The ejaculate microbiota in an avian hybrid system across allo- and sympatry

Simen Fredriksen

Master of Science Thesis

Department of Biosciences

Faculty of Mathematics and Natural Sciences

University of Oslo

01.06.2017

2017

The ejaculate microbiota in an avian hybrid system across allo- and sympatry

Simen Fredriksen

http://www.duo.uio.no/

Print: Reprosentralen, University of Oslo

Abstract

In sexually reproducing taxa, normal sperm function is critical for successful reproduction, and pathogenic bacteria can prevent this. Thus, understanding the role of bacteria in ejaculates can have significant implications for ecology and evolution. Although a few studies have investigated the human ejaculate microbiota, the amount of culture-independent research on other species is limited. By utilizing high-throughput sequencing of 16S rRNA gene amplicons, significant advances can be made in the knowledge of diverse bacterial communities. This study describes and compares the ejaculate microbiota of the house sparrow (*Passer domesticus*), Spanish sparrow (*P. hispaniolensis*) and Italian sparrow (*P. italiae*). These species constitute a hybrid species system in which the Italian sparrow originated through hybridization between the house and Spanish sparrow, and occur in both allo- and sympatry throughout Europe. I found the composition and structure of the sparrow ejaculate microbiota to be highly variable between individuals, and this obscured any species-specific signal. Individuals at different locations did however trend towards being different. I detected a wide range of bacteria belonging to 36 phylum-level classifications, of which Bacteroidetes and Proteobacteria were the most abundant. At the genus level, I found *Flavobacterium* to dominate the avian ejaculate microbiota. Notably, a considerable variety of bacteria classified to unculturable candidate phyla were detected. Overall, I found large overlap with taxa commonly found in the avian gastrointestinal tract and human ejaculate, as well as with previous culture-based studies on avian ejaculate. A wide range of potential pathogens likely to detriment host health or sperm function were detected. It is likely that these cause similar selective pressures on the mating systems in the three species, as no species-specific microbiota is detected. This study presents a significant advance in knowledge on the composition and structure of the avian ejaculate microbiota.

Acknowledgements

First and foremost I would like to thank my supervisors Melissah Rowe, Bastiaan Star, and Glenn-Peter Sætre, as well as PhD student Diana Carneiro. Pål Trosvik and Eric De Muinck were of great help in the lab work, and provided the primers, mock community, and fecal samples used in this study. Caroline Skar Mjønes also contributed to the lab work. Even Sannes Riiser provided help with the bioinformatics, and introduced me to the phyloseq R package. Alfonso Marzal Reynolds and Sergio Magallanes (University of Extremadura) contributed to the fieldwork, including their sampling permit, and are some of the nicest people I have ever met. Samples collected for my project by Mark Ravinet and Camilla Lo Cascio Sætre in Tenerife were unfortunately lost due to a freezer breakdown, but I am equally grateful for their effort nonetheless. Thank you also to everyone else who provided help in smaller ways.

Contents

1 Introduction

Bacteria are among the most abundant organisms on earth and occur virtually everywhere - in the air, soil, and water, in plants and animals, even inhabiting extreme environments such as hydrothermal vents. Host-bacteria interactions are widespread, and many such interactions appear to have a considerable impact on ecological and evolutionary processes, as well as health and wellbeing of the host. For example, commensal gut bacteria in humans are integral in nutrient uptake, and host-bacterial interactions within the gut are thought to help the immune system fight pathogens (Shreiner et al. 2015).

Bacteria are known to occur in ejaculate, and has been found in all animals investigated, e.g. humans (Weng et al. 2014; Mändar et al. 2015), rats (Javurek et al. 2016), boars (Martín et al. 2010), and bees (Andere et al. 2011). Crucially to infected males, pathogens can be among the bacteria inhabiting their ejaculate and reproductive system. Such infections can retard the development of sperm, and decrease sperm motility and ability to fertilize eggs by attaching to them (Diemer et al. 2003). Ultimately, male reproductive tract infection is associated with infertility and reduced fitness (Lockhart et al. 1996; Pellati et al. 2008), and cause approximately 15% of male infertility cases in humans (Diemer et al. 2003).

In addition to the potential impact on males, ejaculate-associated bacteria can be transferred to females during copulation, and these sexually transmitted microbes (STMs) may or may not cause disease (i.e. sexually transmitted disease, STDs). A range of human STDs such as *Chlamydia trachomatis* and *Neisseria gonorrhoeae* are well described, and many are widespread and severely detriment host health (Fung et al. 2007). A range of pathogens are also thought to be sexually transmitted in other animals (Sheldon 1993; Lockhart et al. 1996). The incidence of STMs have been linked to female multiple mating (Sheldon 1993; Poiani and Gwozdz 2002), and ejaculates contain a range of anti-bacterial components (Lung et al. 2001; Rowe et al. 2011; Otti et al. 2013). This indicates that the ejaculate microbiota is likely to be under selective pressure.

Though the adaptive benefit of female multiple mating remains contentious, several theories have been proposed to explain the behaviour. Copulating with additional males could function as bet-hedging by increasing the offspring's genetic diversity, and copulation with males of higher quality than the social partner increases the genetic quality in offspring (Forstmeier et al. 2014). Moreover, allowing copulations with extra-pair males can avoid infanticide and harassment, while securing additional paternal care and resource access (Griffith et al. 2002). These potential benefits are however potentially countered by risk of STD infection, and thus STDs might be of critical importance in the evolution of mating systems (Hamilton and Zuk 1982; Hamilton 1990; Sheldon 1993; Reiber et al. 1995; Poiani and Wilks 2000).

Both the ejaculate and waste pass through the avian cloaca, exposing the ejaculate to the gastrointestinal microbiota, which is known to harbour large bacterial communities including both pathogens and commensals (Kreisinger et al. 2015; Lewis et al. 2016). A range of bacteria have been isolated from the avian cloaca and ejaculate (Lombardo and Thorpe 2000; Stewart and Rambo 2000; Poiani and Gwozdz 2002; Kreisinger et al. 2015), and in theory any of these can be sexually transmitted, and thus be potential STDs. While there is a lack of conclusive evidence linking any particular avian-associated bacteria to sexual transmission and pathology, evidence for sexual transmission of bacteria has been found in birds (Lombardo et al. 1996; Stewart and Rambo 2000; Kulkarni and Heeb 2007; White et al. 2010). Moreover, putative avian STDs such as *Mycoplasma, Salmonella, and Campylobacter* have been found to reduce body mass, fertility and egg production, as well as to cause mortality (Stipkovits et al. 1986; Marius-Jestin et al. 1987; Lockhart et al. 1996; Waldenström et al. 2010). Sexual selection has driven evolution of male secondary sexual characteristics in many avian species, and these have been suggested to signal anti-microbial capability and ejaculate quality, thus enabling females to mate with males less prone to transfer STDs (Able 1996; Poiani 2010; Rowe et al. 2011).

Previous studies on the bacterial flora in avian ejaculates have used culture-based techniques, which do not capture the full diversity or relative abundance of community members (Pace 1997). Thus, a limited number of bacteria are known from avian ejaculates, and knowledge on community structure is lacking. Approaches utilizing next-generation sequencing of 16S rRNA gene amplicons have made it possible to comprehensively characterize bacterial communities (Handelsman 2004), though classification of bacteria to the species level is rarely possible. Studies with this approach have been carried out on the ejaculate microbiota of humans (Weng et al. 2014), and more recently lab mice (Javurek et al. 2016), but data on birds or any wild animal is lacking.

The primary aim of this study was to identify and characterize the ejaculate microbiota of birds using high-throughput sequencing of 16S rRNA gene amplicons. Secondarily, I aimed to gain insight into the relative importance of location and host species as drivers of the composition and structure of the microbiota. To achieve these goals, I made use of a hybrid system consisting of three closely related seed eating *Passer* sparrows, which occur in allopatry and combinations of sympatry throughout Europe. The house sparrow (*Passer domesticus*) is common near human settlements throughout Europe, while the Spanish sparrow *(P. hispaniolensis)* is native to the Mediterranean region but shares most of its range with the house sparrow (Summers-Smith 1988). Hybrids can be found sporadically in sympatric areas (Ait Belkacem et al. 2016) and were observed at low frequencies at the sympatric locations sampled in this study. Past hybridization events have resulted in the formation of the Italian sparrow *(P. italiae*) (Elgvin et al. 2011; Hermansen et al. 2011). This reproductively isolated hybrid species (Trier et al. 2014) is ubiquitous throughout most of Italy (figure 1, Hermansen et al. 2011). The three species are relatively similar in terms of ecology and behaviour, some places nesting side by side (Summers-Smith 1988). Consequently, the system is ideal for investigations of community drivers associated with species or locations.

Several factors could drive bacterial communities to differ between species, as divergent evolutionary history can cause a range of changes to the inner environment of the birds. The seminal fluid proteome of the house and Spanish sparrow is divergent (Rowe, unpublished data), and changes to anti-bacterial peptides affect the microbiota (Franzenburg et al. 2013). Moreover, testes size correlates with levels of multiple mating (Moller 1991; Brown and Brown 2003), and some evidence suggests Spanish sparrows to have larger testes than house sparrows (Moller 1991; Birkhead et al. 1994; Partecke and Schwabl 2008), while preliminary data has indicated Italian sparrows to be intermediate (Rowe, unpublished data). Thus, selection pressures related to spread of bacteria might differ between the species. Finally, reproductive barriers are vital in hybrid speciation (Hermansen et al. 2011), and it is possible that ejaculate-related traits could be involved in these. This could have caused rapid divergence of the ejaculate microbiota in this study system. Species-specific microbiota profiles have been found in several studies, for example in the gut of howler monkeys (Amato et al. 2016) and primates (Yildirim et al. 2010; McCord et al. 2014), skin of amphibians (McKenzie et al. 2012), and the primate vagina (Yildirim et al. 2014).

Environmental factors, often associated with location, have also been found to drive a range of bacterial communities. Examples of this include the primate gut (Moeller et al. 2013; Amato et al. 2016) and saliva (Li et al. 2013), bird gut (Lewis et al. 2016), and mosquito larvae (Coon et al. 2016). Nesting sites and other surfaces are inevitably contaminated with feces from con- and heterospecifics, and high nest density might cause similar effects to high levels of multiple mating between individuals at the site. Thus, the effect of common nesting sites at sympatric locations might largely displace any species-specific microbiota. In addition, different cloacal bacteria are likely available to contaminate the ejaculate at different locations, as the gut microbiota is affected by diet (Pan and Yu 2014) and different seed-bearing plants are available in different habitats. Finally, it is conceivable that a wide range of factors such as temperature, humidity, environmental bacteria, and nest material play a role. Both bacteria and host can face biotic and abiotic challenges from the external environment, and thus community membership and structure might differ over time and between locations. Notably in the study of STDs, ecological interactions within the host might alter infection levels (Belden and Harris 2007), and thus the function of each community member might differ between locations.

In addition to describing the composition and structure of the sparrow ejaculate microbiota, I had a number of additional aims. First, I aimed to identify pathogens and potential STDs known from previous culture-based studies and published literature. Second, I aimed to explore the relationship between the ejaculate microbiota and body condition, i.e. if a particular community composition or structure is associated with birds of putatively high or low condition. Finally, I aimed to investigate if the microbiota changes through the mating season, which could indicate that seasonal factors like temperature or diet plays a role. In a broad perspective, this study aimed to improve the understanding of the avian ejaculate microbiota, as well as the ejaculate microbiota in general, as metagenomics-based studies on wild animals are lacking.

2 Methods

2.1 Sample collection

We collected ejaculate samples from 107 sparrows during the breeding season (March-May) in 2016. We sampled at four locations across Europe (figure 1): Oslo, Norway (house sparrows), Montanari, Italy (Italian sparrows), Lago Salso (LGS), Italy (Spanish and Italian sparrows), and Badajoz, Spain (Spanish and house sparrows). Thus, we sampled house and Italian sparrows at 1 allopatric and 1 sympatric population, and Spanish sparrows in sympatry with the two other species (see appendix 1 for full sampling details). In addition to the samples presented in this study, we collected 8 allopatric Spanish sparrows in Tenerife, Spain, that were lost due to a freezer breakdown. The Lago Salso and Montanari sample sites are located in farmland near the sea, Badajoz is in interior farmland, and the birds from Oslo were sampled in the University Botanical Garden.

Birds were caught using mist nets, and ejaculate samples were collected via cloacal massage (Wolfson 1952; Rowe and Pruett-Jones 2011). Before sampling, birds were evaluated to be healthy, and the exterior of the cloacal protuberance was cleaned with 70% ethanol to avoid sampling of bacteria from the skin or feathers. The ejaculate (average: 1.2 µL, range: 0.2-2.0 µL) was collected with a microcapillary and transferred to a 1.5 mL nunc vial containing 20 µL 30% glycerol to prevent freeze damage to cells. The samples were either immediately put on dry ice and later placed in a -80 °C freezer for storage, or put directly in the -80 °C freezer. We utilized sterile equipment to avoid contamination, and nitrile gloves were cleaned with 70% ethanol before massaging. We collected 1-2 negative sampling controls at each location by pipetting an empty microcapillary into an empty sample tube. These controls were processed in an identical manner to the ejaculate samples. While the ‘true microbiota’ of samples with high bacterial biomass is likely to overwhelm any trace contaminants, samples with decreasing size and bacterial biomass are likely to have contaminants occupy an increasing proportion of reads (Weiss et al. 2014). As sparrow ejaculate samples are small, and it is possible that the bacterial load is low, detecting contaminants was of special importance.

