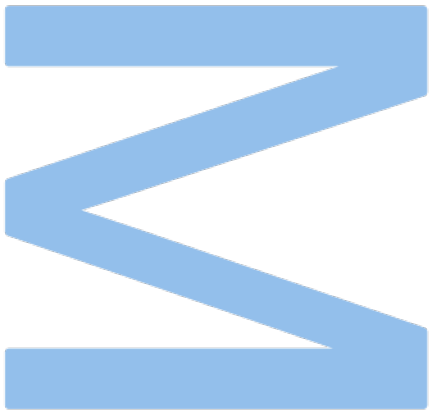
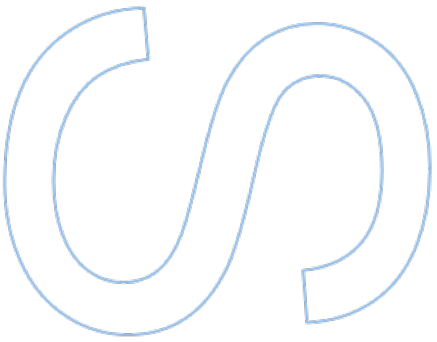
# Using state of the art technologies to characterize the Seminal Microbiome

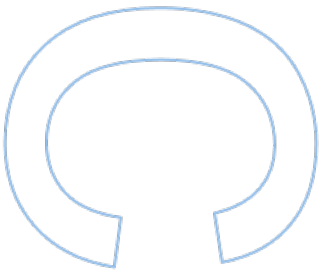
Ivo Manuel Oliveira Pinto
Forensic Genetics Master's Degree
Biology Department
2022

**Supervisor**
Patrícia I. Marques, PhD, Instituto de Investigação e Inovação em Saúde (i3S) e Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP)
**Co-supervisor**
Susana Seixas, PhD, Instituto de Investigação e Inovação em Saúde (i3S) e Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP)


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO







*Dedicated to my father*

# Sworn Statement

I, Ivo Manuel Oliveira Pinto, enrolled in the Master Degree Forensic Genetics at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Ivo Manuel Oliveira Pinto

14th October, 2022

# Agradecimentos

Quero agradecer ao Instituto de Investigação e Inovação em Saúde, por me permitir o desenvolvimento deste trabalho durante o meu segundo ano de mestrado em Genética Forense.

À minha orientadora, Patrícia Isabel Marques, por me integrar tão facilmente com o grupo, e todas as horas dedicadas durante todo o percurso deste trabalho. Além do mais, quero também agradecer pela sua preocupação, sempre disposta a ajudar e pelos conselhos e sugestões que tornaram esta tese possível.

À minha coorientadora, Susana Seixas, por toda a orientação e lições importantes que vou levar para o futuro. Muito obrigado pela dedicação, não só pelo trabalho, mas por todo o conhecimento que espero que me ajudem a moldar e me tornem na pessoa que quero ser.

Ao grupo Genetic Diversity, do i3s, por todo o dinamismo e a facilidade com a qual fui aceite. Particularmente, todas as histórias e anedotas que me contaram, além de serem boas a aliviar a tensão que vem com o trabalho, irão definitivamente ser boas memórias.

Aos meus colegas, professores e amigos do mestrado em genética forense, pelo companheirismo e conhecimento ao longo do caminho.

Aos meus amigos mais próximos, sempre dispostos a ajudar diariamente com o que podem, quer com palavras ou ações: Afonso Mendes, Leonardo Rodrigues, Ricardo Magalhães, Bernardo Francisco, Nuno Monteiro e David Silva. Quero também agradecer em particular à Sílvia Rodrigues, o meu rochedo. Em todas as etapas da minha vida universitária, ela esteve lá, para o bom e para o mau, e eu não seria quem sou sem ela.

Por último, quero agradecer à minha mãe por me proporcionar todas as oportunidades e condições para me moldar como pessoa e estudante.

Obrigado a todos.

# Resumo

O microbioma é atualmente reconhecido por apresentar um importante papel na saúde humana e bem como em várias doenças tais como a doença intestinal inflamatória, a doença de Crohn, a obesidade e o cancro colorretal. No entanto, os estudos sobre o impacto do microbioma no sistema reprodutor e na infertilidade masculina ainda escasseiam e de acordo com o nosso conhecimento, atualmente existem 12 estudos de sequenciação de alta resolução em que se procurou preencher esta lacuna e cujos resultados indicam diferentes repercussões das comunidades bacterianas nos parâmetros de qualidade seminal. É neste âmbito que se enquadra um estudo realizado anteriormente pela nossa equipa em que foi sugerido que o fenótipo de hiperviscosidade seminal pode estar correlacionado com um aumento de Proteobacterias e a oligoastenoteratozoospermia com flutuações das abundâncias relativas de agentes patogénicos e probióticos. No sentido de melhor compreender a ação do microbioma seminal e das suas alterações associadas à doença que justifiquem as diferenças previamente observadas, foi analisada uma coorte de homens estratificada em 15 indivíduos normozoospérmicos (controlos) e 53 casos de infertilidade, incluindo 16 casos de hiperviscosidade seminal (SHV), 24 de oligoastenoteratozoospermia (OAT) e 13 que combinavam ambos os fenótipos (SHV+OAT). Com este objetivo, foi extraído o DNA total de amostras de plasma seminal através do reagentes dos *QIAamp DNA mini kit* e amplificado o gene *16S ribosomal RNA* (*16S rRNA*) por duas metodologias distintas: 1) ensaios de PCR quantitativa (qPCR), implementados para aceder ao conteúdo bacteriano total de cada amostra e estimar o seu *bacterial load*; 2) Experiências de sequenciação de alto rendimento através da utilização do *Ion 16S Metagenomics Kit*, que cobre as regiões hipervariáveis V2, V3, V4, V6-7, V8 e V9, através da plataforma de sequenciação *Ion Torrent S5 XL* and e dos softwares Torrent Suite / Ion Reporter que permitem no final identificar os diferentes *taxa* bacterianos e gerar as correspondentes Operational Taxonomy Units (OTUs). A análise dos dados recolhidos compreendeu a realização de testes estatísticos padrão, de uma curadoria manual das OTUs e de vários outros testes centrados na análise do microbioma disponibilizados através da ferramenta MicrobiomeAnalyst. Em primeiro lugar, os resultados da qPCR mostraram que o conteúdo bacteriano das amostras de plasma seminal é largamente variável, e que pode usado para definir dois grupos tendo em consideração o seu *bacterial load*, alto ou baixo. Surpreendentemente, estes grupos não apresentavam qualquer correlação nem com a condição de infertilidade (casos ou controlos) nem com os fenótipos de infertilidade estudados (SHV, OAT or OAT+SHV).

De seguida, a caracterização taxonómica das amostras revelou que o filo Firmicutes era o mais abundante, independentemente do grupo considerado, casos ou controlos, fenótipo de infertilidade ou *bacterial load*. Por outro lado, os géneros identificados como mais prevalentes foram o *Enterococcus*, *Staphylococcus*, *Streptococcus* e *Facklamia*, embora estes pudessem apresentar elevada variabilidade inter-individual. Além disso as amostras de alto *bacterial load* foram correlacionadas com maiores abundâncias de *Enterococcus*, em contraste com as de baixo *bacterial load*, que foram mais frequentemente associadas com múltiplos taxa, onde era difícil inferir um perfil microbiano comum. Em concordância com estes resultados, os índices de diversidade alfa e beta apenas apresentavam diferenças significativas quanto considerado o *bacterial load* (alto ou baixo). Mais concretamente, amostras com grande abundância de *Enterococcus* e *Staphylococcus* apresentavam uma separação das restantes devido à sua grande similaridade, diversidade reduzida com um alto *bacterial load*. Adicionalmente, algumas diferenças de baixa abundância foram detetadas, nomeadamente relativamente aos géneros *Facklamia* e *Actinobaculum*, que apresentavam um aumento em casos de oligoastenoteratozoospermia. Para o fenótipo de hiperviscosidade seminal, nenhum resultado significativo foi obtido, mas mesmo assim uma tendência de redução de diversidade foi observada, que deveria ser investigada com uma maior amostra.

*O Enterococcus* e *Staphylococcus* são reconhecidos causadores de infeções do trato genitourinário tais como a prostatite, uretrite e epididimite, e também se sabe estarem presentes no sémen de indivíduos saudáveis. A identificação desses géneros em elevadas abundâncias relativas e em amostras de alto *bacterial load* pode sugerir uma situação de disbiose, que, em conjunto com outros fatores, pode contribuir para uma diminuição da qualidade do sémen. Contudo, serão necessários estudos adicionais mais aprofundados para testar esta hipótese de estes taxa causarem efeitos negativos nos espermatozoides e nos diferentes componentes do plasma seminal. Por outro lado, a deteção de uma elevada diversidade microbiana combinada com um baixo *bacterial load* parecem enquadrar-se melhor com a definição de um microbioma seminal normal (eubiose), no qual a perda de qualidade é provavelmente determinada por outras causas.

**Palavras-chave:** microbioma, infertilidade masculina, hiperviscosidade seminal e oligoastenoteratozoospermia, sequenciação de 16S rRNA, *bacterial load*

# Abstract

The microbiome is currently recognized to play an important role in human health and also in several diseases like inflammatory bowel disease, Crohn's Disease, obesity, and colorectal cancer. However, studies of the impact of bacteria in the male reproductive system and in male infertility are still scarce. So far, to the best of our knowledge only 12 high-throughput studies have attempted to fill this gap in which some of their findings point to distinct repercussions of microbial communities in semen quality parameters. In this regard, a study performed by our team suggested that the infertility phenotype of seminal hyperviscosity could be correlated with an increment in Proteobacteria and another, the oligoasthenoteratozoospermia with shifts in the abundance of pathogenic and probiotic genera. To provide a deeper insight about the seminal microbiome and its disease associated changes that could underlie the previously reported differences, we analyzed a well-stratified cohort of male subjects comprising 15 normozoospermic (controls) and 53 infertility cases including 16 individuals with seminal hyperviscosity, 24 with oligoasthenoteratozoospermia and 13 presenting both phenotypes simultaneously. To this end, total DNA was extracted from seminal plasma samples using QIAamp DNA mini kit and amplified for *16S ribosomal RNA* (*16S*) gene using two distinct approaches: 1) Quantitative PCR (qPCR) assays that were implemented to access the total bacterial content and to estimate the samples bacterial load; 2) High throughput sequencing experiments, which were carried out using the *Ion 16S Metagenomics Kit* covering V2, V3, V4, V6-7, V8 and V9 hypervariable regions, a *Ion Torrent S5 XL* sequencing platform and the *Torrent Suite/Ion Reporter* software in order to identify bacterial taxa by generating Operational Taxonomy Units (OTUs). The analysis of collected data comprised standard statistical tests, a manual curation of OTUs and several microbiome-based tests available through the MicrobiomeAnalyst tool. Firstly, the qPCR results demonstrated that the bacterial content of seminal samples was largely variable, thus two bacterial load groups were defined, high and low, respectively. Surprisingly, those did not correlate either with the infertility status (cases or controls) or the studied infertility phenotypes (SHV, OAT and OAT+SHV). Then, the microbiome profiling uncovered Firmicutes as the most abundant phylum in seminal plasma independently of the considered group, cases or controls, phenotype, or high and low bacterial load. At the genus rank *Enterococcus*, *Staphylococcus*, *Streptococcus* and *Facklamia* were the most prevalent taxa in general, but those were found to display a large interindividual variability. Remarkably, the *high bacterial load* samples were found to be correlated with top abundances of *Enterococcus* or *Staphylococcus* contrasting with the *low bacterial*

*load* samples that were more frequently associated with multiple taxa where a common microbial profile was hardly identified. In agreement with these findings alpha and beta diversity indices showed only significant differences when considering the bacterial load (high and low). Precisely, *Enterococcus* or *Staphylococcus* -enriched samples were discriminated from the remaining ones by an impressive similarity, reduced diversity and incremented numbers of bacteria as estimated by bacterial load. Additionally, some differences in low abundant were detected, namely concerning Facklamia and Actinobaculum genera which were found to be augmented in cases displaying oligoasthenoteratozoospermia. For the seminal hyperviscosity phenotype no significant result was achieved, still a trend toward a reduction in diversity was observed that should be investigated with a larger sample.

*Enterococcus* and *Staphylococcus are* known to cause male genitourinary infections like prostatitis, urethritis and epididymitis, and to be present also in the semen of healthy individuals. The identification of these genera at high abundances and at high bacterial load could suggest some kind of dysbiosis that in some individuals together with other factors could contribute to a loss of semen quality. Nevertheless, further studies would be needed to test this hypothesis of a negative effects of these taxa in the spermatozoa and seminal plasma components. Conversely, the detection of some samples with high microbial diversity combined with a low burden of bacteria seem to fit the definition of a regular seminal microbiome (eubiosis) in which the loss of semen quality is most probably due to other causes.

**Keywords:** Microbiome, male infertility, seminal hyperviscosity and oligoasthenoteratozoospermia, *16S rRNA* sequencing, bacterial load

# Table of Contents

# List of Tables

# List of Figures

# List of Abbreviations

*16S* – *16S* ribosomal RNA gene

ART – assisted reproductive techniques

AT – asthenozoospermia

HIV – human immunodeficiency virus

HPV – human papilloma virus

ICMART – International Committee for Monitoring Assisted Reproductive Technology

LDA – linear discriminant analysis

MDP – marker data profiling

MGI – male genitourinary infection

NRM – normozoospermic

OAT – oligoasthenoteratozoospermic

OTU – Operational Taxonomy Unit

PCoA – principal coordinate analysis

PCR – Polymerase Chain Reaction

PSA – prostate specific antigen

qPCR – quantitative Polymerase Chain Reaction

SHV – semen hyperviscosity

STI – sexually transmitted infection

WHO – World Health Organization

# 1. Introduction

## 1.1. Infertility – General information

Infertility is a disorder of the female or male reproductive systems defined by the World Health Organization (WHO) and the International Committee for Monitoring Assisted Reproductive Technology (ICMART) as the inability to achieve a clinical pregnancy after 12 months of regular unprotected sexual intercourse [1, 2]. Infertility affects worldwide about 8-12% of the couples at reproductive age, in which male factors alone are the cause of the disease in up to 30% of cases and are a significant component in additional 20% of patients [3-5]. In Portugal, specifically, infertility was estimated to affect 7.9% of couples [6]. Importantly, the analysis of the prevalence of infertility in an age-standardized mode showed that the disease has been increasing annually by 0.291% in men and for which sperm counts decreased in 50-60% between 1973 and 2011 [3, 7]. Despite the large efforts made over the last decades in the evaluation of male infertility factors, roughly half the patients still have no discernible or identifiable cause [8]. Therefore, these men are classified as idiopathic infertility cases.

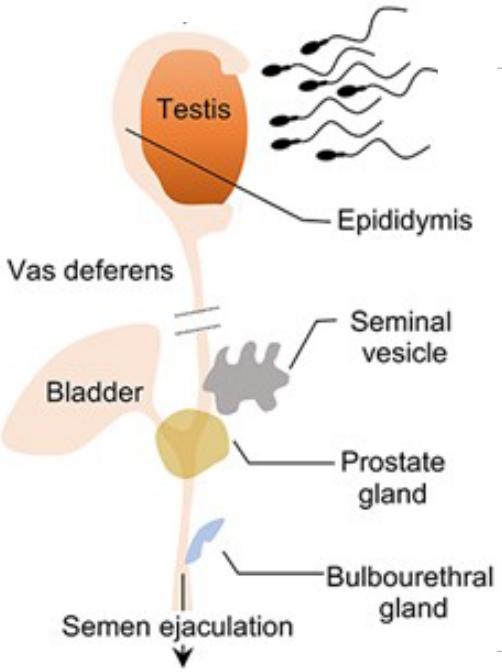
In western societies, reproductive medicine care and assisted reproductive techniques (ART) are offered to infertile couples many times without the identification or the resolution of the cause of the disease [9]. ART were initially developed last century in the 80s and continued to be improved since then, however given that male infertility factors remain frequently not address some men conceived by ART are currently experiencing fertility issues [10].

On the other hand, ART can be physically and emotionally demanding and expensive if not supported by national health systems, leading patients to quit after a few unsuccessful attempts or even not advancing with any treatment [11, 12]. To better understand any study of male infertility factors, a detailed knowledge of the physiology of the male reproductive system is fundamental.

### 1.1.1. The semen and its main components

The semen is a biological fluid composed by a mixture of the secretions from the male accessory glands in which the male germ cells, the spermatozoa, are suspended. While spermatozoa are usually considered the most important fraction of the seminal fluid, it only accounts for 10% of the total semen volume (Figure 1). The liquid fraction of the

semen, the seminal plasma, is mainly composed of fluids produced by the seminal vesicles (~65%) prostate (~25%), epididymis and bulbourethral gland secretions (~5% altogether; Figure 1) [13, 14]. Under healthy conditions, the ejaculate results in >2mL semen volume, with a pH of 7.3 – 7.7 and a spermatozoa content ranging the 150 - 600 million [13].




| % in total ejaculate | Content | Components |
|---|---|---|
| ~10% | Spermatozoa | – |
| ~65% | Seminal vesicle fluid | Semenogelins, fructose |
| ~25% | Prostatic fluid | Lipids, citric acid, proteolytic enzymes, zinc, acid phosphatase |
| ~5% | Epididymis and bulbourethral gland secretions | Epididymosomes mucins |

Figure 1 – Schematic representation of different contributions of male accessory glands in semen composition (adapted from Anamthathmakula *et al.* (2020) [15])

### 1.1.1.1. The spermatozoa fraction

The spermatozoa are the end-product of a complex cell differentiation process called spermatogenesis, which occurs in the testis once reproductive maturity is reached. Anatomically, the testis comprise around 200 to 300 seminiferous tubules imbedded in an interstitial matrix containing the testosterone synthetizing cells - the *Leydig* cells. Conversely, inside the seminiferous tubules the epithelia comprise the germ cells that are sustained throughout their differentiation together with *Sertoli* cells. These are somatic cells that play a critical role in male reproduction by maintaining the unique environment necessary for a successful spermatogenesis including the production of several growth factors and the delivery of distinct nutrients. Furthermore, *Sertoli* cells are

also important in the formation of junctional complexes that provide physical support to spermatozoa differentiation (Figure 2). The original diploid cells, the spermatogonia, are positioned at the base of the epithelium and are divided in two distinct cellular populations: the undifferentiated type A cells and the differentiating type B cells. Only type B cells differentiate into primary spermatocytes, which then undergo meiosis, generating secondary spermatocytes, which are the first type of haploid cells. Later, these undergo a maturation process originating the spermatids in which the tail starts to be assembled. Then, the fully differentiated spermatozoon shows several organelles specialized for fertilization such as the acrosome, a unique nuclear shape and a nearly absent cytoplasm (Figure 3) [16-19].

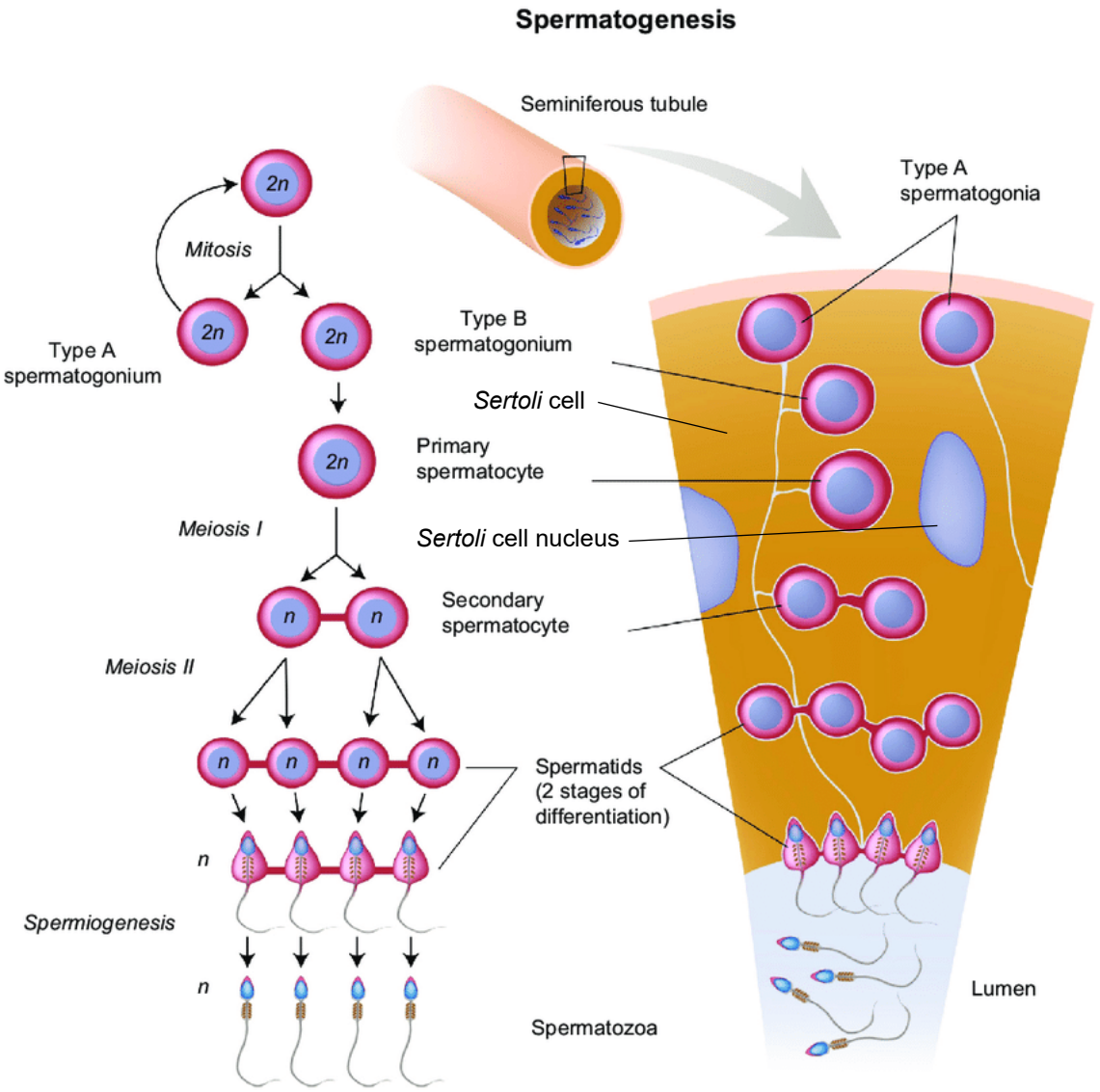



Figure 2 – Schematic representation of the spermatogenesis process (adapted from Levine *et al.* (2013) [20]).

Once the spermatogenesis is completed, the differentiated spermatozoa can be subdivided into two main structures: the head and tail (or flagellum) (Figure 3). Briefly, the head comprises the nucleus, which contains the condensed chromatin and a small layer of cytoplasm surrounded by the plasma membrane envelope. Moreover, in the tip of the head there is the acrosome, which contains several hydrolytic enzymes like hyaluronidase and acrosin fundamental to break down of the outer membrane of the ovum. On the other hand, the tail can be divided into three regions: the midpiece, which encompasses the mitochondrial sheath and the dense fibers; the principal piece that makes up most of the flagellum length, and the terminal piece [14]. The regular organization of these three substructures is fundamental to a proper motility of the spermatozoa and to their capability of swimming through the muco-cervical barrier until reaching the oviduct and the ovum.

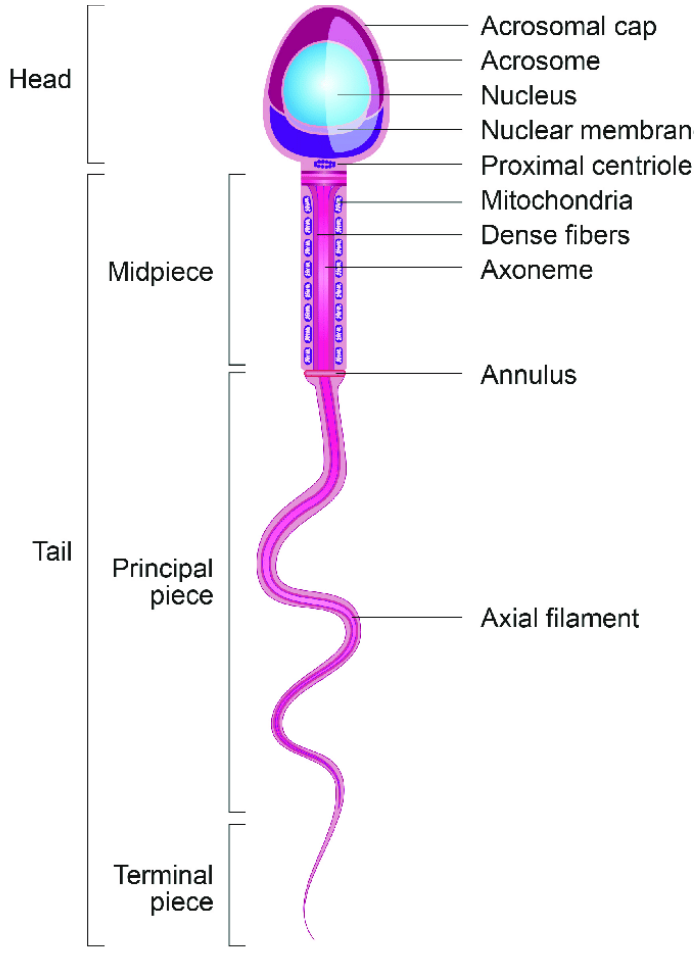



Figure 3 – Schematic representation of the human spermatozoon (adapted from Alves *et al.* (2020) [21]).

The differentiated spermatozoa are then released into the lumen of seminiferous tubules and moved to the epididymis where they are stored until the time of the ejaculation. It is in the epididymis that spermatozoa complete several other steps of their maturation. In short, the spermatozoa undergo significant changes in their nuclear compaction, plasma membrane composition, cytoskeletal structure and protein and non-coding RNA payloads that occur mainly through their direct contact with the epididymal environment [22].

### 1.1.1.2. The seminal plasma fraction

As aforementioned, the seminal plasma is a heterogeneous mixture of the secretions from the testis, epididymis, prostate, seminal vesicles and bulbourethral glands (Figure 1) [23]. In more detail, the seminal vesicles contribute to the seminal plasma with semenogelins, which are the main molecules conferring the tensile strength and thickness (viscosity) to the ejaculate, and fructose that is the main source of energy provided to the spermatozoa [23, 24]. The prostate, on the other hand, contributes with: 1) lipids – essential into the remodeling of the destabilized spermatozoa cell membrane; 2) citric acid – important in maintaining the seminal pH; 3) proteolytic enzymes (e.g. prostate specific antigen – PSA) – fundamental to the cascade of semen coagulum hydrolysis (semenogelin cleaving); 4) zinc – critical in the regulation of proteolytic activity and chromatin stability, and 5) acid phosphatase – relevant in the activation of several growth factors [18, 25, 26]. The epididymis contributes with epididymosomes, containing various proteins with functions such as facilitating the binding of the spermatozoon with the ovum [22]. Finally, the secretion product of the bulbourethral glands contributes with mucins that act as a natural lubricant in semen ejaculation (Figure 1) [23].

Interestingly, in several ART procedures in which the ejaculate is diluted and the spermatozoa separated from the seminal plasma, being also washed, frozen and thawed, losses of function by spermatozoa have been reported [18]. Moreover, in spite of all historically recognized roles of the different elements of the seminal plasma in male reproduction, in the scope of ART they have been neglected. Notwithstanding, recent experiments of spermatozoa supplementation with seminal plasma after their cryopreservation had fewer negative effects when compared with standard ART treatments [9, 18, 26].

