A Shotgun Metagenomic Mining Approach of Human Semen Microbiome

Janaina Aderaldo
Federal University of Rio Grande do Norte
Diego Teixeira Teixeira
Federal University of Rio Grande do Norte
Mychelle Garcia Torres
Federal University of Rio Grande do Norte
Beatriz Albuquerque
Federal University of Rio Grande do Norte
Maryana Oliveira
Federal University of Rio Grande do Norte
Paulo Eduardo Soares
Federal University of Rio Grande do Norte
Lucymara Agnez-Lima
Federal University of Rio Grande do Norte
Ana Rafaela Timoteo
Federal University of Rio Grande do Norte
Rita Silva-Portela
Federal University of Rio Grande do Norte
Daniel Lanza (danielclanza@gmail.com)
Federal University of Rio Grande do Norte

Research Article

Keywords: seminal microbiome, WGS, shotgun, HERV, Plasmodium

Posted Date: March 16th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1220437/v1

Abstract

The testicular environment is immunoprivileged to protect germ cells from autoimmune and anti-inflammatory activities, but, on the other hand, it is susceptible to pathogens. Until now, the works that investigated the microbiological diversity in this environment were restricted to analyzing specific bacteria or viral communities. In this study, we evaluated the diversity in the seminal human microbiome using a whole-genome sequencing approach in a seminal human pool. For this, we collected 50 samples donated by participants from a public reproductive health service in Brazil. We observed a high proportion of the Bacteria domain (71.3%), whose largest groups are *Bacillus, Staphylococcus*, *Mycobacterium*, and *Streptococcus*. The Eukaryotic domain (27.6%) comprises *Plasmodium*, *Trypanosoma*, and *Trichinella*. Viruses (1.1%) are composed of Gammaretrovirus, Herv-K, and Herv-W. These findings expand the current view of microbial diversity in human semen and point out that evaluating uncultivated pathogens could be crucial before concluding reproductive and prophylactic treatments. In addition, the Herv families identified in seminal samples deserve studies with a functional and evolutionary perspective. These data contribute to identifying potential pathogens present in the semen (gamma diversity) and their correlations and opening a new front for research in the diagnosis of fertility-related diseases.

Highlights

- This study is the first to use the WGS technique on seminal human samples, allowing the prospection of gamma diversity focusing on the most relevant pathogens.
- Identifying different taxa from a single analysis is valuable for samples with low amounts of exogenous DNA.
- The approach used here to identify bacterial cluster is compatible with the targeted molecular techniques (16s rRNA), identifying major genera above 0.1% of abundance.
- The presence of *Plasmodium* and *Trypanosoma* indicates potential prophylactic treatment beyond the usual bacterial.
- Identifying retroviruses in the male reproductive system deserves attention and exclusive studies.

Introduction

The mammalian testis is an immunologically privileged environment that protects germ cells from systemic immune attacks and has effective local innate immunity [1]. This immune privilege is mediated by a complex and still unclear combination of cellular structures, hormones, and cytokines that, together with innate immunity, act as a double-edged sword [1–3]. On the other hand, this tissue can be infected by microbial pathogens derived from the circulating blood or ascends to the genitourinary tract, influenced by lifestyle, age at sexual debuts, and geographic localization [4]. It is known that pathogens can cause various disorders such as inflammation of the tissues, obstruction of the genital ducts, epididymitis and orchitis, propensity to cancer, and infertility [5–9].

Compared to other sites in the body, the seminal microbiota has been minimally investigated [10,11]. Research on the potential pathogenic effects on male health has progressed slowly for two main reasons: a) the scarcity of studies with standardized methods to enable comparisons between results of different works [4,10,12,13]; and b) prophylactic antibiotic therapy [14]. Understanding the fundamental aspects described above can provide new insights into developing prevention and treatment approaches for male infertility related to testicular inflammation [3].

The limited number of existing studies on the male reproductive tract microbiome has shown conflicting results. Most studies use techniques that identify only cultivable microorganisms which represent approximately 1% of the microbiota in environments [15]; and all studies published until now focus on pre-defined groups such as bacterial or viral communities [2,4,6,9,16]. This kind of approach gains importance considering that the increase in global infertility, morbidity, and mortality caused by sexually transmitted diseases – (STDs), are assigned to bacteria (350 million new annual cases of four types of curable STDs chlamydia, gonorrhea, syphilis, and trichomonas) and viruses (417 million incurable viral infections like herpes and HPV) [17,18].

Among the modern molecular techniques, next-generation sequencing - NGS allows identifying culture-independent microorganisms more precisely, faster and also providing information about interactions between assemblies of microorganisms and their transcriptomes [7,9,10,13,19,20]. It includes metagenomics analysis (mNGS), a sequencing method that can detect the genetic material of all organisms present in specified samples [15]. This hypothesis-free diagnostic approach can contribute to elucidate specific microbial communities generating significant changes in clinical routine [21].

We characterize the seminal microbiome using the whole-genome sequencing technique (WGS) approach. We analyzed the seminal microbiome of 50 fertility treatment participants at the Reproductive Rights Center of the Januário Cicco Maternity School - MEJC, the university hospital of the Federal University of Rio Grande do Norte - UFRN, Brazil. This technique allowed sequencing all genomes presented, including host-integrated microorganisms [13], providing an unbiased view of the microbiota's phylogeny, regional particularities, and functional composition strata [22]. Our data contribute to identifying potential pathogens present in the semen (gamma diversity) and opening a new front for research in the diagnosis of fertility-related diseases.

Materials And Methods

The national ethics committee (approval number 3.043.526 provided by the Conep - *Comissão Nacional de Ética em Pesquisa*) approved the study protocol, and all participants provided written informed consent. The research has been performed following the Declaration of Helsinki.

Samples

Seminal samples were collected from 50 (fifty) participants who performed spermogram in the Reproductive Rights Center of the Januário Cicco Maternity School - MEJC, a university hospital of Federal University of Rio Grande do Norte - UFRN, Brazil. A spermogram for each sample was performed to investigate marital infertility and include results that show azoospermia to normospermia (Supplementary Table S1). A 1.5 mL aliquot of semen from each patient was stored in liquid nitrogen by slow freezing without cryopreservation.