All fieldwork was done with permission from the appropriate authorities at all locations. In Norway, under permit issued by the Norwegian Directorate for Nature Management (ref#

2016/87), in Italy, under permit issued by the Istituto Superiore per la Protezione e la Ricerca Ambientale (ref# 7612/2015), and in Spain, under permit issued by the Consejería de Medio Ambiente y Rural, Políticas Agrarias y Territori (ref# CN0009/16/AAN). We adhered to the Nagoya protocol when transporting samples out of Italy and Spain. One human fecal sample was used in this study, and this was collected under approval by the Regional Ethics Committee of South-East Norway (ref# 2014/656).

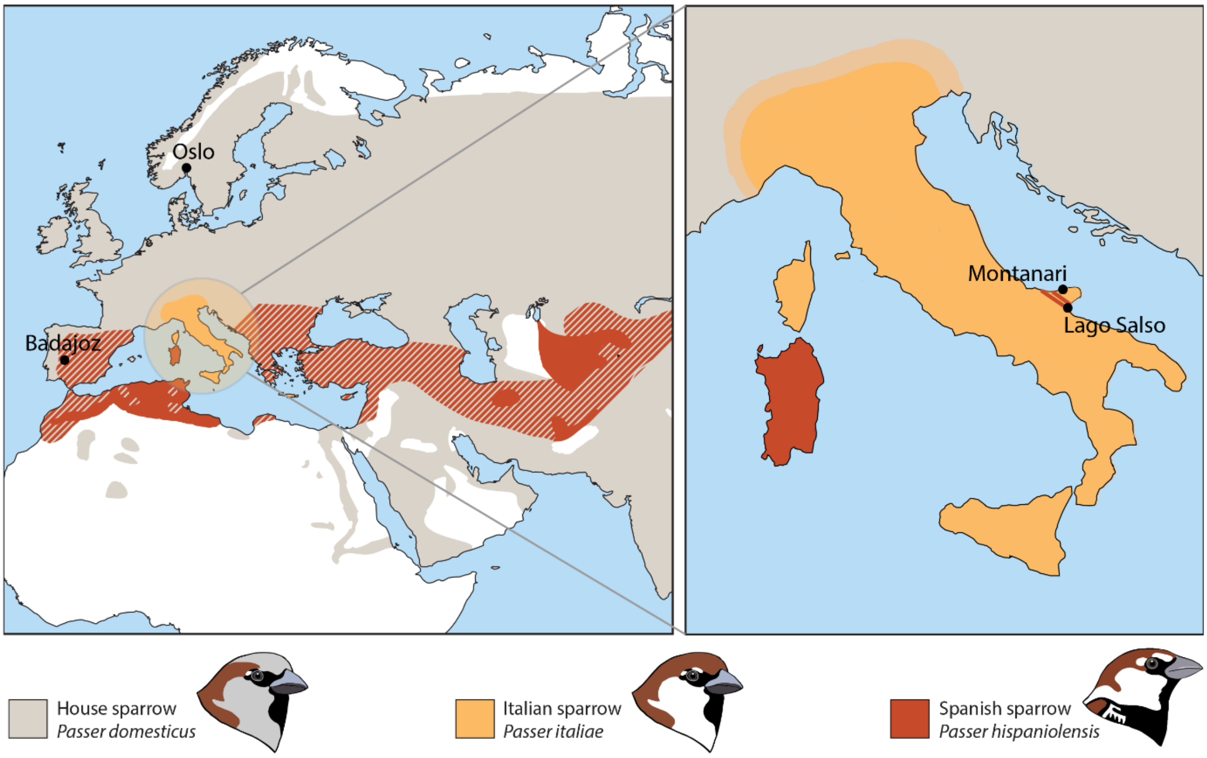

Figure 1 Distribution map. This map shows the European distribution of species in the study system, where the colors represent species, hatched areas sympatry, and sampling sites are indicated with black dots.

2.2 Laboratory protocol

2.2.1 DNA isolation

Due to the unusual sample material, a novel DNA isolation method modified from a previous study (Yuan et al. 2012) was used (see appendix 2 for details on the methods development). The frozen samples were first thawed on ice and homogenized in 250 µL sterile water. Subsequently, they were mixed with an enzymatic lytic cocktail containing 25 µL lysozyme (10 mg/mL), 3 µL mutanolysin (6250 U/mL) and 1.5 µL lysostaphin (4000 U/mL). After one hour of incubation at 37 °C, the lysate was transferred to a FastPrep tube containing 250 mg ≤106 µm acid-washed silica beads (Sigma-Aldrich, St. Louis, MO, USA). It was then subjected to one minute of bead beating at 4.0 m/s. Samples were spun down at 13 000 RPM for 1 minute, and 170 µL of supernatant was transferred to a new sterile tube. The supernatant was then treated according to the standard protocol of the DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA) "Purification of Total DNA from Animal Tissues (Spin-Column Protocol)", except for the amount of proteinase K being increased from 20 to 25 µL. Successful DNA isolation was visually confirmed on a 1% agarose gel.

2.2.2 Library preparation

The 253 bp (base pair) long V4 region of the 16S rRNA gene was amplified with the universal primers 515f (5'-GTGYCAGCMGCCGCGGTAA-3') and 806r (5'-GGACTACNVGGGTWTCTAAT-3') (Caporaso et al. 2012). We used a triple indexing approach (Muinck et al. 2017), where index sequences identifying the sample are added with each of the 515f-806r primers along with heterogeneity spacers. The DNA concentration of all samples are then normalized, and samples on the same plate pooled. A separate PCR with 2F and 2R primers adds Illumina adapters containing a third index identifying the plate. This method facilitates the multiplexing of large numbers of samples in each sequencing run. During development of the library preparation protocol (appendix 2), PCR amplification proved challenging with this protocol. This was solved by changing polymerase, increasing the number of PCR cycles, and by dividing the PCR into three PCR steps. In PCR 1 the 515f-806r region was amplified with non-indexed primers, in PCR 2 indexed primers were used, and in PCR 3 Illumina adapters were added. Both protocols result in the same amplicon structure (table 1).

In PCR 1, each reaction contained 0.25 µL Q5 High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA, USA) (2000 U/mL), 5 µL Q5 Reaction Buffer, 5 µL 515F primer (1 µM), 5 µL 806R primer (1 µM), 2.5 µL dNTPs (2 µM), 5.25 µL H_2O, and 2 µL template DNA. The following thermocycler conditions were used: 30 sec at 98 ºC, followed by 35 cycles [10 sec at 98 ºC, 30 sec at 53 ºC, 45 sec at 72 ºC], then 2 min at 72 ºC. Successful amplification was visually confirmed on a 1% agarose gel, and PCR product was cleaned with AMPure XP (Agencourt Bioscience Corporation, Beverly, MA, USA) following kit instructions. In PCR 2, each reaction contained 10 µL 5Prime Hot MM (Quantabio, Beverly, MA, USA), 2.5 µL 515F primer (1 µM), 2.5 µL 806R primer (1 µM), 5 µL H_2O, and 5 µL template DNA. The following thermocycler conditions were used: 3 min at 98 ºC, followed by 10 cycles [30 sec at 94 ºC, 30 sec at 50 ºC, 45 sec at 72 ºC], then 10 min at 72 ºC. Samples were cleaned and normalized to equal DNA concentrations (20 µL of ~1 ng/µL) using a SequalPrep Normalization Plate Kit (Invitrogen, Carlsbad, CA, USA) following kit instructions. In PCR 3, each reaction contained 10 µL 5Prime HotMaster Taq, 5 µL 2F primer (1 µM), 5 µL 2R primer (1 µM), 10 µL H_2O, and 10 µL template DNA. The following thermocycler conditions were used: 3 min at 98 ºC, followed by 5 cycles [30 sec at 94 ºC, 30 sec at 50 ºC, 45 sec at 72 ºC], then 10 min at 72 ºC. Samples were then pooled before being cleaned with AMPure XP. The final amplicon quantity was measured with a Qubit 2.0 Fluorometer (Invitrogen) using a dsDNA HS Assay. Each of the 2 plates used contained 2 negative controls, 3 bacterial mock community samples, and 3 standardized fecal samples. All sampling controls were put on the same plate.

Table 1 Example amplicon. Functional components of a theoretical amplicon from the current study. Total amplicon length should be between 416 to 430 bp, but a few species deviate from normal V4 region length of 253 bp.

Component	Sequence
Forward Illumina adapter (36 bp)	5’ - TCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Heterogeneity spacer (0-7 bp)	TTAACTG
Index (12 bp)	GAAGCCCTGTGG
515f (19 bp)	GTGYCAGCMGCCGCGGTAA
16S rRNA V4 region (253 bp)	TACGTAGGG[...]GCAAACAGG
806r (20 bp)	ATTAGAWACCCBNGTAGTCC
Index (12 bp)	TATCAGGCATCT
Heterogeneity spacer (0-7 bp)	TACG
Reverse Illumina adapter (part 1, 34 bp)	AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Illumina index (6 bp)	ATCACG
Reverse Illumina adapter (part 2, 24 bp)	ATCTCGTATGCCGTCTTCTGCTTG - 3’

2.2.3 Sequencing

The samples were paired-end sequenced at the Norwegian Sequencing Centre (NSC). Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) using Rapid mode and a v2 500 cycle kit with a 10% Phix spike-in produced a 250 bp sequence for each amplicon end.

2.3 Data processing

Demultiplexing was done using Trimmomatic (Bolger et al. 2014) and Cutadapt (Martin 2011). The paired end reads were joined into contigs using FLASH v1.2.11 (Magoč and Salzberg 2011), accepting overlaps between 90 and 250 bp. The contig files were converted from FASTQ to FASTA format, and primers were removed (see appendix 3). Further processing was performed in Mothur v1.38.0 (Schloss et al. 2009) (see appendix 4). Singletons and sequences longer than 263 bp or shorter than 243 bp were discarded. Sequences were aligned to the V4 region in SILVA database SSU v123 (Quast et al. 2013), and those aligning outside the appropriate range were cropped or discarded. VSEARCH (Rognes et al. 2016) was used for de-novo chimera detection. Sequences found to be chimeric in one sample were removed from the entire dataset. Sequence taxonomy was classified against the SILVA database using the Wang naive Bayesian method (Wang et al. 2007) with a 80% bootstrap cutoff, and sequences not classified to bacteria or archaea were discarded. Operational taxonomic units (OTUs, phylogenetic units based solely on sequence similarity independent of assigned taxonomic classification) were clustered at 97% identity using VSEARCH distance-based greedy clustering (DGC). This heuristic method assigns sequences to the existing OTU with which centroid it shares the highest sequence similarity with, or as the centroid of a new OTU, if no OTUs are within the 97% threshold (Westcott and Schloss 2015; Rognes et al. 2016). OTU taxonomy was determined with a 51% consensus cutoff. Neighbor-joining of the most abundant sequence in each OTU created a phylogenetic tree of OTUs. Both ejaculate and control samples were included in the pipeline used for the main analysis. See appendix 5 for analysis of the filtered reads from the main analysis.

The OTU table, OTU taxonomy, OTU phylogenetic tree, and sample metadata was imported to R using the package Phyloseq v1.19.1 (McMurdie and Holmes 2013). OTUs with total abundance of 50 sequences or less were discarded. Samples were rarefied to 42049 reads, compromising between the number of reads and samples retained (figure 2a). This rarefied dataset was used for all following analysis. As the 3 largest OTUs were identified as potential contaminants the analysis was rerun without them. In this dataset, samples were rarefied to 30984 reads (figure 2b).

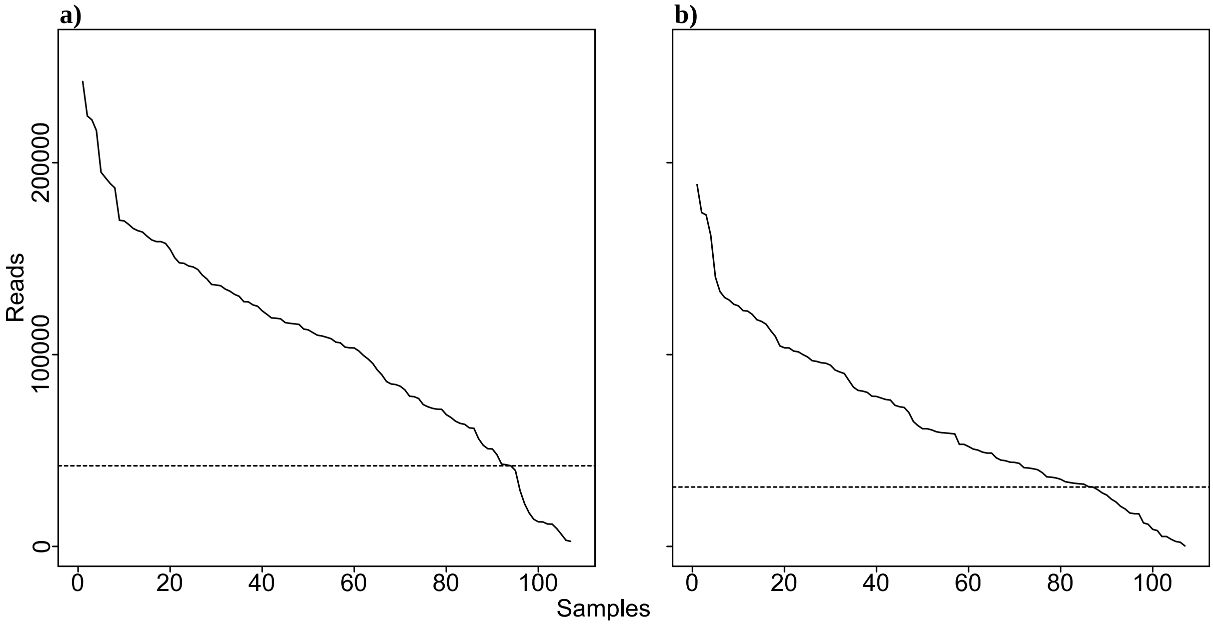

Figure 2 Reads per sample. Reads per sample after filtering for a) the full dataset (rarefied to 42049 reads, 13 samples were discarded) and b) the dataset with OTU 1-3 removed (rarefied to 30984 reads, 21 samples were discarded). The rarefaction threshold is indicated with dashed lines.