## 1.1.2. The diagnosis of male infertility

Traditionally, to achieve a diagnosis of male infertility two steps are needed. The first consists of a thorough physical examination of reproductive organs, as well as an assessment of the individual medical history to identify major causes and/or risk factors leading to the disease. The second relies in a semen analysis – the spermiogram, in which the semen quality is evaluated according to the fulfillment of a series of parameters as determined by the WHO [2, 27, 28].

### 1.1.2.1. Clinical study, major causes and infertility risk factors

In the clinical study of infertile men, the physical examination is fundamental to exclude any congenital malformations and other alterations of the reproductive organs known to cause infertility such cryptorchidism and varicocele (Table 1: testicular factors) or the obstruction of the epididymis and vas deferens (Table 1: post-testicular factors). Simultaneously, the assessment of the medical history and the health status of an individual are fundamental to rule out other infertility causes, namely physiological ones correlated with hormonal dysregulations like hypothyroidism and several hypothalamic disorders (Table 1: hormonal factors) and distinct genetic conditions resulting from either major chromosomal anomalies, including Klinefelter syndrome (XXY karyotype) and Y chromosome microdeletions or monogenic disorders such as cystic fibrosis and Kallman syndrome (Table 1: genetic abnormalities).

Notwithstanding, in the clinical evaluation it is also relevant to investigate a wide-range of lifestyle and environmental factors possible contributing to male infertility or subfertility. Those may include for example specific medications (e.g. steroids), toxins (e.g. alcohol, tobacco, drugs) and diet (e.g. poor intake of zinc and vitamin C), which are all recognized to have an impact in semen quality. The presence of varied infections of the male genitourinary system such as prostatitis, epididymitis, orchitis, and urethritis or even sexually transmitted infections (STIs), like gonorrhea and chlamydia infection, should be considered and treated since these are reported to affect men reproductive capacity [29, 30].

Table 1 – Major causes and risk factors implicated in male infertility (adapted from Brookings *et al.* (2013) [29])

| **Major types** | **Causes and risk factors for male infertility** |
|---|---|
| Lifestyle and Environmental Factors | Age<br>Obesity<br>Stress<br>Poor dietary intake zinc and vitamin C |
| Toxins | Alcohol<br>Tobacco smoking<br>Recreational drugs-marijuana<br>Anabolic steroid use |
| Medications | Chemotherapy<br>Phenytoin<br>Spironolactone<br>Sulfasalazine |
| Genetic Abnormalities | Genetic defects on the Y chromosome<br>Y chromosome microdeletions<br>Klinefelter syndrome<br>Cystic fibrosis<br>Kallman syndrome |
| Hormonal Factors | Hypothalamic disorder<br>Hyperprolactinemia<br>Primary hypogonadism<br>Hypothyroidism |
| Testicular Factors | Cryptorchidism<br>Varicocele<br>Trauma<br>Hydrocele<br>Testicular cancer<br>Idiopathic |
| Post-testicular Causes | Epididymal obstruction<br>Vas deferens obstruction<br>Hypospadias<br>Retrograde ejaculation |
| Sexual Dysfunction | Erectile dysfunction<br>Premature ejaculation<br>Ejaculatory incompetence |

### 1.1.2.2. The spermiogram analysis and the evaluation of semen quality

Significant changes in semen composition, affecting either the spermatozoa or the seminal plasma, are known to have serious consequences into male fertility.

Broadly speaking, the parameters evaluated in a spermiogram include biological (e.g. spermatozoa features), chemical (e.g. pH) and physical (e.g. viscosity) properties of the semen, defined according to reference values (Table 2) [2, 11, 12]. However, those thresholds have been revised by the WHO over the years, giving rise in some instances to abnormal semen parameters with narrower intervals and consequently, to a different classification of infertility patients [2, 27, 28]. To be more accurate, the application of the WHO 2010 values (Table 2), which were inferred from the distribution of semen parameters based on a wide population study of fertile fathers (known time-to-pregnancy below the 12 months), resulted in samples previously showing an infertility phenotype to be considered as “normal” [31]. This, together with other concerns regarding the selection of the individuals, the fathers, as representatives of a worldwide population sample, raised several criticisms to the 5$^{th}$ Edition of the WHO manual published in 2010 [32]. In the latest edition, released in 2021 some of these criticisms were already addressed [28], namely the under-representation of various areas of the globe, such as Sub-Saharan Africa and South America. Moreover, this new edition includes new 5$^{th}$ percentile values of some semen parameters, such as the threshold of sperm concentration and motility, while maintaining others concerning seminal viscosity and spermatozoa morphology (Table 2) [33].

If one or more of the semen parameters evaluated through the spermiogram do not fulfil the recommended cut-off, the patient is classified into different infertility phenotypes. In this respect, it is fundamental to acknowledge which WHO guidelines for spermiogram analysis are being employed.

Table 2 – Threshold values for selected semen parameters evaluated through routine spermiograms.

| | **Guidelines for laboratorial processing of human semen** | | |
|---|---|---|---|
| **Parameter** | WHO (1999) [27] | WHO (2010) [2] | WHO (2021) [28] |
| Volume | ≥ 2.0 mL | ≥ 1.5 mL | ≥ 1.4 mL |
| pH | ≥ 7.2 | ≥ 7.2 | ≥ 7.2 |
| Viscosity | ≤ 2cm thread length | ≤ 2cm thread length | ≤ 2cm thread length |
| Liquefaction | Complete until 60 minutes | Complete until 60 minutes | Complete until 60 minutes |
| Total sperm count | ≥ 40 million per ejaculate | ≥ 39 million per ejaculate | ≥ 39 million per ejaculate |
| Sperm concentration | ≥ 20 million per mL | ≥ 15 million per mL | ≥ 16 million per mL |
| Sperm Motility | ≥ 50% with progressive (rapid and slow) motility Or ≥ 25% with rapid progressive motility | ≥ 40% total (progressive and non-progressive) motility Or ≥ 32% progressive (rapid and slow motility) | ≥ 42% total (progressive and non-progressive) motility Or ≥ 30% progressive (rapid and slow motility) |
| Sperm Morphology | ≥ 14% with normal forms (recommended) | ≥ 4% with normal forms | ≥ 4% with normal forms |
| White blood cells | ≤ 1 million per mL | ≤ 1 million per mL | ≤ 1 million per mL |

In accordance with the selected reference values shown in Table 2 the phenotypes most frequently used in the literature are related with:

1) Spermatozoa concentration, motility and morphology where the following definitions correspond to:
    a) Azoospermia – Absence of spermatozoa in the ejaculate;
    b) Oligozoospermia – Reduced number of spermatozoa;
    c) Asthenozoospermia – Immotile spermatozoa or with decreased motility;
    d) Teratozoospermia – Low percentage of morphologically normal spermatozoa below the defined threshold;
2) Semen viscosity and liquefaction:
    a) Semen Hyperviscosity (SHV) – Increased semen viscosity as indicated by the formation of a long thread;
    b) Delayed Liquefaction – Slower thinning of the semen sample – longer time until the liquefaction process is completed;
3) Presence of other cell types:
    a) Leukocytospermia – Increased number of white blood cells;
    b) Bacteriospermia – The presence of bacteria in the semen.

Regarding spermatozoa parameters, if a sample shows abnormal concentration, motility and morphology it can be labelled as oligoasthenoteratozoospermic (OAT), whereas if no alteration is observed in spermatozoa or in semen as a whole, the sample is designated as normozoospermic (NRM).

### 1.1.2.3. Pathogenic agents associated with sexually transmitted and others male urogenital infections

A multitude of bacteria, some parasitic eukaryotes and several viruses have been reported to cause STIs and other male genitourinary infections (MGIs), with diverse implications in reproductive health and fertile potential. In general, MGI and STI can affect the spermatogenesis causing damage to spermatozoa DNA and cell dead, disturb sperm motility through modifications of flagellum and the agglutination of sperm cells and also inhibit acrosome function [29, 34, 35]. The most prevalent STIs worldwide comprise the bacterial infections caused by *Chlamydia trachomatis* (chlamydia infection), *Neisseria gonorrhoeae* (gonorrhea), *Treponema pallidum* (syphilis), *Mycoplasma*

*genitalium, Ureaplasma urealyticum* and *Ureaplasma parvum*; the infection by the parasite *Trichomonas vaginalis* (trichomoniasis) and the viral infections by human immunodeficiency virus (HIV) and human papilloma virus (HPV) [29, 36].

STIs can be associated with varied clinical manifestations. More specifically, *Chlamydia* has been reported to infect different male accessory glands causing orchitis, epididymitis and prostatitis [37]; gonorrhea was connected with urethritis and epididymo-orchitis [38] and syphilis was described to cause epididymitis, where the obstruction of the epididymis can occur together with multiple chronic lesions [29].

Concerning the effects of STIs in semen quality, the pathogens *C. trachomatis, M. genitalium, U. urealyticum* and *U. parvum* have all been associated with a decrease in sperm counts, with *M. genitalium* also having adverse effects on motility, morphology and DNA condensation [29, 39]. Given their negative impact in semen quality, the STIs agents have been also correlated with different spermiogram phenotypes such as azoospermia, oligozoospermia, asthenozoospermia and teratozoospermia. On the other hand, the connection of trichomoniasis in male infertility have been controversial. While sperm motility reduction has been attributed to *Trichomonas vaginalis* infection, these consequences seem to be reversible in a relatively short period of time [29, 36, 39]. The presence of these pathogens has also been correlated with an increased concentration of leukocytes, which may also be associated with seminal hyperviscosity [40].

In cases of bacteriospermia, when an unusual number of bacteria is detected during the spermiogram analysis, semen cultures are performed to identify the underlying infectious agent. However, these are often correlated with other MGIs rather than the pathogens known to cause STIs. In this regard, Domes *et al.* (2012) reported that *Enterococcus faecalis* was the most prominent bacteria in bacteriospermic samples, but with no significant correlation between its overrepresentation and poor sperm quality [41]. Similar results were obtained by Vilvanathan *et al.* (2016), which reported *E. faecalis* together with *Staphylococcus aureus* and *Escherichia coli* to be the most common bacteria associated with bacteriospermia cases [42] Other taxa like *Klebsiella pneumoniae*, *Proteus* sp. And *Citrobacter* sp. Were also observed but less frequently. Again, in this study no correlation between bacteriospermia and a poor quality of the semen was registered [42]. Although it has been found in the male reproductive tract in apparently healthy conditions, *Enterococcus* has been described to also cause MGIs and to be connected with prostatitis, orchitis and epididymitis [43]. Therefore, the findings concerning the impact of *Enterococcus* genus in male fertility are still contradictory.

Like for *Enterococcus*, different *Staphylococcus* taxa were described as MGIs agents and linked to epididymitis, orchitis, prostatitis and urethritis. [42]. However, for some *Staphylococcus* species, such as *Staphylococcus saprophyticus and Staphylococcus epidermidis*, a negative impact of these bacteria in sperm motility and morphology was reported [44, 45].

### 1.1.3. The microbiome – its importance in human health and disease

The human microbiome, or the communities of microorganisms including bacteria, fungi, parasitic eukaryotes and virus inhabiting the human body, was incipiently investigated for decades with light microscopy, bacteria Gram staining and observation of cell morphology. These allowed the identification of different microbes based on certain characteristics, such as the presence or absence of a bacterial cell wall composed of peptidoglycans, which discriminate Gram positive and negative bacteria [46].

Later, the development of microbiological studies through *in vitro* based culture methods, which consisted of growing collected biological samples in different bacterial growth media, together with the evaluation of several biochemical characteristics such as phosphatase activity, permitted the identification of additional bacterial groups. Despite being very cheap, these methods are time consuming and have low sensitivity, selecting only for those taxa capable of growing under a limited repertoire of nutritional and physiological conditions found in laboratories. Consequently, the detected bacteria are not necessarily the most abundant, nor a negative result means the absence of bacteria. Indeed, this bias was shown to be especially problematic when attempting to grow anaerobic taxa, as their laboratory propagation is slow, and require special conditions to successfully expand in *in vitro* culture [46, 47]. Other bacteria are not cultivable at all, leading to a considerable loss of information regarding their identification as part of the human microbiome [48].

From the beginning of the 1990s, new methods with higher degree of sensitivity began to be applied to the identification of bacterial taxa. These started to use Polymerase Chain Reaction (PCR) to amplify different target regions within the bacterial genome, being firstly applied to samples collected from different macroenvironments and only later to biological samples when conventional bacteriological techniques failed to identify microorganisms [49].

At that time, the *16S rRNA* gene (*16S*) emerged as the gold standard region for bacteria identification. This is a housekeeping gene common to all bacteria that has the peculiarity

of containing 9 regions largely varying across different taxa – the hypervariable regions V1-V9 – flanked by highly conserved regions (Figure 4). This organization of *16S* gene revealed to be very advantageous for microbiome studies because many universal primers were designed to hybridize within the conserved sequences and then to amplify the different hypervariable regions. To date, there is already a long list of universal primers with different efficiencies and specificities on taxa identification [46, 50].

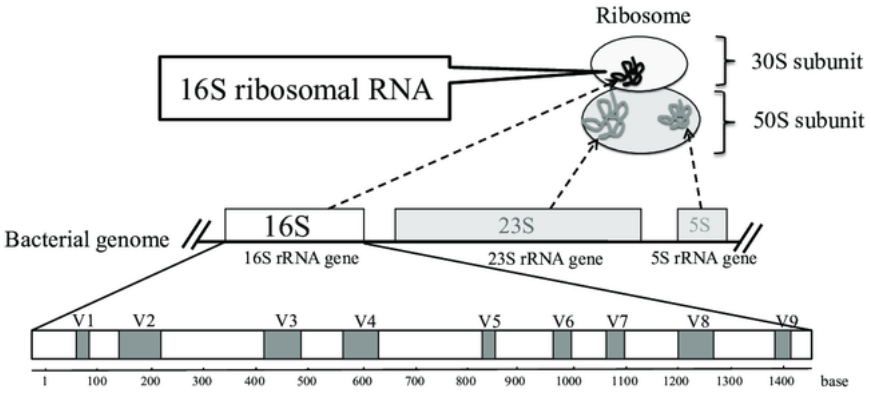



Figure 4 – *16S* gene organization, showing the positioning of the V1 to V9 hypervariable regions (adapted from Fukuda *et al.* (2016) [51]).

Initially, this *16S* amplification approach was used to construct multiple clone libraries followed by Sanger sequencing of each clone, exhibiting more phylogenetically diverse results than the previous established methodology of culturing and microscopical analyses [52-54].

More recently, the replacing of Sanger sequencing by state-of-the-art high-throughput methods dismissed the laborious cloning task allowing to identify multiple taxa at once in a single sample. These technological advances contributed largely to the boost of microbiome studies and to the understanding that very few (or no) body sites are sterile. In fact, the human body is heavily colonized by bacteria, coexisting in the same order of magnitude as the human cells (~$3.8x10^{13}$ and ~$3.0x10^{13}$ cells, for bacteria and human, respectively) [55].

Currently, the best characterized microbiome is the one of the lower tract of the gastrointestinal system, more frequently named as the gut microbiome. This is believed to have the highest bacterial load within the human body and it has been shown to not only influence its own niche but also almost every organ in the human body. For example, in a healthy state, gut microbes present an interspecies balance (eubiosis) and are reported to positively affect the host in various ways [56]. Precisely, it has been described to modulate the immune system as well as several neurobehavioral and metabolic traits.

Some of those benefits are correlated with an improvement in the lipid metabolism, production of several antioxidants and short-chain fatty acids [57].

On the other hand, the gut microbiome has also been associated with negative effects, mainly when the bacteria present in the intestine show significant changes in their abundance and composition (dysbiosis), and which are usually correlated with a loss of microbial diversity [56]. Negative effects are often correlated with augmented inflammatory states and with the onset of several multifactorial diseases like inflammatory bowel disease, Crohn's Disease, obesity, psoriasis, and colorectal cancer [57-62]. Indeed, Marichanh *et al.* (2006) has reported that patients with Crohn's Disease present a lower microbial diversity due to a reduction in the abundance of the Firmicutes phylum [63].

### 1.1.3.1. The seminal microbiome

The seminal microbiome is relatively understudied in comparison to the gut one, or even to its vaginal counterpart [34]. Nevertheless, in the recent years, some studies have been implemented in order to assess the composition of the microbiome of the seminal plasma and its impact in human health and infertility [34, 46, 64-70]. One of the major breakthroughs retrieved by those studies is that bacteria can be found in the semen in the absence of any symptom of STIs or MGIs. Nowadays, it is accepted that bacteria are detected in the semen of healthy individuals with a normal spermiogram [70].

So far, previous studies of the seminal microbiome have shown that fertile patients exhibited a higher abundance of Proteobacteria and Actinobacteria phyla, contrasting with the high abundance of Firmicutes and Bacteroidetes phyla observed in infertile patients with azoospermia [66, 71].

The most common genera of bacteria identified in the semen according to a recent review study of high-throughput sequencing by Farahani *et al.* (2021) [5] are the following: *Ralstonia*, *Lactobacillus*, *Prevotella*, *Corynebacterium*, *Pseudomonas*, *Streptococcus*, *Staphylococcus*, *Ureaplasma*, *Clostridiales*, *Atopobium*, *Anaerococcus*, *Gardnerella*, *Rhodanobacter*, *Finegoldia*, *Haemophilus*, *Planococcaceae* and *Burkholderia* [34, 65, 69, 70].

Among those studies Monteiro *et al.* (2018), was the only one to document a high prevalence of *Enterococcus* and *Staphylococcus* in consistence with findings from semen culture works [34]. Moreover, Monteiro *et al.* (2018), as previously suggested by Weng *et al.* (2014), also reported a decrease in *Lactobacillus* abundance in male

infertility cases, suggesting that this genus might exert a protective effect on semen quality [34, 65]. Together with this microbiome change, Monteiro *et al.* (2018) reported an augmented prevalence of Proteobacteria in men presenting with the hyperviscosity phenotype and increased abundances of *Pseudomonas, Klebsiella*, *Aerococcus*, *Actinobaculum*, *Neisseria*, and *Enterobacteriaceae* in the same hyperviscosity group or in oligoasthenozoospermia cases [34]. Interestingly, an enrichment in *Pseudomonas* was proposed to have a negative effect particularly when the abundance of *Lactobacillus* is low and a low abundance of *Lactobacillus* was also correlated with azoospermia [65].

It is important to note that the results of the different studies are not always concordant. For example, *Streptococcus* and *Anaerococcus,* which have been described as part of the normal seminal flora have been also observed in individuals with low semen quality [65, 69, 71]. Furthermore, the parameter of sperm motility was described to be impacted by the presence of several genera and species such as *Ureaplasma*, *Bacteroides*, *Finegoldia* and *Acinetobacter iwoffii* [72]. A very recent study conducted by Venneri *et al.* (2022) has further described *Enterococcus faecalis*, *Escherichia coli*, *Ureaplasma urealyticum* and *Streptococcus agalactiae* as markers for poor sperm motility, which contradicts, in part, the study by Weng *et al.* (2014), who had previously observed no correlation between *E. coli* with infertility parameters [65, 73]. Several other genera and species have been described as affecting sperm morphology when present in a higher abundance. These include *Ureaplasma*, *Enterococcus*, *Mycoplasma*, *Prevotella* and *Bacteroides urealyticus* [73].

## 1.2. Applications of the seminal microbiome to Forensic Genetics

The Forensic community has expressed an interest in microbiomics as early as in 2001 regarding its potential applications in the investigation of bioterrorism [74]. More recently, this interest has expanded considerably to other areas such as individual identification, geolocation and post-mortem interval establishment [74]. This interest is related to certain aspects of the microbiome, namely its high dynamicity, or its sensitivity to daily activities such diet, age, sex, geographical location and even interactions with other microbe communities. These characteristics render the microbiome some profiling potential that might be useful in the scope of murder, missing belongings and sexual assault investigations [75]. Although this field is still in its infancy, some studies are showing promising results. For example, the skin microbiome was already used to link several belongings to their owners with a reported success rate of 93% [76].

Furthermore, in a sexual assault approach, a fraction of the female microbiome seemed to be derived from the aggressor, suggesting a microbiological analysis as useful when the list of potential suspects is small [77, 78]. Dobay *et al.* (2019) has carried out a study to identify different body fluids, in which semen samples were included. Briefly, 12 semen samples were tested for their microbiome after sample collection and again 30 days after their exposure to an indoor environment. Dobay *et al.* (2019) results also uncovered several differences in abundances of determined taxa, but no definitive conclusions could be drawn since the sample size was too small [79, 80].

So far, this is still a poorly explored area of forensic sciences and there are many doubts about its future application as a routine practice in crime resolution. In this field, the transfer of microbiome between individuals and other microbe communities, such as the soil, is still not well understood, and many more factors play a role in its composition, such as soil characteristics, environmental exposure and lifestyle practices, resulting in possible temporal mismatches [74, 81].

# 2. Aims

In contrast to other areas of microbiome knowledge, the male reproductive tract and the semen have been poorly investigated in spite of its potential impact in human fertility. According to a first exploratory study performed by of our team, Monteiro *et al*. (2018), two phenotypes of male infertility diverged from controls in their bacterial communities present in the seminal plasma. While SHV cases differ from controls and other phenotypes by an increased abundance of Proteobacteria, the OAT cases could be separated by a decreased prevalence of probiotic genera combined with an augmented proportion of known pathogenic bacteria. Therefore, to further explore the contribution of the seminal microbiome in the male reproductive health and in the quality of the semen, we performed a quantitative and qualitative characterization of the bacteria found in the semen of Portuguese cases, with or without SHV and OAT, and controls. To achieve this main goal, we used different methodological approaches to address the following specific points:

1) Quantitative PCR assays for the amplification of *16S* gene were implemented to estimate the bacterial content of each semen sample.
2) High-throughput sequencing of multiple hypervariable regions of the *16S* gene to identify the most abundant bacteria present in seminal samples.
3) Analyses of the generated Operational Taxonomy Units (OTUs) were performed through the MicrobiomeAnalyst package to evaluate semen samples microbial profiling, alpha and beta diversity as well as to identify statistically significant taxa differing between groups defined by infertility status (cases and controls), phenotypes (NRM, OAT, SHV, OAT+SHV) and bacterial content (high or low bacterial load).

# 3. Material and Methods

## 3.1. Semen samples and DNA extraction

In this study, 68 semen samples were analyzed for their microbiome composition. Those were selected from a large cohort previously obtained from *Centro de Genética da Reprodução Prof. Alberto Barros*. The sample collection was carried out by masturbation after a period of at least 3 days of sexual abstinence, passed urine and upon genitals and hands washed with soap. For the propose of our study the following data and spermiogram results were retrieved: patient nationality, age, seminal viscosity, spermatozoa concentration, motility and morphology. Fifteen Portuguese men with normal semen parameters (NRM) were considered as controls (Table 3). The remaining 53 samples were infertility cases that were stratified into three groups according to their phenotypes. These comprised 24 individuals showing abnormal spermatozoa concentration, motility and morphology and therefore labeled as oligoasthenoteratozoospermia (OAT), 16 subjects presenting only seminal hyperviscosity (SHV) and 13 patients combining the two previous phenotypes OAT+SHV (Table 3). This division in these phenotype groups is based on a previous study of our group conducted by Monteiro *et al*. (2018) [34].

All samples were separated into a cellular fraction and seminal plasma by centrifugation at 7000*g* for 10 minutes and stored at -80°C. Then, for the propose of seminal microbiome studies total DNA (human and non-human) was extracted from 50-200 µL of seminal plasma using the *QIAamp DNA mini kit* (Qiagen) following the recommended protocol for DNA Purification from Body Fluids. Additionally, a negative control for the DNA extraction was also included in our study to oversee environmental and reagent contaminants. Basically, the same DNA extraction protocol was used without adding any biological sample material. This will be named from this point forward as *DNA extraction control*.

Table 3 – Sample composition according to analyzed cases and control groups.

| Infertility Status Group | Phenotype Group* | Sample Size (%) |
|---|---|---|
| Cases (n=53) | SHV<br>Seminal Hyperviscosity<br>(≥ 2cm thread length) | 16 (23.5%) |
| | OAT<br>Oligoasthenoteratozoospermia<br>(≤ 15 million spermatozoa per mL; ≤ 40% total motility or ≤ 32% progressive and ≤ 4% with normal forms) | 24 (35.3%) |
| | OAT+SHV<br>Oligoasthenoteratozoospermia with Seminal Hyperviscosity | 13 (19.1%) |
| Controls (n=15) | NRM<br>Normozoospermia | 15 (22.1%) |

*The different groups were defined according to the 2010 WHO guidelines [2]

## 3.2. Bacterial load measurement

To quantify the bacterial load, or the total amount of bacteria per sample, several quantitative Polymerase Chain Reaction (qPCR) assays were performed. Initially, two segments of the *16S* gene covering distinct hypervariable regions were selected to perform qPCR experiments. The first segment spanned from V1 to V2 and it was based in the work of Sulaiman *et al.* (2021), who had previously carried qPCR assays to evaluate the total bacterial content of their samples [82]. The second segment targeted the V3-V4 region and used selected primers from Klindworth *et al.* (2013) described to have a good performance in terms of bacteria (Eubacteria) identification. Additionally, the latter primers were already proven to be successful in *16S* amplification by Monteiro *et al.* (2018) and also produced a shorter fragment than the ones from V1-V2 [34, 50] (Table 4). Although an effective amplification of the *16S* gene was achieved for both

fragments, we selected the V3-V4 segment because aside from their aforementioned advantage it also provided slight better amplification yields (Annex I: Figure 17).

Table 4 – Primers used in the amplification of *16S* by quantitative Polymerase Chain Reaction (qPCR), as well as its conditions.

| Primer Pairs | qPCR Target Regions | qPCR cycling conditions |
|---|---|---|
| **S-D-Bact-0049-a-S-21 – 68F** TNANACATGCAAGTCGRRCG | 294 bp fragment V1 – V2 regions: Based on Sulaiman *et al.* (2021) | 5'' at 95°C<br>10'' at 50°C<br>30'' at 72°C<br>40 cycles |
| **S-D-Bact-0343-a-A-15 – R357** CTGCTGCCTYCCGTA | | |
| **S-D-Bact-0564-a-S-15 – 520F** AYTGGGYDTAAAGNG | 221 bp fragment V3 – V4 regions: Based on Monteiro *et al.* (2018) | |
| **S-D-Bact-0785-a-A-21 – 805R** GACTACHVGGGTATCTAATCC | | |

The *16S* qPCR reactions were performed in a 7500 Fast Real-Time PCR System using a mixture containing 5µL of 2X SensiFAST SYBR Lo-ROX Mix (Meridian Bioscience); 1µM of each primer; 1µL of DNA and RT-PCR Grade $H_2O$ (Ambion) up to a final volume of 10µL. The cycling conditions were as indicated in table 4.

In each qPCR experiment, up to 23 samples of seminal plasma DNA were analyzed simultaneously with 5 standards derived from *Escherichia coli* genomic DNA (Thermo Fisher – kindly donated by the i3s GenCore Platform) and a negative qPCR control. All of them were run in triplicate.