DNA extraction

The extraction of genomic DNA was carried out in triplicate by preparing a single "pool" of 50 seminal samples. For this, a volume of 200 µl of each sample was used to obtain a final volume of 10 mL. Thus, 1 mL of the pool was centrifuged at 1,600 rpm for 3 minutes, and the precipitate obtained was directly used for extraction with the PureLink ™ Genomic DNA kit (Invitrogen) to the manufacturer's instructions. To increase the final concentration of the extracted DNA, the elution step was modified using only 25 µL of the eluting reagent.

DNA quantification was performed by fluorimetry, using the Quantus™ Fluorometer (Promega), which indicated a concentration equal to 123 ng/µL. The DNA sample was analyzed in the Genomics Nucleus - NUGEN of the Molecular and Genomic Biology Laboratory - LBMG for shotgun sequencing performed on Ion Torrent system (Ion PGM).

Ion Torrent PGM sequencing

The total DNA was quantified with Qubit Fluorometer (Thermo Fisher). The library was generated according to the Ion Xpress Plus Fragment Library kit protocol (Thermo Fisher) with some adjustments. The total DNA (193ng) was fragmented by the Ion Shear Plus Reagents Kit (Thermo Fisher). After this enzymatic lysis, the sample was linked to the P1 and Ion Xpress barcode adapters, concomitantly to a nick repaired to connect the adapter and DNA insert. The selection of fragments by size was performed using the E-Gel SizeSelect II (Thermo Fisher).

The amplification reaction was performed using the following cycling profile: 95°C for 5 minutes; followed by ten cycles of 95°C for 15s, 58°C for 15s and 70°C for 60s, in a final volume of 130µL. We used magnetic beads on Agencourt® AMPure® XP Kit (Beckman Coulter) for the purification step. The template preparation was conducted using the Ion PGM™ Template OT2 Kit (Thermo Fisher), and 25pM of the library was added to an aqueous master mix containing Ion Sphere Particles (ISPs) at the manufacturer's specified proportions.

The emulsion was produced by Ion OneTouch^{TM} 2 System (Thermo Fisher). Then, DNA-positive ISPs were recovered, enriched, and annealed to the sequencing primer, and the sequencing polymerase was added before loading ISPs into Ion 318 sequencing chip, according to the manufacturer's protocol. The sequencing procedure was conducted according to the Ion PGM™ 400 Sequencing Kit Protocol (Thermo

Fisher) including 850 nucleotide flows to deliver 400 base read lengths, on average. The library was sequenced on Ion Torrent Personal Genome Machine.

Data Analysis

First, we performed a quality assessment, trimmed poor bases, and removed low-quality reads using FastQC [66] and Trimmomatic [67] Softwares. Trimmomatic ran under the single-end mode, keeping reads 75bp longer with a sliding window of 4 bases, cutting out those with Phred-Score below 20. We then blasted all reads against the NCBI non-redundant protein database (downloaded at 2020/Aug) using the Diamond [68] software applying the blastx algorithm under default parameters. Next, we imported the hits table into the MEGAN (v.6.18.0) [69] using the Megan-map file dated 2020/Jul, under default Last Common Ancestor - LCA algorithm parameters. We followed with a series of filtering steps to remove hits that had matched with Amniota, Viridiplantae, Arthropoda, and Archea, using scripts written in R. After the filtering process, a new hit table was imported again into the MEGAN to recalculate the number of assigned reads for each taxon, recalculate LCA and to build.

We opted to represent the taxonomic classification analysis at the gender level to ensure greater accuracy in the results presented. We used percentages instead of the absolute number of mapped reads to minimize the artifacts generated in the sequencing process for community comparison purposes. The data are available in Bio Project ID PRJNA741509. To validate the identity of bacterial reads, we compared our results to five reference studies that studied the seminal microbiota using the 16S technique. To validate the identity of eukaryotic reads, we mapped them in NCBI RefSeq.

To validate the identity of viral reads, we used two complementary methods: map to reference and pairwise alignment in Refseqs based on contigs (De novo generated). In cases where the methods returned different IDs, the returned strings were aligned for similarity assessment. In cases of discrepancy, these were aligned to read, and the most similar was considered. The main results were checked manually using BlastX on their Refseq or chromosome assembly. All *in silico* steps was performed in the High Performance Processing Center of the Federal University of Rio Grande do Norte - NPAD/UFRN.

Results

In total, 1,624,764 reads had hits to any protein from NCBI's nr database; from these, most hits were attributed to the genus Homo (~ 99%). The upset plot (Figure 1) demonstrates the classification of the reads found in this study. The most significant number of reads that did not show redundancies between taxonomic groups belong to the Bacteria group, followed by Aconoidasida, Saccharomycetes, and viruses. The reads of the Mammalia and Actinopteri groups were disregarded from this point of the analysis.

A higher proportion of hits for the Bacteria domain was observed (71.3%), followed by the Eukarya domain (27.6%) and viruses (1.1%), considering only identifications with high homology and well-

annotated strings (Figure 2A). The phyla that showed the greatest diversity of genera were Firmicutes, Proteobacteria, Actinobacteria, and Chlamydiae. The most abundant genera comprise four bacterial and one protist (Figures 2B and C).

Considering only bacterial genera (n = 87), the most abundant groups were *Bacillus* (8.2%) and *Staphylococcus* (7.2%), followed by *Mycobacterium* (6.5%), *Streptococcus* (6.1%), *Escherichia* (5.1%) and *Enterococcus* (5.1%) (Figure 2C). Most of the groups found (> 60%) comprise organisms previously reported as pathogens (Figure 3A).

We compare our results with similar studies to evaluate the total number of genera representing above 0.1% (Figure 3B). This comparison revealed that the most genera identified by our approach presented abundance greater than 0.1%. This value is similar to that observed by Wang et al. [20] and higher than that observed in other studies. Even considering the different approaches used for target identification (16SrRNA V1-V2, V3-V6, and V4), the largest number of genera with representation greater than 0.1% was observed using WGS. (Figure 3C).