2.4 Analysis

Analysis was done in R v3.3.3 (R Core Team 2017), except for detection of the core microbiota, which was done by exporting the filtered dataset back into Mothur and using the get.coremicrobiome function. Species diversity (Simpson's diversity index) was calculated with the estimate_richness and plot_richness functions in Phyloseq, and Mann-Whitney U-test and Kruskal-Wallis rank sum test was used to test for significant differences between the populations. Levene's test as implemented in the Rcmdr package (Fox 2005) was used for testing for significant differences in variance. The package Vegan v2.4.2 (Oksanen et al. 2017) was used to make OTU accumulation curves with the function specaccum and options "random" and 10000 permutations.

Bray-Curtis dissimilarities between the samples was calculated with the vegdist function in Vegan, and weighted- and unweighted UniFrac with the UniFrac function in Phyloseq. Vegan was used for statistical comparisons with ANOSIM (Analysis of similarities) and Adonis (Permutational Multivariate Analysis of Variance Using Distance Matrices) using 10000 permutations. Neighbor-joining with the function nj in the package Ape v4.1 (Paradis et al. 2004) was used for creating Newick tree of the samples that was subsequently plotted in MEGA v7.0.14 (Tamura et al. 2007). DESeq2 v1.14.1 (Anders and Huber 2010) was used to identify taxa with significantly different levels of abundance between populations with Wald testing and a significance threshold of $p < 0.05$. The function ordinate in phyloseq was used to make PCoA plots using the built in distance measurement calculations. Plots were made with the default plot function or ggplot2 v2.2.1 (Wickham 2017). See appendix 6 for the most essential R code used.

For comparison between birds of putatively high and low condition, body condition was calculated as residuals of a linear model of log transformed tarsus length and body mass. I then tested for a correlation between Simpson's diversity index and body condition using a linear model. The birds were divided into two groups defined as high condition (i.e. those over the regression line) and low condition (i.e. those under the regression line) for analysis with DESeq2 and Adonis. This analysis was only possible for Lago Salso birds, as we had insufficient data from the other populations. To investigate seasonal variation of the ejaculate microbiota I utilized samples collected in Badajoz. These samples were chosen because it was the only location with sufficient spread in sampling dates. The samples were divided into two groups; those sampled in March (between 17.03.2016 and 29.03.2016) and those sampled in April (between 17.04.2016 and 21.04.2016). The origin of interesting OTUs was investigated in the NCBI nucleotide collection (Ncbi Resource Coordinators 2016) using BLAST+ (Camacho et al. 2009).

2.5 Bacterial mock community and reference sample

In order to verify my laboratory methods, 6 replicates a bacterial mock community and a standardized fecal sample was included in the study. By comparing the results from these with the known bacterial mock community composition and the results from a previous study (Muinck et al. 2017) the bias and reproducibility of the laboratory protocols used in the present study could be investigated.

The bacterial mock community consisted of plasmids containing near full length 16S rRNA gene sequences from 33 species (Pinto and Raskin 2012). These 33 species have a distinct phylogenetic diversity, and the sequences cover a wide range of GC (guanine-cytosine)-content, known to affect the efficiency of PCR amplification (Polz and Cavanaugh 1998). Plasmids were mixed in at equal proportions, and approximately $2.5e^6$ molecules were used as template in each reaction. The 6 mock samples were analysed in a separate pipeline together with a subset of 6 samples from Muinck et al. (2017) that were amplified with in total 40 cycles PCR with 5Prime HotMaster Taq, but otherwise treated identically to those from the present study. The samples were processed with an identical pipeline to the ejaculate samples, but in addition the correct full-length sequences were extracted from the dataset, and the relative abundance to the other correct-full length calculated per sample.

The standardized fecal sample contained homogenized feces from a human infant, and was processed together with the ejaculate samples from the DNA isolation stage. Biases detected in these samples should therefore reflect the full laboratory protocol used. The 6 samples were compared with a subset of 6 samples from Muinck et al. (2017), who used a PowerSoil 96 well DNA isolation kit (MO BIO Laboratories Inc., Carlsbad, CA, USA) with identical bead beating as the present study, but without the lytic enzyme cocktail step. The 6 standardized fecal samples were treated in the same pipeline as outlined for the ejaculate samples.

3 Results

3.1 Community composition

3.1.1 Community members

The 107 ejaculate samples contained 11227481 reads (average 104930, standard deviation ±53838) after filtering. After rarefaction to 42049, each bird contained between 51 and 195 OTUs, totalling 1292 unique OTUs. The OTUs were classified to 36 phyla, 77 classes, 134 orders, 249 families and 290 genera (only 72% of reads were assigned to OTUs classified at the genus level). See appendix 7 for the full list of phyla detected, and appendix 8 for the most abundant order- and genus-level classifications. 173 of the OTUs were unclassified to the phylum level, but despite being taxonomically unknown, most of these still had > 97% identity matches to 16S sequences found in previous studies when compared against the NCBI nucleotide collection using BLAST. Among abundant OTUs with poor taxonomic resolution, the "most unknown" were OTU 9 (unclassified Proteobacteria) and OTU 55 (unclassified bacteria) with 92% and 93% identity BLAST hits respectively.

At the phylum level (appendix 7), the avian ejaculate microbiota is dominated by Bacteroidetes (41%). This is mostly due to *Flavobacterium* being the dominant genus (35%), but a wide range of other genera were also detected at lower abundance. Proteobacteria was less abundant (26%), but comprised far more diversity. Firmicutes (11%) and the candidate phylum Parcubacteria (6%) were also relatively abundant. Actinobacteria, Verrucomicrobia, the candidate phylum Gracilibacteria, Spirochaetae, and Microgenomates also contributed more than 1% of the dataset. Only negligible amounts of archaeal sequences were detected.

There was large variation in membership and relative abundances between individuals, both at the OTU-level, but also when comparing phyla (figure 3). A significant part of the between-individual variation was caused by OTUs that were unique to one bird, or were only shared with a few others. Often, these rare OTUs occurred in considerable abundance in the few samples they were present. Due to this pattern, each bird sampled introduces several novel OTUs to the dataset (figure 4). In addition to this, rarefaction curves did not fully saturate (figure 5) for all birds, and thus the sequencing effort of this study underrepresents the full community composition, both within individuals and within populations.

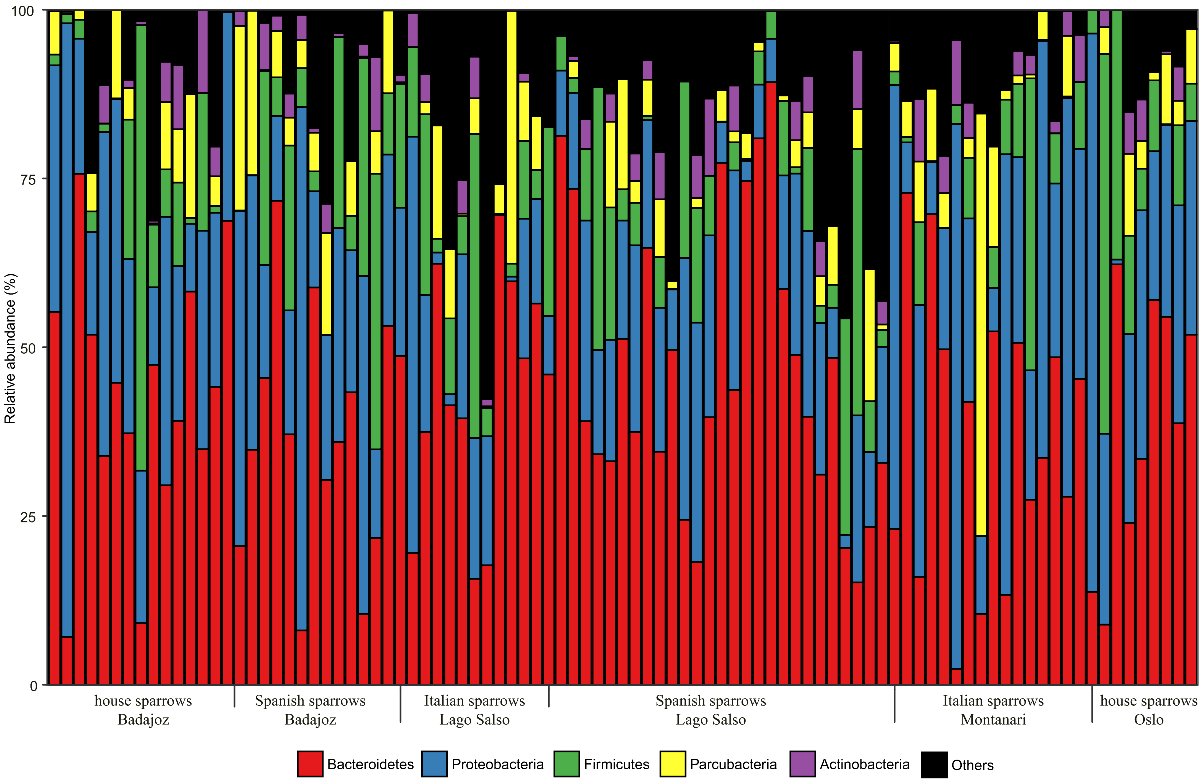

Figure 3 Phylum-level composition. Phylum-level composition of the sparrow ejaculate microbiota. Phyla less abundant than 3% are grouped into the ‘Others’ category.

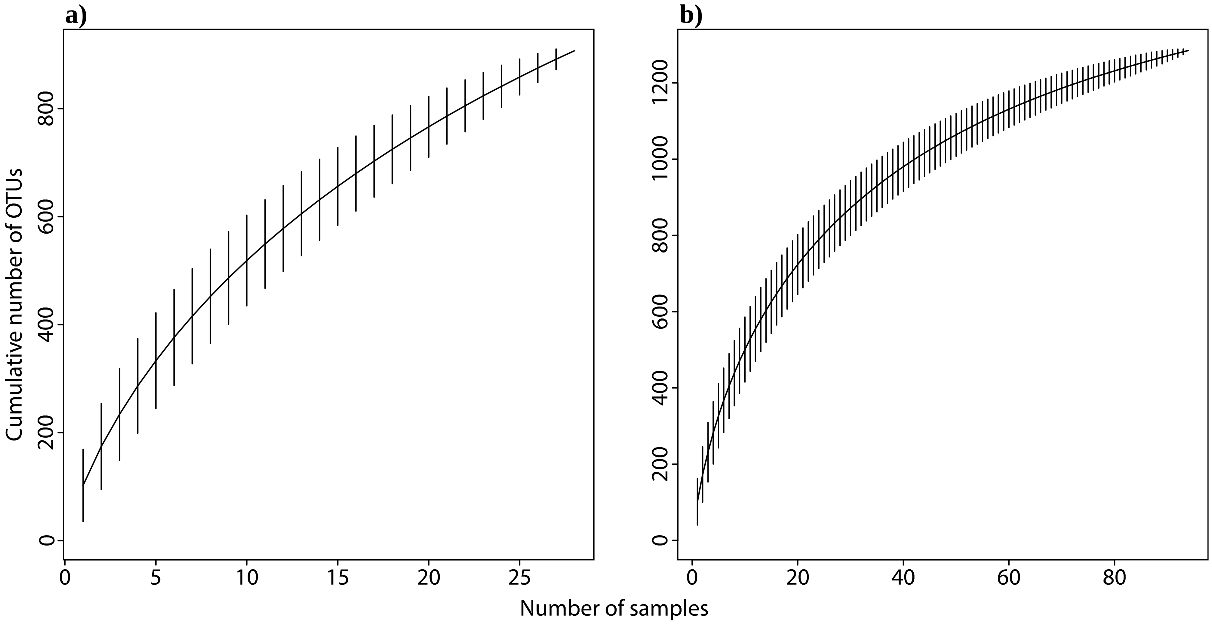

Figure 4 Species accumulation curves. Mean OTU richness per number of sparrows sampled for a) Spanish sparrows from Lago Salso (the largest population), and b) all individuals from all populations. The whiskers indicate 1 SD from 10000 permutations of random ordering of samples.