To generate qPCR standard curves, serial dilutions of *E. coli* stock solution comprising 30 ng/µL of DNA were prepared according to the formula below (Equation 1). The number of *16S* copies used in most experiments were: $10^6$, $10^5$, $10^4$, $10^3$, and $10^2$ copies. However, to achieve accurate bacterial load values for samples exceeding $10^6$ or under

$10^3$ an extra experiment was carried out with two additional dilutions of $5x10^6$ and $5x10^2$. The *DNA extraction control* was also quantified, in order to establish a *background noise* value. This means that samples containing a value lower than this *background noise* were removed from further analysis.

Equation 1 – Equation used to calculate *16S* copy number. (According to Whelan *et al.* (2003) [83].

$$\text{DNA (copy)} = \frac{6.02 \times 10^{23}(\text{copy/mol}) \times \text{DNA amount(g)}}{\text{DNA length(dp)} \times 660(\text{g/mol/dp})}$$

To examine if the bacterial load differed between infertility status or among phenotype groups several t-tests were employed, in which the two-tailed p-value was observed.

## 3.3. High-throughput sequencing of *16S* gene

The sequencing of the *16S* gene was performed by the GenCore Platform at i3S. Briefly, the 68 seminal DNA samples together with the *DNA extraction control* were submitted to a quality control using the *Qubit® dsDNA HS Assay kit* and the *Qubit® 3.0* fluorometer (Invitrogen). Subsequently, seven *16S* hypervariable regions were amplified with *Ion 16S™ Metagenomics Kit* (Thermo Scientific), which uses two primer sets: one multiplex targets regions V2, V4 and V8 and the other one generates amplicons for V3, V6-7 and V9. The PCR reactions were carried out according to manufacturer's instructions and using 4µL of total DNA input, which corresponded to a concentration range of 0.52 to 57.5 ng/µL for seminal DNA samples and an unquantifiable amount for the *DNA extraction control*. For each independent experiment of *16S* amplification a *negative PCR control* was also included. These were later pooled and treated as a single *PCR control* sample. All amplified targets were then subjected to another quality control step using the *Agilent 2200 TapeStation* (*HS D1000 Screen Tape;* Agilent Technologies). In order to achieve the minimum required DNA quantity for library preparation, samples were divided into three categories according to their PCR product concentrations as measured by the *Agilent 2200 TapeStation* and the volumes used as input (Table 5).

Next, the *Ion Plus Fragment Library Kit* and the *Ion Xpress™ Barcode Adapters* (Ion Torrent, Thermo Fisher Scientific) were used for library construction.

Table 5 – Volume of *16S* PCR products used as inputs to create DNA libraries for high-throughput sequencing.

| PCR products concentration | Volume used in library construction (µL) |
|---|---|
| > 20 ng/µL | 1 |
| 2 – 20 ng/µL | 5 |
| < 2 ng/µL | 8 |

A third quality control check was carried out with *Agilent 2200 TapeStation – HS D1000 Screen Tape* to verify the constructed libraries. Finally, template preparation and Semiconductor sequencing were done using the Ion Torrent S5 XL System – Ion 530 chip kit (Ion Torrent, Thermo Fisher Scientific).

Afterwards, data from the Ion Torrent S5 XL System was processed with the Ion Torrent platform specific pipeline software Torrent Suite v5.12, to generate the sequence reads, trim the adapter sequences, filter and remove poor signal reads and split the reads according to the barcode.

The obtained FASTQ and Bam files were then imported to the *Ion Reporter™ Specific pipeline for Metagenomics analysis*, which was used to align the results against two reference databases of *16S* – Greengenes and MicroSEQout *16S* reference library. This pipeline also removes non-specific amplifications, identifies and filters PCR chimeras, clusters sequences into Operational Taxonomic UnitoutOTU) per individual sample and per hypervariable region and performs their taxonomic assignment.

## 3.4. Statistical analysis

The statistical analysis was done using the Marker Data Profiling (MDP) module of MicrobiomeAnalyst software [84], which was designed for the analysis of *16S* data and requires four different types of input files [84, 85]:

1) A metadata file (.txt), containing all relevant information for statistical analysis processing. Precisely, in our study, this file included for each sample its grouping according to the following variables: bacterial load (high or low), infertility status (case or control) and phenotype (NRM, SHV, OAT or OAT+SHV);
2) A samples OTUs abundance file (.txt), which was generated by merging the individual OTUs files outputted from the *Ion Reporter™ Specific pipeline for Metagenomics analysis*;
3) A taxonomy mapping table (.txt) ranking each OTUs from phylum to its lower identified taxonomic rank. Although in most instances it was possible to identify the genus rank (identified genera), in some cases only the family rank or a list of possible genera was retrieved. When this list of possible genera contained two those were globally labeled as *ambiguous genera*, and when it exceeded two, it would merge with all OTUs belonging to the same family with no genus identification into a soutle OTU and designated as *unidentified genera*;
4) A phylogenetic tree (.tre) representing the evolutionary relationships among the identified OTUs. To generate this file, a *16S* sequence was selected as representative of each genus, aligned using the MAFFT online version software (https://mafft.cbrc.jp/alignment/server/) [86, 87] with default parameters, and then ran through FastTree (version 2.1.11) tool to create the phylogenetic tree [88]. Once output files were created, the resulting phylogenetic trees were inspected for large inconsistencies. Only minor taxonomic shifts across some genera were detected.

Files types 2 to 4 were generated only for the V3, V4 and V6-7 hypervariable regions.

Next, several criteria were defined to trim collected data prior to downstream analysis. First, a low count cut-off was applied to remove OTUs with less than 0.1% mean abundance across all sequenced seminal samples. This step was implemented to remove less biologically significant taxa, as well as taxa present due to sample contamination. This filter was applied together with a prevalence in samples filter of 10%. This means that if a taxon is present in over 10% of samples with an equivalent or bigger number of reads than the 0.1% threshold, it will not be removed.

Then, to allow more meaningful comparisons of collected data, it was normalized by the total sum scaling method [84, 85]. In this work, normalized read counts will be referred as taxon abundance.

Finally, the collected bacterial abundances were used in: 1) the taxonomic profiling of seminal samples, with Z-scores being used to test taxa abundance differences between

groups; and 2) in the analysis of alpha-diversity (within-sample diversity) through Chao1, Shannon and Simpson diversity indices, and of beta-diversity (between-sample diversity) through Bray-Curtis, Jensen-Shannon, Jaccard, unweighted and weighted UniFrac distances. The Mann-Whitney test was used in the comparison of alpha-diversity indices between sample groups as defined by bacterial load, infertility status and phenotype to visualize significant associations between these sample groups and taxon abundances, while beta-diversity analyses were performed with PCoA ordination method and Permutational MANOVA statistical method to explore dissimilarity between samples.

Furthermore, to facilitate the identification of samples with similar microbiome profiles a heatmap using genera abundance was generated by Euclidean distance and the Ward clustering algorithm.

In addition, to detect with high statistical power the taxa differing in their abundances according to tested variables (bacterial load, infertility status and phenotype) the edgeR (log2 of Fold Change – log2FC and log of Counts per million – logCPM statistics), DESeq2 (log2FC and log2FC Standard Error – lfcSE statistics) and linear discriminant analysis (LDA) effect size (LEfSe) packages were implemented through the MicrobiomeAnalyst software [89].

# 4. Results and Discussion

## 4.1. Total bacterial content per sample and groups

For each sample to quantify their microbial content, bacterial load estimates were obtained based on *E. coli* standard curves. Those values showed a large variability in the amount of bacterial DNA across samples, ranging from 345 to 1,796,043 copies of *16S* gene (Figure 5A).

First, by taking into account the number of *16S* copies achieved for the *DNA extraction control* a baseline of contaminants was established and labeled here as *background noise*. Thus, a sample falling below that baseline was removed from the following analyses (Figure 5A). Then, the other samples were divided into *high* and *low bacterial load* according with an intermediate value of 30,000 copies of the *16S gene*. This number was selected not only because it was close to the average of bacterial load estimates as it represented also a good compromise with the separation into the three sample input groups used for the library construction (see section 3.4). Basically, whereas samples using a *1 uL volume* of *16S* amplicons tended to be above the 30,000 copies threshold, the ones with an input of a *8 uL* volume were all below (Figure 5A).

In overview, if a sample presented more than 30,000 copies it was labeled as *high bacterial load*, while in the opposite scenario it would be defined as a *low bacterial load* sample.

To investigate if an augmented content of bacteria could be related with male infertility in general, or with any specific phenotypes, several comparisons were made between cases and controls (infertility status) and across phenotypes. No significant differences were obtained between groups that might indicate a correlation between a male infertility condition and an augmented content of bacteria as asserted by *high bacterial load* (T-test: $P>0.05$ Figure 5B and 5C). Nevertheless, given the large variability observed in the total bacterial content, from this step forward the *high*- or *low bacterial load* was considered in the study as an independent variable.

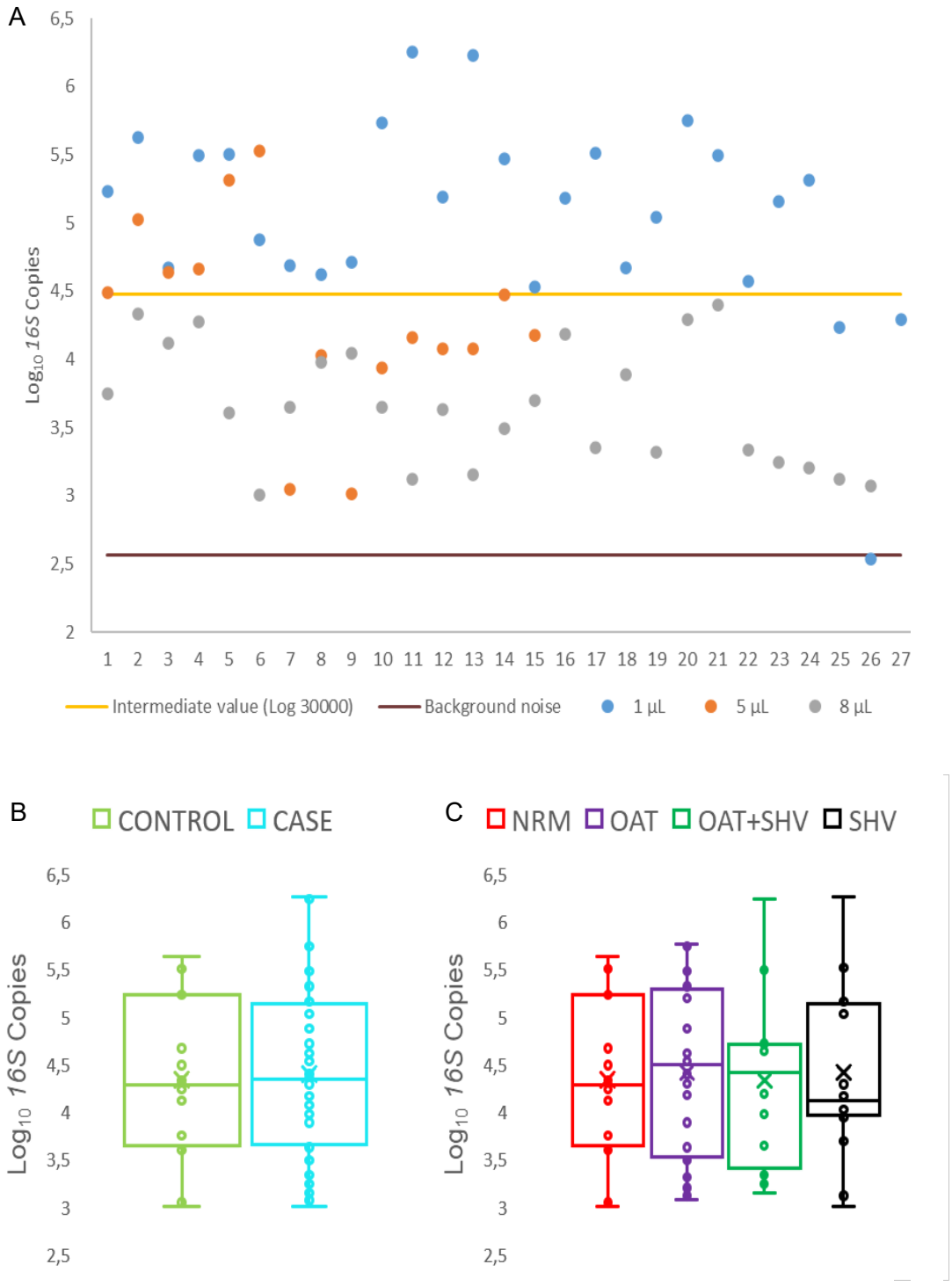



Figure 5 – Bacterial load estimates. A) Scatter plot of the bacterial load per individual sample. The color code indicates the volume of *16S* PCR products used in the library preparation (see section 3.2; Table 5). The yellow line indicates the value of *background noise* as determined by the amplification of the *DNA extraction control* and the burgundy line represents the 30,000 copy threshold used in the classification of *high-* and *low bacterial load* samples. B) Plot of the bacterial load estimates for cases and controls (infertility status). C) Plot of the bacterial load estimates for the studied phenotypes. The lines in each boxplot represent the median while the X indicate the mean. No statistically significant differences between groups were observed (T-Test: Two-Sample Assuming Unequal Variances: p-values > 0.05).

The seminal microbiome has been described to have a low bacterial content when compared with other systems, unless in cases of bacteriospermia and/or acute infection [90, 91]. So far, a single study by Baud *et al.* (2019) has attempted to evaluate the bacterial load of semen samples in infertile cases and controls. Similarly to our findings they did not detect any evidence for a direct link between an increased bacterial load and a poor quality of the semen [70].

## 4.2. Overview of the *16S* sequencing

Sequence reads for the seven hypervariable regions screened by the *Ion 16S™ Metagenomics* kit (V2, V3, V4, V6-7, V8 and V9) were obtained in all seminal samples, as well as in the *DNA extraction control* and the *negative PCR control*. However, the different regions displayed very different amplifications yields as indicated by the average number of reads shown in Table 6. These results are partially concordant with the evaluation of the *Ion 16S™ Metagenomics* kit conducted by Barb *et al.* (2016), which reported the highest and lowest number of reads being assigned to regions V3 and V9, respectively, but with a much higher number of reads in the V8 region than what our study detected [92]. Moreover, the same study also found that V2, V4 and V6-7 hypervariable regions provided a better taxonomic resolution at family and genus rank [92].

Table 6 – Average number of reads obtained per screened *16S* hypervariable region.

| ***16S* Hypervariable Region** | **Average number of reads per sample (Standard deviation)** |
|---|---|
| V2 | 7112 (±3715) |
| V3 | 26563 (±15082) |
| V4 | 11521 (±5276) |
| V6-7 | 12761 (±7220) |
| V8 | 8939 (±5551) |
| V9 | 1197 (±3520) |

Taking all of this into consideration, in the present work, regions V2, V8 and V9 were excluded and only V3, V4 and V6-7 regions were used in the downstream analyses. Concerning the OTUs assignment, some differences were also observed across the three selected hypervariable regions, such as the number of identified OTUs and its resolution at genus rank (Table 7).

Table 7 – Evaluation of the taxonomic resolution achieved at the genus rank for V3, V4 and V6-7 hypervariable regions*

| | | **V3** | **V4** | **V6-7** |
|---|---|---|---|---|
| Genera* | Identified genera | 36 (61%) | 32 (74%) | 21 (75%) |
| | Ambiguous genera[a] | 6 (10%) | 4 (9%) | 3 (11%) |
| | Unidentified genera[b] | 17 (29%) | 7 (17%) | 4 (14%) |
| | Total | 59 | 43 | 28 |

* The shown numbers correspond to the genera that passed the > 0.001 frequency and 10% prevalence filtering criteria (section 3.3).
[a] – OTUs that showed 2 likely genera.
[b] – OTUs that presented 3 or more possible genera and were merged into single OTUs, or taxa identification terminated at the family rank.

Although the V3 region provided the highest number of genera identified with no ambiguities (N= 36), it was the V4 that delivered the better results when considering the *ambiguous*, *unidentified* and *identified* taxonomic groups at the genus rank (74%). These results are consistent with the previous mentioned evaluation of Barb *et al.* (2016), in which it was demonstrated the V4 region shows a higher taxonomic resolution power than V3. Furthermore, our results are also in agreement with a former investigation performed by Yang *et al.* (2016), in which the taxonomic resolution of each hypervariable region was appraised, showing the V4 region as the best one displaying an increased accuracy (highest sensitivity) and thus being considered the best marker for taxonomic characterization and phylogenetic analysis of bacteria [93].

For simplicity, the following sections (4.3 and 4.4) only contemplate the results obtained for the V4 hypervariable region. The remaining data generated for the V3 and V6-7 regions is included in the Annex section (Annexes IV and V, respectively).

Along with the reads generated for seminal samples, which for V4 region averaged the 11,521 as shown in Table 6; 15,203 and 193 reads were obtained in *DNA extraction control* and *Negative PCR control* samples, respectively (Annex II: Table 13). While the numbers achieved for the *Negative PCR control* did not raise any concerns, the same could not be applied to the *DNA extraction control* given some seminal sample delivered lower numbers, as low as 963 reads. Although *a single sample* was excluded based in the qPCR analyses, here we used the *DNA extraction control* to perform also a qualitative analysis of possible contaminants (Annex II: Table 13).

## 4.3. Semen Microbiome composition

### 4.3.1. Microbiome Profiling

The analysis of seminal microbiome composition uncovered Firmicutes as the most abundant phylum reaching a frequency of 77.7% when considering either cases or controls (Table 8 and Figure 6) and ranging from 70 to 87.5% if accounting the distinct infertility phenotypes (Table 8). The remaining identified phyla, namely Proteobacteria, Actinobacteria and Bacteroidetes all displayed frequencies below the 18%. These phyla presented similar frequencies in both cases and controls (Proteobacteria: 12.3 vs 15.3%; Actinobacteria: 7.5 vs 3.9%, respectively and Bacteroidetes ranging 2-3% in both groups; Table 8) and also per phenotype (Proteobacteria: 9.6-15.3%; Actinobacteria: 3.9-17.7% and Bacteroidetes at ~3% in all groups except SHV – 1.1%) (Table 8). Overall, these results agree with the Monteiro *et al.* (2018) work [34] carried out by our group, in which a ION sequencing chemistry was also employed and where equivalent abundances for Firmicutes (~70%) were observed in 3 out of the 4 analyzed groups - Controls, Asthenoteratozoospermic and OAT. Similar abundances were detected in these 3 groups as well for Proteobacteria (10-15%), Actinobacteria (8-12%) and Bacteroidetes (4-8%). The single exception to this common pattern registered by the Monteiro *et al.* (2018) study was verified in a SHV group, which displayed a much lower frequency of Firmicutes (~50%) accompanied by an augmented abundance of Proteobacteria (~25%) [34]. The current results of this study show the opposite trend – an increase in the abundance of Firmicutes (87.5%), and a reduction of Proteobacteria (9.6%). These results will be discussed with further detailed in a downstream section centered in the comparative analysis of male infertility phenotypes (section 4.4)

Table 8 – Frequencies for the observed phyla and the ten more abundant genera found in semen samples.

| Taxon | Relative Frequencies (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Infertility status | | Phenotype | | | Bacterial load | |
| | Case | Control | OAT | OAT+SHV | SHV | High | Low |
| Phyla | | | | | | | |
| Firmicutes | 77.7 | 77.7 | 73.9 | 70.8 | 87.5 | 95.6 | 54.9 |
| Proteobacteria | 12.3 | 15.3 | 14.3 | 12.6 | 9.6 | 0.7 | 28.5 |
| Actinobacteria | 7.5 | 3.9 | 8.5 | 13.6 | 17.7 | 2.8 | 11.7 |
| Bacteroidetes | 2.4 | 3.0 | 3.2 | 3.0 | 1.1 | 0.8 | 4.8 |
| Genera | | | | | | | |
| *Enterococcus* | 55.6 | 58.2 | 48.0 | 55.4 | 65.6 | 73.5 | 33.9 |
| *Staphylococcus* | 12.7 | 12.0 | 13.5 | 2.4 | 12.6 | 15.8 | 8.2 |
| *Cupriavidus* | 7.9 | 10.7 | 10.4 | 9.4 | 4.1 | 0.2 | 19.1 |
| *Streptococcus* | 2.1 | 1.0 | 3.2 | 0.5 | 1.7 | <0.1 | 4.2 |
| *Facklamia* | 2.1 | 0 | 4.6 | 0.4 | <0.1 | 2.8 | <0.1 |
| *Corynebacterium* | 2.0 | 0.8 | 1.2 | 5.4 | 0.7 | 0.5 | 3.4 |
| *Actinobaculum* | 2.0 | <0.1 | 3.1 | 2.3 | 0.4 | 1.4 | 1.8 |
| *Peptoniphilus* | 1.7 | 0.9 | 2.4 | 2.9 | <0.1 | 1.4 | 1.7 |
| *Escherichia/Shigella* | 1.5 | 0.3 | 0.2 | 0.4 | 3.9 | <0.1 | 2.8 |
| *Finegoldia* | 1.4 | 1.3 | 0.9 | 4.2 | 0.1 | 0.6 | 2.4 |

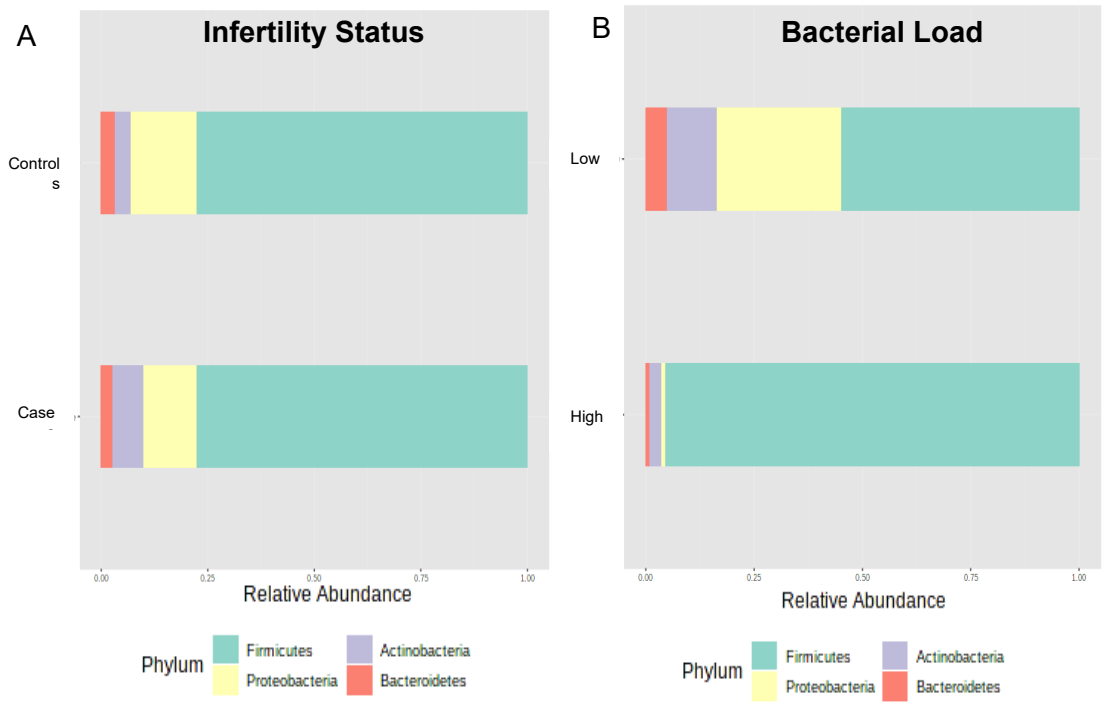



Figure 6 – Microbiome profiles of most abundant phyla (*f*>0.1%) according to V4 hypervariable region of 16S gene. (A) infertility status (B) bacterial load.

It is known that the proportions of the different bacterial taxa may vary across studies. For example, Mändar *et al.* (2017) reports a lower prevalence of Firmicutes (~50%) and an increased abundance of Bacteroidetes (20%) in men with or without prostatitis [67]. Conversely, Chen *et al.* (2018) described Proteobacteria as more common in controls (~42%) than in cases with obstructive or non-obstructive azoospermia (22 and 18%, respectively), and the opposite tendency for Firmicutes (~30% in controls and ~45% in azoospermic samples). The same study also showed for control samples a higher abundance of Actinobacteria (~20%) combined with a reduced prevalence of Bacteriodetes (3%), while in azoospermic samples the Bacteroidetes superseded the Actinobacteria (15-20% versus 5-8%, respectively) [66]. Nonetheless, Altmäe *et al.* (2019), in a review paper that combines 8 high-throughput studies of the seminal microbiome indicates the prevalence of Firmicutes to be around 50% and of Proteobacteria to be ~35% in healthy men [94]. These discrepancies could be due to differences in studied populations with different genetic backgrounds [34, 65, 69, 70], as well as methodological issues, such as differences in the selected *16S* primers, or the use of different high-throughput sequencing platforms [69, 95-98].

At this point, no statistically significant differences were obtained for the group comparisons, neither for infertility status (cases versus controls) nor per phenotype (Table 8). However, when discriminating the samples according to their bacterial content, a significantly higher frequency of Firmicutes was detected for the *high bacterial load* (~95%) in contrast with the *low bacterial load* (~55%; Table 8 and Figure 6B) as depicted by the low p-value on a Z-score test (p-value: 0.00026). Indeed, the other common phyla Proteobacteria, Actinobacteria and Bacteroides where nearly absent from the *high bacterial load* samples, differing from the *low bacterial load* group where those phyla showed prevalence's of ~28%, ~11%, and ~5%, respectively.

The depicting of the bacterial communities of the semen at the genus rank showed that Firmicutes abundances are mainly explained by *Enterococcus,* which represents more than 50% of all taxa found in cases and control samples (Table 8 and Figure 7A). Although *Enterococcus* is commonly found in the seminal microbiome and identified as a common cause of bacteriospermia [42], no other study to date reported such high prevalence for this genus. In a general profiling of the seminal microbiome, together with *Enterococcus* are several other genera reaching relative abundances above the 0.1%, those comprise *Staphylococcus, Cupriavidus, Streptococcus, Facklamia, Corynebacterium, Actinobaculum, Peptoniphilus, Escherichia/Shigella* (ambiguous) and *Finegoldia* (Table 8). From these genera the prevalence of *Cupriavidus* might be overestimated given it was identified as the most abundant genus in both negative controls, the *negative PCR control* and the *DNA extraction control* (Annex 2: Table 12), suggesting this taxon as common contaminant in our laboratory. Nonetheless, it cannot be neglected that controls showed in both instances extreme low yields of *16S* amplification and maximum amplicon quantities were used as input for library construction prior to the high throughput sequencing.

At this point, no statistically significant differences were detected between cases and controls for the most prevalent genera, as well as between phenotype groups. Again, the sample stratification in *high* or *low bacterial load* produced different results, where the observed Firmicutes increment is apparently explained by an overdominance of *Enterococcus* and *Staphylococcus*. *Enterococcus* presents a significantly higher proportion in *high bacterial load* samples (73.5%) than in the *low bacterial load* group (33.9%), as depicted by the low p-value of the Z-score test (p-value: 0.00112), while the p-value for the same test for *Staphylococcus* proved to be nonsignificant. Curiously, both taxa have been previously correlated with a dysbiosis of the seminal microbiome [42].

Concerning the bacterial pathogens known to cause STIs none was detected in our study, or at least it did not pass our filtering criteria. Notwithstanding, *Neisseriaceae,* the family rank to which *N. gonorrhea* belongs*,* was pointed out as an OTU present in *low bacterial load* samples.