The reads from eukaryotes comprise six genera: the two most abundant, *Plasmodium* (64.4%) *and Trichinella* (21.2%), and other less represented, *Hanseniospora* (5.4%), *Caenorhabditis* (5.3%), *Trypanosoma* (3.2%), and *Trichuris* (0.6%). The relative abundance of identified Eukarya genera was presented in Figure 2A. These findings were confirmed *in silico* by assembling the reads into reference sequences (Table 1).

Table 1
Additional validation of eukaryotic and viral reads

Identity read	Sequences hits	Max sequence lengths	Reference sequence	Reads assembled to the reference sequence	% of mapped reads
Eukaryotic					
Plasmodium	76,581	385	NC_009919.1	53,061	69,29%
Trichinella	25,163	379	JYDN01000002.1	22,987	91,35%
Hanseniospora	6,370	376	NG_021422.1	6,028	94,63%
Caenorhabditis	6,281	374	MH590409.1	5,572	88,71%
Trypanosoma	3,809	372	HE573027.1	2,187	57,42%
Viral					
Herv-K	1,508	369	Y17832.2	1,275	84,55%
Herv-W	1,703	375	AF009668.1	1,082	63,54%

Viruses represented the group with the lowest abundance of genera (1.1% of the total amount of hits) comprising Human endogenous retroviruses (HERVs) and Gammaretrovirus. HERV-K, HERV-W were

identified with high percentual coverage (Figure 4). These results were also confirmed *in silico* by mapping in reference sequences (Table 1).

Discussion

The high depth sequencing and the proposed pipeline allow the identification of different taxonomic groups from only one analysis, allowing inferences about the presence of several pathogenic microorganisms in a shorter time. This approach allowed us to observe a broader range of taxonomic groups, indicating the occurrence of the most common pathogens or others that so far had not been described in human semen.

In addition to the diversity of pathogenic bacteria (Figure 3A), the number of reads for *Plasmodium*, *Trypanosoma*, and *Trichinella* caught our attention. Although we have exhausted *in silico* confirmations, additional *in vitro* analyzes are usually required to confirm whether these eukaryotic pathogens are present in the samples. As seen in Figure 1, the reads of some eukaryotic groups showed redundancy with other taxonomic groups. This is common in this type of analysis, given the complexity of these eukaryotic genomes. In any case, we believe that our objective of using WGS as a prospective technique has been met since obtaining many results from a few analyzes speeds up the diagnosis. To achieve this same number of taxa adopting other approaches, combining techniques and most samples would be necessary.

An unexpected finding was that bacterial genera identified by WGS were better represented when compared to other classical approaches for bacterial identification. As no other studies are using WGS in seminal human samples, the most recent meta-analysis of the seminal human microbiome identified four studies were used 16s rRNA techniques [10,20,23,24], and, although not the same we used, they are culture-independent and seek to identify non-target organisms within the proposed group [9]. From these, just three presented data for comparison [20,23,24]. Our pipeline identified a significant number of bacterial genera above 0.1% of the total genera identified (Figures 3B-C). The coverage and similarity of genera identified by WGS concerning bacterial-target techniques indicate that the gain in identifying the other phyla outweighs the loss of target-specificity. In this way, despite the predominance of the abundance of nucleic acids dominated by the host's background [21], WGS could be an interesting strategy for bacterial analysis.

It is noteworthy that differences in results are expected even when participants and techniques are maintained due to the natural dynamic population balance. The results are consistent regarding technique, even respecting this premise, as prospecting from primers covering the V3-V4 region has confirmed more organisms than the primers for the V1-V2 region [25]. The disadvantage of lower coverage must be balanced with the advantage of avoiding false positives depending on the objective of each study. In this intuit, we adopt rigorous processing of reads, as Phred Score 20.0 and reads with references to exclusive groups were adopted, as shown in the upset plot.

The identification Eukarya group was specifically more challenging due to the homology between gene sequences. For this reason, we confirmed the findings for Eukarya and virus groups by aligning to reference sequences in BlastX in order to decrease the risk of false positives due to sequence homology. This step is crucial since a higher number of different regions of the microorganism's genome reduces the chance of a false positive.

Reports on eukaryotic pathogens in human semen are scarce, with only reviews available [26,27]. Our results presented two dominant genera: *Plasmodium* (64.4%), *Trichinella* (21.2%).

The *Plasmodium* was the organism that presented the higher number of hits identified, already described as potential agents in infertility [27]. The first report of the destructive potential of *Plasmodium* in male human fertility was published in 1987 [28] and, since then, researches have shown a reduction in testosterone levels, an increase in cortisol, decreased in the ratio of T-helper to T-suppressor cytotoxic cells, decreased sperm motility[29], decreased sperm count and also adverse effect to antimalarial drugs [27,28,30]. The impact of this genus in reproductive capacity is considered harmful in mice and may cause congenital transmission, lower pregnancy, reduced fertility, increased abortion, increased neonatal mortality, overproduction of inflammatory cytokines (tumor necrosis factor - α), and degeneration testicular [27,31,32].

About *Trichinella*, as much as they belong to different taxonomic families, helminths share the ability to modulate the host's immune response directed at themselves and bystander antigens, such as vaccines and allergens, with both advantageous and disadvantageous consequences [33,34]. The coexistence of parasites affects the host organism uniquely [35], and, like the microbiome, this joint action requires further studies. *Trichinella* is the only known genus of the family Trichinellidae, and a minor human nematode parasite and the largest of intracellular parasites [36], with larvae already identified in body fluid and organ of the body, including lymph nodes, urine, placenta, mammary gland tissue, milk, skin, and virtually every tissue [37,38]. Pawlowski [39] claims that "many aspects of clinical trichinosis remain unknown or vague due partly to the limited possibilities for studying trichinosis in man."

Hanseniaspora represented 5.4% of the eukarya identified; however, their pathogenesis and behavior remain unclear [40]. Despite rare findings, *H. uvarum* has already been identified in nails [41], oral cavity [42], and epithelial lesion [40]. Jankowski *et al.* [40] cite a finding in vaginal discharge, but we have not retrieved the data reliably. Batista *et al.* [43] identified *Hanseniospora valbyensis* in a patient's appendectomy sample and a report of onychomycosis by *Hanseniospora* in 1928 in the German medical literature. Using advanced detection/identification methods, the list of emerging opportunistic infections by unusual fungi is expanding rapidly worldwide [41].