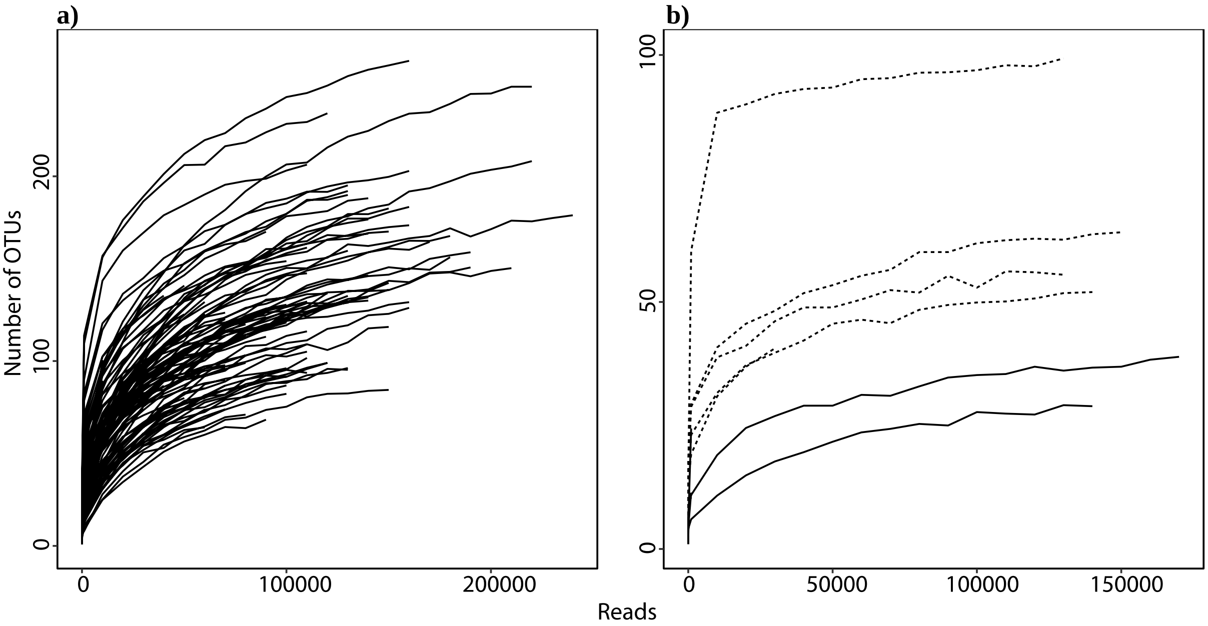

Figure 5 Rarefaction curves. Rarefaction curves for a) the ejaculate samples, and b) for PCR controls (full lines, n = 4) and sampling controls (dashed lines, n = 5).

Table 2 Genera known from culture-based studies. Abundance and commonness in the present study of genera known from previous culture-based studies on the avian cloaca or ejaculate. A genus was counted as present in a sample if it had a relative abundance of 0.1% or more. *OTUs classified to the family, but unresolved at the genus level were detected. These could not be further classified as several genera have identical 16S V4 regions.

Genus	Abundance avg. % (max. %)	Prevalence	Source and study
Acinetobacter	4.071 (45)	72	Ejaculate (Westneat and Rambo 2000)
Streptococcus	2.558(41)	28	Ejaculate (Hupton et al. 2003)
Campylobacter	1.824 (62)	9	Cloaca (Lombardo et al. 1996)
Staphylococcus	1.407 (18)	49	Ejaculate (Hupton et al. 2003)
Yersinia	0.909 (12)	32	Ejaculate (Lombardo and Thorpe 2000)
Pseudomonas	0.810 (20)	25	Ejaculate (Hupton et al. 2003)
Lactobacillus	0.549 (35)	14	Ejaculate (Lombardo and Thorpe 2000)
Micrococcus	0.218 (8)	14	Ejaculate (Hupton et al. 2003)
Escherichia/Shigella	0.068(5)	3	Ejaculate (Lombardo and Thorpe 2000)
Aeromonas	0.010(1)	1	Ejaculate (Hupton et al. 2003)
Bacillus	0*	0	Ejaculate (Hupton et al. 2003)
Chlamydia	0	0	Cloaca (Poiani and Gwozdz 2002)
Enterobacter	0*	0	Ejaculate (Hupton et al. 2003)
Enterococcus	0	0	Ejaculate (Hupton et al. 2003)
Ewingella	0*	0	Ejaculate (Westneat and Rambo 2000)
Gardnerella	0	0	Ejaculate (Hupton et al. 2003)
Listeria	0	0	Ejaculate (Hupton et al. 2003)
Salmonella	0*	0	Cloaca (Stewart and Rambo 2000)
Serratia	0*	0	Ejaculate (Hupton et al. 2003)
Vibrio	0	0	Ejaculate (Lombardo and Thorpe 2000)

Table 3 The most abundant OTUs. Classification and abundance of the 20 most abundant OTUs after rarefaction.

OTU	Abundance	Phylum	Order (family)	Genus
Otu0001	26.06 %	Bacteroidetes	Flavobacteriales	*Flavobacterium*
Otu0002	6.12 %	Bacteroidetes	Flavobacteriales	*Flavobacterium*
Otu0003	3.22 %	Proteobacteria	Pseudomonadales	*Acinetobacter*
Otu0004	2.34 %	Proteobacteria	Campylobacterales	*Helicobacter*
Otu0005	1.95 %	Proteobacteria	Burkholderiales (Comamonadaceae)	
Otu0006	1.80 %	Proteobacteria	Campylobacterales	*Campylobacter*
Otu0007	1.64 %	Firmicutes	Lactobacillales	*Streptococcus*
Otu0008	1.41 %	Firmicutes	Bacillales	*Staphylococcus*
Otu0009	1.28 %	Proteobacteria	Proteobacteria_unclassified	
Otu0010	0.93 %	Proteobacteria	Oceanospirillales	*Halomonas*
Otu0011	0.93 %	Gracilibacteria	Gracilibacteria_unclassified	
Otu0012	0.96 %	Proteobacteria	Rhodobacterales (Rhodobacteraceae)	
Otu0013	1.17 %	Spirochaetae	Spirochaetales	*Borrelia*
Otu0014	0.91 %	Proteobacteria	Enterobacteriales	*Yersinia*
Otu0015	0.88 %	Parcubacteria	Parcubacteria_unclassified	
Otu0016	0.65 %	Proteobacteria	Burkholderiales	*Polynucleobacter*
Otu0017	0.68 %	Proteobacteria	Campylobacterales	*Helicobacter*
Otu0018	0.78 %	Firmicutes	Lactobacillales	*Streptococcus*
Otu0019	0.47 %	Bacteroidetes	Bacteroidales	*Bacteroides*
Otu0020	0.69 %	Proteobacteria	Pseudomonadales	*Acinetobacter*

3.1.2 Core microbiota

I defined the core microbiota as all OTUs present in all birds at 0.1% relative abundance (i.e. 42 reads after rarefaction) or more. As between-individual differences dominate the dataset, and the vast majority of OTUs were only present in a small number of samples (figure 6), no strong core microbiota can be found in the sparrow ejaculate. Only OTU 1 (*Flavobacterium*) was present in all 94 samples at 0.1% abundance or more, and few other OTUs were nearly as common. The OTUs present in more than 50% of the birds were OTU 2 (*Flavobacterium*, 88 samples), OTU 3 (*Acinetobacter*, 68 samples), OTU 5 (Comamonadaceae, 65 samples), OTU 8 (*Staphylococcus*, 48 samples), and OTU 16 (*Polynucleobacter,* 56 samples). In addition, OTU 7 (*Streptococcus*), OTU 10 (*Halomonas*), OTU 12 (*Rhodobacteraceae*), and OTU 14 (*Yersinia*) were present at approximately 33% of the samples. At the phylum level, Bacteroidetes, Proteobacteria, Firmicutes, Parcubacteria, and Actinobacteria were present in all samples. Verrucomicrobia, Planctomycetes, Spirochaetae, Cyanobacteria, Gracilibacteria, Candidate division SR1 and Saccharibacteria were present in half the samples or more.

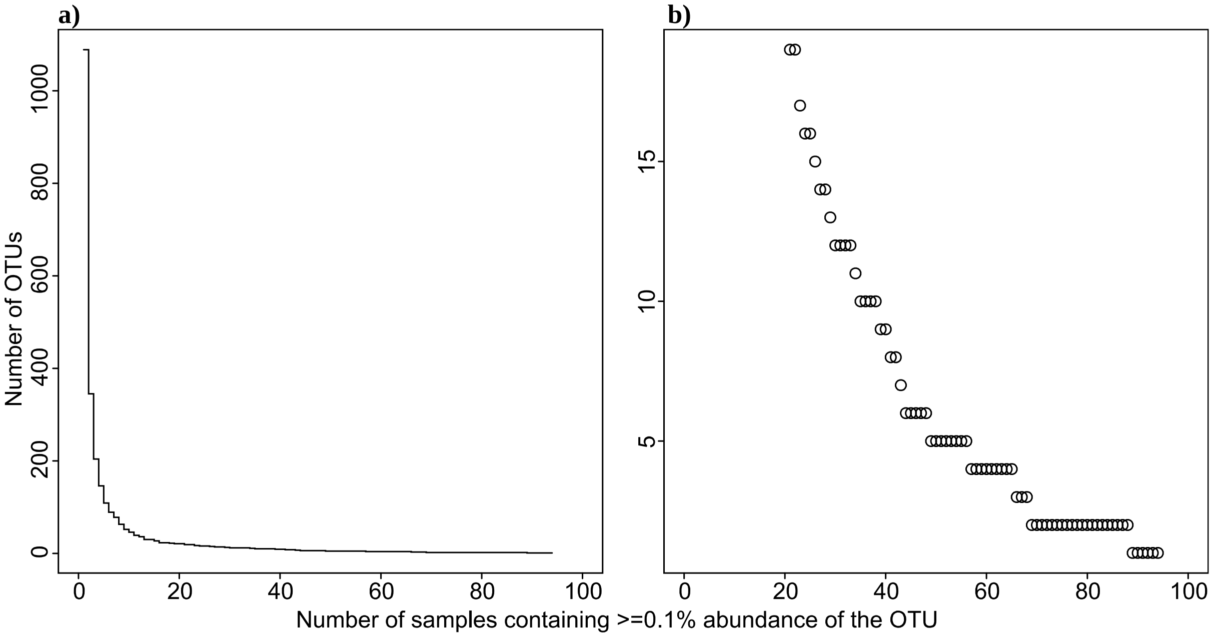

Figure 6 Commonness of OTUs. Number of OTUs comprising 0.1% or more of the microbiota for n number of samples in a) the full dataset, b) OTUs present in 20 samples or more (i.e. the right side of plot a)).

3.1.3 Potential contaminants

Two out of four PCR controls and 4/5 sample controls had enough reads to be considered positive (i.e. over the rarefaction threshold (figure 5). Two PCR controls (one from each plate) were among the 4 samples with the least reads, which might indicate that low levels of bacterial template DNA were present. The 3 controls discarded by rarefaction did however cluster together with their respective control types in a PCoA plot (not shown), although the clusters were weak, and when clustered with the ejaculate samples they did not form distinct groups.

A considerable proportion of OTUs were present in one or more control sample at some level of abundance. OTU 1 (*Flavobacterium*, 26% of all reads) and OTU 2 (*Flavobacterium*, 6% of all reads) were abundant in the ejaculate samples, disproportionally abundant in sample controls, but not present in PCR controls, making them potential DNA isolation associated contaminants. OTU 3 (*Acinetobacter*, 3% of all reads) was present in most sample control and ejaculate samples, but disproportionally abundant in the PCR controls. Thus, it is a possible library preparation contaminant. Several other OTUs were sporadically present in the controls (figure 7), but these were found only sporadically or at lower abundances than in ejaculate samples. There was little overlap between OTUs present in controls and the standardized fecal and mock community samples, which did not contain more of the 3 largest OTUs than what could be expected from index bleeding.

To test the effect of sample amount, which should correlate with bacterial biomass, I compared samples with 1 µL (n = 63) and 2 µL (n = 22) of collected ejaculate. Samples with more bacteria are predicted to have less contaminants (Salter et al. 2014; Weiss et al. 2014), but DESeq2 did not find any such pattern. Moreover, Adonis on Bray-Curtis dissimilarities did not find the groups to be significantly different ($R^2 = 0.01$, p = 0.74), and diversity was also similar (p = 0.14, Mann-Whitney U-test on Simpson's diversity index).

Analyses was rerun without the main suspected contaminant OTUs (1-3), but this caused no significant changes to the overall results and conclusions (data not shown).