These results do not entirely comply with previous studies from our group. Although Monteiro *et al*. (2018) indicated *Enterococcus* and *Staphylococcus* as the most prevalent genera, their abundances across the different studied groups vary between 22-32% and 6-15%, respectively [34]. These results are even more striking when compared with the ones from other groups carried in samples from other populations and using mainly Illumina sequencing technologies. Despite the lack of consensus across the different independent studies, other most frequently identified genera in seminal samples include *Ralstonia, Lactobacillus, Prevotella, Corynebacterium, Pseudomonas, Streptococcus, Finegoldia* and *Anaerococcus*, which were detected in our study at a relative abundance below 2%, as well as *Ureaplasma, Clostridiales, Atopobium, Gardnerella, Rhodanobacter, Haemophilus, Planococcaceae* and *Burkholderia* that were not found in our samples [5, 65, 69]. Notably, a very recent study conducted by Yao *et al.* (2022) also reported *Enterococcus* as a prominent genus in semen samples together with *Veillonella*, *Acinetobacter*, *Rhodococcus* and *Peptoniphilus* [99].


Figure 7 – Microbiome profiles of most abundant genera (f>0.1%) according to V4 hypervariable region of 16S gene. (A) infertility status. (B) bacterial load. Entries such as f__Comamonadaceae:g_ represent the grouping of when the genus was not discriminated.

Similar findings with slight variations in some taxa were obtained for the V3 and V6-7 hypervariable regions (Annex IV: Figures 20 and 21; Table 13 and Annex V: Figures 26 and 27; Table 14)

To provide further insights about the interindividual variability of the semen microbiome, the taxa abundance was then analyzed per sample (Figure 8). Given that samples mainly diverged in their composition according with the bacterial load (high or low), the sample plotting in Figure 8 takes this stratification into account. Equivalent plots for infertility status and phenotype groups are provided in the Annex section (Annex III: Figures 18 and 19).

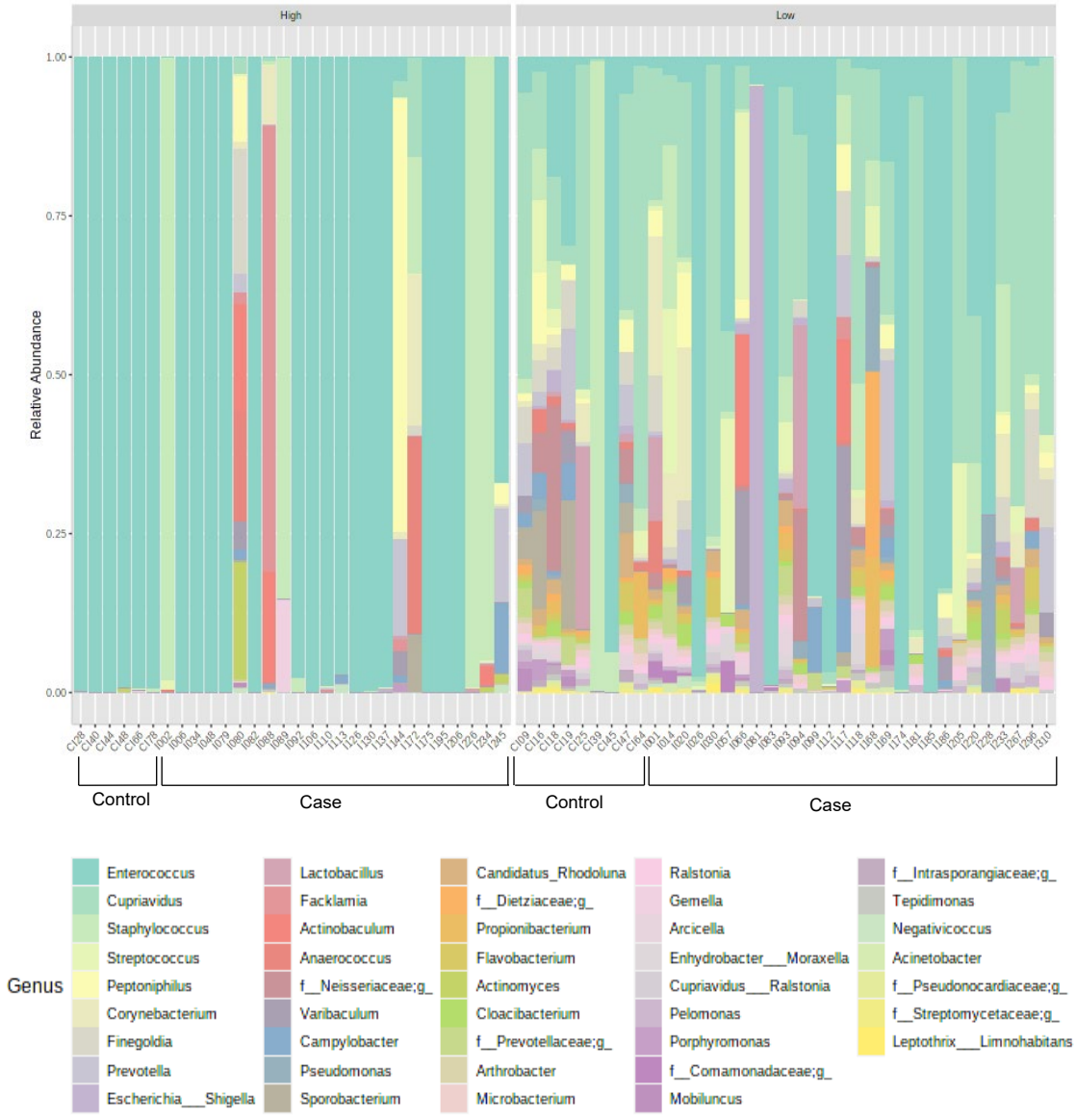



Figure 8 – Microbiome profiles of most abundant genera (f>0.1%) according to V4 hypervariable region of 16S gene scattered per individual sample and divided by bacterial load grouping (high or low). Samples corresponding to cases and controls are indicated.

These results show that there is a large interindividual variability in the composition of the seminal microbiome, in which as indicated by previous findings, *Enterococcus* basically dominates most of the *high bacterial load* samples, being in some instances replaced by *Staphylococcus* (Figure 8). Although these results seem to agree with the principle of dysbiosis when one or few genera grow substantially in total amount often overcoming other taxa previously present in the sample [100], this hypothesis may be contradicted by an associations with both cases and controls (Annex III: Figure 18) without reported evidence of infection or bacteriospermia .

On the other hand, the *low bacterial load* samples show a larger taxonomic richness where in most subjects more than 10 genera were identified. Even though, *Enterococcus* and *Staphylococus* persist as the most prevalent genera in this bacterial load group, indeed 5 samples are almost completely dominated by *Enterococcus* and 1 by *Staphylococcus* (Figure 8; Table 8).

### 4.3.2. Sample discrimination into distinct microbial groups

To investigate how the seminal samples clustered based on their microbial composition and how those could be correlated with the three main analyzed variables (infertility status, phenotype and bacterial load) a heatmap was generated using the relative abundances of the identified genera (Figure 9). This analysis uncovered a stratification of the samples in four clades (or groups) as shown in the top cladogram of the heatmap. The first clade, the *clade 1,* showed a high similarity across samples therein while largely diverging from the remaining clades Furthermore, this clade could be correlated with the high prevalence of the *Enterococcus* genus and in most instances with *high bacterial load* samples (Figure 9). The *clade 3* also displayed a high similarity across samples, but in this turn associated with the overdominance of *Staphylococcus* and a *high bacterial load*, too (Figure 9).

Conversely, the *clade 2* and *clade 4* both grouped more heterogenous samples composed by multiple genera. However, while the *clade 2* could be related with a low bacterial load and a sharing of the same genera like *Cupriavidus* (which could indicate environmental contamination, as this genus is found in high abundances in the *DNA extraction control*) *Microbacterium* and *Ralstonia*; the *clade 4* combined the most divergent samples largely varying in their taxonomic composition and also showed a less clear trend toward a *low bacterial load*.

These results meet the previous findings illustrated in the Figure 8 where the *high bacterial load* samples are in most circumstances associated with an overrepresentation of *Enterococcus* or *Staphylococcus* genera.

No pattern of clustering according to infertility status or phenotype was observed.

Similar clustering patterns were observed for the v3 and v6-7 regions (Annex IV: Figure 22 and Annex V: Figure 28).

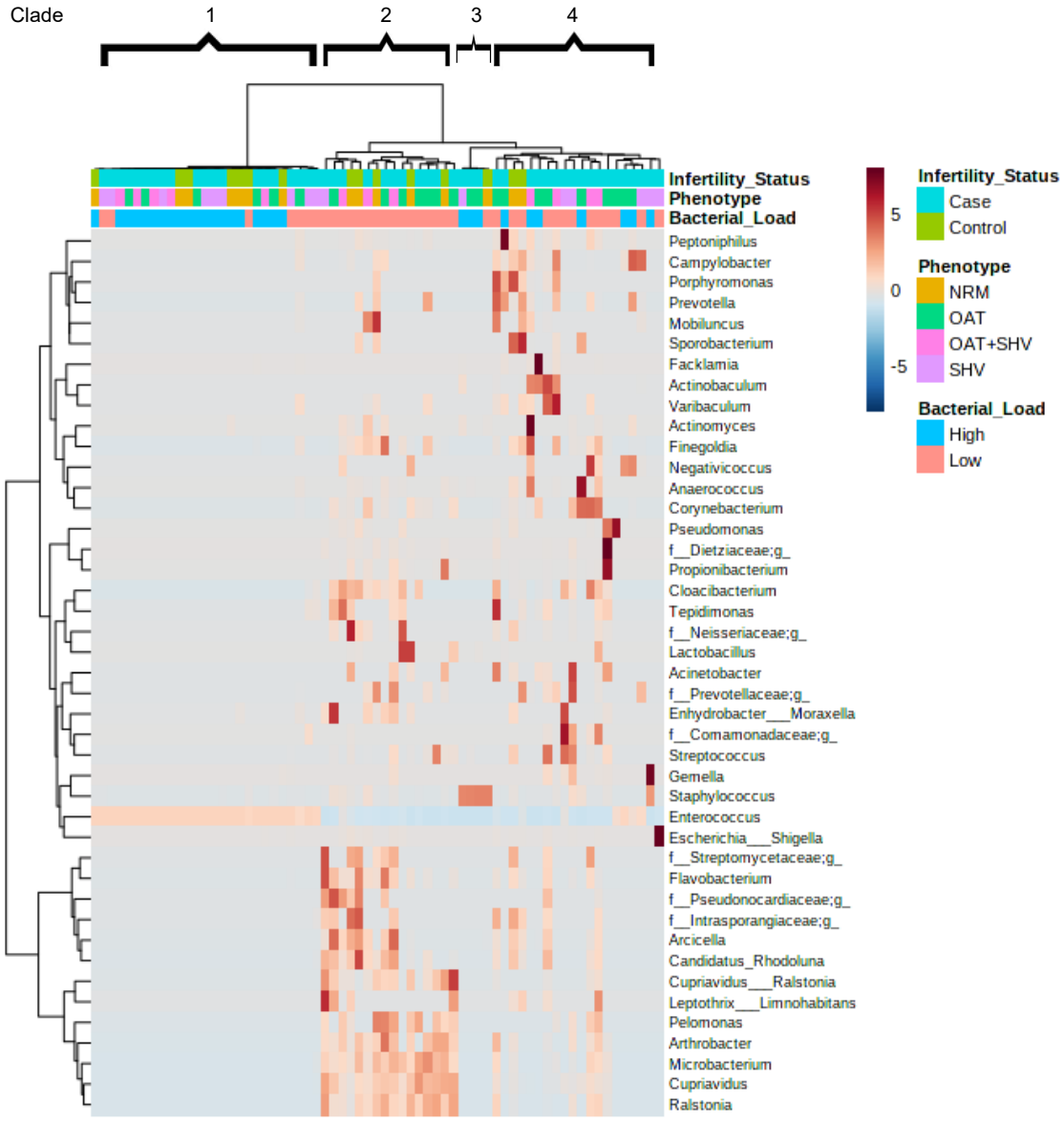



Figure 9 – Heatmap of the identified genera (>0.1%) and their relative abundances (V4 region data).

### 4.3.3. Characterization of the seminal microbiome diversity

To further evaluate the composition of the seminal microbiome and its microbial richness several alpha diversity indices were calculated. These analyses showed that independently of the considered index, the *high bacterial load* samples always displayed a statistically significant lower diversity than *low bacterial load* samples (Figure 10A, B and C). These results provide support to the previous conjectures based strictly on the inspection of individual microbial profiles.

On the other hand, and as expected from the previous findings, no statistically significant differences were observed for any of the alpha diversity indices calculated when considering infertility status (Figure 10D, E and F) or phenotypes (see section 4.4.). Congruent results were observed for V3 and V6-7 hypervariable regions (Annex IV: Figure 23 and Annex V: Figure 29).

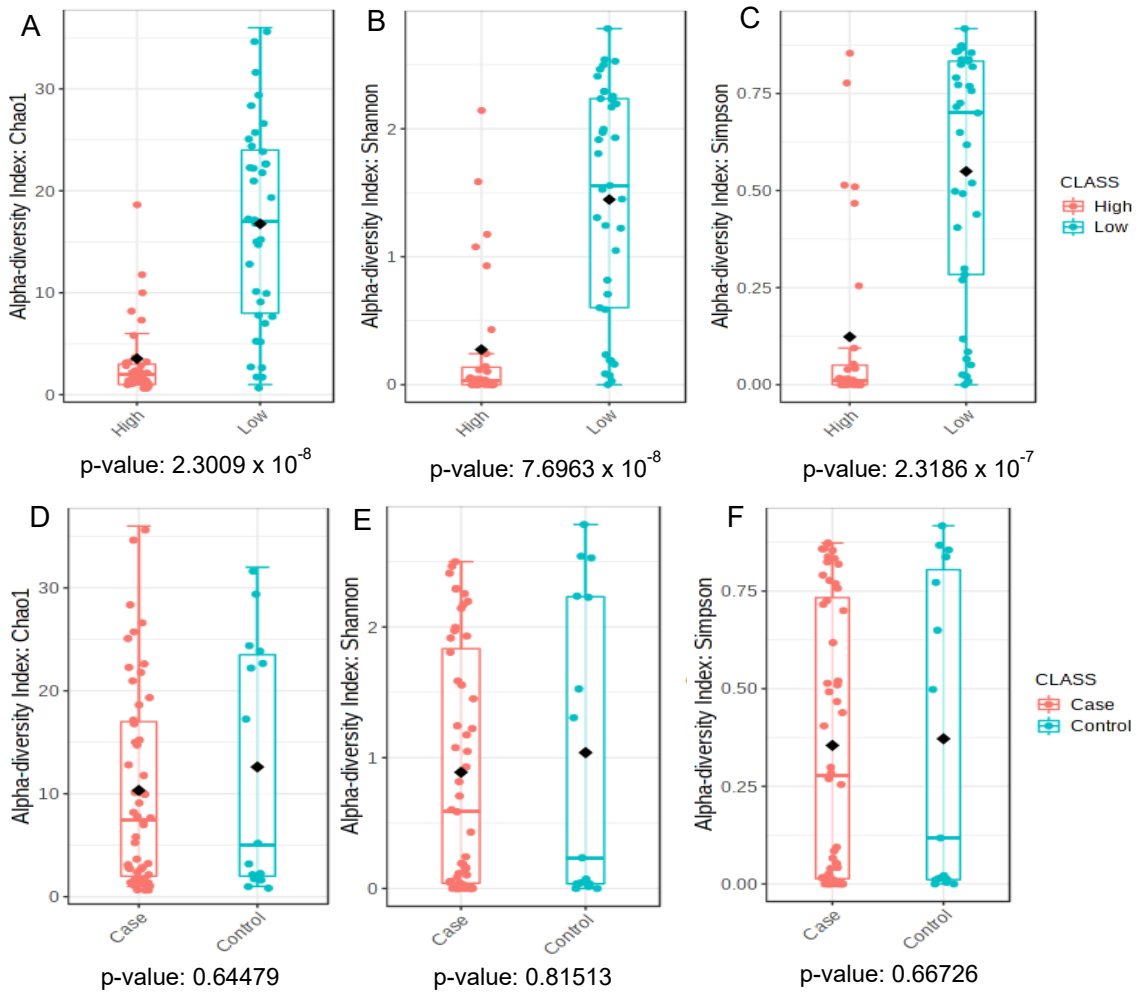



Figure 10 – Alpha diversity (Chao1, Shannon and Simpson indices) of seminal samples based in the V4 hypervariable region. Stratification of the samples according with bacterial load (A, B, C) and infertility status (D, E, F).

In order to appraise the similarities across the microbial profiles observed for each sample, distinct beta diversity indices were calculated, and distances plotted in PCoA (Bray-Curtis, Jensen-Shannon, Jaccard, Unweighted and Weighted UniFrac; Figure 11). The Bray-Curtis index, as well as Jensen-Shannon and Jaccard, uncovered that seminal microbiome samples separated into three distinct clusters (Figure 11A, B and C). The *Cluster 1*, displayed minimal distances between samples and was far related to the remaining clusters; the *Cluster 2*, showed also a high similarity between samples as indicated by their short distances and finally the *Cluster 3*, was associated with a lower relatedness of microbial profile. The UniFrac distance indices revealed a differing clustering pattern, with only 2 clusters for the Unweighted UniFrac distance, and 3 for the Weighted UniFrac index, but where the most compact cluster replicates the high microbial similarity registered in the *Cluster 1* from the Bray-Curtis index (Figure 11D and E).

Statistically significant differentiation of microbial profiles according to bacterial load was registered with Bray-Curtis, Jensen-Shannon, Jaccard, Unweighted and Weighted UniFrac distance indices (p-value < 0.001 for all indices; Figure 11), thus supporting an association of the *high bacterial load* with more alike microbiomes as the ones as the *Cluster 1* and *Cluster 2*, and the low bacterial load with dissimilar microbial profiles of *Cluster 3*. Overall, beta diversity clustering for the V3 and V6-7 hypervariable regions seemed to align with the V4 region results (Annex IV: Figure 24; Annex V: Figure 30) Once again, no differentiation was observed while considering infertility status or phenotypes (not shown).

Given the previously observed aggregation of *high bacterial load* samples into clades 1 and 3 was connected with the high prevalence of *Enterococcus*, or *Staphylococcus,* respectively (Figure 9), we also tested the discrimination of microbial profiles based on these taxa. Statistically significant results were obtained for the Bray-Curtis (Figure 12) and for the 4 remaining indices (not shown) and as it is shown in the PCoA of Bray-Curtis (Figures 12), whereas the *Enterococcus* enriched samples perfectly matched the former identified *Cluster 2*, the *Staphylococcus* overrepresented samples exactly fited the *Cluster 1*. Conversely, in the *cluster* 3 these taxa were in most instances inexistent or found at very low proportions (Figure 12).

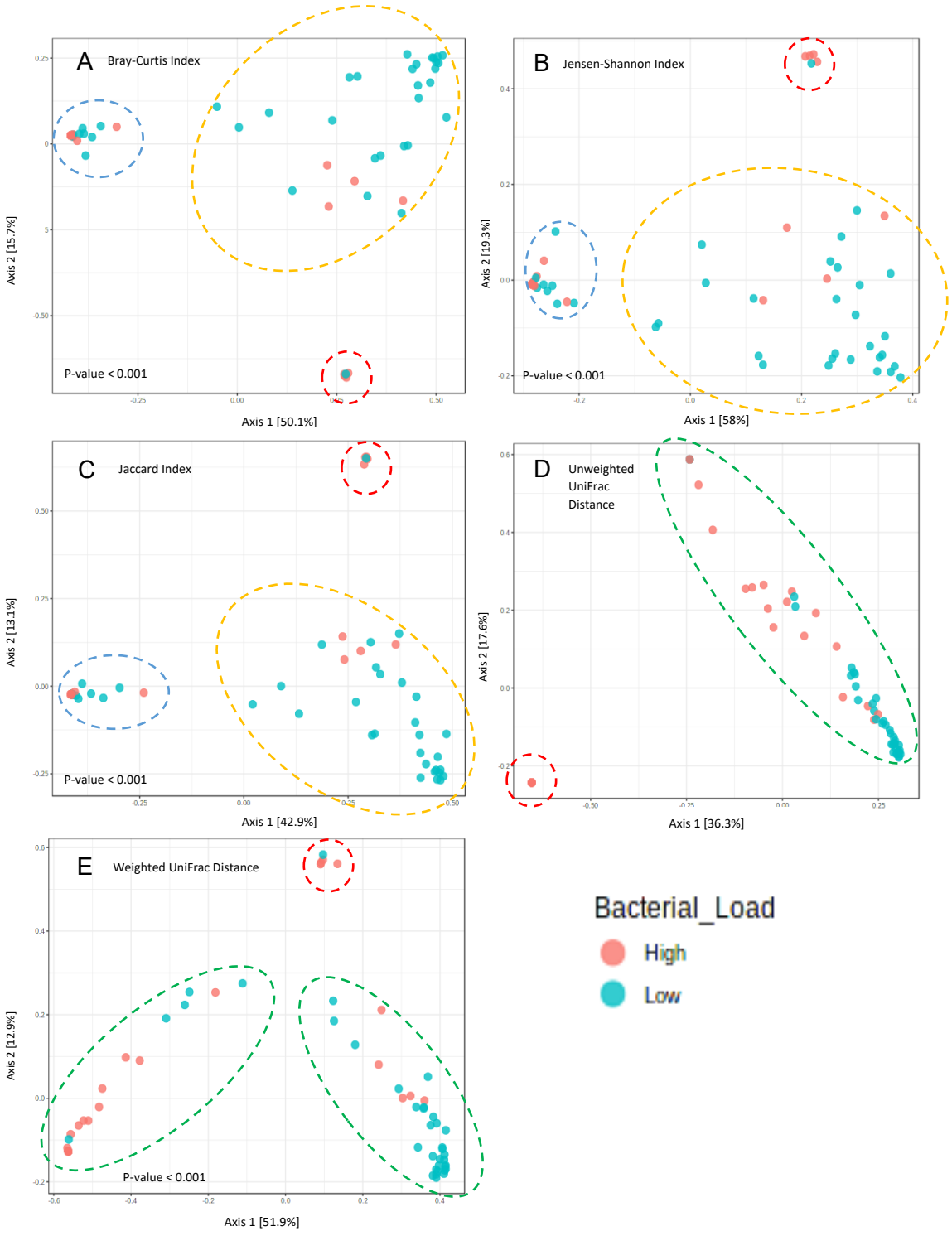



Figure 11 – Beta diversity as shown by principle coordinate analysis (PCoA) of Bray-Curtis (A), Jensen-Shannon (B), Jaccard (C), Unweighted UniFrac (D) and Weighted UniFrac (E) distances. Sample stratification according to Bacterial load groups is shown. In red, *Cluster 1*, while the blue is *Cluster 2* and in yellow, *Cluster 3*. In green, the different clustering of the UniFrac distance indices is demonstrated.

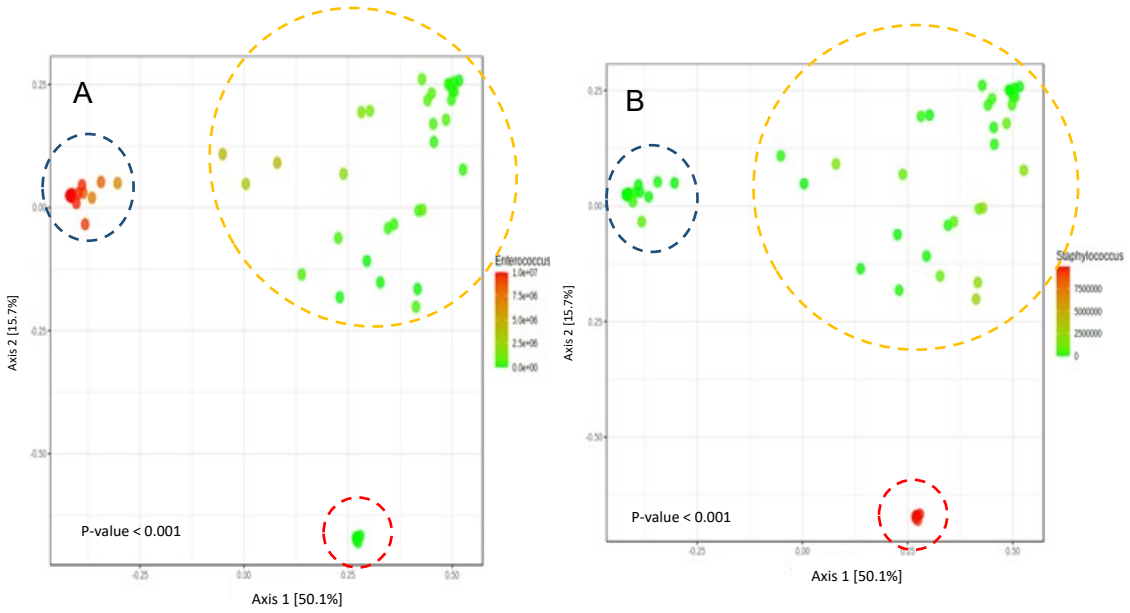



Figure 12 – Displayed microbial distances between samples using the Bray-Curtis index, according to abundance of *Enterococcus* (A) and *Staphylococcus* (B). Clustering pattern is the same is in Figure 11A.

## 4.3.4. Identifying taxa differing between groups

To further explore if any taxa besides *Enterococcus* and *Staphylococcus* could be correlated with bacterial load and/or infertility status, we performed several differential discriminating analysis tests implemented through LEfSe (LDA scores), edgeR (log2FC and logCPM statistics) and DESeq2 (log2FC and lfcSE statistics) software packages. In a first analysis of the data, LDA scores (LEfSe) indicated 33 significant taxa differing according to *bacterial load* (*high* or *low*), but none was pointed out if considering infertility status (cases and controls). Notably, 29 of those taxa were found in *low bacterial load* samples as illustrated in Figure 13, whereas *Enterococcus, Staphylococcus, Peptinophilus. and Anaerococcus* were the genera associated with a *high bacterial load*.

In a second analysis using log2FC and logCPM scores as estimated through edgeR, 36 genera were identified as significantly differing between *high* and *low bacterial load* samples and 12 as discriminating cases and controls (Table 9). Again, aside from being associated with most instances with *low bacterial load* samples those taxa were often found at reduced abundances and present in very few samples (Table 10). Conversely, *Enterococcus* and *Staphylococcus* (Table 10 and Figure 14) together with *Facklamia* and *Actinomyces* were found to be correlated with *high bacterial load* (Table 10).

Regarding the infertility status, we highlight *Facklamia, Actinobaculum and Escherichia_Shigella* taxa, which showed stronger significant results for both log2FC and

logCPM scores (EdgeR algorithms) and given their abundances those bacteria appear as more promising for the discrimination of cases from controls. Precisely, those genera varied between 1.5-2% in cases against 0-0.3% in controls (Table 9; Figure 15), Nonetheless, in an in-depth analysis of the cases showed that those three taxa were found in restricted samples, never exceeding the 12 observed in *Escherichia_Shigella* (Figure 15).