The genus *Caenorhabditis* is a genus of taxonomy still under construction, with at least eight species described only in the last decade [44]; some are not even formally classified [45]. *C. elegans*, one of the most used models in research directed to human health for having 30-60% of the orthologous genes or strong

homology with mammals [46], are identified. However, as this nematode has a free life, we consider the finding an artifact because most of the hits for amniotic have been removed and allocated to this species.

We consider the identification of *Trypanosoma* relevant, despite containing only 3.2% of the reads for Eukarya. Although few studies exist in humans, these point out that during trypanosomiasis infection, sterility or infertility, menstrual disorder, loss of libido, impotence, amenorrhea, and degeneration of seminiferous tubules and testis may occur [47,48]. There is evidence that the pathogen also causes specific damage to the hypothalamic-pituitary axis [27,48]; however, the mechanisms of action are unknown.

The considerable number of reads associated with retroviruses (5,051) draws attention. It leads us to reflect on the potential incorporation of the viral genome into the germ cell genome (and the consequent risk of transmission to offspring), already commented on by Dejucq & Jégou as worrisome [6]. We consider the hypothesis that it is possible to have genetic variability benefits in this incorporation. It is important to note that the classification was performed by sequence homology comparison against a public database, whose data on virus-related genomes are incomplete and with a large number of "unknown" sequences since despite being the most abundant biological entities on the planet, it is estimated that only 1% of viral sequences are recorded in the reference databases [49,50].

About 8% of the human genome comprises a material remnant from viral infections known as endogenous human retroviruses (HERVs) [51]. These elements were acquired during the evolution process by vertical inheritance, reaching ~203,000 copies in the human genome, and it is assumed that about 30% of this material is active and transcribed elements [52]. It is common to infer that HERVs only cause harm to the host. However, it is necessary to remember that they are expressed at low levels in all human tissues and can provide potential benefits to their hosts, such as the hypothesis that their cooptation by vertebrates prevents infection by other related exogenous viruses [53]. A relevant example is the cooptation of an endogenous retrovirus envelope gene that started to form the syncytiotrophoblast during pregnancy [54] and its active expression in embryogenesis [51,52]. Particularly the testicles and the placenta appear to be privileged tissues for the expression of HERV [55].

The sequence mapped to Y17832.2, HERV –K (C7), which is believed to be an allelic variant (YIDD-to-CIDD mutation) of a proviral sequence that carries all the ORFs that supposedly express a non-functional RT[56]. The HERV-K family is the most recently acquired [54] and the most active transcription [51]. Almost a third of the proviruses in this family represent specific human inserts, of which 48% are polymorphisms [52]. The high expression of HERV-K envelope proteins in placental cytotrophoblast cells suggests their potential involvement in placentogenesis and pregnancy [54,57–59], with many copies with full open reading frames (ORFs) transcribed and translated, especially in the initial embryogenesis [52] and differentiation of human fetal tissues [57]. Unlike the HERV-K family, open reading frames encode functional proteins, including a fusogenic glycoprotein attributed to normal placental development. [60–62]. However, it is unknown whether it represents an exogenous retrovirus with closely related endogenous elements or an

endogenous replication-competent, virion-producing provirus [63]. Also, the mechanisms by which neuroinflammation occurs are still unclear [64]. The Gammaretrovírus presented 28% of identified viral reads. It is also known as onco-retroviruses for its leukemia-inducing properties and transducing properties in stem cells and progenitors [65].

Our findings allow considering the use of prophylactic protocols not only for bacteria but also for eukaryotes. We consider that the conclusions about favorable or unfavorable seminal microbiotas need more attention because a) the actual microbiota potentially present in the semen is unknown, and b) there is no statistically solid evidence or methods reproduced in a multicenter way. We believe that our efforts, coupled with multicenter prospective efforts added to machine learning, will allow for the elucidation of the functional microbiome of the male reproductive system and will bring a new view to fertility parameters and clinical practice.

Declarations

ACKNOWLEDGMENTS

We want to thank all sample donors for participating in this study in the Januário Cicco Maternity Hospital School of the Federal University of Rio Grande do Norte – MEJC/UFRN. We want to thank the High Performance Processing Center of the Federal University of Rio Grande do Norte - NPAD/UFRN to allow us to access their computer facilities.

AUTHOR CONTRIBUTIONS STATEMENT

Conceived and designed the experiments: JFA, DCFL. Performed the experiments: JFA, MMGT BHDRA, MTFCO. Analyzed the data: JFA, DGT, PETS, DCFL. Contributed reagents/materials/analysis tools: LFAL, ARST, RCBSP. Wrote the paper: JFA, DGT, DCFL. All authors reviewed the manuscript.

FUNDING

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

COMPETING INTERESTS STATEMENT

The author(s) declare no competing interests.

References

1. Zhao, S., Zhu, W., Xue, S. & Han, D. Testicular defense systems: Immune privilege and innate immunity. Cellular and Molecular Immunology vol. 11 428–437 (2014).
2. Fijak, M., Bhushan, S. & Meinhardt, A. Immunoprivileged sites: the testis. Methods in molecular biology (Clifton, N.J.) vol. 677 459–470 (2011).