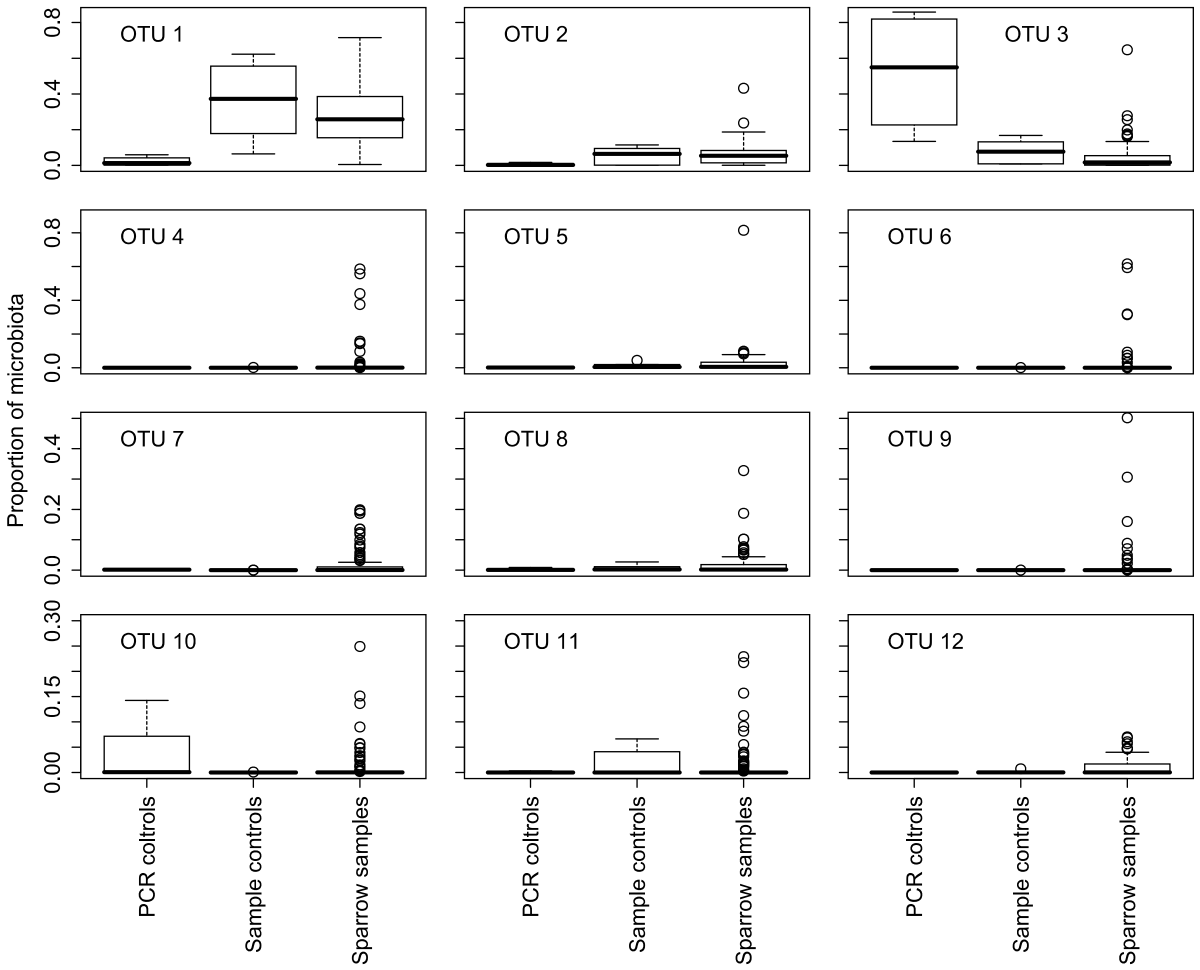

Figure 7 OTU abundance in samples and controls. Comparison of relative abundance of OTUs in different sample types for all samples in the unrarefied dataset. OTU 10 (*Halomonas*) appears to be an abundant contaminant in this figure, but was only present in 1 of the 4 PCR controls.

3.2 Community structure

3.2.1 Location and host species as community drivers

I found no support for species as a driver of the ejaculate-associated bacterial community (Adonis $R^2 = 0.02$, p = 0.82, ANOSIM R = 0.02, p = 0.66 on Bray-Curtis dissimilarity). When comparing the sympatric populations, ANOSIM found the two species in Badajoz to not differ (R = -0.02, p = 0.75), while two populations in Lago Salso were non-significantly different (R = 0.02, p = 0.37). In contrast, location appears to have a weak effect on the ejaculate microbiota. Adonis found location to have a stronger, but still non-significant, effect ($R^2 = 0.04$, p = 0.07), while ANOSIM found the effect to be significant (R = 0.09, p = 0.04). No Adonis pairwise comparisons of populations on Bray-Curtis dissimilarity were significantly different when multiple testing correction was applied (table 4), but both Adonis and ANOSIM trended towards the sympatric populations in Badajoz and Lago Salso being the most similar to each other, while Oslo was the most diverged from the other populations. Results with weighted UniFrac were similar.

Table 4 Pairwise comparison of populations. Adonis pairwise comparisons of populations using Bray-Curtis dissimilarity and 10000 permutations. Some p-values were below 0.05, but none of these were significant when sequential Bonferroni correction was applied.

	Badajoz		**Lago Salso**		**Montanari**
	Spanish	**house**	**Spanish**	**Italian**	**Italian**
Badajoz	$R^2 = 0.027$	-			
house	p = 0.900				
Lago Salso	$R^2 = 0.035$	$R^2 = 0.024$	-		
Spanish	p = 0.049	p = 0.041			
Lago Salso	$R^2 = 0.038$	$R^2 = 0.035$	$R^2 = 0.023$	-	
Italian	p = 0.478	p = 0.574	p = 0.617		
Montanari	$R^2 = 0.030$	$R^2 = 0.021$	$R^2 = 0.035$	$R^2 = 0.037$	-
Italian	p = 0.654	p = 0.984	p = 0.038	p = 0.400	
Oslo	$R^2 = 0.036$	$R^2 = 0.037$	$R^2 = 0.040$	$R^2 = 0.050$	$R^2 = 0.047$
house	p = 0.858	p = 0.694	p = 0.057	p = 0.384	p = 0.250

Due to the high between-individual variation, neither neighbor-joining (figure 8) nor PCoA (figure 9) produced any well defined clusters. Oslo house 13 and Badajoz house 08 cluster somewhat due to being dominated by *Streptococcus* and *Yersinia*. LGS Spanish 25, 28, and 31 cluster due to sharing *Borrelia* as a dominant community member. Badajoz house 02 and

Montanari Italian 12 and 16 clusters due to sharing *Campylobacter* as the dominant community member, composing 57, 59 and 59% of the microbiota respectively.

The large between-individual variation is largely caused by most birds containing several rare OTUs. For instance, one house sparrow from Badajoz was dominated by OTU 18 (*Streptococcus*) and OTU 114 (*Lachnoclostridium*), another by OTU 24 (*Bacteroides*) and OTU 96 (unclassified Microgenomates), while a Spanish sparrow at the same location was dominated by OTU 97 and OTU 191 (both unclassified Proteobacteria), OTU 7 (*Streptococcus),* and OTU 264 (*Ruminococcaceae*). One Italian sparrow from Montanari was dominated by OTU 3 and 20 (both *Acinetobacter*), another sampled the same day was dominated by OTU 45 and OTU 31 (both unclassified Parcubacteria), OTU 98 (*Bradyrhizobium*) and OTU 186 (*Leptotrichia*).

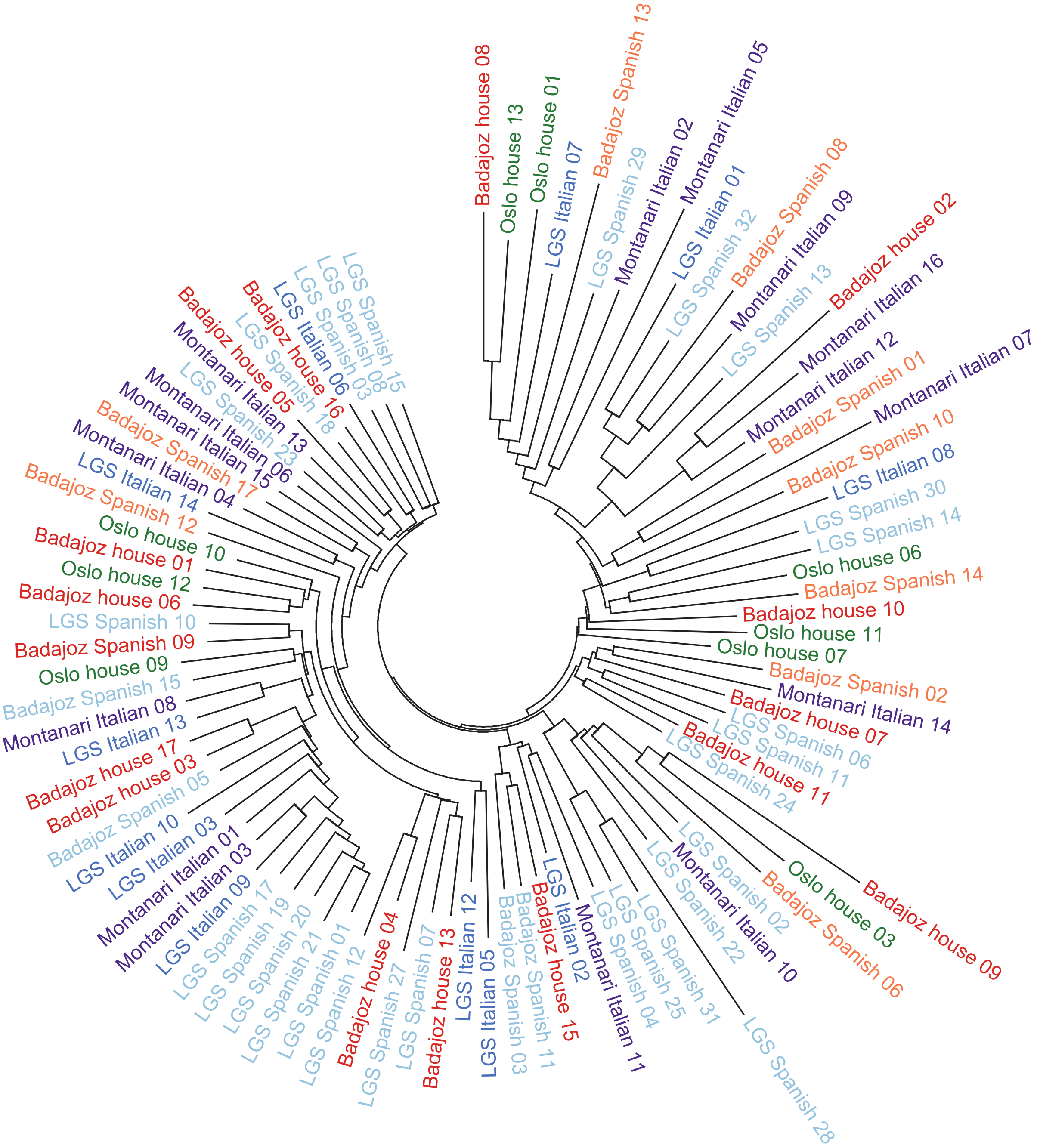

Figure 8 Neighbor-joining tree of samples. Neighbor-joining tree based on weighted UniFrac distances between all samples. The number at the end of each identifier represents the order in which the birds were sampled. LGS = Lago Salso.

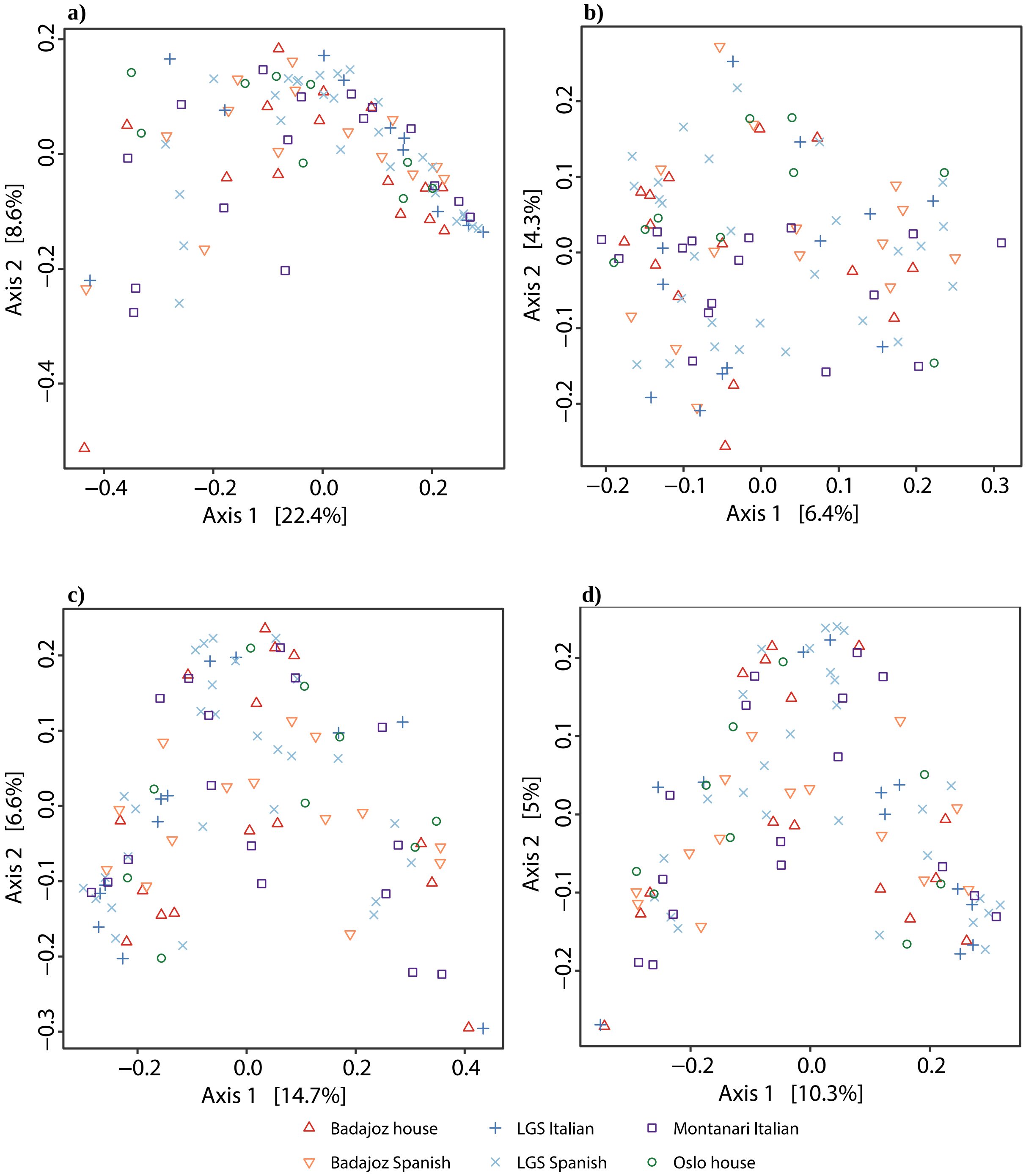

Figure 9 Ordination. PCoA plots of all samples, using a) weighted UniFrac distances, b) unweighted UniFrac distances, c) Bray-Curtis dissimilarity, d) Jaccard index. Eigenvalues are shown in brackets. LGS = Lago Salso.