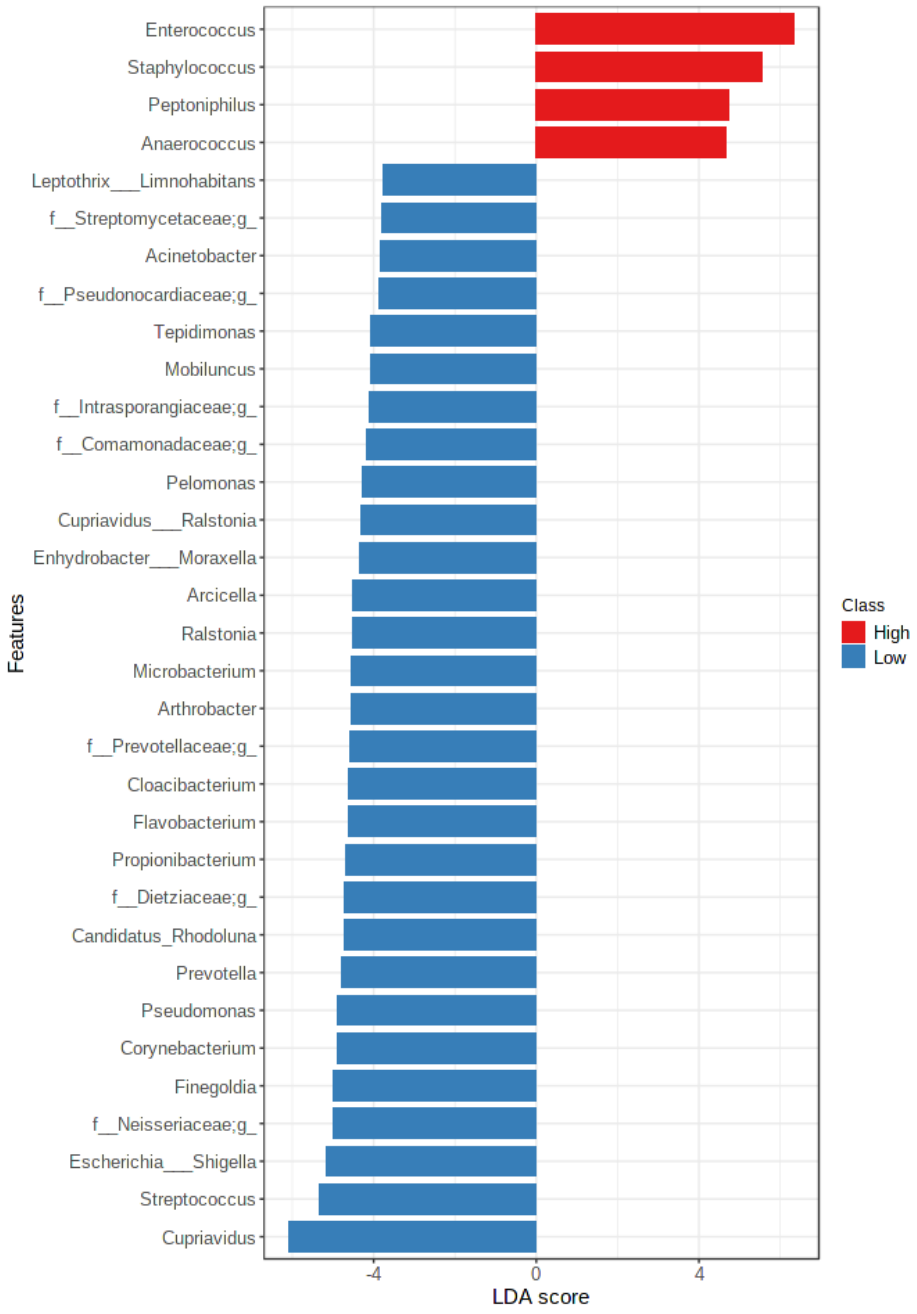



Figure 13 – Microbial differentiation of seminal samples according with the linear discriminant analysis (LDA) effect size (LEfSe) algorithm for *high* and *low bacterial load*.

Table 9 – List of taxa significantly differing according to Infertility Status based on Log2FC, LogCPM as implemented through edgeR algorithms. Relative abundances are indicated.

| Genus | Case | Control | Log2FC | LogCPM | P-value | FDR* |
|---|---|---|---|---|---|---|
| *Pseudomonas* | 0.7% | 0.2% | -4.6456 | 13.483 | 4.7514e-04 | 0.0040862 |
| *Facklamia* | 2.0% | 0% | -7.9617 | 14.915 | 7.8231e-06 | 3.061e-4 |
| *Escherichia___Shigella* | 1.5% | 0.3% | -5.7276 | 14.616 | 8.3142e-05 | 0.0011917 |
| *f__Comamonadaceae* | 0.2% | <0.1% | -3.0382 | 10.068 | 0.0010034 | 0.0071911 |
| *Leptothrix__Limnohabitans* | <0.1% | <0.1% | -1.9803 | 9.6538 | 0.0083111 | 0.029781 |
| *Gemella* | 0.5% | <0.1% | -3.9855 | 13.093 | 0.002447 | 0.011769 |
| *Peptoniphilus* | 1.7% | 0.9% | -3.1477 | 14.188 | 0.0075144 | 0.029375 |
| *Actinobaculum* | 2% | 0.1% | -6.7988 | 14.02 | 1.4237e-05 | 3.061-4 |
| *Tepidimonas* | 0.1% | <0.1% | -2.0602 | 9.5904 | 0.0025636 | 0.011769 |
| *Sporobacterium* | <0.1% | 2.1% | 2.212 | 10.702 | 3.4416e-04 | 0.0036997 |
| *Varibaculum* | 1.2% | 0.4% | -3.4251 | 12.824 | 0.0027369 | 0.011769 |
| *Campylobacter* | 0.9% | 0.6% | -3.618 | 13.615 | 0.0026193 | 0.011769 |

* FDR – False Discovery Rate

Table 10 – List of taxa significantly differing according to Bacterial Load based on Log2FC and LogCPM scores as implemented through edgeR algorithms. Relative abundances are indicated.

| Genus | High | Low | Log2FC | LogCPM | P-values | FDR* |
|---|---|---|---|---|---|---|
| *Microbacterium* | 0% | 0.5% | 3.9058 | 11.147 | 3.1926e-14 | 1.3728e-12 |
| *Ralstonia* | 0% | 0.5% | 4.1922 | 11.408 | 4.8709e-13 | 9.3196e-12 |
| *Arthrobacter* | 0% | 0.6% | 4.1411 | 11.361 | 6.502e-13 | 9.3196e-12 |
| *Candidatus_Rhodoluna* | 0% | 1.0% | 4.0191 | 11.249 | 9.3991e-13 | 1.0104e-11 |
| *Flavobacterium* | 0% | 0.8% | 4.3934 | 11.595 | 1.27e-12 | 1.0922e-11 |
| *Pseudomonas* | 0% | 1.3% | 6.3557 | 13.483 | 6.7264e-12 | 4.8206e-11 |
| *f__Neisseriaceae* | 0% | 1.7% | 4.1403 | 11.36 | 4.2015e-11 | 2.5809e-10 |
| *Cupriavidus* | 0.2% | 19.1% | 5.0857 | 16.533 | 8.4502e-11 | 4.3067e-10 |
| *Propionibacterium* | 0% | 0.7% | 4.7714 | 11.95 | 9.014e-11 | 4.3067e-10 |
| *Cupriavidus___Ralstonia* | 0% | 0.4% | 4.0161 | 11.246 | 1.4676e-10 | 6.3106e-10 |
| *f__Dietziaceae* | 0% | 0.6% | 4.7597 | 11.939 | 4.3184e-10 | 1.6881e-9 |
| *Facklamia* | 2.8% | <0.1% | -5.7291 | 14.915 | 3.3408e-09 | 1.1971e-8 |
| *Pelomonas* | 0% | 0.3% | 3.1767 | 10.501 | 1.0156e-08 | 3.3593e-8 |
| *f__Prevotellaceae* | <0.1% | 0.8% | 4.0276 | 11.826 | 1.1292e-08 | 3.4683e-8 |
| *Escherichia___Shigella* | <0.1% | 2.8% | 5.2941 | 14.616 | 3.3186e-08 | 9.5132e-8 |
| *Cloacibacterium* | <0.1% | 0.7% | 2.6602 | 10.791 | 6.523e-08 | 1.7531e-7 |
| *Arcicella* | <0.1% | 0.6% | 2.8503 | 10.562 | 9.0124e-08 | 2.2796e-7 |
| *Streptococcus* | <0.1% | 4.2% | 4.195 | 13.842 | 3.4816e-07 | 8.3172e-7 |
| *Actinomyces* | 0.7% | 0.2% | -3.8504 | 12.84 | 5.2914e-07 | 1.1975e-6 |
| *Enhydrobacter___Moraxella* | <0.1% | 0.3% | 2.8701 | 10.855 | 3.6549e-06 | 7.858e-6 |
| *f__Comamonadaceae* | 0% | 0.3% | 2.6602 | 10.068 | 9.9246e-06 | 2.0322e-5 |
| *Leptothrix___Limnohabitans* | 0% | 0.1% | 2.1256 | 9.6538 | 0.00010858 | 2.1223e-4 |
| *Lactobacillus* | <0.1% | 2.0% | 3.4247 | 13.244 | 0.0001313 | 2.4548e-4 |
| *Gemella* | 0.6% | 0.3% | -3.0772 | 13.093 | 0.00017057 | 2.9869e-4 |
| *f__Pseudonocardiaceae* | 0% | 0.1% | 2.0268 | 9.5794 | 1.7365e-04 | 2.9869e-4 |
| *Negativicoccus* | <0.1% | 0.1% | -2.1244 | 10.55 | 0.00033631 | 5.562e-4 |
| *f__Intrasporangiaceae* | 0% | 0.2% | 1.6527 | 9.3121 | 4.085e-04 | 6.5057e-4 |
| *Peptoniphilus* | 1.4% | 1.7% | -2.5052 | 14.188 | 0.0010688 | 0.0016414 |
| *Actinobaculum* | 1.4% | 1.8% | -2.6814 | 14.02 | 0.0021262 | 0.0031526 |
| *Tepidimonas* | <0.1% | 0.2% | 1.4379 | 9.5909 | 0.0025981 | 0.003724 |
| *Enterococcus* | 73.5% | 33.9% | -2.053 | 20.246 | 0.0035055 | 0.0048624 |
| *Anaerococcus* | 0.6% | 0.9% | -2.0452 | 12.874 | 0.0040829 | 0.0054865 |
| *f__Streptomycetaceae* | 0% | 0.1% | 1.3765 | 9.1375 | 0.0046443 | 0.0060516 |
| *Acinetobacter* | <0.1 | 0.1 | 1.2657 | 9.8233 | 0.018706 | 0.023657 |
| *Sporobacterium* | <0.1% | 1.1% | 1.4317 | 10.704 | 0.022711 | 0.027706 |
| *Staphylococcus* | 15.8% | 8.3% | -2.0067 | 17.976 | 0.023196 | 0.027706 |

* FDR – False Discovery Rate

However, most of these findings did not hold true when applying the log2FC statistics as implemented through the DESeq2 package (Table 11).

Table 11 – List of taxa significantly differing according to Bacterial Load (top lines) and Infertility status (bottom lines) based on Log2FC, lfcSE scores as implemented through DESeq2 algorithms. Relative abundances are indicated.

| Genus | High | Low | Log2FC | lfcSE | P-values | FDR* |
|---|---|---|---|---|---|---|
| *f__Neisseriaceae;g_* | 0% | 1.7% | 26.803 | 1.5718 | 3.35979e-65 | 1.44469e-63 |
| *Pseudomonas* | 0% | 1.3% | 25.534 | 1.6683 | 7.08059e-53 | 1.5223e-51 |
| *f__Comamonadaceae;g_* | 0% | 0.3% | 24.458 | 2.1057 | 3.4431e-31 | 4.935e-30 |
| *Mobiluncus* | <0.1% | 0.2% | 23.832 | 2.1414 | 9.04269e-29 | 9.7209e-28 |
| *Microbacterium* | 0% | 0.5% | 7.6658 | 0.7612 | 7.4481e-24 | 6.4054e-23 |
| *Ralstonia* | 0% | 0.5% | 7.5815 | 0.85031 | 4.8282e-19 | 3.4602e-18 |
| *Arthrobacter* | 0% | 0.6% | 7.6531 | 0.92496 | 1.2951e-16 | 7.9554e-16 |
| *Lactobacillus* | <0.1% | 2.0% | 24.122 | 2.927 | 1.7036e-16 | 9.1567e-16 |
| *Flavobacterium* | 0% | 0.8% | 7.8853 | 1.0125 | 6.8013e-15 | 3.2495e-14 |
| *Candidatus_Rhodoluna* | 0% | 1.0% | 8.2528 | 1.0959 | 5.0553e-14 | 2.1738e-13 |
| *Cupriavidus* | 0.2% | 19.1% | 5.8159 | 0.86466 | 1.7407e-11 | 6.8044e-11 |
| *Streptococcus* | <0.1% | 4.2% | 8.6733 | 1.3372 | 8.8147e-11 | 3.1586e-10 |
| *Arcicella* | <0.1% | 0.6% | 7.5368 | 1.2174 | 5.9756e-10 | 1.9765e-9 |
| *Cupriavidus___Ralstonia* | 0% | 0.4% | 6.9577 | 1.1329 | 8.1832e-10 | 2.5134e-9 |
| *Escherichia___Shigella* | <0.1% | 2.8% | 6.9991 | 1.2551 | 2.4528e-8 | 7.0314e-8 |
| *Propionibacterium* | 0% | 0.7% | 7.1725 | 1.3057 | 3.9456e-8 | 1.0604e-7 |
| *Pelomonas* | 0% | 0.3% | 6.7731 | 1.2549 | 6.7589e-8 | 1.7096e-7 |
| *Finegoldia* | 0.6% | 2.4% | 5.3946 | 1.0285 | 1.56e-7 | 3.7267e-7 |
| *f__Prevotellaceae;g_* | <0.1% | 0.8% | 7.6505 | 1.6346 | 2.8638e-6 | 6.4812e-6 |
| *Cloacibacterium* | <0.1% | 0.7% | 4.7242 | 1.1136 | 2.211e-5 | 4.7537e-5 |
| *f__Intrasporangiaceae;g_* | 0% | 0.2% | 6.2161 | 1.6066 | 1.092e-4 | 2.236e-4 |
| *f__Dietziaceae;g_* | 0% | 0.6% | 6.2715 | 1.7148 | 2.5482e-4 | 4.9806e-4 |
| *Corynebacterium* | 0.5% | 3.4% | 3.7893 | 1.1454 | 9.385e-4 | 0.0017546 |
| *Porphyromonas* | <0.1% | 0.2% | 6.601 | 2.2545 | 0.0034128 | 0.0061145 |
| *Tepidimonas* | <0.1% | 0.2% | 4.3903 | 1.5368 | 0.0042785 | 0.0073591 |
| *f__Streptomycetaceae;g_* | 0% | 0.1% | 5.2102 | 1.9681 | 0.0081127 | 0.012689 |
| *f__Pseudonocardiaceae;g_* | 0% | 0.1% | 5.5036 | 2.0834 | 0.0082489 | 0.012689 |
| *Enhydrobacter___Moraxella* | <0.1% | 0.3% | 4.2417 | 1.606 | 0.0082627 | 0.012689 |
| *Leptothrix___Limnohabitans* | 0% | 0.1% | 5.3215 | 2.1152 | 0.011876 | 0.017609 |
| | | | | | | |
| **Genus** | **Case** | **Control** | **Log2FC** | **lfcSE** | **P-values** | **FDR score** |
| *Actinobaculum* | 2% | 0.1% | -23.174 | 3.5015 | 3.6289e-11 | 1.5604e-9 |

* FDR – False Discovery Rate

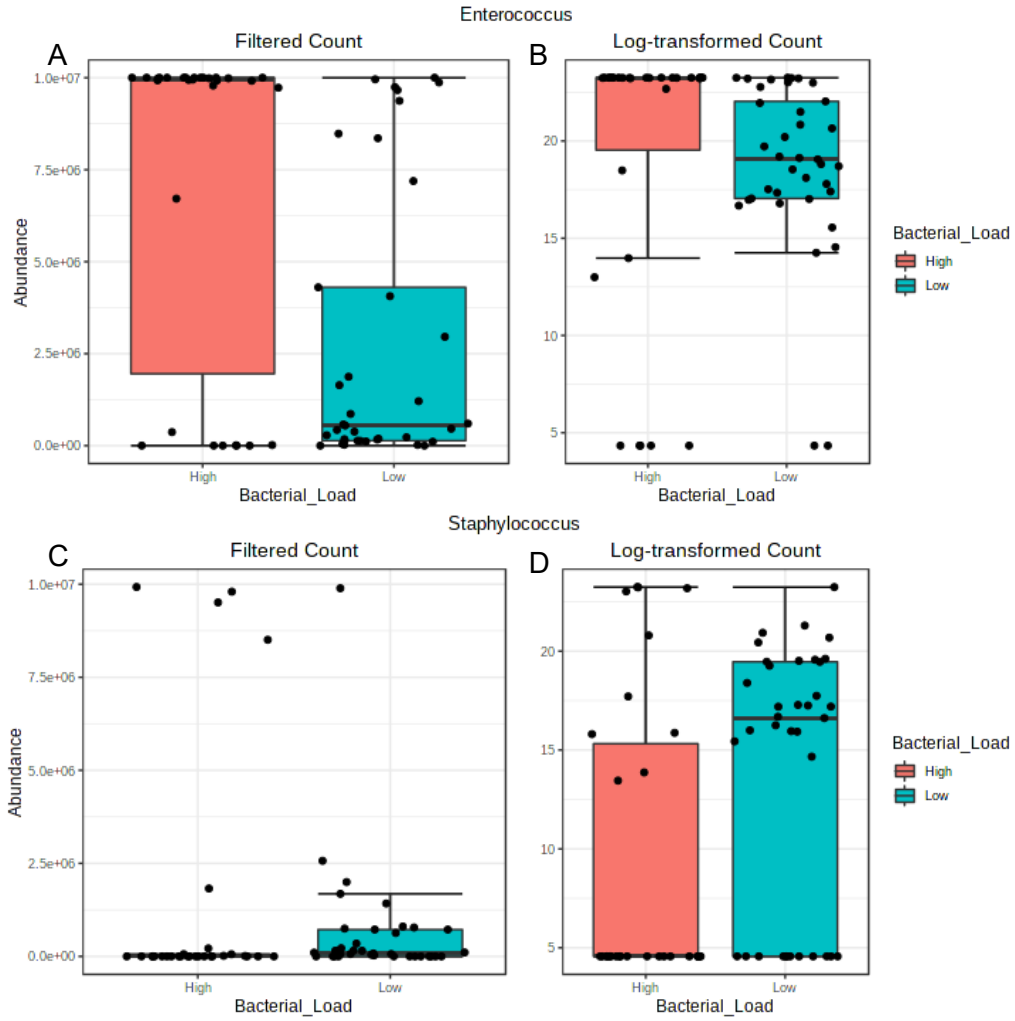



Figure 14 – Scatter plots of *Enterococcus and Staphylococcus* abundances according with high and low bacterial load samples (A and C) Filtered based counts (B and D) log-transformed data as calculated through edgeR algorithm.

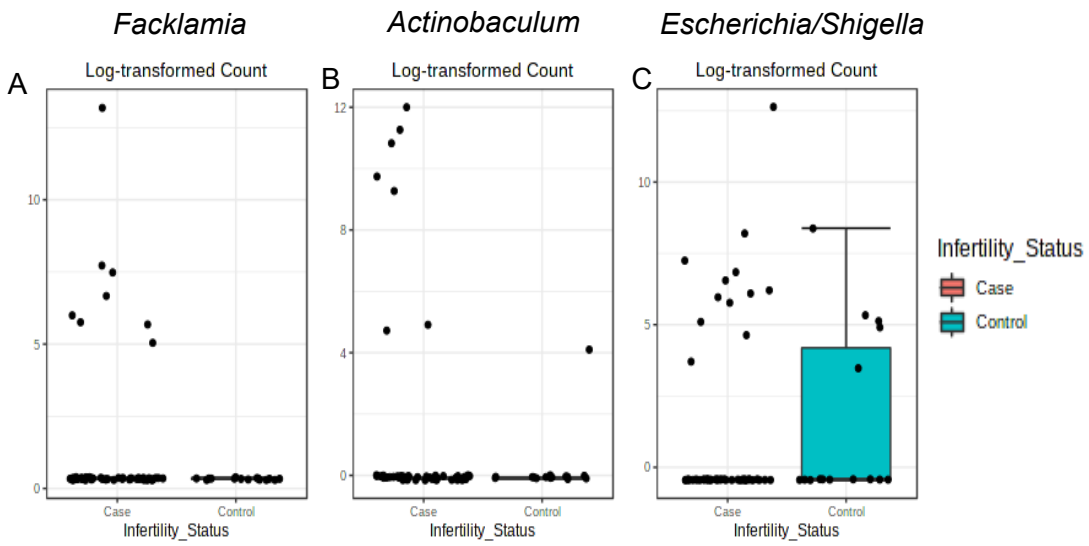



Figure 15 – Scatter plots of *Facklamia* (A), *Actinobaculum* (B) and *Escherichia/Shigella* (C) log-transformed abundances in cases and control samples as calculated through edgeR algorithm.

In our study, we found significant and robust evidence for a correlation of *Enterococcus* and *Staphylococcus* genera with an increased bacterial content (*high bacterial load)*. While *Staphylococcus* has been previously reported as a predominant genus in healthy seminal samples by multiple studies, the same does not apply to *Enterococcus* which is more frequently associated with cases of bacteriospermia confirmed by semen cultures [34, 65, 69, 70]. To the best of our knowledge, only Baud *et al.* (2019) has analyzed the bacterial load of seminal samples from infertility cases and controls, where higher bacterial load were connected with *Prevotella*-enriched communities instead *of Enterococcus* and *Staphylococcus* [70]. Notably, in their study *Staphylococcus* correlated with normal seminal parameters. Conversely, our results indicate *Prevotella* as a genus more prevalent in *low bacterial load* samples. As for the taxa previously correlated with a higher abundance in cases (*Facklamia*, *Actinobaculum* and *Escherichia/Shigella*), only *Escherichia/Shigella* was associated also with *low bacterial load* samples, while the other 2 showed no significant differences between the high and low bacterial load groups.

These analyses support on one hand the association of *Enterococcus* and *Staphylococcus* with a high bacterial load, or in other words with an increased content of bacteria in male reproductive tract of these men. Although none of the analyzed seminal samples showed signs of bacteriospermia, it is interesting to note that subjects with a *high bacterial load* present also top abundances of two genera known to cause MGI such as prostatitis, urethritis and epididymitis [67, 101-103]. In this respect, *Enterococcus* has been also correlated with abnormal sperm quality parameters (motility, morphology and concentration), changes of the chromatin integrity (global DNA damage, double-stranded DNA breaks and DNA protamination status) and increase oxidative stress levels [95].

Given that *Enterococcus* and *Staphylococcus* are reported to be part of a healthy microbiome [34, 65, 69, 70, 99] and also found at reduced abundances in low bacterial load samples, it is attractive to conjecture about a dysbiotic microbiome where *Enterococcus* or *Staphylococcus* growth in numbers overriding the prevalence of other taxa. Oddly, those microbes were found in similar proportions (55.6-58.2% and 12.0-12.7%) in both cases and controls, which are all normozoospermic followed at an infertility clinic. The opposite, the identification of several cases with a reduced bacterial content connected with lower abundances of *Enterococcus* or *Staphylococcus* and high bacterial diversity, is not unexpected because male infertility is a multifactorial disorder where many other factors are likely to contribute to loss of semen quality. In this regard, we cannot exclude the role of the host immunity in controlling infection and any negative effect of bacterial proliferation (*Enterococcus* or *Staphylococcus).* To explore better any

of these hypotheses, it would be desirable to perform some laboratorial experiences to assess the impact of *Enterococcus* or *Staphylococcus* in semen quality.

The *Facklamia*, *Actinobaculum* and *Escherichia/Shigella* genera found to be slight augmented in cases, in general were already described by our team (Monteiro *et al.,* 2018) and others as associated with male infertility [34]. From these, *Facklamia* which can cause different types of infection in human body sites was the only one described as increased in SHV (2.6%) and OAT (1.0%) when compared with controls (0.1%) [34]. *Actinobaculum* and *Escherichia/Shigella*, as well as their higher taxonomic level, *Enterobacteriaceae*, did not show any significant changes in this same work [34].

## 4.4. Male infertility phenotypes and microbiome

Monteiro *et al.* (2018) previously found evidence for a differential microbial composition of SHV in comparison to controls and to the male infertility phenotypes, asthenozoospermia (AT) and OAT. Precisely, the SHV cases were associated with an increase in Proteobacteria together with a decrease of Firmicutes [34]. Additionally, Monteiro *et al.* (2018) also reported some shifts in the abundance of several known pathogenic and probiotic genera in both SHV and OAT when compared to controls. Whereas the *Pseudomonas, Klebsiella*, *Aerococcus*, and *Neisseria*, as well as undefined genus of *Enterobacteriaceae* (possibly *Escherichia*) were found to be augmented in cases, *Lactobacillus* and *Propionibacterium* genera were reduced [34]. Taking into account that the present study was designed as a follow-up of the former findings of Monteiro *et al.* (2018), in which microbial associations to SHV and OAT were explored, a detailed assessment of these phenotypes is performed in this section.

The overall abundance of Firmicutes in cases was estimated as 77.7%. However, if stratifying per phenotype some variation could be detected ranging from 70.8% in OAT+SHV to 87.5% in SHV but without reaching significance by standard statistical tests (Table 8 in section 4.3.1.). The same applied to the proportions of Proteobacteria, Actinobacteria and Bacteriodetes, suggesting at this point a lack of correlation between the composition of the semen microbiome and male infertility phenotypes (Table 8). Although these results may seem to contradict the former work by Monteiro *et al.* (2018), it is interesting to note that OAT+SHV show a similar trend of reduction in Firmicutes phylum this time associated with an increment in Actinobacteria. Indeed, the OAT+SHV group is probably more alike to the SHV of Monteiro *et al.* (2018) than our current SHV group because those combined cases with or without other abnormal parameters

together with seminal hyperviscosity [34]. Moreover, considering our current findings of a large interindividual variability it is not completely unexpected to obtain contrasting results with different samples, especially when the sample size is relatively small (<25 per phenotype). In this regard, other authors such as Chen *et al.* (2018) also suggested a higher prevalence of Proteobacteria but in normozoospermic samples [66].

Concerning the genus rank, some variation in the abundance of the most prevalent taxa was observed when dividing cases according to OAT, SHV or OAT+SHV (Figure 16). For example, *Enterococcus* (48-65%), *Staphylococcus* (2.4-13.5%), *Cupriavidus* (4.1-10.4%) and *Streptococcus* (0.5-3.2%; Table 8). Yet, no significant differences were detected again by standard statistical tests.