3. Li, N., Wang, T. & Han, D. Structural, cellular and molecular aspects of immune privilege in the testis. Frontiers in Immunology vol. 3 152 (2012).
4. Tomaiuolo, R., Veneruso, I., Cariati, F. & D'argenio, V. Microbiota and human reproduction: The case of male infertility. High-Throughput vol. 9 (2020).
5. Kurscheidt, F. A. *et al.* Effects of Herpes Simplex Virus Infections on Seminal Parameters in Male Partners of Infertile Couples. Urology **113**, 52–58 (2018).
6. Dejucq, N. & Jégou, B. Viruses in the Mammalian Male Genital Tract and Their Effects on the Reproductive System. Microbiol. Mol. Biol. Rev. **65**, 208–231 (2001).
7. Le Tortorec, A. *et al.* From Ancient to Emerging Infections: The Odyssey of Viruses in the Male Genital Tract. Physiol. Rev. **100**, 1349–1414 (2020).
8. Gimenes, F. *et al.* Male infertility: a public health issue caused by sexually transmitted pathogens. Nat. Rev. Urol. **11**, 672–687 (2014).
9. Farahani, L. *et al. The semen microbiome and its impact on sperm function and male fertility: A systematic review and meta-analysis*. *Andrology* andr.12886 (Blackwell Publishing Ltd, 2020). doi:10.1111/andr.12886.
10. Baud, D. *et al.* Sperm Microbiota and Its Impact on Semen Parameters. Front. Microbiol. **10**, 234 (2019).
11. Altmäe, S., Franasiak, J. M. & Mändar, R. The seminal microbiome in health and disease. Nature Reviews Urology vol. 16 703–721 (2019).
12. Duffy, J. M. N. *et al.* A protocol developing, disseminating and implementing a core outcome set for infertility. *Hum. Reprod. Open* 2018, (2018).
13. Koedooder, R. *et al.* Identification and evaluation of the microbiome in the female and male reproductive tracts. Hum. Reprod. Update **25**, 298–325 (2019).
14. Kroon, B., Hart, R. J., Wong, B. M., Ford, E. & Yazdani, A. Antibiotics prior to embryo transfer in ART. Cochrane Database Syst. Rev. (2012) doi:10.1002/14651858.cd008995.pub2.
15. Mardanov, A. V., Kadnikov, V. V. & Ravin, N. V. Metagenomics: A Paradigm Shift in Microbiology. in *Metagenomics: Perspectives, Methods, and Applications* 1–13 (Elsevier Inc., 2018). doi:10.1016/B978-0-08-102268-9.00001-X.
16. Kumar, N. & Singh, A. Trends of male factor infertility, an important cause of infertility: A review of literature. Journal of Human Reproductive Sciences vol. 8 191–196 (2015).
17. Boivin, J., Bunting, L., Collins, J. A. & Nygren, K. G. International estimates of infertility prevalence and treatment-seeking: Potential need and demand for infertility medical care. Hum. Reprod. **22**, 1506–1512 (2007).
18. World Health Organization. WHO | Global health sector strategy on Sexually Transmitted Infections, 2016-2021. *WHO* (2016).
19. Chiu, C. Y. & Miller, S. A. Clinical metagenomics. Nature Reviews Genetics vol. 20 341–355 (2019).

20. Weng, S.-L. L. *et al.* Bacterial Communities in Semen from Men of Infertile Couples: Metagenomic Sequencing Reveals Relationships of Seminal Microbiota to Semen Quality. PLoS One **9**, e110152 (2014).
21. Gu, W., Miller, S. & Chiu, C. Y. Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. Annu. Rev. Pathol. Mech. Dis. **14**, 319–338 (2019).
22. Hozzein, W. N. *Metagenomics - Basics, Methods and Applications*. *Metagenomics - Basics, Methods and Applications* (IntechOpen, 2019). doi:10.5772/intechopen.78746.
23. Hou, D. *et al.* Microbiota of the seminal fluid from healthy and infertile men. Fertil. Steril. **100**, 1261 (2013).
24. Monteiro, C. *et al.* 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**, (2018).
25. Graspeuntner, S., Loeper, N., Künzel, S., Baines, J. F. & Rupp, J. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. Sci. Reports 2018 *81* **8**, 1–7 (2018).
26. Martínez-García, F. *et al.* Protozoan infections in the male genital tract. J. Urol. **156**, 340–349 (1996).
27. Nourollahpour Shiadeh, M., Niyyati, M., Fallahi, S. & Rostami, A. Human parasitic protozoan infection to infertility: a systematic review. Parasitology Research vol. 115 469–477 (2016).
28. Singer, R. *et al.* Decreased semen quality in a male infected with malaria. Int. J. Androl. **10**, 685–689 (1987).
29. Wiwanitkit, V. Malaria and semen quality: An issue in reproductive health? J. Hum. Reprod. Sci. **4**, 58 (2011).
30. Zei, G., Lisa, A. & Astolfi, P. Fertility and malaria in sardinia. Ann. Hum. Biol. **17**, 315–330 (1990).
31. Vincendeau, P. & Bouteille, B. Immunology and immunopathology of African trypanosomiasis. An. Acad. Bras. Cienc. **78**, 645–665 (2006).
32. Schuster, J. P. & Schaub, G. A. Experimental chagas disease: The influence of sex and psychoneuroimmunological factors. Parasitol. Res. **87**, 994–1000 (2001).
33. Gazzinelli-Guimaraes, P. H. & Nutman, T. B. Helminth parasites and immune regulation [version 1; peer review: 2 approved]. *F1000Research* vol. 7 1685 (2018).
34. Maizels, R. M. & McSorley, H. J. Regulation of the host immune system by helminth parasites. J. Allergy Clin. Immunol. **138**, 666–675 (2016).
35. Supali, T. *et al.* Polyparasitism and its impact on the immune system. International Journal for Parasitology vol. 40 1171–1176 (2010).
36. Dick, T. A. Species, and Infraspecific Variation. in *Trichinella and Trichinosis* 31–73 (Springer US, 1983). doi:10.1007/978-1-4613-3578-8_2.
37. Weatherly, N. F. Anatomical Pathology. in *Trichinella and Trichinosis* 173–208 (Springer US, 1983). doi:10.1007/978-1-4613-3578-8_5.