3.2.2 Population-specific bacteria

DESeq2 was used to compare each location and species with all the others. No OTUs or phyla were associated with any of the species, but some location-specific patterns were apparent. OTU 13 (*Borrelia*) was only present in Spanish sparrows in Lago Salso, and OTU 4 (*Heliobacter*) was detected only in Lago Salso except for 1 bird from Badajoz, in which it was dominant. OTU 26 (*Catellicoccus*) was associated with the Italian locations, but was also present at low abundance in Oslo and Badajoz. Finally, OTU 6 (*Campylobacter*), OTU 9 (Unclassified Proteobacteria), and OTU 22 (*Lactococcus*) were present at all locations except Oslo, though I note that the sample size for Oslo is relatively small and thus suggest caution in drawing conclusions.

At the phylum level, Oslo was associated with low relative abundance of Candidate division SR1 and Gracilibacteria, and Lago Salso was associated with high relative abundance of Gracilibacteria, Bacteroidetes, Verrucomicrobia, and Spirochaetae. In contrast, Badajoz and Montanari were not associated with any specific phyla. When comparing the sympatric populations in Badajoz, no significant differences were found. In Lago Salso, Spirochaetae (*Borrelia* contributed near all abundance of this phylum) and Candidate division SR1 were significantly more abundant in the Spanish Sparrows.

3.2.3 Diversity analysis

Alpha diversity varied widely between individuals in the study, and I found no significant differences between populations (figure 10, p = 0.7 for Simpson's Diversity Index, p = 0.6 for observed OTU number, Kruskal-Wallis one-way analysis of variance). Most of the sparrows with the lowest Simpson’s diversity index were dominated by OTU 1, 2, and 3, although often in combination with significant amounts of other less common OTUs. However, not all birds with low diversity contained significant amounts of the top 3 OTUs. For example, one bird with Simpson’s diversity index 0.55 was dominated by 58% *Helicobacter* and 32% *Campylobacter*, while another individual with Simpson’s diversity index 0.68 was dominated by 44% *Yersinia* and 32% *Fluviicola*. The most diverse birds had between 30 and 35 OTUs composing 0.5% or more of the microbiota, but like in the low-diversity birds OTU 1 was commonly the most abundant. While some of the low-diversity birds were dominated by potential pathogens, these were also found in the high-diversity birds at lower abundances.

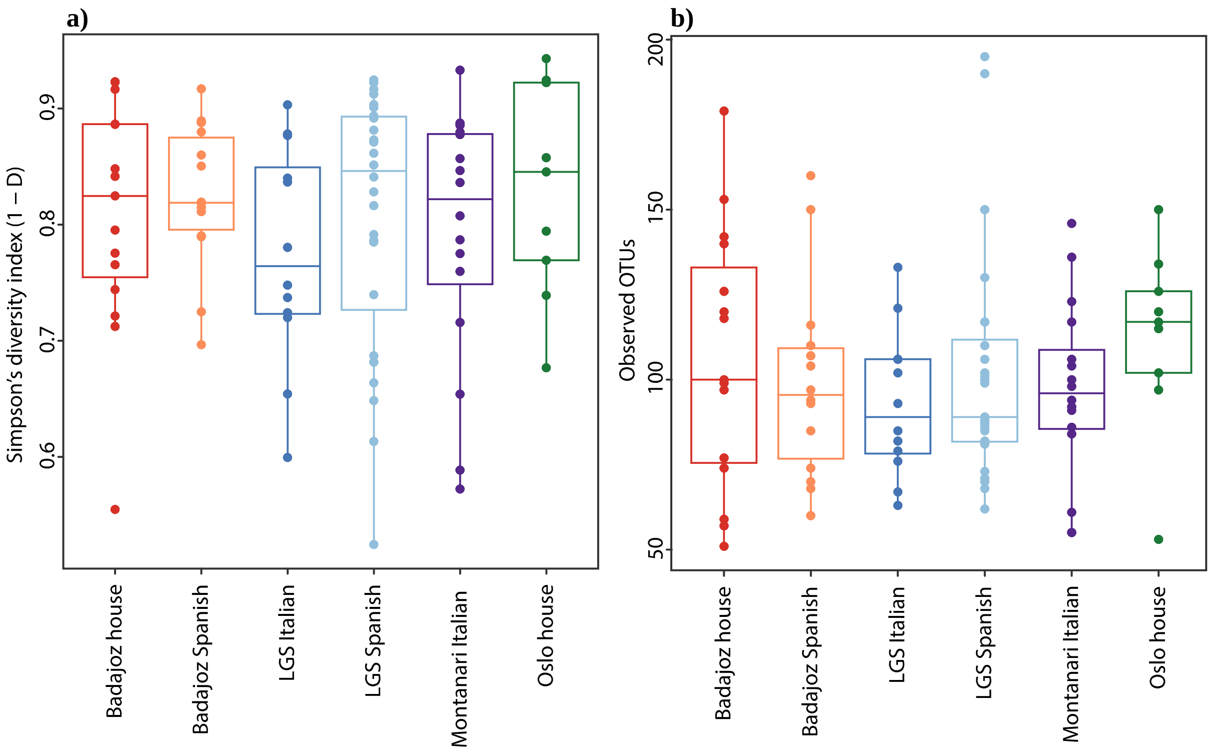

Figure 10 Alpha diversity. Boxplot comparing populations a) Simpson's Diversity Index (1 - D, i.e. higher values equals higher diversity) and b) number of OTUs observed.

3.3 Relationship with body condition

No significant differences were found between birds of high and low body condition. Simpson's diversity index was not correlated with body condition (linear model, t = -0.775, p = 0.45), and no OTUs were associated with either high or low condition birds. Moreover, Adonis on Bray-Curtis dissimilarities showed the groups not to be significantly different (R^2 = 0.04, p = 0.36). Analysis on birds with body condition of either extreme (1 SD or more from the regression line) gave similar results.

3.4 Temporal variation

I found no significant differences between birds sampled early versus late in the season. Specifically, no OTUs were associated birds sampled in either March or April, and Adonis on Bray-Curtis dissimilarity found no significant differences between the two groups (R^2 = 0.03, p = 0.79). Similarly, the groups did not significantly differ in Simpson's diversity index (p = 0.43, Mann-Whitney U-test). Interestingly, however, the majority of late season samples had similar alpha diversity values, but due to several outliers the variance in Simpson's diversity index was not significantly lower (Levene's test, p = 0.12).

3.5 Bacterial mock community

Most of the 33 sequences in the bacterial mock community deviated considerably from the expected relative abundance (3%), but deviations largely followed the same trends as in the samples from Muinck et al. (2017) (Spearman's rank correlation between the studies = 0.34). On average, the 33 sequences deviated from the expectation by 2.3% (compared to 1.3% in Muinck et al. (2017)). Amplification of 3 GC-rich sequences failed completely in all samples from the present study, including *Thermomicrobium roseum* (70% GC-content), *Thermotoga neapolitana* (64% GC-content), and Uncultured Gemmatimonadetes (63% GC-content). Four other GC-rich sequences only amplified a low number of samples. In contrast, sequences with low GC-content, most notably the least GC-rich sequence (uncultured cyanobacterium, 43% GC-content), were strongly overrepresented (figure 11). This bias against high and for low GC-content, is considerably stronger in the present study than in Muinck et al. (2017) (linear model: $R^2 = 0.59$ vs 0.16). Significantly higher spread in between-samples variation was also observed in the present study (figure 12, mean standard deviation: 1.57% vs 0.66%, $p < 0.0001$, Wilcoxon signed-rank test), and the mean distance between samples were considerably larger (mean Bray-Curtis dissimilarity: 0.31 vs 0.11, mean weighted UniFrac distance: 0.29 vs 0.11). One sample in the present study performed particularly poorly, as it failed to amplify 14 of the 33 sequences, but removal of this did not cause significant changes to the results.

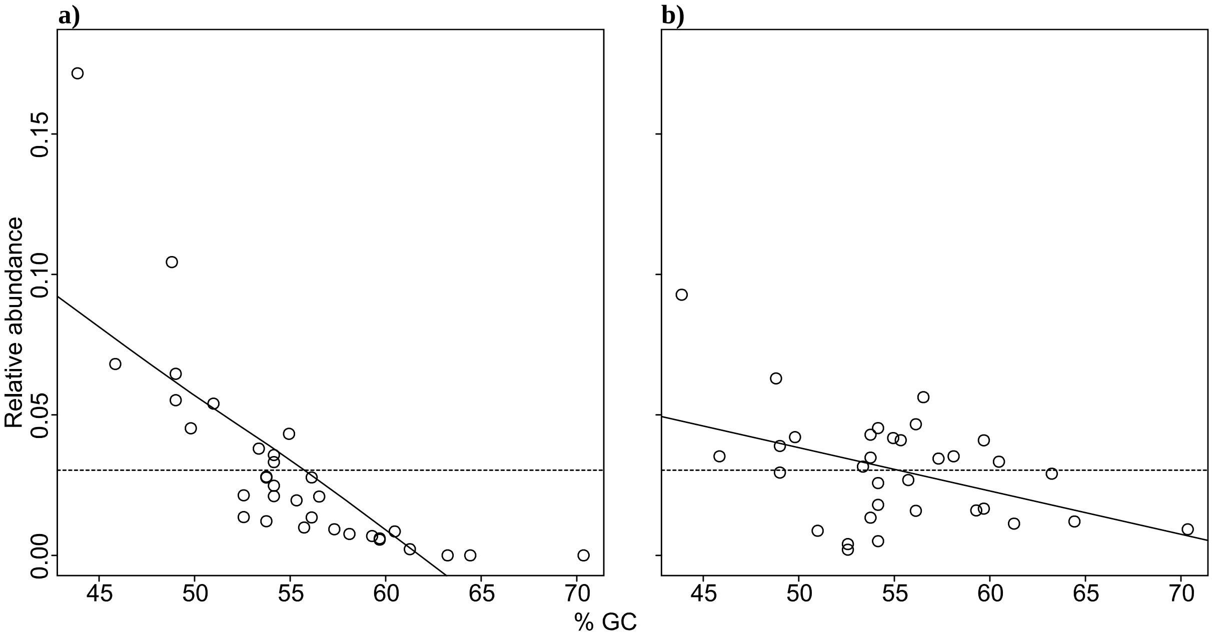

Figure 11 GC-content of mock sequences plotted vs relative abundance. Relative sequence abundance in the present study plotted vs the V4 region GC-content for a) the present study, and b) Muinck et al. (2017). Expected relative abundance (3%) is shown as a dashed line, the linear model as a full line.