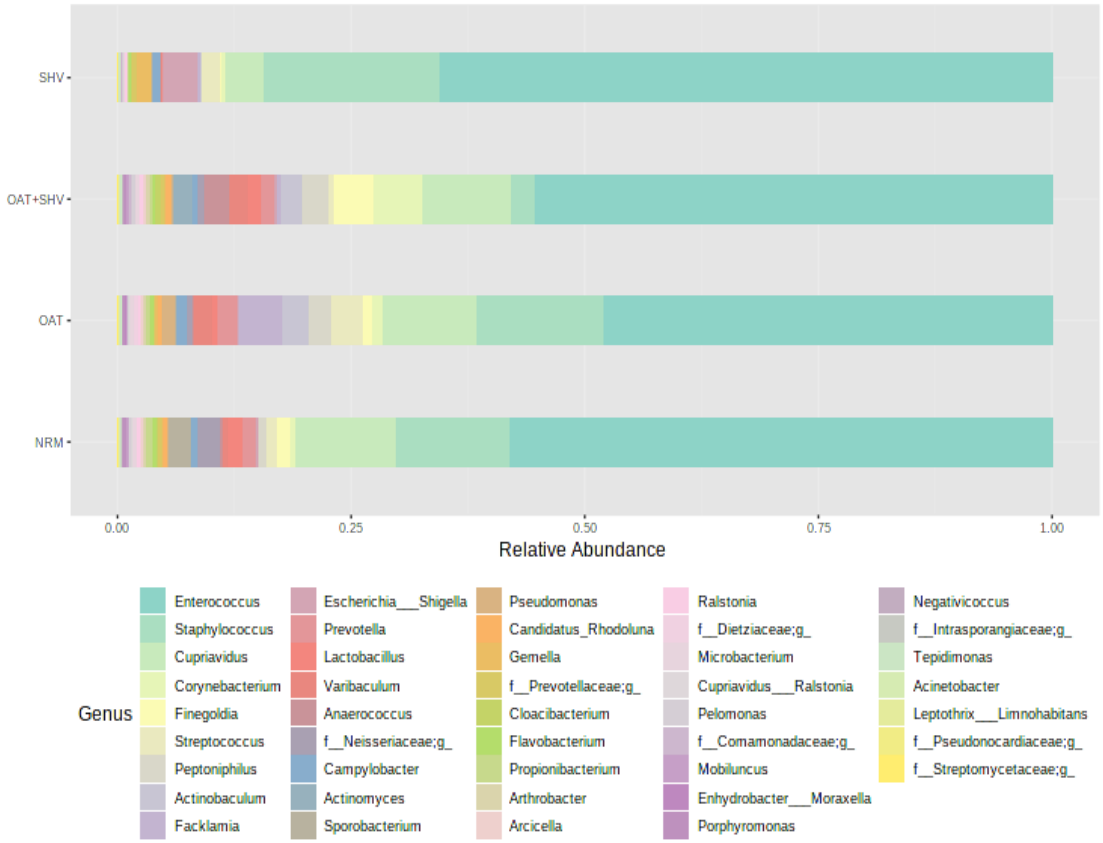



Figure 16 – Microbiome profiles of the most abundant genera (>0.1%) per individual sample divided according to their phenotype groups.

The analysis of alpha and beta diversity did not uncover any significant results as previously mentioned in section 4.3.3. Nonetheless, a closer inspection of alpha diversity plots shows a trend for a lower diversity of SHV when compared with other phenotypes and controls (Figure 17) Therefore, to test the divergence of SHV from the other groups

several pairwise tests were performed, alpha and beta diversity for all the previously viewed indices but still no significant p-value was found (not shown).

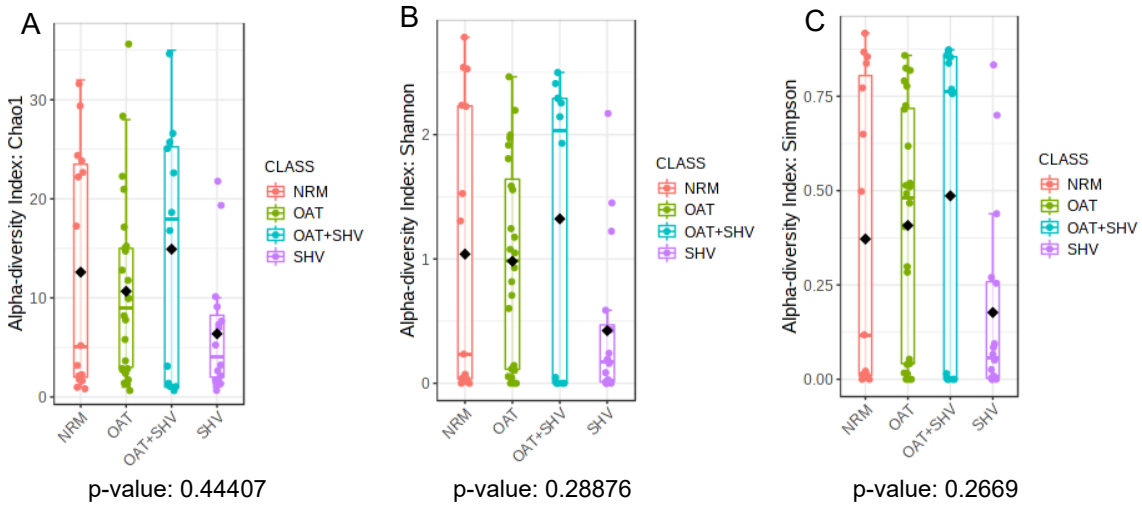



Figure 17 – Alpha diversity according to phenotypes: (A) Chao1 index; (B) Shannon index; (C) Simpson index.

Nevertheless, some statistically significant results were detected between groups when performing discriminant analyses through log2FC and logCPM as implemented in edgeR and log2FC in DESeq2 (Table 12). However, no taxa were found to differ between groups when applying LDA scores (LEfSe algorithm). Firstly, on a phylum level, Actinobacteria was suggested as significant through the edgeR package, presenting a much higher abundance in OAT and OAT+SHV groups than in control or SHV groups (Table 8). Then, a genus approach revealed 12 significant genera. Among those significant taxa were *Facklamia*, which was augmented in the OAT group and *Actinobaculum* which was increased in OAT and OAT+SHV, respectively (Table 12). The overlap of these results with the ones obtained in the case versus controls comparisons indicate that the former statically significance of *Facklamia* and *Actinobaculum* was derived mostly from OAT and/or OAT+SHV, in contrary to the *Escherichia_Shigella* which is probably transversal to the different phenotypes.

Table 12 – List of taxa significantly differing according to phenotype based in Log2FC and LogCPM scores as implemented through edgeR or DESeq2 algorithms. Relative abundances are indicated.

| Genus | Frequency Abundance (%) | | | | EdgeR | | | | DESeq2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NRM | OAT | OAT+SHV | SHV | Log2FC | logCPM | P-value | FDR score | Log2FC | lfcSE | P-value | FDR score |
| *Facklamia* | 0 | 4.6 | 0.4 | <0.1 | 9.037 | 14.915 | 7.969E-8 | 3.4267E-6 | NS* | NS | NS | NS |
| *Pseudomonas* | 0.2 | 1.5 | 0.1 | 0 | 5.731 | 13.482 | 6.3692E-7 | 1.3694E-5 | -21.638 | 3.2791 | 4.1451E-11 | 3.5648E-10 |
| *Actinobaculum* | <0.1 | 3.1 | 2.3 | 0.4 | 6.8054 | 14.02 | 1.4474E-5 | 2.0747E-4 | NS | NS | NS | NS |
| *Tepidimonas* | <0.1 | 0.2 | 0.1 | <0.1 | 2.4488 | 9.591 | 4.8826E-4 | 0.0043805 | NS | NS | NS | NS |
| *Peptoniphilus* | 0.9 | 2.4 | 2.9 | <0.1 | 3.8812 | 14.188 | 5.0936E-4 | 0.0043805 | -7.0518 | 2.1685 | 0.0011463 | 0.0070419 |
| *Sporobacterium* | 2.2 | <0.1 | 0.1 | 0 | -2.3217 | 10.705 | 0.0018621 | 0.013345 | -27.435 | 3.5542 | 1.1718E-14 | 2.5194E-13 |
| *Varibaculum* | 0.4 | 1.8 | 1.8 | 0 | 3.0394 | 12.824 | 0.0023633 | 0.014517 | -23.315 | 3.0676 | 2.9543E-14 | 4.2345E-13 |
| *Campylobacter* | 0.7 | 1.0 | 0.6 | 0.8 | 3.6949 | 13.615 | 0.0027261 | 0.014653 | -8.1984 | 2.3485 | 4.8146E-4 | 0.0034505 |
| *Leptothrix/Limnohabitans* | 0.1 | <0.1 | <0.1 | 0.1 | 2.1371 | 9.6537 | 0.00566 | 0.026536 | NS | NS | NS | NS |
| *Prevotella* | 1.3 | 2.0 | 1.2 | <0.1 | 3.2404 | 13.645 | 0.0061711 | 0.026536 | NS | NS | NS | NS |
| *f__Pseudonocardiaceae;g_* | 0.1 | 0.1 | <0.1 | <0.1 | 1.9806 | 9.5795 | 0.0079632 | 0.031129 | NS | NS | NS | NS |
| *f__Dietziaceae;g_* | 0.2 | 0.6 | <0.1 | 0.1 | 2.646 | 11.939 | 0.0089384 | 0.032029 | NS | NS | NS | NS |
| *f__Neisseriaceae;g_* | 2.3 | 0.4 | 0.5 | <0.1 | NS | NS | NS | NS | -24.913 | 2.8969 | 7.9783E-18 | 3.4307E-16 |
| *f__Prevotellaceae;g_* | 0.5 | 0.1 | 0.4 | 0.6 | NS | NS | NS | NS | -24.184 | 3.5504 | 9.6591E-12 | 1.0384E-10 |

*NS signifies a nonsignificant p-value

Overall, this in-depth phenotype centered approach showed that strict SHV cases tend to show a reduced diversity when compared with any other group, including controls (NRM) and other cases (OAT and OAT+SHV). However, the differences are not enough to reach statistically significance probably due to the limited sample size of SHV, which comprises only 16 individuals. Although no significant correlation was disclosed for *Enterococcus and Staphylococcus* in the analysis per phenotype, the evaluation of relative abundances estimated in SHV as a group suggests that those taxa might be responsible for the decay in diversity. Additional studies with larger sample sizes would be necessary to corroborate this hypothesis. Moreover, this investigation centered in phenotypes also uncovered differences in several taxa found at reduced abundances including not only *Facklamia and Actinobaculum* but also other genera previously correlated with abnormal semen parameters. Those include *Prevotella* and *Pseudomonas,* which were previously associated with poor semen quality when samples were enriched in any of these taxa and depleted in *Lactobacillus* [65]. Still according to our results only *Pseudomonas* seems to be correlated with a loss of semen quality as indicated by its increased prevalence in OAT. On the other hand, *Prevotella* appers to diverge by their near absence in SHV and concerning *Lactobacillus* no statistically significant differences were detected between groups.

Genera like *Haemophilus, Sneathia*, *Lysobacter* and *Solibacillus* which were connected previously with male infertility cases and oligozoospermia, asthenozoospermia, or azoospermia [34, 65, 69, 70], were not identified in our study or did not pass our filtering criteria (>0.1% and 10% prevalence) Only *Anaerococcus* that was connected before with OAT phenotype [5] was detected in our study at low prevalence, not showing any significant result in comparisons per phenotype .

## 4.5. Implication into Forensic sciences

Seminal microbiome studies may be of particular interest for forensic sciences in the context of individual identification, geolocation and post-mortem interval establishment. The understanding that the seminal microbiome shows a large interindividual variability, in which a fraction of subjects exhibits diverse and unique microbiome profiles represents an advantage or a promising feature. Briefly, the application of this methodology can identify taxa present in the microbiome with a high level of sensitivity, and therefore help provide a more certain suspect identification. Furthermore, it may also allow for the comparison of samples taken in different time frames, showing how these

microorganisms' abundances and diversity fluctuates, possibly providing an aid to infer suspect's activities and whereabouts over a larger period of time. These usages of the seminal microbiome may assist the investigation of crime scenes, particularly those related with sexual assaults.

# 5. Conclusions

Following a previous evaluation of the seminal microbiome composition by Monteiro *et al.* (2018), in this study we performed an in-depth analysis of the bacterial communities in selected infertility cases and controls. This work employed qPCR assays for 16S gene to determine the total bacterial content of analysed samples, the high-throughput sequencing of 7 hypervariable regions of *16S* to generate OTUs for bacterial identification and the statistical analysis of collected data to pinpoint significant microbial changes with potential links to the loss of semen quality. Globally, this study shows:

**I.** There is a large variability across individuals in the number of bacteria present in the genitourinary system as it was inferred by the bacterial load of seminal samples. These estimates of *high* or *low bacterial load* have no correlation with the infertility status (cases or controls) nor with the tested phenotypes of oligoasthenoteratozoospermia (OAT) and/or seminal hyperviscosity (SHV).

**II.** The bacterial communities found in the semen are, independently of the reproductive health status, dominated by Firmicutes which contributes with more than 75% to the seminal microbiome, far followed by Proteobacteria with less than 15%, Actinobacteria and Bacteroidetes. These proportions are attributed to a high prevalence of *Enterococcus* that exceeds 50% and by other less abundant genera like *Staphylococcus*, *Streptococcus*, *Facklamia*, *Corynebacterium*, *Actinobaculum*, *Peptoniphilus*, *Escherichia/Shigella* and *Finegoldia*.

**III.** Microbial profiles are mainly stratified according with samples bacterial content, in which *high bacterial load* ones are more homogenous (less diverse) and characterized by an enrichment in *Enterococcus* or *Staphylococcus* genera. In contrary, the *low bacterial load* samples are more heterogenous and diverse and

not defined by a single genus but rather by the presence of several low abundance taxa.

**IV.** The observation of *Enterococcus* and *Staphylococcus,* which are recognized agents of male genitourinary infections (e.g. prostatitis and epidydimitis), as the dominating genera in samples showing a *high bacterial content* indicates a possible condition of dysbiosis due to the overgrowth of these taxa. In contrary, the detection of diverse microbial communities at a lower bacterial baseline fits better the previous concepts of a healthy microbiome.

**V.** The lack of correlation between *Enterococcus* and *Staphylococcus* enriched samples and male infertility suggests that other host factors may protect affected subjects from the negative effects of these taxa in semen quality.

**VI.** *Facklamia* and *Actinobaculum* are low abundant genera found to be augmented in infertility cases in general, and more specifically in oligoasthenoteratozoospermia.

# 6. References

1. Zegers-Hochschild, F., et al., *The International Committee for Monitoring Assisted Reproductive Technology (ICMART) and the World Health Organization (WHO) Revised Glossary on ART Terminology, 2009.* Hum Reprod, 2009. **24**(11): p. 2683-7.
2. WHO, *WHO laboratory manual for the examination and processing of human semen.* . Vol. 5th edition. 2010.
3. Agarwal, A., et al., *Male infertility.* Lancet, 2021. **397**(10271): p. 319-333.
4. Ombelet, W., et al., *Infertility and the provision of infertility medical services in developing countries.* Hum Reprod Update, 2008. **14**(6): p. 605-21.
5. Farahani, L., et al., *The semen microbiome and its impact on sperm function and male fertility: A systematic review and meta-analysis.* Andrology, 2021. **9**(1): p. 115-144.
6. Carvalho J, S.A., *Estudo AFRODITE: Caracterização da Infertilidade em Portugal. Estudo na comunidade. .* 2009.
7. Levine, H., et al., *Temporal trends in sperm count: a systematic review and meta-regression analysis.* Hum Reprod Update, 2017. **23**(6): p. 646-659.
8. Lundy, S.D., et al., *Functional and Taxonomic Dysbiosis of the Gut, Urine, and Semen Microbiomes in Male Infertility.* Eur Urol, 2021. **79**(6): p. 826-836.
9. Schjenken, J.E. and S.A. Robertson, *The Female Response to Seminal Fluid.* Physiol Rev, 2020. **100**(3): p. 1077-1117.
10. Wang, J. and M.V. Sauer, *In vitro fertilization (IVF): a review of 3 decades of clinical innovation and technological advancement.* Ther Clin Risk Manag, 2006. **2**(4): p. 355-64.
11. Hasanpoor-Azghdy, S.B., M. Simbar, and A. Vedadhir, *The emotional-psychological consequences of infertility among infertile women seeking treatment: Results of a qualitative study.* Iran J Reprod Med, 2014. **12**(2): p. 131-8.
12. Petok, W.D., *Infertility counseling (or the lack thereof) of the forgotten male partner.* Fertil Steril, 2015. **104**(2): p. 260-6.
13. Mesiano S., J.E.E., *The Male Reproductive System*, in *Medical Physiology*. 2017, Elsevier. p. 1092-1107.
14. Plant T., Z.A., *Knobil and Neill's Physiology of Reproduction*. 4 ed. 2014: Academic Press.

15. Anamthathmakula, P. and W. Winuthayanon, *Mechanism of semen liquefaction and its potential for a novel non-hormonal contraceptiondagger.* Biol Reprod, 2020. **103**(2): p. 411-426.
16. Griswold, M.D., *The central role of Sertoli cells in spermatogenesis.* Semin Cell Dev Biol, 1998. **9**(4): p. 411-6.
17. Stukenborg, J.B., et al., *Male germ cell development in humans.* Horm Res Paediatr, 2014. **81**(1): p. 2-12.
18. Rodriguez-Martinez, H., et al., *Seminal Plasma: Relevant for Fertility?* Int J Mol Sci, 2021. **22**(9).
19. Teves, M.E., et al., *Sperm Differentiation: The Role of Trafficking of Proteins.* Int J Mol Sci, 2020. **21**(10).
20. Levine, J., Fernbach, & Stahl, *Evaluation and preservation of fertility in adolescent and young adult patients with testicular cancer.* Clinical Oncology in Adolescents and Young Adults, 2013(29).
21. Alves, M.B.R., E.C.C. Celeghini, and C. Belleannee, *From Sperm Motility to Sperm-Borne microRNA Signatures: New Approaches to Predict Male Fertility Potential.* Front Cell Dev Biol, 2020. **8**: p. 791.
22. James, E.R., et al., *The Role of the Epididymis and the Contribution of Epididymosomes to Mammalian Reproduction.* Int J Mol Sci, 2020. **21**(15).
23. Du Plessis, S.S., S. Gokul, and A. Agarwal, *Semen hyperviscosity: causes, consequences, and cures.* Front Biosci (Elite Ed), 2013. **5**(1): p. 224-31.
24. Fallah, A., A. Mohammad-Hasani, and A.H. Colagar, *Zinc is an Essential Element for Male Fertility: A Review of Zn Roles in Men's Health, Germination, Sperm Quality, and Fertilization.* J Reprod Infertil, 2018. **19**(2): p. 69-81.
25. Toragall, M.M., et al., *Evaluation of Seminal Fructose and Citric Acid Levels in Men with Fertility Problem.* J Hum Reprod Sci, 2019. **12**(3): p. 199-203.
26. Szczykutowicz, J., et al., *The Potential Role of Seminal Plasma in the Fertilization Outcomes.* Biomed Res Int, 2019. **2019**: p. 5397804.
27. WHO, *WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction*. 1999.
28. WHO, *WHO laboratory manual for the examination and processing of human semen Sixth Edition*. 6th ed. 2021, Geneva. 276.
29. Brookings, C., D. Goldmeier, and H. Sadeghi-Nejad, *Sexually transmitted infections and sexual function in relation to male fertility.* Korean J Urol, 2013. **54**(3): p. 149-56.

30. Stevenson, E.L., P.E. Hershberger, and P.A. Bergh, *Evidence-Based Care for Couples With Infertility.* J Obstet Gynecol Neonatal Nurs, 2016. **45**(1): p. 100-10; quiz e1-2.
31. Cooper, T.G., et al., *World Health Organization reference values for human semen characteristics.* Hum Reprod Update, 2010. **16**(3): p. 231-45.
32. Esteves, S.C., et al., *Critical appraisal of World Health Organization's new reference values for human semen characteristics and effect on diagnosis and treatment of subfertile men.* Urology, 2012. **79**(1): p. 16-22.
33. Boitrelle, F., et al., *The Sixth Edition of the WHO Manual for Human Semen Analysis: A Critical Review and SWOT Analysis.* Life (Basel), 2021. **11**(12).
34. 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, 2018. **79**(6): p. e12838.
35. Pilmis, B., et al., *Enterococcus faecalis-related prostatitis successfully treated with moxifloxacin.* Antimicrob Agents Chemother, 2015. **59**(11): p. 7156-7.
36. Organization, W.H. *Sexually transmitted infections (STIs)*. 07/09/2022]; Available from: https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis).
37. Goulart, A.C.X., et al., *HIV, HPV and Chlamydia trachomatis: impacts on male fertility.* JBRA Assist Reprod, 2020. **24**(4): p. 492-497.
38. Punjani, N., et al., *Genitourinary Infections Related to Circumcision and the Potential Impact on Male Infertility.* World J Mens Health, 2022. **40**(2): p. 179-190.
39. Gdoura, R., et al., *Assessment of Chlamydia trachomatis, Ureaplasma urealyticum, Ureaplasma parvum, Mycoplasma hominis, and Mycoplasma genitalium in semen and first void urine specimens of asymptomatic male partners of infertile couples.* J Androl, 2008. **29**(2): p. 198-206.
40. Moubasher, A., et al., *Impact of leukocytospermia on sperm dynamic motility parameters, DNA and chromosomal integrity.* Cent European J Urol, 2018. **71**(4): p. 470-475.
41. Domes, T., et al., *The incidence and effect of bacteriospermia and elevated seminal leukocytes on semen parameters.* Fertil Steril, 2012. **97**(5): p. 1050-5.
42. Vilvanathan, S., et al., *Bacteriospermia and Its Impact on Basic Semen Parameters among Infertile Men.* Interdiscip Perspect Infect Dis, 2016. **2016**: p. 2614692.

43. Shahroodian, S., et al., *Association between virulence factors and biofilm formation in Enterococcus faecalis isolated from semen of infertile men.* Am J Reprod Immunol, 2022. **88**(1): p. e13561.
44. Dutta, S., et al., *Staphylococcal infections and infertility: mechanisms and management.* Mol Cell Biochem, 2020. **474**(1-2): p. 57-72.
45. Ghasemian, F., S. Esmaeilnezhad, and M.J. Mehdipour Moghaddam, *Staphylococcus saprophyticus and Escherichia coli: Tracking from sperm fertility potential to assisted reproductive outcomes.* Clin Exp Reprod Med, 2021. **48**(2): p. 142-149.
46. Koedooder, R., et al., *Identification and evaluation of the microbiome in the female and male reproductive tracts.* Hum Reprod Update, 2019. **25**(3): p. 298-325.
47. Grice, E.A. and J.A. Segre, *The skin microbiome.* Nat Rev Microbiol, 2011. **9**(4): p. 244-53.
48. Bursle, E. and J. Robson, *Non-culture methods for detecting infection.* Aust Prescr, 2016. **39**(5): p. 171-175.
49. Sontakke, S., et al., *Use of broad range16S rDNA PCR in clinical microbiology.* J Microbiol Methods, 2009. **76**(3): p. 217-25.
50. Klindworth, A., et al., *Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies.* Nucleic Acids Res, 2013. **41**(1): p. e1.
51. Fukuda, K., et al., *Molecular Approaches to Studying Microbial Communities: Targeting the 16S Ribosomal RNA Gene.* J UOEH, 2016. **38**(3): p. 223-32.
52. Verhelst, R., et al., *Cloning of 16S rRNA genes amplified from normal and disturbed vaginal microflora suggests a strong association between Atopobium vaginae, Gardnerella vaginalis and bacterial vaginosis.* BMC Microbiol, 2004. **4**: p. 16.
53. Smith, B.C., et al., *The cervical microbiome over 7 years and a comparison of methodologies for its characterization.* PLoS One, 2012. **7**(7): p. e40425.
54. Yatera, K., S. Noguchi, and H. Mukae, *The microbiome in the lower respiratory tract.* Respir Investig, 2018. **56**(6): p. 432-439.
55. Sender, R., S. Fuchs, and R. Milo, *Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans.* Cell, 2016. **164**(3): p. 337-40.
56. Al-Rashidi, H.E., *Gut microbiota and immunity relevance in eubiosis and dysbiosis.* Saudi J Biol Sci, 2022. **29**(3): p. 1628-1643.

57. Valdes, A.M., et al., *Role of the gut microbiota in nutrition and health.* BMJ, 2018. **361**: p. k2179.
58. Bilen, M., *Strategies and advancements in human microbiome description and the importance of culturomics.* Microb Pathog, 2020. **149**: p. 104460.
59. Miyake, M., et al., *Prostate diseases and microbiome in the prostate, gut, and urine.* Prostate Int, 2022. **10**(2): p. 96-107.
60. DiBaise, J.K., et al., *Gut microbiota and its possible relationship with obesity.* Mayo Clin Proc, 2008. **83**(4): p. 460-9.
61. Gao, Z., et al., *Substantial alterations of the cutaneous bacterial biota in psoriatic lesions.* PLoS One, 2008. **3**(7): p. e2719.
62. Kostic, A.D., R.J. Xavier, and D. Gevers, *The microbiome in inflammatory bowel disease: current status and the future ahead.* Gastroenterology, 2014. **146**(6): p. 1489-99.
63. Manichanh, C., et al., *Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.* Gut, 2006. **55**(2): p. 205-11.
64. Nguyen, P.V., et al., *Innate and adaptive immune responses in male and female reproductive tracts in homeostasis and following HIV infection.* Cell Mol Immunol, 2014. **11**(5): p. 410-27.
65. Weng, S.L., et al., *Bacterial communities in semen from men of infertile couples: metagenomic sequencing reveals relationships of seminal microbiota to semen quality.* PLoS One, 2014. **9**(10): p. e110152.
66. Chen, H., et al., *Seminal bacterial composition in patients with obstructive and non-obstructive azoospermia.* Exp Ther Med, 2018. **15**(3): p. 2884-2890.
67. Mandar, R., et al., *Seminal microbiome in men with and without prostatitis.* Int J Urol, 2017. **24**(3): p. 211-216.
68. Liu, C.M., et al., *The semen microbiome and its relationship with local immunology and viral load in HIV infection.* PLoS Pathog, 2014. **10**(7): p. e1004262.
69. Hou, D., et al., *Microbiota of the seminal fluid from healthy and infertile men.* Fertil Steril, 2013. **100**(5): p. 1261-9.
70. Baud, D., et al., *Sperm Microbiota and Its Impact on Semen Parameters.* Front Microbiol, 2019. **10**: p. 234.
71. Brandao, P., M. Goncalves-Henriques, and N. Ceschin, *Seminal and testicular microbiome and male fertility: A systematic review.* Porto Biomed J, 2021. **6**(6): p. e151.
72. Yang, H., et al., *Potential Pathogenic Bacteria in Seminal Microbiota of Patients with Different Types of Dysspermatism.* Sci Rep, 2020. **10**(1): p. 6876.

73. Venneri, M.A., et al., *Human genital tracts microbiota: dysbiosis crucial for infertility.* J Endocrinol Invest, 2022. **45**(6): p. 1151-1160.
74. Haarkotter, C., et al., *Usefulness of Microbiome for Forensic Geolocation: A Review.* Life (Basel), 2021. **11**(12).
75. Metcalf, J.L., et al., *Microbiome Tools for Forensic Science.* Trends Biotechnol, 2017. **35**(9): p. 814-823.
76. Diez Lopez, C., A. Vidaki, and M. Kayser, *Integrating the human microbiome in the forensic toolkit: Current bottlenecks and future solutions.* Forensic Sci Int Genet, 2022. **56**: p. 102627.
77. Garcia, M.G., et al., *Impact of the Human Microbiome in Forensic Sciences: a Systematic Review.* Appl Environ Microbiol, 2020. **86**(22).
78. Williams, D.W. and G. Gibson, *Classification of individuals and the potential to detect sexual contact using the microbiome of the pubic region.* Forensic Sci Int Genet, 2019. **41**: p. 177-187.
79. Yao, T., et al., *Effect of indoor environmental exposure on seminal microbiota and its application in body fluid identification.* Forensic Sci Int, 2020. **314**: p. 110417.
80. Dobay, A., et al., *Microbiome-based body fluid identification of samples exposed to indoor conditions.* Forensic Sci Int Genet, 2019. **40**: p. 105-113.
81. Salzmann, A.P., et al., *Transcription and microbial profiling of body fluids using a massively parallel sequencing approach.* Forensic Sci Int Genet, 2019. **43**: p. 102149.
82. Sulaiman, I., et al., *Microbial signatures in the lower airways of mechanically ventilated COVID-19 patients associated with poor clinical outcome.* Nat Microbiol, 2021. **6**(10): p. 1245-1258.
83. Whelan, J.A., N.B. Russell, and M.A. Whelan, *A method for the absolute quantification of cDNA using real-time PCR.* J Immunol Methods, 2003. **278**(1-2): p. 261-9.
84. Chong, J., et al., *Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data.* Nat Protoc, 2020. **15**(3): p. 799-821.
85. Dhariwal, A., et al., *MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data.* Nucleic Acids Res, 2017. **45**(W1): p. W180-W188.
86. Kuraku, S., et al., *aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W22-8.