38. Nuzzolo-Shihadeh, L. *et al.* Human trichinosis mimicking polymyositis. International Journal of Infectious Diseases vol. 92 19–20 (2020).
39. Pawłowski, Z. S. Clinical Aspects in Man. in *Trichinella and Trichinosis* 367–401 (Springer US, 1983). doi:10.1007/978-1-4613-3578-8_11.
40. Jankowski, M., Jagielski, T., Misiak, G. & Czajkowski, R. Hand dermatitis with Hanseniaspora uvarum as a plausible causative agent. Postepy Dermatologii i Alergologii vol. 35 641–643 (2018).
41. Karimi, L., Mirhendi, H., Khodadadi, H. & Mohammadi, R. Molecular identification of uncommon clinical yeast species in Iran. Curr. Med. Mycol. **1**, 1–6 (2015).
42. Emmanouil-Nikoloussi, E. *et al.* 'Hanseniaspora uvarum' the ultrastructural morphology of a rare ascomycete, isolated from oral thrush. Bull. du Group. Int. pour la Rech. Sci. en Stomatol. Odontol. **37**, 13–17 (1994).
43. Batista, A. C., Coêlho, R. P. & Vieira, J. R. Da ocorrência de Hanseniaspora valbyensis Kloecker em apêndice cecal humano e lesões epidérmicas. Mycopathol. Mycol. Appl. **12**, 185–188 (1960).
44. Félix, M. A., Braendle, C. & Cutter, A. D. A streamlined system for species diagnosis in caenorhabditis (Nematoda: Rhabditidae) with name designations for 15 distinct biological species. PLoS One **9**, (2014).
45. Kiontke, K. C. *et al.* A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. BMC Evol. Biol. **11**, 339 (2011).
46. Apfeld, J. & Alper, S. What Can We Learn About Human Disease from the Nematode C. elegans? Methods Mol. Biol. **1706**, 53–75 (2018).
47. Bouteille, B. & Buguet, A. The detection and treatment of human African trypanosomiasis. Res. Rep. Trop. Med. **3**, 35 (2012).
48. Ikede, B. O., Elhassan, E. & Akpavie, S. O. Reproductive disorders in African trypanosomiasis: A review. Acta Tropica vol. 45 5–10 (1988).
49. Edwards, R. A. & Rohwer, F. Viral metagenomics. Nat. Rev. Microbiol. **3**, 504–510 (2005).
50. Bzhalava, Z., Tampuu, A., Bała, P., Vicente, R. & Dillner, J. Machine Learning for detection of viral sequences in human metagenomic datasets. BMC Bioinformatics **19**, 336 (2018).
51. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature **522**, 221–246 (2015).
52. Garcia-Montojo, M., Doucet-O'Hare, T., Henderson, L. & Nath, A. Human endogenous retrovirus-K (HML-2): a comprehensive review. Critical Reviews in Microbiology vol. 44 715–738 (2018).
53. Moelling, K. & Broecker, F. The reverse transcriptase-RNase H: From viruses to antiviral defense. Ann. N. Y. Acad. Sci. **1341**, 126–135 (2015).
54. Hohn, O., Hanke, K. & Bannert, N. HERV-K(HML-2), the best preserved family of HERVs: Endogenization, expression, and implications in health and disease. *Frontiers in Oncology* vol. 3 SEP (2013).
55. Pérot, P. *et al.* Microarray-based sketches of the HERV transcriptome landscape. PLoS One **7**, (2012).

56. RR, T. *et al.* Genome-wide screening, cloning, chromosomal assignment, and expression of full-length human endogenous retrovirus type K. J. Virol. **73**, 9187–9195 (1999).
57. Andersson, A. C. *et al.* Developmental expression of HERV-R (ERV3) and HERV-K in human tissue. Virology **297**, 220–225 (2002).
58. Kämmerer, U., Germeyer, A., Stengel, S., Kapp, M. & Denner, J. Human endogenous retrovirus K (HERV-K) is expressed in villous and extravillous cytotrophoblast cells of the human placenta. J. Reprod. Immunol. **91**, 1–8 (2011).
59. Kristensen, M. K. & Christensen, T. Regulation of the expression of human endogenous retroviruses: elements in fetal development and a possible role in the development of cancer and neurological diseases. APMIS **129**, 241–253 (2021).
60. Blond, J.-L. *et al.* Molecular Characterization and Placental Expression of HERV-W, a New Human Endogenous Retrovirus Family. J. Virol. **73**, 1175–1185 (1999).
61. Sverdlov, E. D. Retroviruses and primate evolution. Bioessays **22**, 161–171 (2000).
62. An, D. S., Xie, Y. & Chen, I. S. Y. Envelope Gene of the Human Endogenous Retrovirus HERV-W Encodes a Functional Retrovirus Envelope. J. Virol. **75**, 3488 (2001).
63. Perron, H. *et al.* Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 7583 (1997).
64. Wang, X. *et al.* Human endogenous retrovirus W family envelope protein (HERV-W env) facilitates the production of TNF-α and IL-10 by inhibiting MyD88s in glial cells. Arch. Virol. **166**, 1035–1045 (2021).
65. Hunter, J. E., Ramos, L. & Wolfe, J. H. Viral vectors in the CNS. in *The Curated Reference Collection in Neuroscience and Biobehavioral Psychology* 179–188 (Elsevier Science Ltd., 2016). doi:10.1016/B978-0-12-809324-5.02446-9.
66. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research **7**, 1338 (2018).
67. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**, 2114–2120 (2014).
68. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. Nature Methods vol. 12 59–60 (2014).
69. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. Genome Res. **17**, 377–386 (2007).

