):266.

Tomaiuolo R, Veneruso I, Cariati F, D'argenio V. 2020. Microbiota and human reproduction: The case of male infertility. High-Throughput. 9(2):10.

Weidner W, Jantos C, Schiefer HG, Haidl G, Friedrich HJ. 1991. Semen parameters in men with and without proven chronic prostatitis. Syst Biol Reprod Med. 26(3):173–183.

Weng SL, Chiu CM, Lin FM, Huang WC, Liang C, Yang T, Yang TL, Liu CY, Wu WY, Chang YA, et al. 2014. Bacterial communities in semen from men of infertile couples: Metagenomic sequencing reveals relationships of seminal microbiota to semen quality. PLoS One. 9(10):e110152.

Woźniakowski G, Samorek-Salamonowicz E. 2015. Animal herpesviruses and their zoonotic potential for cross-species infection. Ann Agric Environ Med. 22(2):191–194.

Yazdi M, Bouzari M, Ghaemi EA. 2019. Genomic analyses of a novel bacteriophage (VB_PmiS-Isfahan) within Siphoviridae family infecting Proteus mirabilis. Genomics. 111(6):1283–1291.

Figure Captions

Figure 1. Sample homogeneity analysis based on their microbiome profile.

A) Dendrogram of the hierarchical clustering of the samples based on the OTU (Operational Taxonomic Unit) profiles of the identified species. Y-axis of the dendrogram represents the Euclidean distance between clusters.

B) Principal Component Analysis (PCA) of the samples based on the OTU profiles of the identified species. Principal Component 1 (PC1) refers to the first dimension of the analysis (X axis of the PCA) and Principal Component 2 (PC2) represents the second dimension (Y axis of the PCA).

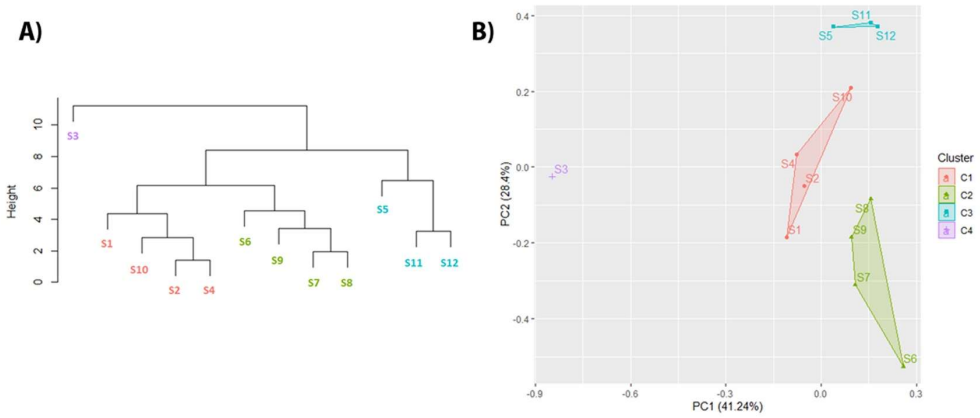
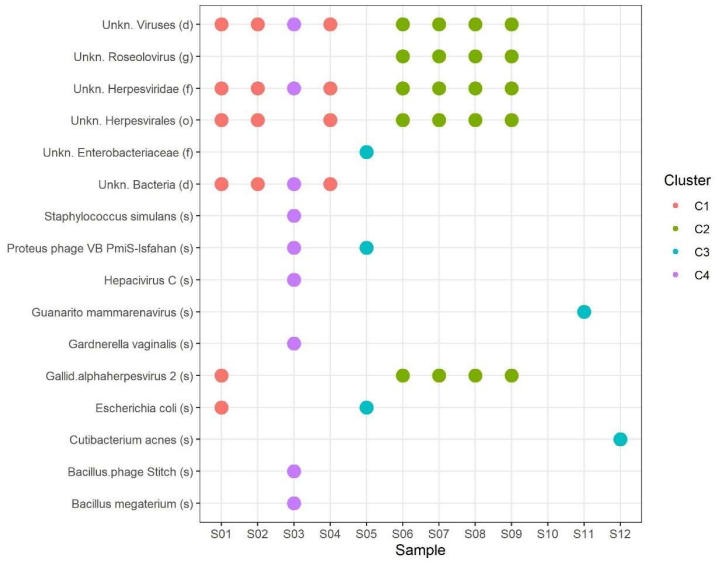


Figure 2. Microbiome profiling of the samples included in each cluster

Species present in the microbiome of each sample displaying a minimum OTU (Operational Taxonomic Unit) of 1% that allow visualizing the similarities in the composition of each cluster.

o = order; f = family; g = genus; s = species.



Supplemental Material Captions

Supplemental Table 1. Number and percentage of reads corresponding to each sample before GAIA alignment, and after their aligning to the corresponding taxonomic lineage.

Table 1. RNA abundance of the 16 Operational Taxonomic Units (OTU) detected in the 12 analysed sperm RNA samples (S).