87. Katoh, K., J. Rozewicki, and K.D. Yamada, *MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization.* Brief Bioinform, 2019. **20**(4): p. 1160-1166.
88. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
89. Fischer, R.A., *THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS.* Annals of Eugenics, 1936. **7(2)**: p. 179–188. .
90. Wang, H., et al., *The Microbiome, an Important Factor That Is Easily Overlooked in Male Infertility.* Front Microbiol, 2022. **13**: p. 831272.
91. Alqawasmeh, O., et al., *The microbiome and male infertility: looking into the past to move forward.* Hum Fertil (Camb), 2022: p. 1-13.
92. Barb, J.J., et al., *Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples.* PLoS One, 2016. **11**(2): p. e0148047.
93. Yang, B., Y. Wang, and P.Y. Qian, *Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis.* BMC Bioinformatics, 2016. **17**: p. 135.
94. Altmae, S., J.M. Franasiak, and R. Mandar, *The seminal microbiome in health and disease.* Nat Rev Urol, 2019. **16**(12): p. 703-721.
95. Perez-Losada, M., K.A. Crandall, and R.J. Freishtat, *Two sampling methods yield distinct microbial signatures in the nasopharynges of asthmatic children.* Microbiome, 2016. **4**(1): p. 25.
96. D'Amore, R., et al., *A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling.* BMC Genomics, 2016. **17**: p. 55.
97. Clooney, A.G., et al., *Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis.* PLoS One, 2016. **11**(2): p. e0148028.
98. Hahn, A., et al., *Different next generation sequencing platforms produce different microbial profiles and diversity in cystic fibrosis sputum.* J Microbiol Methods, 2016. **130**: p. 95-99.
99. Yao, Y., et al., *Semen microbiota in normal and leukocytospermic males.* Asian J Androl, 2022. **24**(4): p. 398-405.
100. Yang, I., et al., *The Infant Microbiome: Implications for Infant Health and Neurocognitive Development.* Nurs Res, 2016. **65**(1): p. 76-88.
101. Magri, V., E. Marras, and G. Perletti, *Chronic bacterial prostatitis: enterococcal disease?* Clin Infect Dis, 2011. **53**(12): p. 1306-7; author reply 1307-8.

102. Grande, G., et al., *Semen Proteomics Reveals the Impact of Enterococcus faecalis on male Fertility.* Protein Pept Lett, 2018. **25**(5): p. 472-477.

103. Esmailkhani, A., et al., *Assessing the prevalence of Staphylococcus aureus in infertile male patients in Tabriz, northwest Iran.* Int J Reprod Biomed, 2018. **16**(7): p. 469-474.

# 7. Annexes

Annex I – Testing of the two pairs of primers used in this work.

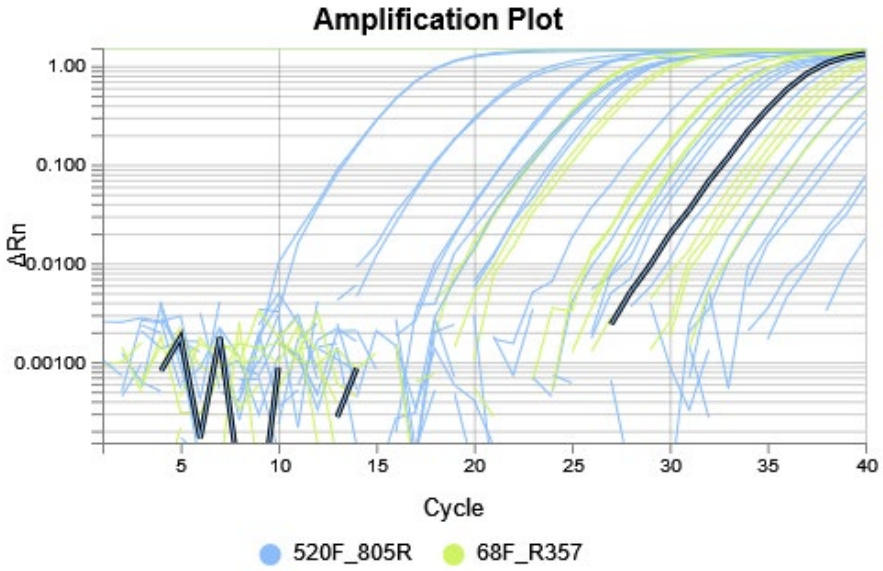



Figure 18 – Amplification plot of a qPCR test run with both primer pairs.

Annex II – Microbial composition of the *negative PCR control* and the *DNA extraction control* for the V4 hypervariable region.

Table 13 – Microbial composition of the most notable genera present in the *negative PCR control* and the *DNA extraction control* for the V4 region.

| **Genus** | **Frequency (absolute number of reads)** | |
|---|---|---|
| | ***Negative PCR control*** | ***DNA extraction control*** |
| *Cupriavidus* | 193 | 8666 |
| *Ralstonia* | 0 | 763 |
| *Sphingomonas* | 0 | 380 |
| *Arthrobacter* | 0 | 349 |
| *Cupriavidus/Ralstonia* | 0 | 340 |
| *Pelomonas* | 0 | 332 |
| *Acinetobacter* | 0 | 281 |

# Annex III – Remaining analyses for the V4 hypervariable region


Figure 19 – Microbiome profiles of the most abundant genera (f>0.1%) per individual sample divided according to their infertility status grouping (case and control).

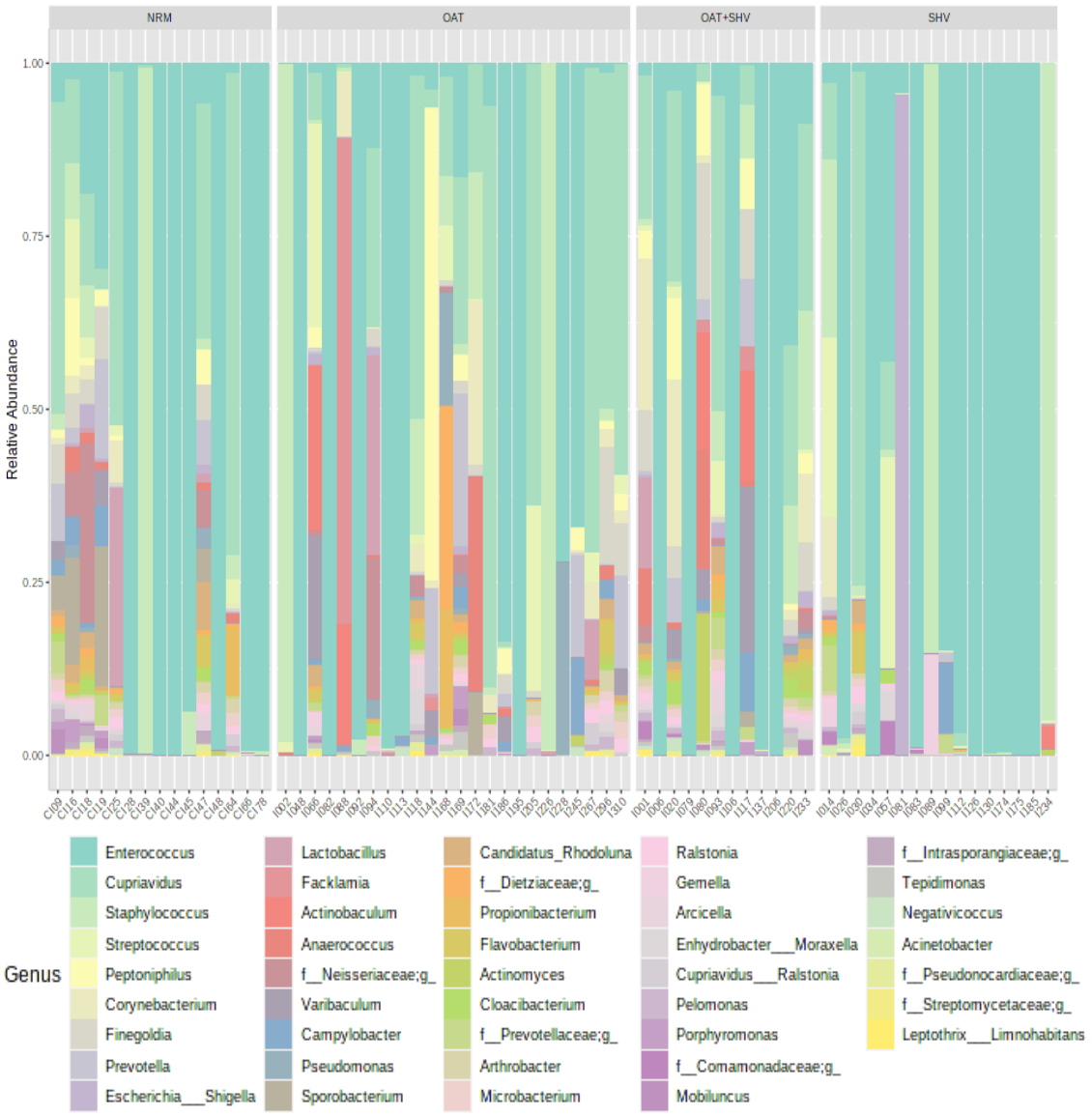



Figure 20 – Microbiome profiles of the most abundant genera (f>0.1%) per individual sample divided according to their phenotype grouping (NRM, OAT, OAT+SHV and SHV).

## Annex IV – Results for the V3 hypervariable region

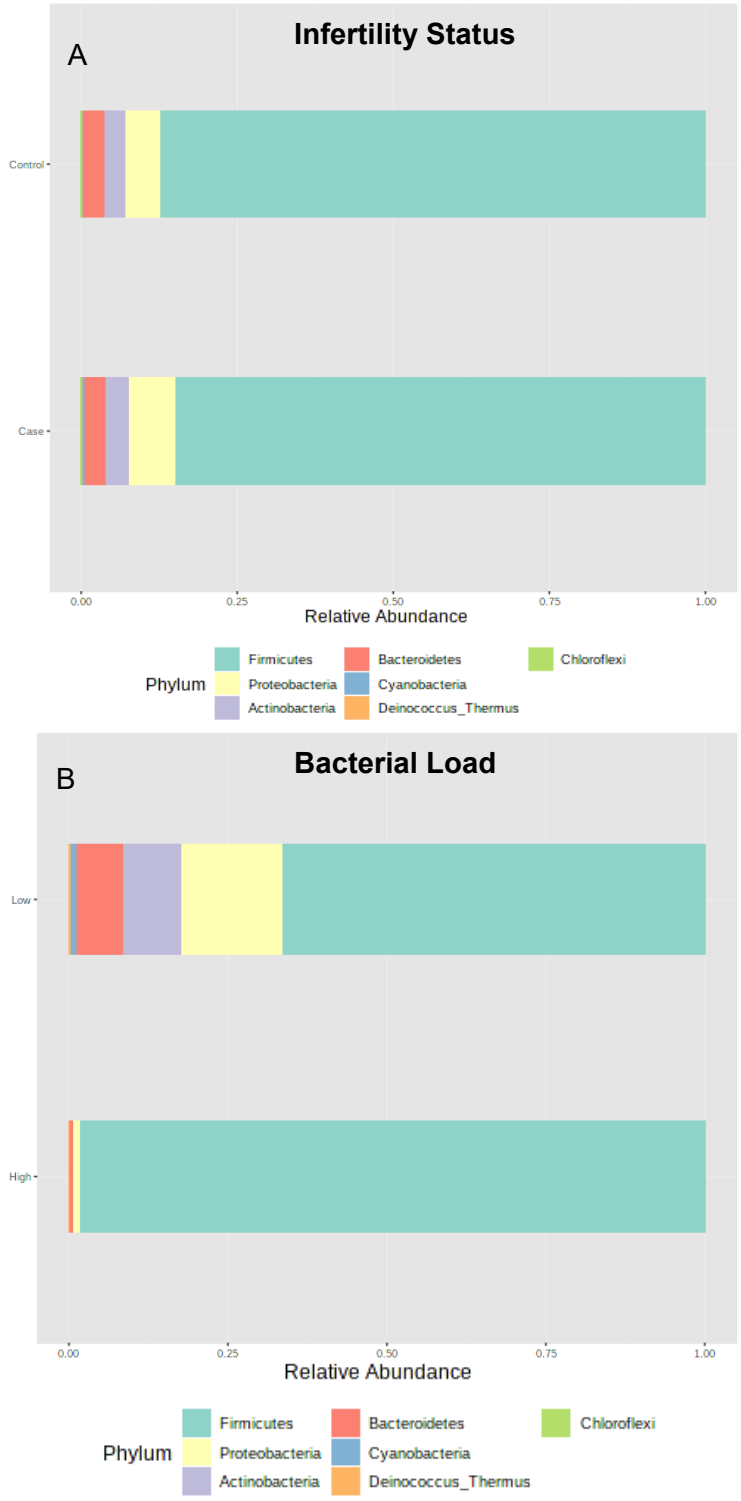



Figure 21 – Microbiome profiles of most abundant phyla (f>0.1%) according to infertility status (A) and bacterial load (B) for the v3 hypervariable region.

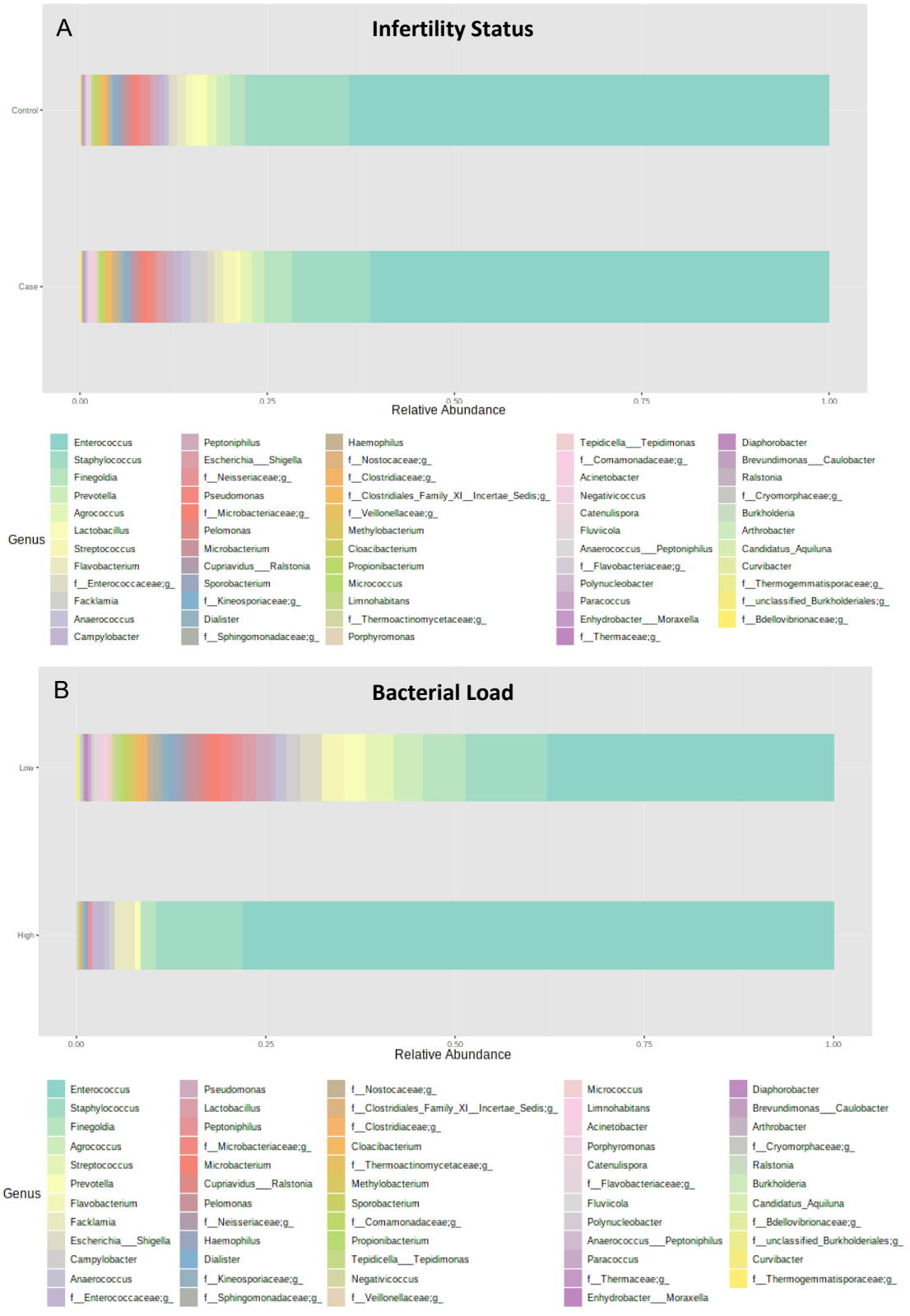



Figure 22 – Microbiome profiles of most abundant genera (f>0.1%) according to infertility status (A) and bacterial load (B) for the v3 region. Entries such as f__Enterococcaceae:g_ represent the grouping of when the genus was not discriminated.

Table 14 – Frequencies for the observed phyla and the ten more abundant genera found in semen samples for the v3 hypervariable region.

| Taxon | Relative Frequencies (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Infertility status | | Phenotype | | | Bacterial load | |
| | Case | Control | OAT | OAT+SHV | SHV | High | Low |
| Phyla | | | | | | | |
| Firmicutes | 85.0 | 87.4 | 82.2 | 82.8 | 89.8 | 98.3 | 66.5 |
| Proteobacteria | 7.3 | 5.4 | 8.3 | 6.4 | 6.5 | 8.6 | 15.8 |
| Actinobacteria | 3.8 | 3.6 | 4.5 | 5.6 | 1.9 | <0.1 | 9.3 |
| Bacteroidetes | 3.4 | 3.3 | 4.3 | 4.4 | 1.7 | 7.5 | 7.3 |
| Cyanobacteria | 0.4 | 0.3 | 0.6 | 0.5 | <0.1 | <0.1 | 0.8 |
| Deinococcus_Thermus | <0.1 | <0.1 | <0.1 | 0.2 | <0.1 | 0 | 0.2 |
| Chloroflexi | <0.1 | <0.1 | 0.1 | <0.1 | <0.1 | <0.1 | <0.1 |
| Genera | | | | | | | |
| *Enterococcus* | 61.3 | 64.1 | 53.6 | 62.9 | 69.5 | 78.1 | 37.9 |
| *Staphylococcus* | 10.3 | 13.8 | 10.1 | 2.6 | 15.3 | 11.4 | 10.5 |
| *Finegoldia* | 3.9 | 2.0 | 4.6 | 8.2 | 3.2 | 2.0 | 5.7 |
| *Facklamia* | 2.1 | <0.1 | 4.7 | 0.5 | 0.1 | 2.6 | 0.2 |
| *Streptococcus* | 1.7 | 0.6 | 2.2 | 0.3 | 2.0 | <0.1 | 3.7 |
| *Agrococcus* | 1.6 | 1.4 | 1.9 | 2.4 | 0.8 | <0.1 | 3.9 |
| *Prevotella* | 1.5 | 1.6 | 2.3 | 1.4 | 0.6 | 0.7 | 2.8 |
| *Escherichia/Shigella* | 1.4 | 0.3 | 0.3 | 0.5 | 3.2 | <0.1 | 2.8 |
| *Anaerococcus* | 1.2 | 0.7 | 1.0 | 3.5 | 0.1 | 0.8 | 1.6 |
| *Campylobacter* | 1.2 | 0.6 | 1.5 | 1.2 | 0.8 | 0.6 | 1.8 |


Figure 23 – Heatmap measuring over and under expression of each of the identified genera for the v3 hypervariable region.

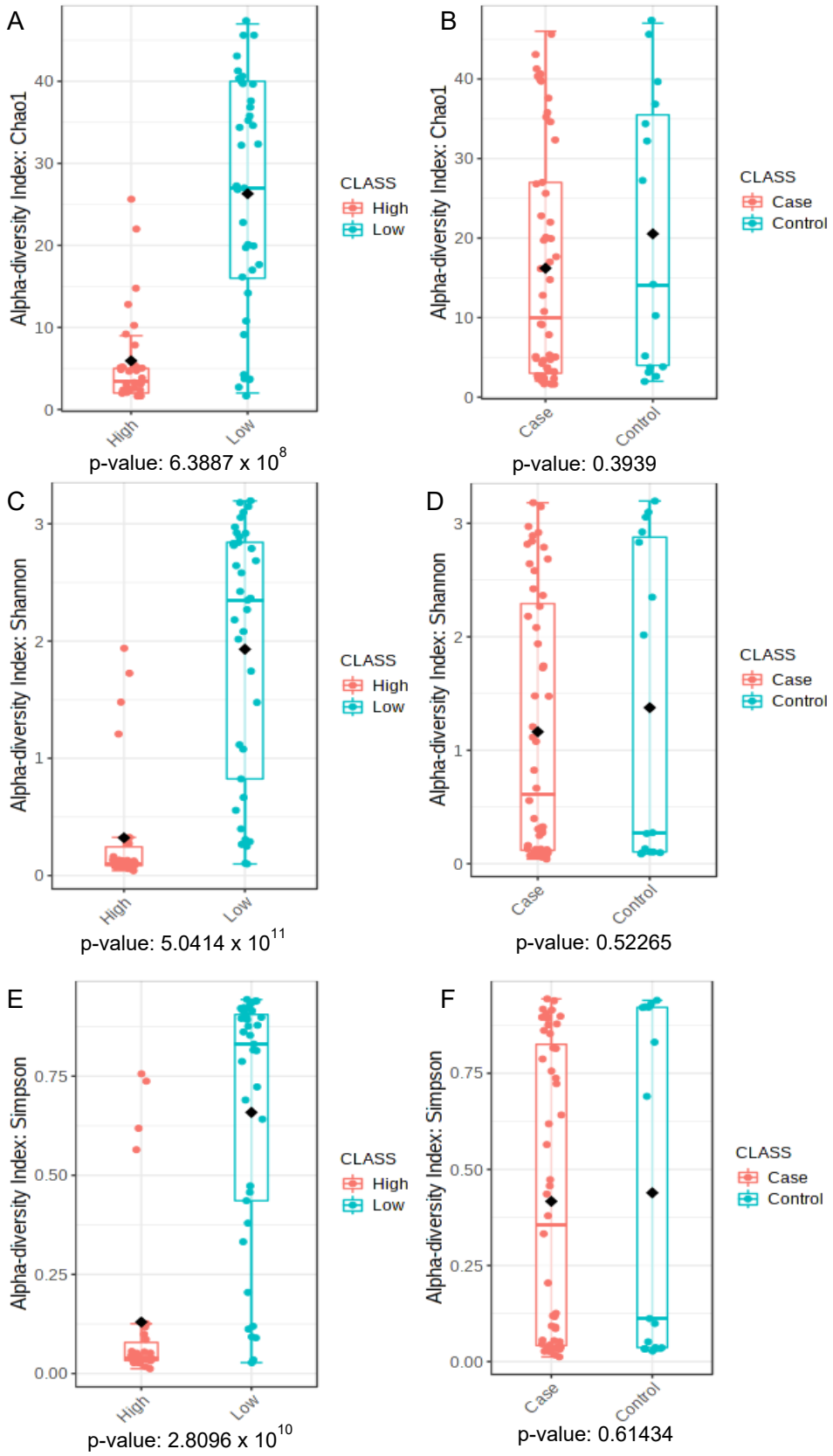



Figure 24 – Alpha diversity (Chao1 index) of bacterial load (A) and infertility status (B) groups; Alpha diversity (Shannon index) of bacterial load (C) and infertility status (D) groups; Alpha diversity (Simpson index) of bacterial load (E) and infertility status (F) groups. These graphs are for the v3 hypervariable region.

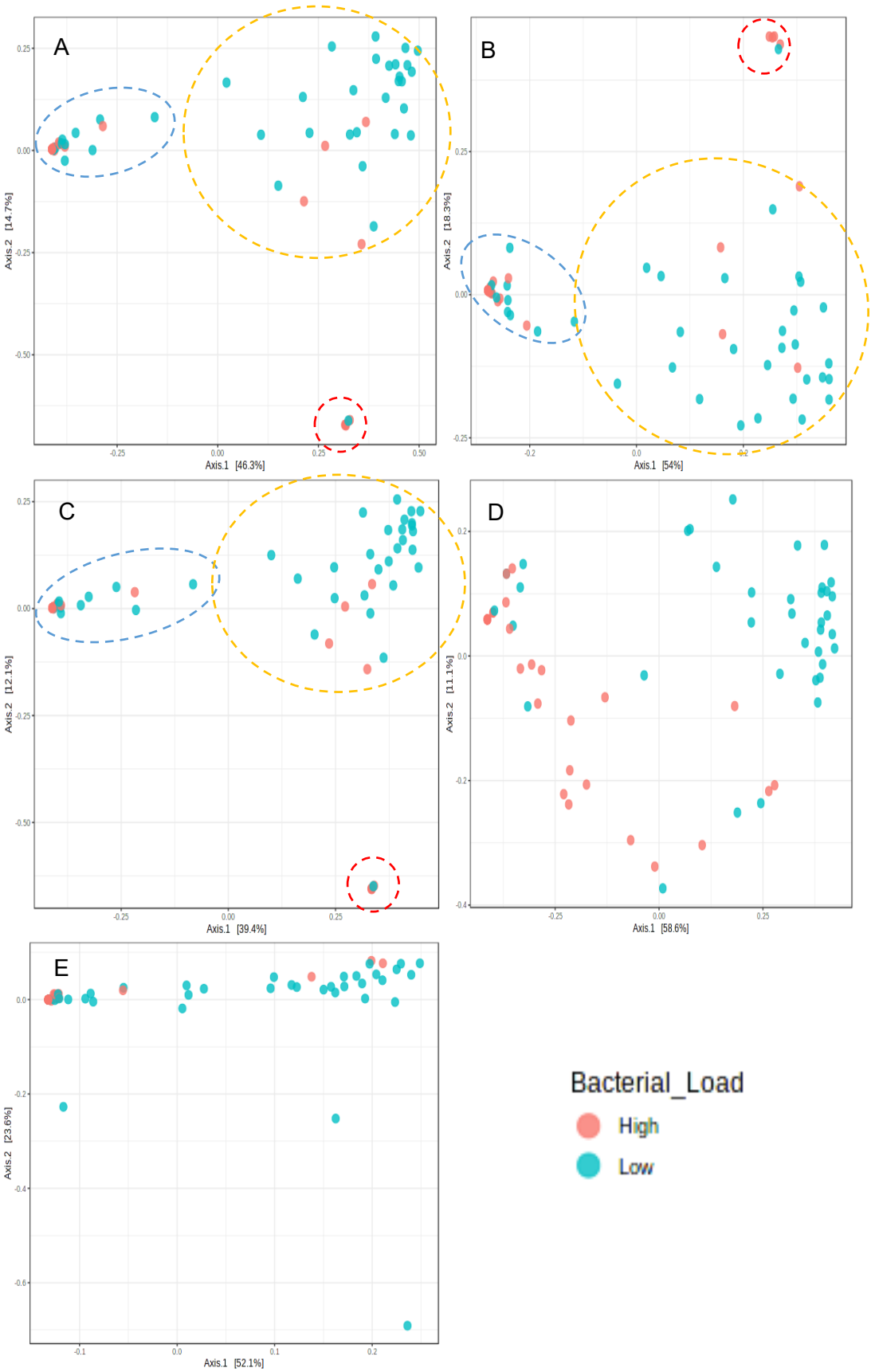



Figure 25 – Displayed microbial distances between samples using the Bray-Curtis (A), Jensen-Shannon (B), Jaccard (C), Unweighted UniFrac (D) and Weighted UniFrac (E) indices, according to bacterial load for the v4 hypervariable region. Clusters are the same as in Figure 11.