Figures

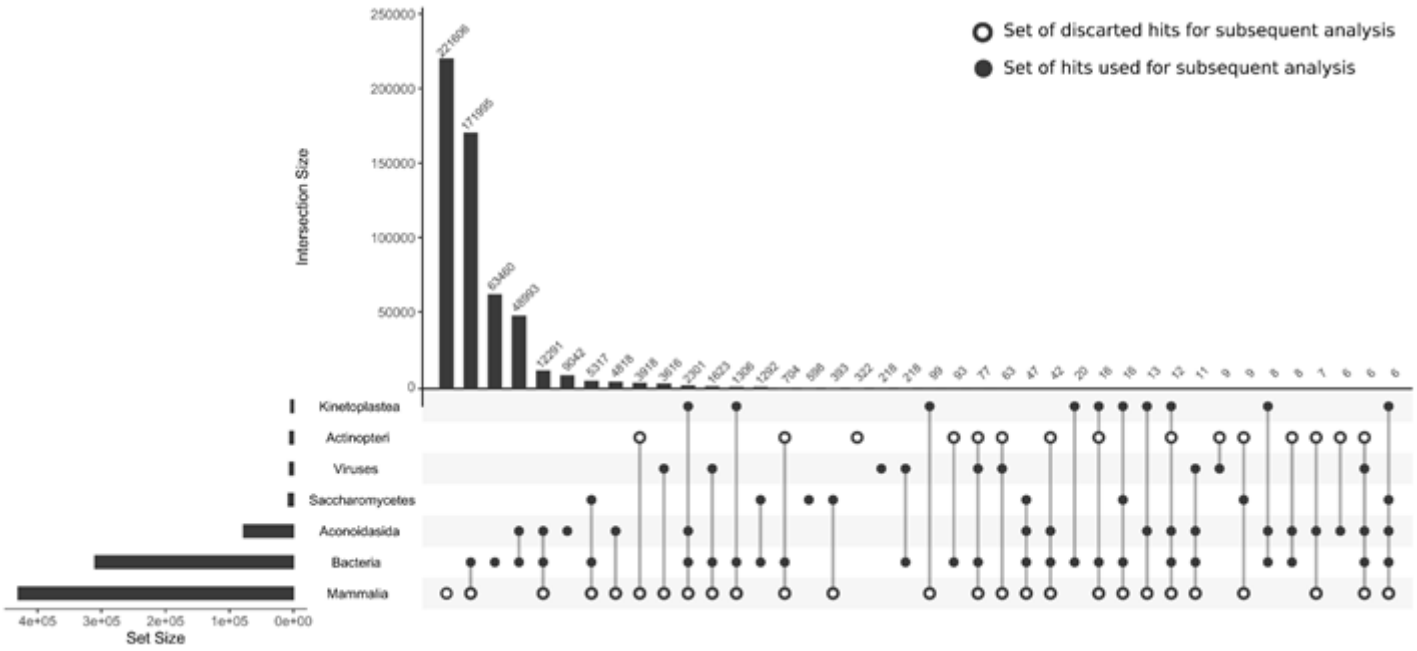

Figure 1

Upset plot showing the number of reads with hits between different taxa.

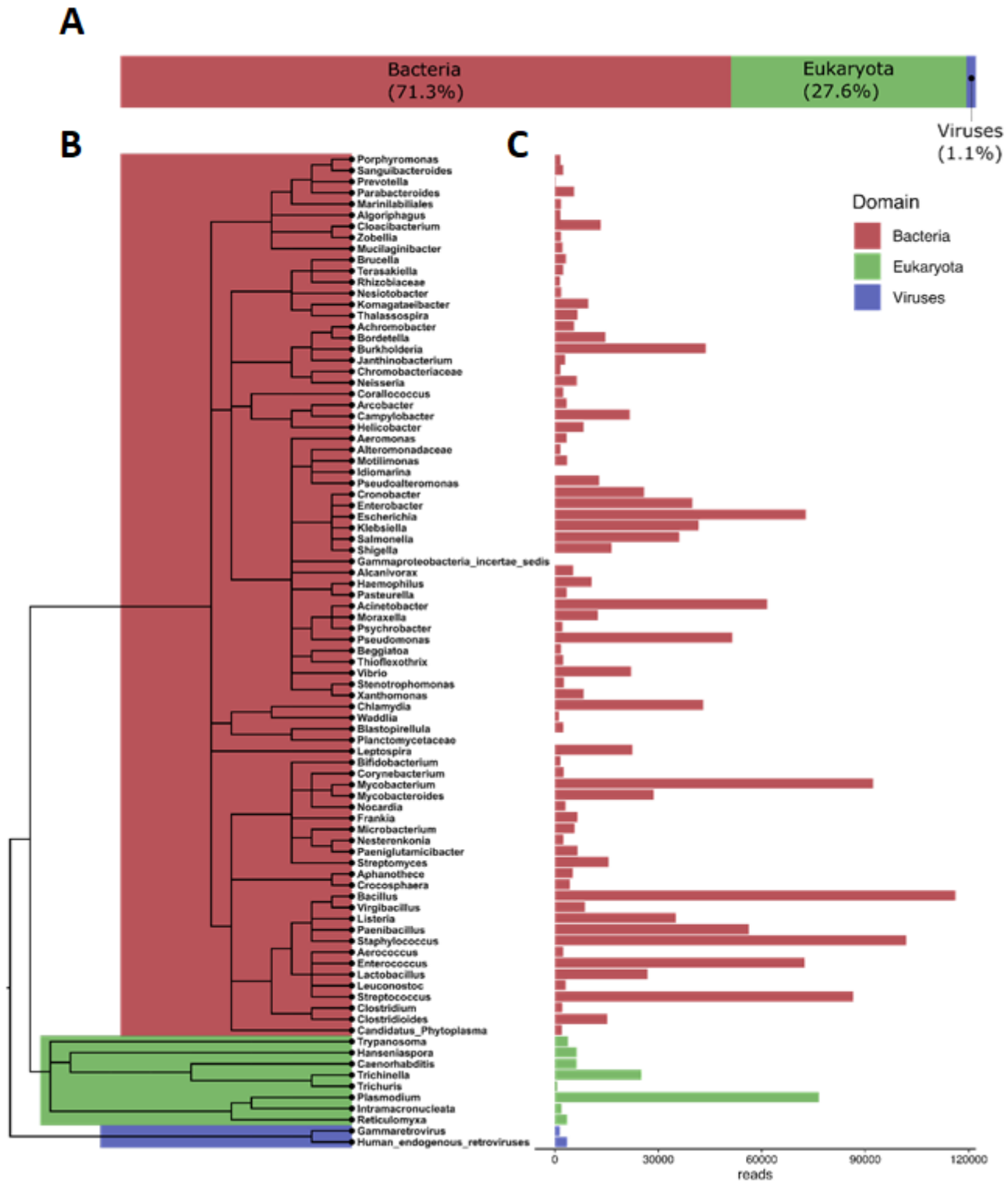

Figure 2

Organism composition into seminal metagenomic data analysis. A) Relative abundance of reads mapped into the Bacteria and Eukaryota domains and Viruses; B) Phylogenetic three indicating the genus of the organisms identified in the sample; C) Number of reads for each organism identified. Images generated by the LCA algorithm into the MEGAN software.