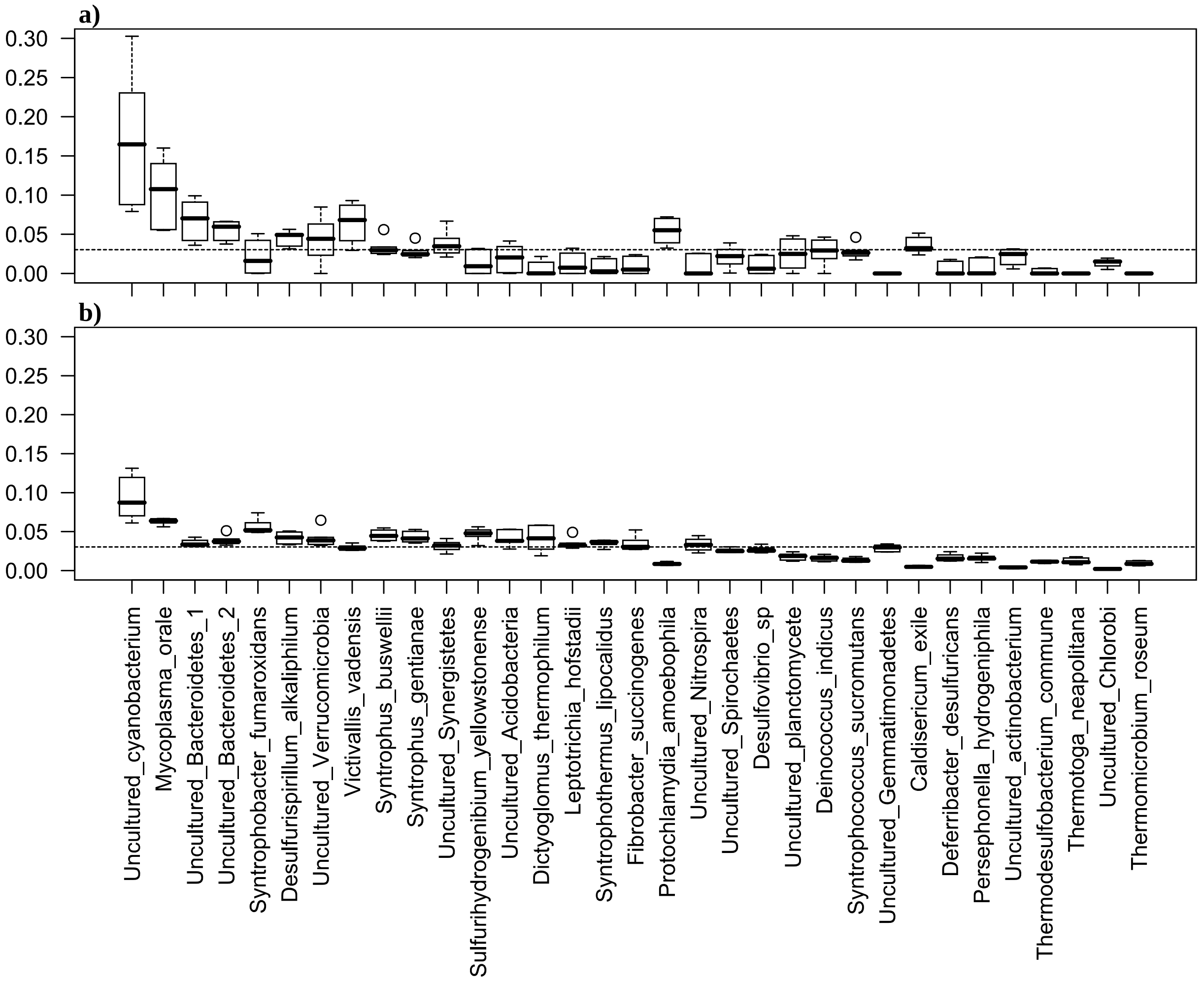

Figure 12 Relative abundances of the mock sequences. Proportion of sequences with 100% full length match to the correct mock community sequences in a) the present study (n = 6), and b) subset of 6 mock samples from Muinck et al. (2017). True community proportion is 3% (dashed line) for all sequences.

3.6 Standardized fecal sample

Dominant phyla in the standardized fecal samples from the present study were Firmicutes (74%), Euryarchaeota (16%), and Bacteroidetes (8%), while the samples from Muinck et al. (2017) were dominated by 55% Bacteroidetes and 35% Firmicutes. In both studies, most OTUs had intermediate GC-content (i.e. 49-54%), suggesting that PCR biases did not have a large impact. Most notable at the OTU level, OTU 1 (*Prevotella*) was dominant in Muinck et al. (2017), but of medium abundance in the present study (44% vs 6% of reads), while OTU 2 (*Methanobrevibacter,* an archaeon), was vastly more abundant in the present study (16% vs 0.001% of reads). Interestingly, OTU 11 (*Bifidobacterium,* GC-content: 59%) classified to Actinobacteria (another gram-positive phylum), was highly abundant in one of the samples in the present study, but not present at all in the others (figure 13). DESeq2 found gram-positives Firmicutes and Tenericutes, as well as Euryarchaeota, to be significantly associated with the present study, while gram-negatives Bacteroidetes and Proteobacteria were more abundant in Muinck et al. (2017).

While significantly fewer OTUs were detected in the present study (mean: 181 vs 197, p = 0.03, Mann-Whitney U-test), Simpson's diversity index was significantly higher (mean: 0.95 vs 0.87, p < 0.01, Mann-Whitney U-test). The mean distance between samples was considerably higher in the present study (Bray-Curtis dissimilarity: 0.26 vs 0.16, weighted UniFrac distances: 0.17 vs 0.10). The average relative OTU abundance correlated somewhat better between the studies than what was found for the mock community (Spearman's rank correlation on the 33 most abundant OTUs = 0.52).

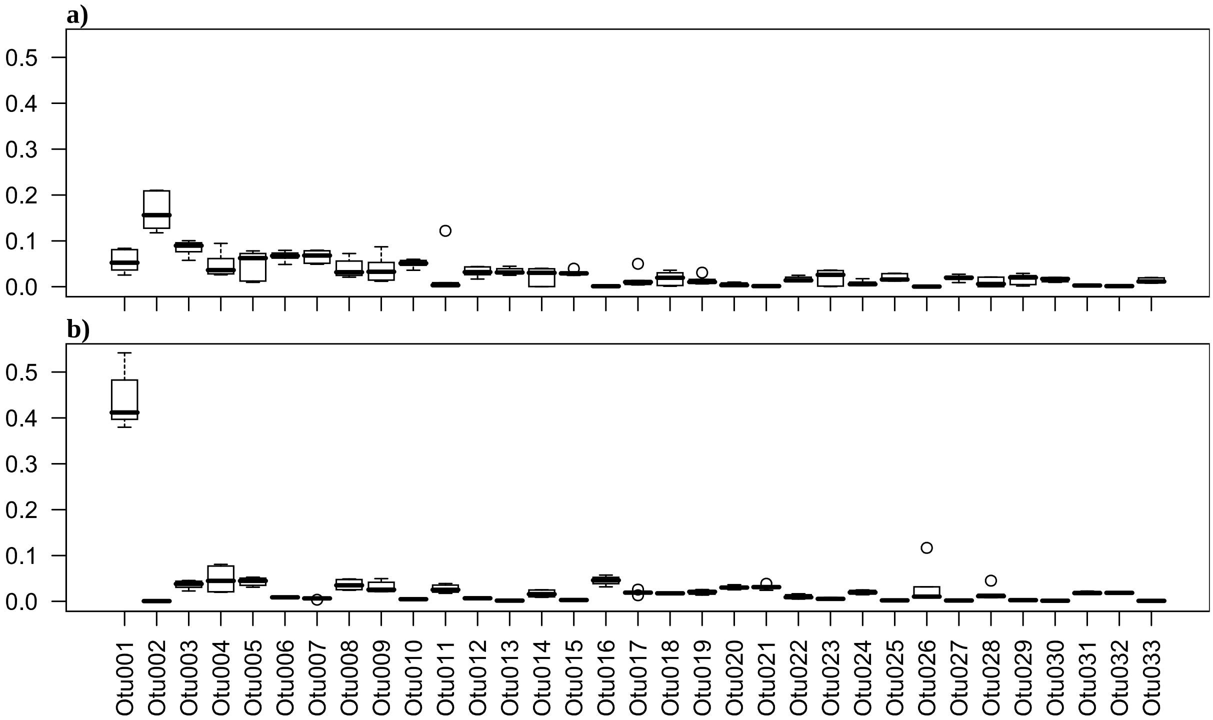

Figure 13 Relative OTU abundances in the standardized fecal samples. Relative abundances of the 33 most abundant OTUs in the standardized fecal samples in a) the present study (n = 6), and b) a subset of 6 samples from Muinck et al. (2017). DESeq2 found 23 of these 33 OTUs to significantly (p < 0.05) differ between the studies.

4 Discussion

Using 16S amplicon sequencing, I characterized the composition and structure of the avian ejaculate microbiota and investigated the role of host species and location as community drivers. Examination of the ejaculate microbiota from three species of *Passer* sparrows showed these communities to be highly diverse with large between individual variability in both bacterial composition and diversity, which is consistent with studies of the avian fecal and cloacal microbiota. This work identified a number of bacterial taxa that have previously been identified in culture-based studies, as well as a range of taxa that have not previously been reported. Notably, I detected a large variety of OTUs classified to candidate phyla, of which little is known about function and ecology. Importantly, the observed high between-individual variation appear to mask potential differences between the species. In contrast, there was a small effect of location on community composition, but not diversity. Finally, I found no associations with neither body condition nor temporal variation, suggesting that factors other than those investigated here are important in explaining variation in avian ejaculate bacterial communities.

4.1 The avian ejaculate-associated microbiota

I found Bacteroidetes, Proteobacteria, and Firmicutes to be the three most abundant phyla, but a wide range of other phyla also contributed significant abundance and diversity. These phyla are the same as found in studies on the human ejaculate (Hou et al. 2013; Weng et al. 2014; Mändar et al. 2015) and the seminal fluid of mice (Javurek et al. 2016; Javurek et al. in press). There was also strong overlap in community members at the genus level, as the human ejaculate microbiota has been found to be dominated by *Prevotella*, *Lactobacillus* and *Pseudomonas* (Hou et al. 2013; Weng et al. 2014), which is consistent with findings in the present study. Likewise, the most abundant genera identified in this study, such as *Flavobacterium*, *Helicobacter*, *Streptococcus* and *Campylobacter* have also been detected in humans (Hou et al. 2013).

Studies on the microbiota of the avian gastrointestinal tract have largely found Firmicutes and Proteobacteria to be the dominant phyla, with Actinobacteria also being abundant. Bacteroidetes is found in most gastrointestinal samples, but at low abundance (Banks et al. 2009; Xenoulis et al. 2010; Videnska et al. 2013; Mirón et al. 2014; Kreisinger et al. 2015;

Lewis et al. 2016). Two studies on passerine feces additionally found Fusobacteria (Ryu et al. 2014) and Tenericutes and Chlamydiae (Kropáčková et al. in press) at considerable abundance. While I also found these taxa in this study, the abundance was relatively low. At the lower taxonomic levels, abundant taxa shared with the present study include Enterobacteriaceae, *Campylobacter, Helicobacter, Staphylococcus, Bacteriodes, Lactococcus,* and *Lactobacillus* (Videnska et al. 2013; Lewis et al. 2016; Kropáčková et al. in press). Thus, my results mirror those reported for both the mammalian ejaculate and the avian gastrointestinal tract.

Nonetheless, significant differences were also observed. While a wide range of genera I detected in the sparrow ejaculate are also present in mammalian ejaculate and the avian gastrointestinal tract, relative abundances vary widely. Most notably, the sparrow ejaculate microbiota is dominated by *Flavobacterium*, which has only been found at low abundances elsewhere. At the phylum level, I found Bacteroidetes to compromise on average 42% the sparrow ejaculate microbiota, far more than what has been detected in previous studies on the avian gastrointestinal tract (Lewis et al. 2016; Kropáčková et al. in press), while Firmicutes is less abundant. *Flavobacterium* contributed most of the abundance of Bacteroidetes in this study, but a wide range of other Bacteroidetes genera such as *Bacteroides, Chryseobacterium, Prevotella, Pseudarcicella, Fluviicola,* and *Hymenobacter,* known to be associated with the avian gastrointestinal tract (Videnska et al. 2013; Kreisinger et al. 2015), were present in notable abundance. As there are large differences between studies on the avian gastrointestinal tract and human ejaculate, no doubt due to both methodology and biological differences, inferring more nuanced differences is difficult.

Proteobacteria was the most diverse phylum found in this study, and a range of the genera detected, such as *Campylobacter, Helicobacter,* and *Escherichia,* are considered pathogens both in humans, birds, and a range of other animals (Skirrow 1994; Mukhopadhya et al. 2012; Ryu et al. 2014). However, while Proteobacteria contains a range of bacteria known to be pathogenic, these are often found to be common and persistent community members in the avian gastrointestinal tract. In fact, Proteobacteria has been found to be the most abundant phylum in some studies on the avian gastrointestinal tract (Kreisinger et al. 2015; Lewis et al. 2016). In chickens, symptoms of experimental *Campylobacter* infection appear to be less severe than what is observed in humans (Black et al. 1988; Alemka et al. 2010; Waldenström et al. 2010). It has been suggested that the divergent chicken intestinal mucus in combination with other factors inhibits *Campylobacter* infection (Alemka et al. 2010). *Campylobacter* has even been suggested to have a commensal role in the avian gastrointestinal tract (Young et al. 2007). Thus, the role of Proteobacteria appears be different in birds than humans. This warrants caution in interpretation of the ecology and function of Proteobacteria, as well as other putative pathogens like *Streptococcus* and *Staphylococcus* in the avian ejaculate.

Notably, I found considerable abundance and diversity of OTUs classified to candidate phyla. Parcubacteria contributed 6% of reads and 14% of OTUs, while Gracilibacteria contributed 1% of reads and OTUs. Candidate divisions TM7 (Saccharibacteria) and SR1 among others were also detected, but at lower abundance. These phyla have largely not been reported in the studies on mammalian ejaculate or avian gastrointestinal tracts, except for low abundances of Parcubacteria being found in some species in Kropáčková et al. (in press). Although biases in DNA isolation and assignment of sequence classification might have prevented detection in some studies, this might suggest Parcubacteria and Gracilibacteria to be particularly associated with the avian reproductive tract.