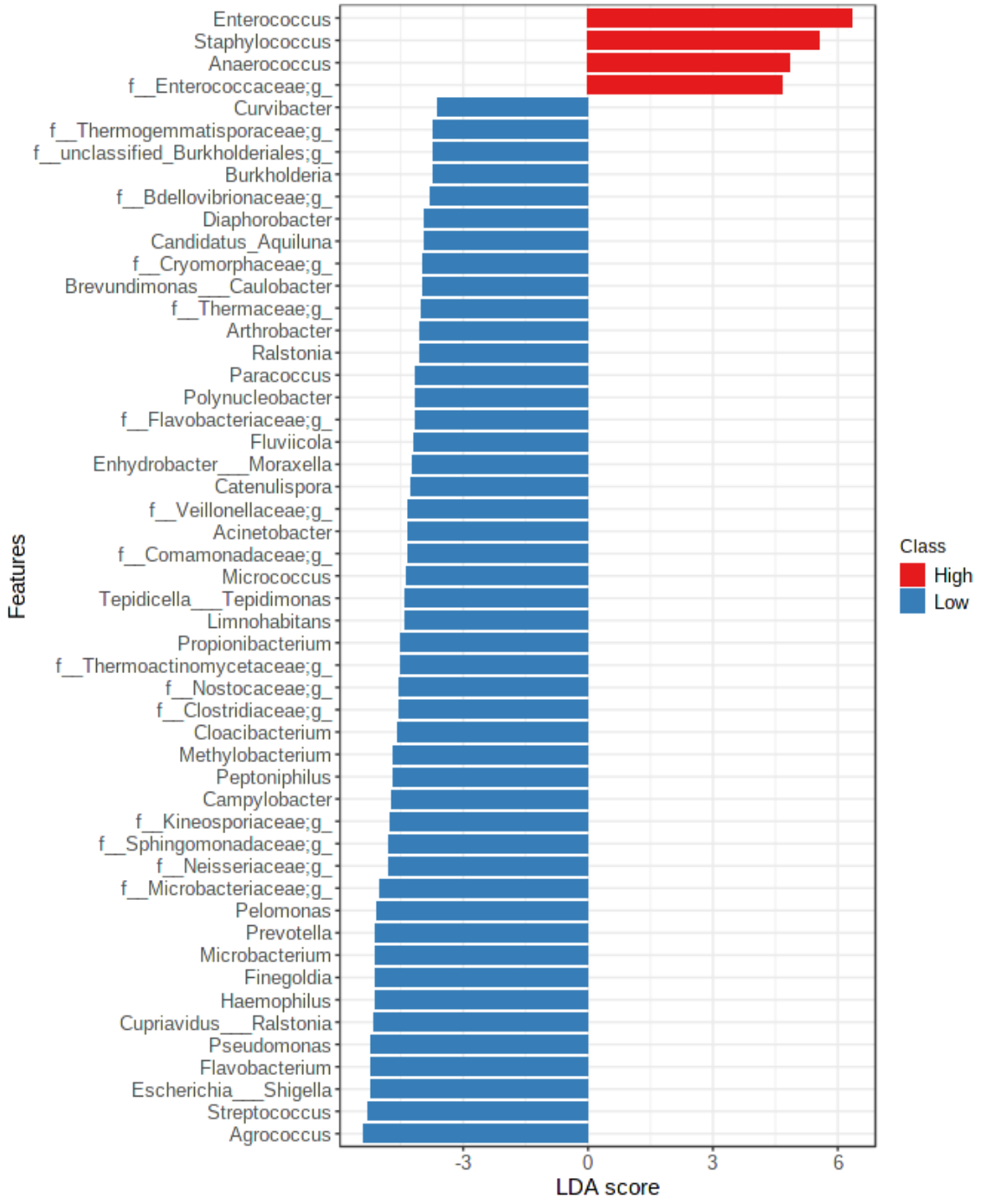



Figure 26 – Microbial differentiation of samples according to high and low bacterial load and linear discriminant analysis (LDA) effect size (LEfSe) algorithm for the v3 hypervariable region.

## Annex V – Results for the V6-7 hypervariable region

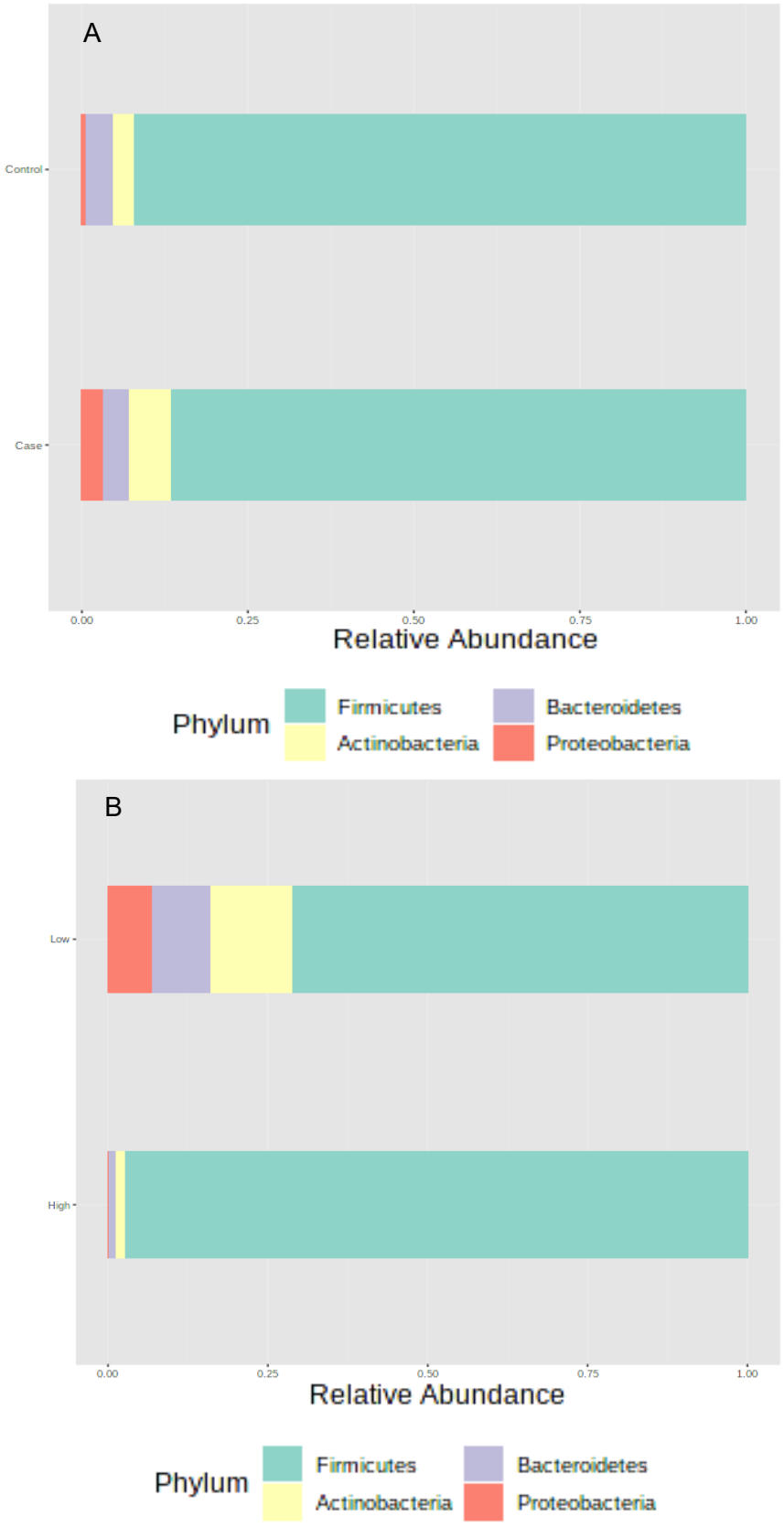



Figure 27 – Microbiome profiles of most abundant phyla (f>0.1%) according to infertility status (A) and bacterial load (B) for the v6-7 hypervariable region.


Figure 28 – Microbiome profiles of most abundant genera (f>0.1%) according to infertility status (A) and bacterial load (B) for the v6-7 region. Entries such as f__Cytophagaceae:g_ represent the grouping of when the genus was not discriminated.

Table 15 – Frequencies for the observed phyla and the ten more abundant genera found in semen samples for the v6-7 hypervariable region.

| Taxon | Relative Frequencies (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Infertility status | | Phenotype | | | Bacterial load | |
| | Case | Control | OAT | OAT+SHV | SHV | High | Low |
| Phyla | | | | | | | |
| Firmicutes | 86.5 | 92.0 | 87.4 | 79.0 | 90.0 | 97.2 | 71.2 |
| Bacteroidetes | 6.5 | 3.4 | 4.8 | 4.6 | 2.4 | 1.0 | 0.9 |
| Actinobacteria | 3.9 | 4.0 | 5.0 | 15.2 | 3.1 | 1.8 | 13.0 |
| Proteobacteria | 3.1 | 0.5 | 2.8 | 1.2 | 4.6 | <0.1 | 6.8 |
| Genera | | | | | | | |
| *Enterococcus* | 65.6 | 71.1 | 61.6 | 64.7 | 70.6 | 77.6 | 47.9 |
| *Staphylococcus* | 11.5 | 12.2 | 11.3 | 2.3 | 17.3 | 12.6 | 10.0 |
| *Corynebacterium* | 2.9 | 1.2 | 1.5 | 7.7 | 1.5 | 0.5 | 5.9 |
| *Peptoniphilus* | 2.8 | 1.7 | 4.4 | 4.4 | <0.1 | 2.1 | 3.5 |
| *Facklamia* | 2.0 | <0.1 | 4.4 | 3.3 | <0.1 | 2.3 | 0.1 |
| *Escherichia / Shigella* | 1.9 | 0.2 | 2.9 | 0.7 | 4.5 | <0.1 | 4.1 |
| *Prevotella* | 1.6 | 1.6 | 2.9 | 1.6 | 1.3 | 1.0 | 2.8 |
| *Actinobaculum* | 1.3 | 0 | 1.5 | 2.5 | 0.3 | 0.8 | 1.3 |
| *Finegoldia* | 1.1 | 0.8 | 0.7 | 3.4 | <0.1 | 0.4 | 2.0 |
| *Propionibacterium* | 1.0 | 1.2 | 0.6 | 2.1 | 0.9 | <0.1 | 2.9 |

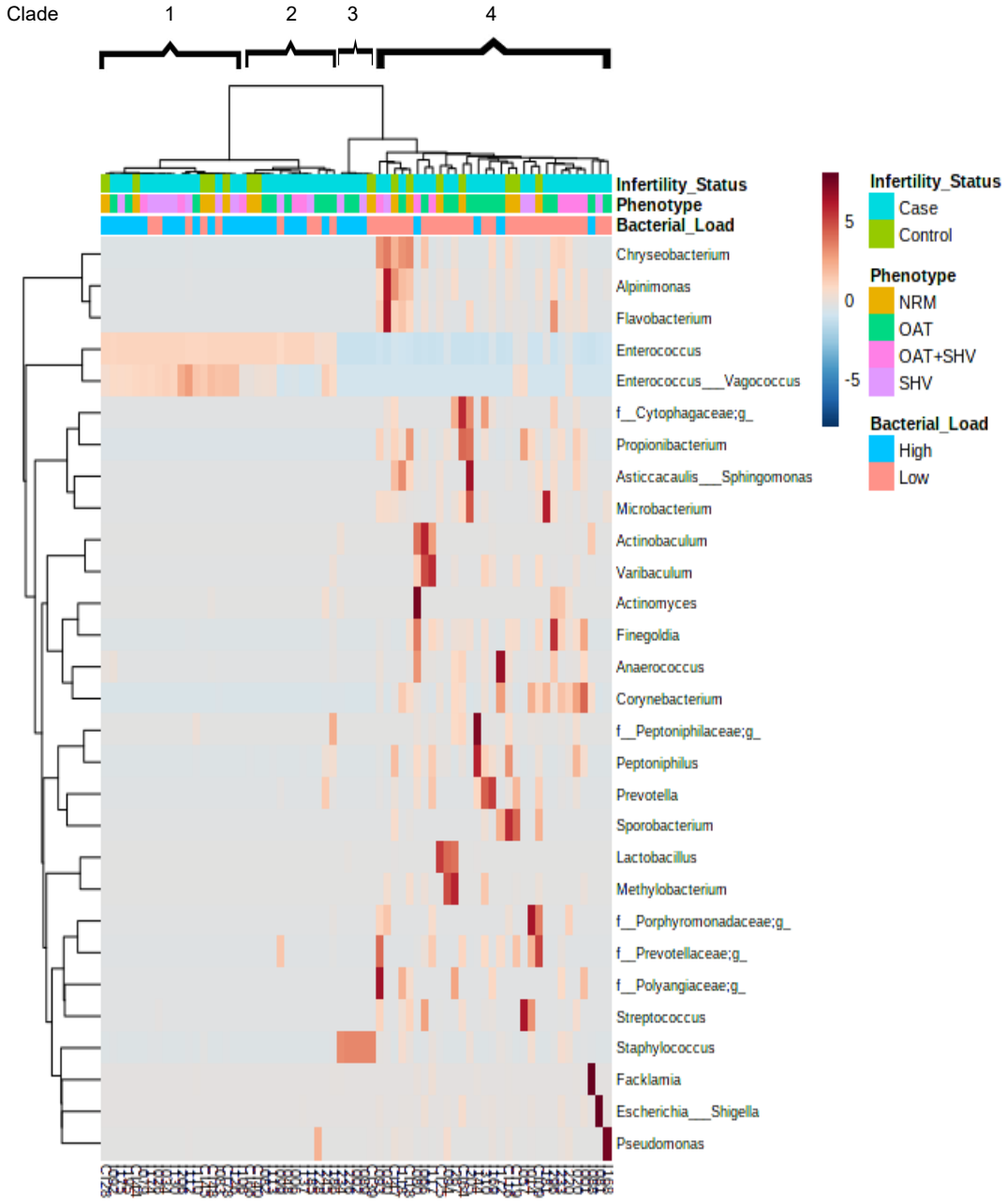



Figure 29 – Heatmap measuring over and under expression of each of the identified genera for the v6-7 hypervariable region.

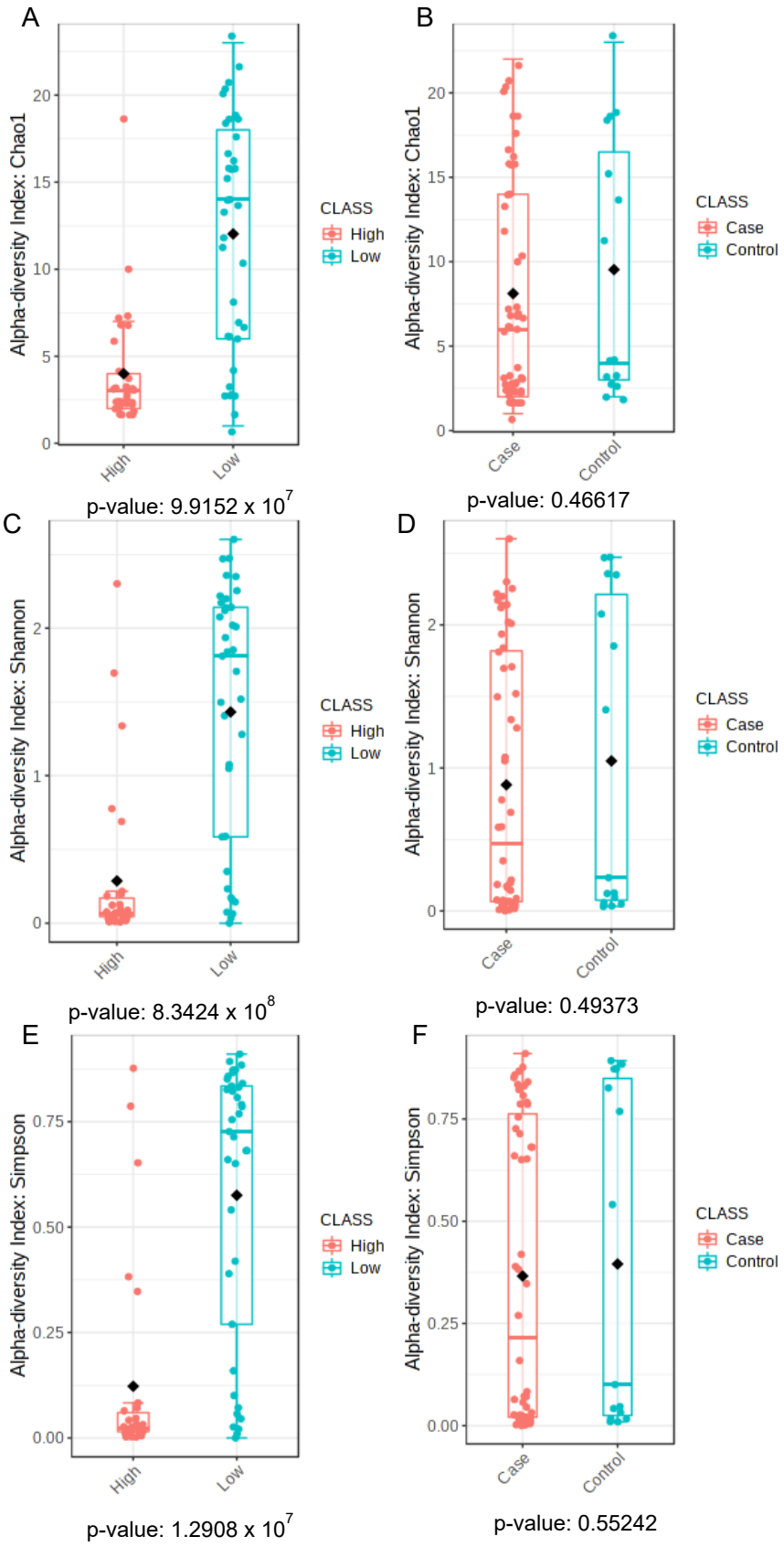



Figure 30 – Alpha diversity (Chao1 index) of bacterial load (A) and infertility status (B) groups; Alpha diversity (Shannon index) of bacterial load (C) and infertility status (D) groups; Alpha diversity (Simpson index) of bacterial load (E) and infertility status (F) groups. These graphs are for the v6-7 hypervariable region.

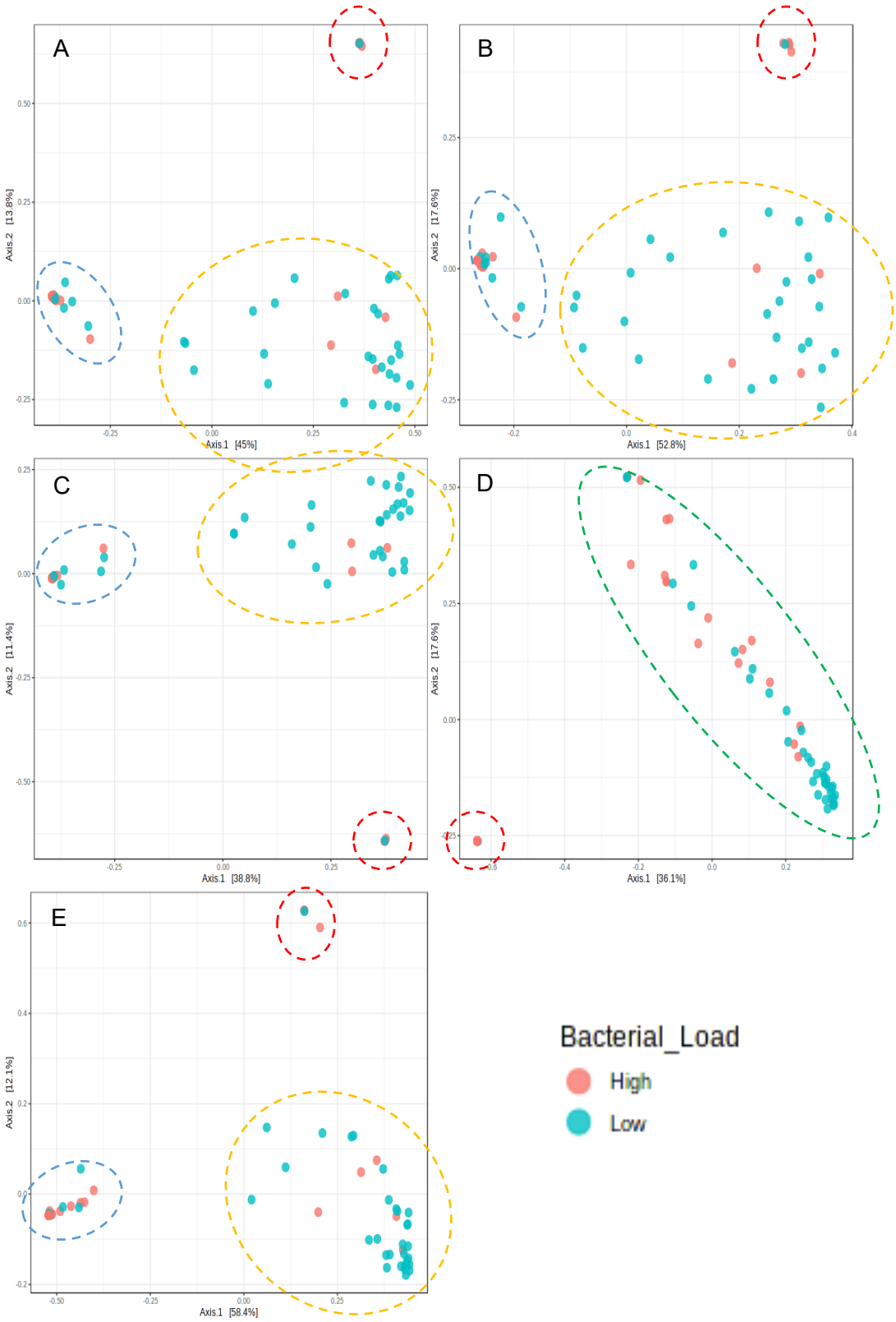



Figure 31 – Displayed microbial distances between samples using Bray-Curtis index (A), Jensen-Shannon (B), Jaccard (C), Unweighted UniFrac (D) and Weighted UniFrac (E) for the V6-7 hypervariable region, according to bacterial load. The clusters are the same as Figure 11.

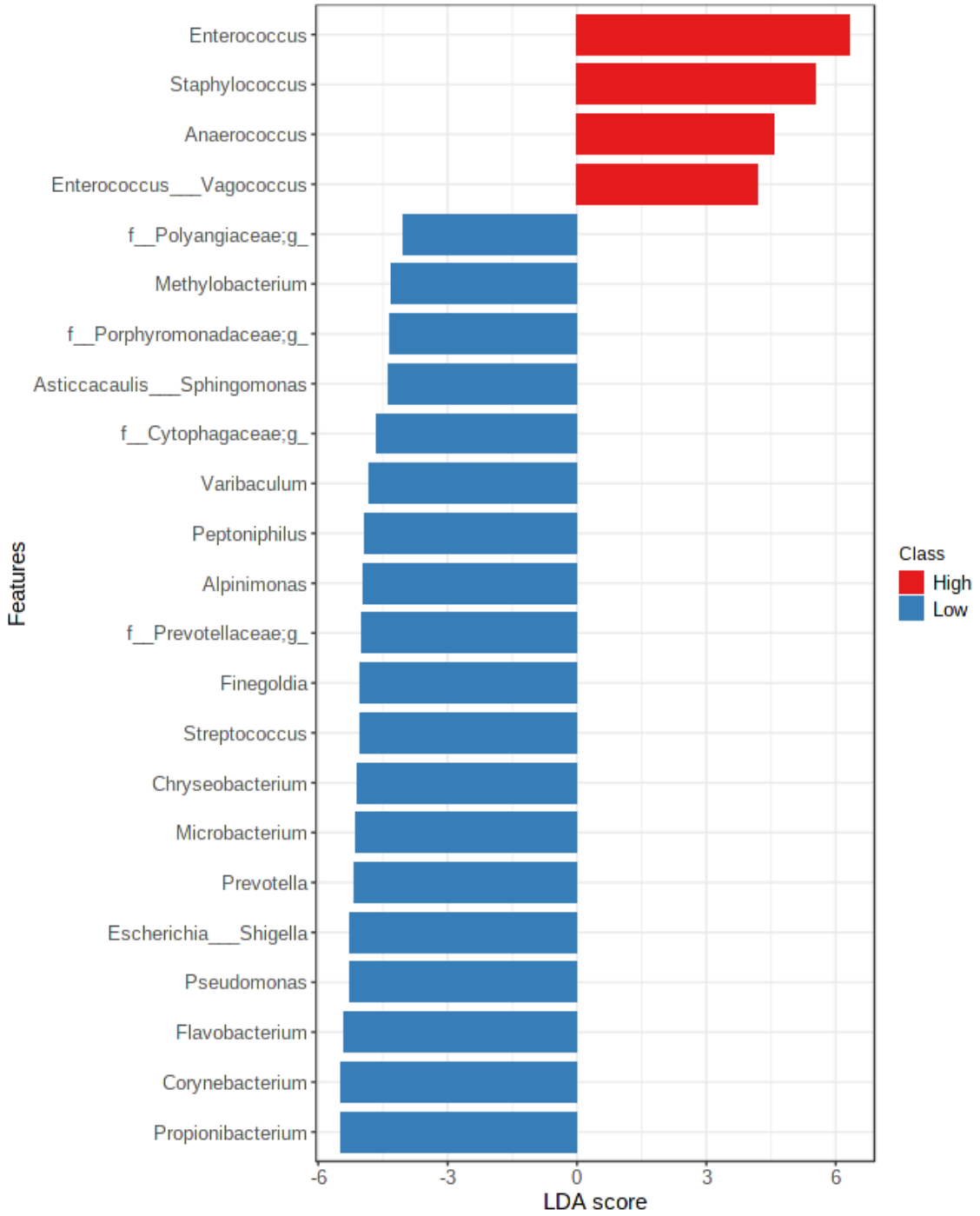



Figure 32 – Microbial differentiation of samples according to high and low bacterial load and linear discriminant analysis (LDA) effect size (LEfSe) algorithm for the v6-7 hypervariable region.