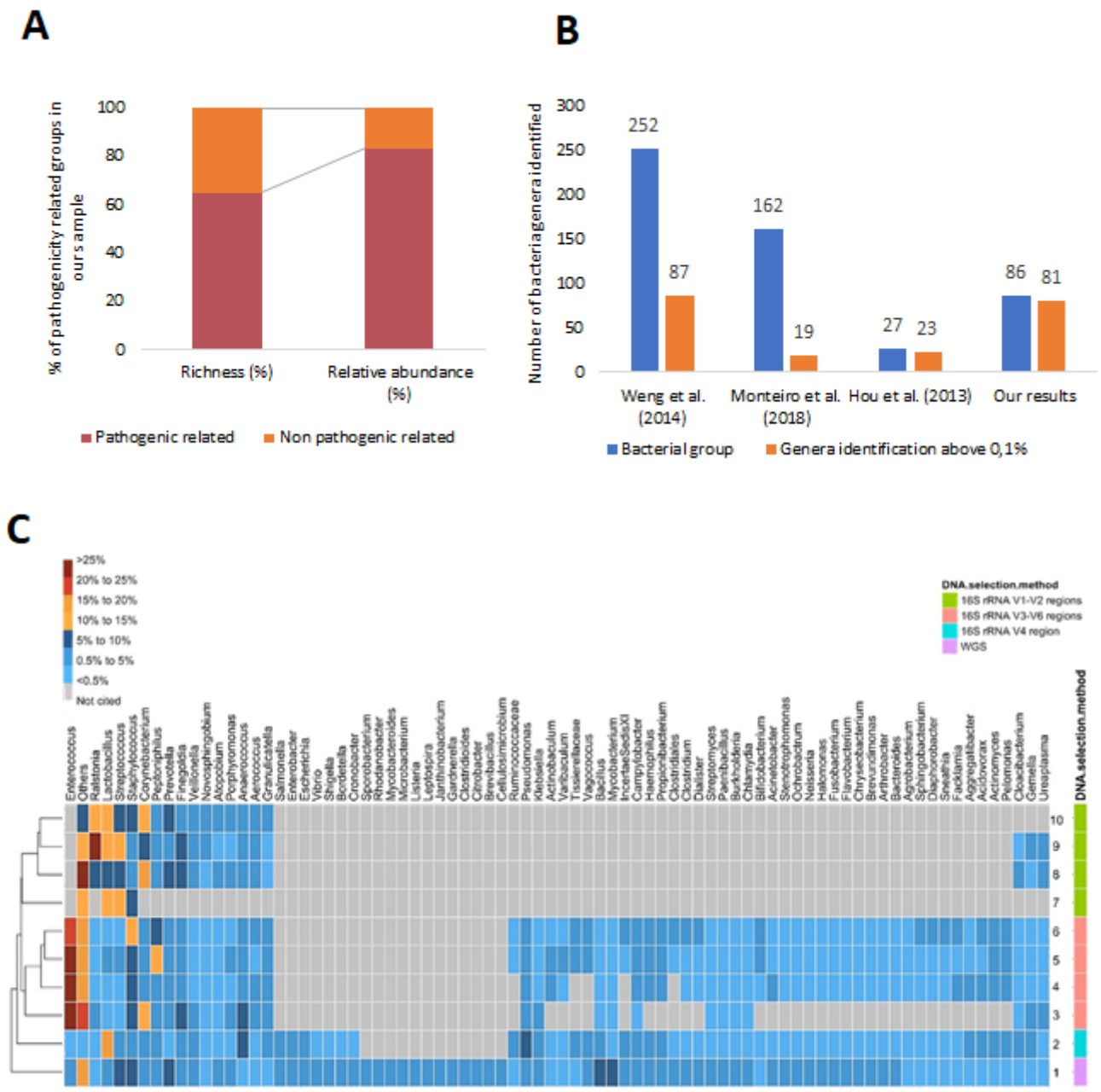

Figure 3

Bacterial metagenomic data. A) Number of pathogenic bacterial groups related in sample. B) Number of bacterial groups identifed above 0.1%. C) Heatmap for % of bacterial genera in reference studies considering different target approaches. 1 - Our data; 2 – Weng *et al.*; 3 – Monteiro *et al.* (H); 4 – Monteiro *et al.* (OAT); 5 – Monteiro *et al.* (C); 6 – Monteiro *et al.* (AT); 7 – Hou *et al.* (AZ); 8 – Hou *et al.* (C); 9 – Hou *et al.* (AT); 10 – Hou *et al.* (OAT).

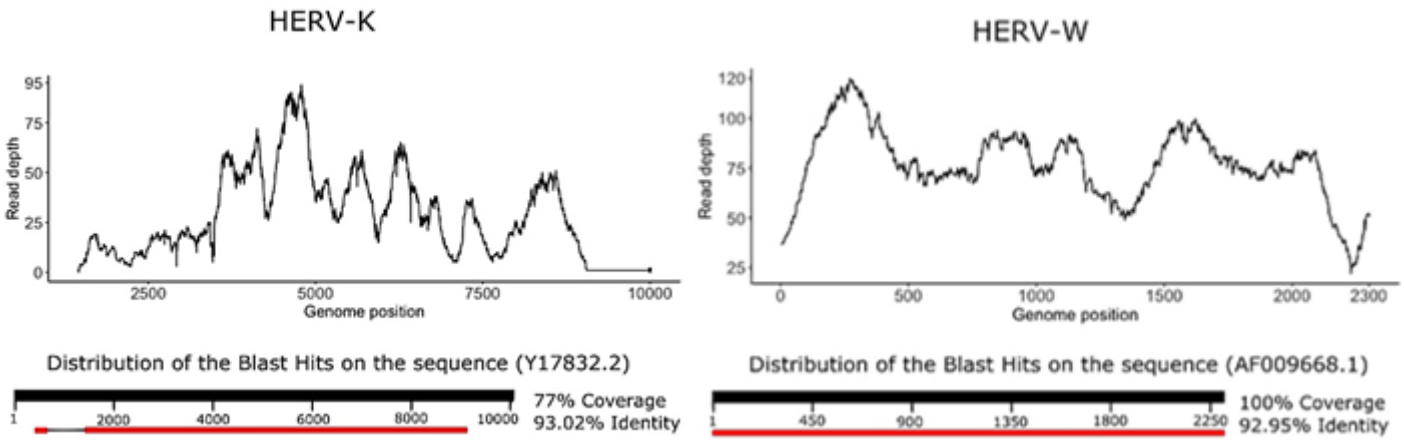

Figure 4

Distribution of the Blast hits on the HERV-K and HERV- W reference sequences.

Supplementary Files

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

- SupplementaryTableS1.Generaldataofcohortselection200122.docx