Database	OTU	S1 %OTU	S2 %OTU	S3 %OTU	S4 %OTU	S5 %OTU	S6 %OTU	S7 %OTU	S8 %OTU	S9 %OTU	S10 %OTU	S11 %OTU	S12 %OTU	Mean %OTU	SD %OTU
Virus	Unknown *Herpesvirales* (o)	20.27	13.11	0.24	3.33	0.01	40.68	36.48	23.18	9.6	0.15	0	0	12.25	14.79
Virus	Unknown *Herpesviridae* (f)	5.59	3.26	2.38	4.09	0	10.13	7.96	5.27	21.61	0.18	0	0	5.04	6.18
Virus	Unknown Viruses (d)	4.57	3.44	2.48	2.15	0.09	11.75	3.47	1.75	2.89	0.34	0.01	0.07	2.75	3.22
Prokaryote	Unknown Bacteria (d)	1.64	1.26	4.04	1.53	0.16	0.33	0.92	0.21	0.79	0.47	0.02	0.04	0.95	1.13
Virus	*Gallid alphaherpesvirus 2* (s)	1.37	0.87	0.17	0.35	0	2.84	1.93	1.34	1.36	0.02	0	0	0.85	0.93
Virus	Unknown *Roseolovirus* (g)	0.97	0.52	0.02	0.35	0	2.08	1.64	1.06	1.1	0	0	0	0.64	0.72
Virus	*Proteus phage VB PmiS-Isfahan* (s)	0.2	0.35	2.89	0.09	1.52	0.17	0.16	0.47	0.18	0	0.17	0.31	0.54	0.84
Prokaryote	*Escherichia coli* (s)	2.17	0.14	0.45	0.18	2.19	0.04	0.11	0.11	0.12	0.02	0	0.01	0.46	0.81
Virus	*Guanarito mammarenavirus* (s)	0.16	0.22	0.49	0.11	0.22	0.04	0.09	0.12	0.46	0	1.81	0.86	0.38	0.51
Prokaryote	Unknown *Enterobacteriaceae* (f)	0.01	0	0.09	0.03	4.09	0.01	0	0.04	0.02	0	0.01	0.01	0.36	1.18
Prokaryote	*Bacillus megaterium* (s)	0.84	0.56	1.7	0.32	0.04	0.08	0.32	0.06	0.21	0.13	0	0	0.36	0.49
Virus	*Hepacivirus C* (s)	0.56	0.44	1.34	0.54	0.03	0.11	0.33	0.12	0.28	0.08	0	0	0.32	0.38
Virus	*Bacillus phage Stitch* (s)	0.53	0.33	1.41	0.55	0.03	0.08	0.19	0.08	0.3	0.07	0	0	0.3	0.4
Prokaryote	*Cutibacterium acnes* (s)	0.27	0.11	0.08	0.01	0.01	0.47	0.06	0.09	0	0.02	0.46	1.83	0.28	0.52
Prokaryote	*Staphylococcus simulans* (s)	0.41	0.4	1.05	0.29	0.06	0.05	0.22	0.09	0.13	0.08	0	0	0.23	0.29
Prokaryote	*Gardnerella vaginalis* (s)	0	0	1.49	0	0	0	0	0	0	0	0.02	0.01	0.13	0.43

Abundance is expressed in percentages (%) of OTU. The mean and standard deviation (SD) of the % OTU are indicated.

o = order; f = family; g = genus; s = species.

Table 2. RNA abundance of the 16 Operational Taxonomic Units (OTU) detected in each sample cluster.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	Total	
Sample	Mean	SD	Mean	SD	Mean	SD	Mean	Mean	SD
Unknown *Herpesvirales* (o)	9.21	9.21	27.49	14.06	0.00	0.00	0.17	12.25	14.79
Unknown *Herpesviridae* (f)	3.28	2.28	11.24	7.19	0.00	0.00	0.09	5.04	6.18
Unknown Viruses	2.63	1.81	4.97	4.58	0.06	0.05	0.02	2.75	3.22
Unknown Bacteria	1.23	0.53	0.56	0.34	0.07	0.08	0.24	0.95	1.13
Gallid alphaherpesvirus2 (s)	0.65	0.59	1.87	0.70	0.00	0.00	1.70	0.85	0.93
unkn. *Roseolovirus* (g)	0.46	0.40	1.47	0.49	0.00	0.00	0.08	0.64	0.72
Proteus phage VB PmiS-Isfahan (s)	0.16	0.15	0.24	0.15	0.66	0.74	1.05	0.54	0.84
Escherichia coli (s)	0.63	1.03	0.10	0.03	0.73	1.26	2.38	0.46	0.81
Guanarito mammarenavirus (s)	0.12	0.09	0.18	0.19	0.97	0.80	1.41	0.38	0.51
unkn. *Enterobacteriaceae* (f)	0.01	0.01	0.02	0.02	1.37	2.36	0.45	0.36	1.18
Bacillus megaterium (s)	0.46	0.31	0.17	0.12	0.01	0.02	4.04	0.36	0.49
Hepacivirus C (s)	0.40	0.22	0.21	0.11	0.01	0.02	1.34	0.32	0.38
Bacillus phage Stitch (s)	0.37	0.22	0.16	0.10	0.01	0.02	2.89	0.30	0.40
Cutibacterium acnes (s)	0.10	0.12	0.16	0.21	0.77	0.95	2.48	0.28	0.52
Staphylococcus simulans (s)	0.29	0.15	0.12	0.07	0.02	0.04	0.49	0.23	0.29
Gardnerella vaginalis (s)	0.00	0.00	0.00	0.00	0.01	0.01	1.49	0.13	0.43

Mean and standard deviation (SD) of the RNA abundance (expressed in percentages of OTU) are displayed.

o = order; f = family; g = genus; s = species.