Due to lacking a seminal vesicle (Birkhead and Møller 1992), which produces a range of substances amounting to approximately 70% of the human ejaculate (Aumüller and Riva 1992), the ejaculate microbiota of passerines could be expected to differ considerably from humans, mice, and most other animal taxa. Components excreted from the seminal vesicle, for instance fructose, can be metabolized by bacteria (Javurek et al. 2016), and might provide niches for bacterial growth. However, this potentially large difference in ejaculate composition appears not to have caused considerable differences. One possible reason for this is that while the bacterial community has been found to differ between the ejaculate and cloaca (Hupton et al. 2003), bacteria might not grow freely in the ejaculate itself or metabolize its content. Ejaculate contains anti-bacterial substances (Poiani 2006; Rowe et al. 2013) that might prevent this. It is possible that the majority of the microbiota associated with the ejaculate and male reproductive tract grows attached to the seminal duct epithelium or mucus, and are protected by extracellular structures. This surface is far greater than that of the part of the cloaca traversed by the ejaculate (Birkhead and Møller 1992), and thus might contribute a far greater proportion of the persistent microbiota in contact with the ejaculate. It should also be considered that cloacal flushes and swabs used in studies on the cloacal microbiota might be far more invasive and thorough than the passage of ejaculate through the cloaca, thus causing only a subset of the full cloacal microbiota to be incorporated. It is possible that the large variation observed in community composition and structure is largely caused by transient bacteria in the cloaca and the lower reproductive tract.

In the human ejaculate and vagina, *Lactobacillus* has been found to be dominant in most individuals, and is thought to be a probiotic (Weng et al. 2014; Yildirim et al. 2014). A suggested benefit of *Lactobacillus* in the vagina is to create an acidic environment which prevents colonization of pathogens (Yildirim et al. 2014). If *Lactobacillus* colonization of the reproductive tract is beneficial, selection should drive hosts to facilitate its presence, which should cause it to have the persistent ubiquitous and abundant pattern of distribution found. If growth is not facilitated in some way, it would be unlikely to be so persistently abundant due to the high between-individual variation observed in the remaining microbiota. *Lactobacillus* has also been found to be dominant in the chicken gut (Videnska et al. 2013), and has been suggested as an avian ejaculate probiotic (Lombardo et al. 1999), although more recently it has been found to severely retard sperm function in chickens (Haines et al. 2013). In this study I did not find *Lactobacillus* to share the abundant distribution pattern found in human ejaculate or the chicken gastrointestinal tract. Instead, both the abundant *Flavobacterium* OTUs fits the putative beneficial bacteria distribution pattern, and it is possible that it colonizes the male reproductive tract. It is not clear what benefits *Flavobacterium* colonization of the male reproductive tract could confer, but as it is ubiquitous, it does at the very least appear to be tolerated, and not a pathogen actively fought by the host immune system.

I found the community structure of the sparrow ejaculate microbiota to be dominated by between-individual variation, and this appears to mask potential community drivers. The ejaculate microbiota is also highly diverse within individuals. It is thought that high diversity characterizes stable communities (Lozupone et al. 2012), but as ejaculate is an transient fluid likely to collect bacteria from several separate communities before exiting the cloaca, such analysis might not be warranted here. Large between-individual variation in the ejaculate microbiota is however not unexpected, as it has also been found in previous studies on the avian cloaca and ejaculate (Poiani and Gwozdz 2002; Kreisinger et al. 2015). Large between-individual variation has also been found in the gastrointestinal microbiota of other passerines such as thrushes, catbirds (Lewis et al. 2016), and zebra finches (Benskin et al. 2010). The studies on the avian gastrointestinal microbiota did however find a stronger core microbiota than what I found in the ejaculate.

While my results suggest that host species do not drive the ejaculate microbiota, and that differences between individual populations were small and non-significant, I found location to have a small effect. Thus, some trends can be inferred. I found the birds from Oslo, the sampling location most diverged from the others both in geography and ecology, to have the most diverged ejaculate microbiota. The two sympatric populations at Badajoz, which have the most similar environment due to nesting side by side in a storks nest, were the most similar. The populations in Lago Salso are also sympatric, but they have separate nesting sites. Thus, their environment is less shared than in Badajoz. Bacteria experimentally introduced to feathers have been shown to colonize the avian cloaca (Kulkarni and Heeb 2007), and closer sympatry likely facilitates more transmission of ejaculate-associated bacteria both sexually and indirectly via the environment.

House and Spanish sparrows are thought to have diverged approximately 1 million years ago, while the Italian sparrow originated less than 10000 years ago (Sætre et al. 2012). This is not a long time in an evolutionary perspective, and it is evident that their similar way of life has not created divergent selective pressures strong enough to shape the ejaculate microbiota in different directions through co-evolution with bacteria or otherwise. As the primers used in this study do not allow exhaustive classification of the detected bacteria, it is however possible that some host-specific species or strains exist. Studies that have found host species to significantly drive bacterial communities have most often found this between hosts diverged at the genus- or family-level. For example, a study on the bat intestinal microbiota found species from different families to differ (Phillips et al. 2012). Similarly, the amphibian skin microbiota has been found to differ between host species at high taxonomic levels (salamanders, frogs and toads; Kueneman et al. 2014), but in a study comparing 3 different genera, differences were not significant (Vences et al. 2015). It should also be noted that studies investigating species-specificity most often do not investigate species living in sympatry, or have a far wider definition of sympatry than what is used in this study. Thus, it is not clear to what extent host-specific drivers shape the microbiota relative to environmental factors.

No birds sampled in this study were visually observed to be unhealthy, but infection from pathogenic bacteria such as *Campylobacter*, which was detected in several of the investigated birds, can cause decreased body mass without strong visible signs of sickness (Waldenström et al. 2010). A range of other possible pathogens were also present, but I did not find any correlations between body condition and the ejaculate microbiota. It is possible that most pathogens detected in the ejaculate are too transient to cause significant changes to body mass, or that the infected birds incur costs not affecting body mass in the short term. Notably, 3 of the 4 highest body condition birds have high abundance of *Borrelia*, which have been experimentally shown to not reduce body mass or fat storage (Olsen et al. 1996).

House sparrow pairs are known to have convergent bacterial communities (Stewart and Rambo 2000), and an experimental study has found sexual contact to increase the diversity of the kittiwake cloacal microbiota (White et al. 2010). In the populations investigated in the present study, a range of novel bacteria should be introduced to the birds during each extra-pair copulation due to the high between-individual variation. Thus, the ejaculate microbiota might be expected to converge in composition and structure over time as more mating occurs. However, there does not appear to be a significant change in community composition in the ejaculate microbiota associated with temporal variation. I did however find the birds sampled late in the season to trend towards having less between-individual variation in diversity, but this trend was not significant. Thus, it is possible that while the mean community composition remains the same, sexual transmission or other factors stabilizes the community structure over time.

4.2 Putative STDs

As ejaculate is transferred to the female reproductive tract, any ejaculate-associated bacteria will be transferred to the female. To be a STD, the bacteria must also be 1) able to colonize the female, and 2) be pathogenic. It is likely that some bacteria detected in the present study were not alive when incorporated into the ejaculate, e.g. if they originated from digested food or otherwise upstream in the gastrointestinal tract. Additionally, some bacteria were likely transient, i.e. not able to competitively grow in the male or female cloaca or reproductive tract, and thus not classifiable as STDs. Some bacteria might be able to colonize males, but not females. These might however still be detrimental to the male due to an increased female immune response to the ejaculate killing more sperm than usual (Poiani 2010), or by directly harming sperm cells. Five of the bacteria detected in this study have been found to be detrimental to the sperm of domestic chicken (Haines et al. 2013) and turkey (Triplett et al. 2016). These include *Bifidobacterium*, *Campylobacter*, *Clostridium*, *Escherichia*, and *Lactobacillus*.

In general, STDs are thought to have low virulence (Lombardo 1998; Poiani 2010). Bacteria are not sexually transmitted outside the mating season, and thus STDs are dependent on avoiding the host immune system for long periods (Lombardo 1998). It has been suggested that the fitness of STMs are linked to that of the host, as a decrease in host health might decrease attractiveness to potential mating partners, leading to less sexual transmission (Poiani 2010). It is also thought that an ability to live intracellularly is an advantage (Poiani 2010). Most potential STDs found in this study are able to live intracellularly, for instance *Chlamydia* (Stephens et al. 1998), *Borrelia* (Ma et al. 1991), *Helicobacter* (Tang et al. 2012), *Mycoplasma* (Dallo and Baseman 2000), and *Campylobacter* (Watson and Galán 2008). It is possible that some potential STDs found in low abundance in this study, for instance *Mycoplasma*, are important in the populations investigated in this study, but were not in a stage of abundant colonization in any of the sampled birds at the time of sampling.

The mammalian immune system is compartmentalized, enabling the same bacteria to be fought and facilitated at different locations in the same animal (Macpherson and Uhr 2004; Belkaid and Naik 2013), and this is likely also the case in birds. As many putative pathogenic genera found in this study are also putative commensals in the avian gastrointestinal tract, it is possible that these are opportunistic pathogens in the reproductive tract, not specialized STDs. In humans, urinal tract infections are thought to transmit both sexually and via fecal contamination (Foxman 2010), and most of the common urinal tract infections in humans, such as Enterobacteriaceae, *Streptococcus*, *Staphylococcus*, and *Pseudomonas* (Foxman 2010), are among the most abundant taxa found both in Passerine feces (Lewis et al. 2016) and the present study. In humans, these infections are largely transient, but recurring, and strains transferred to the urinal tract from the gastrointestinal microbiota of the same individual are rapidly cleared (Foxman 2010). In this study, I observe these taxa to be more common and less abundant than the putatively more virulent pathogens *Campylobacter* and *Helicobacter*. Thus, it is possible their role is similar in humans and birds.

If a bacterium colonizing the reproductive tract is a virulent pathogen, the host immune system should attack it, meaning it is unlikely to be common and abundant across time and populations. It is however possible that they are virulent and highly abundant infections in the short term. Notably among bacteria found in this study, *Campylobacter, Helicobacter*, and *Borrelia* were rare, but highly abundant where present. This might indicate the birds containing these genera were severely infected at the time of sampling. *Campylobacter* infections have been found to last on average only one week in the avian gut (Waldenström et al. 2010), while *Borrelia* infection has been found to be present after 4 weeks in feces, but far longer in internal organs (Olsen et al. 1996). Thus, it is possible that while most birds are susceptible to these pathogens, and many are persistently infected at low levels, only these few birds were infected at sufficient levels to be detected at the time of sampling. *Campylobacter* is perhaps the strongest candidate for being an avian STD, as it can retard sperm motility (Haines et al. 2013), decrease body mass (Waldenström et al. 2010), and is commonly detected in ejaculates (Sheldon 1993; Poiani 2010). *Helicobacter* is a known pathogen in a range of animals (Skirrow 1994), and has been found in the avian gut (Ryu et al. 2014), but has not previously been associated with the ejaculate microbiota. Similarly, *Borrelia* is a known avian pathogen (Anderson 1988; Olsen et al. 1996) not previously known to be ejaculate-associated. It has however been suggested to be a STD in humans (Middelveen et al. 2015). As these three genera stand out as well-known pathogens, and were detected in abundance with a distribution pattern separate from most other abundant bacteria, they are strong STD candidates.

The candidate phylum Parcubacteria has been found to have severely reduced metabolic function, and likely depends on symbiotic or parasitic interactions with other organisms (Nelson and Stegen 2015). In the present study, Parcubacteria is also of special interest due to its unique pattern of distribution, as it is represented by a large number of low-abundance OTUs. At the phylum level, I found Parcubacteria to be a part of the core microbiota of the sparrow ejaculate, and thus it might have considerable ecological and evolutionary significance. It is possible that they do not interact directly with the host, but with other members of the microbial community instead. As most Parcubacteria OTUs have strong BLAST hits to sequences found in environments with no avian or other animal hosts, this is perhaps more likely. One Parcubacteria species has been found to live intracellularly in an alveolate (Gong et al. 2014), whose distant relatives were detected in this study (see appendix 5). Due to the large diversity of Parcubacteria detected, it is conceivable that it contains both commensals and pathogens colonizing the male reproductive tract, and that some or all of these are sexually transmitted.

4.3 Methodological considerations

My results come with a range of caveats in line with what is expected when characterizing a potential low-biomass microbiota (Salter et al. 2014; Weiss et al. 2014; Glassing et al. 2016). While not biologically unexpected, the discrete pattern of OTU abundance between samples is similar to the variability caused by PCR bias observed in the standardized fecal- and mock community samples. The 10 additional PCR cycles, and possibly also the different polymerase, employed in this study compared to the original protocol from Muinck et al. (2017) likely contribute to the significantly larger amount of between-sample variation, similarly to what has been found in other studies (Gohl et al. 2016). The biases observed in the mock community might be even larger in the ejaculate samples if DNA template amounts are low (Kennedy et al. 2014). The DNA template amount is however unknown due to an unknown ratio of host to bacterial DNA in measured samples.

GC-bias strongly affects several mock community sequences, and the most GC-extreme sequences completely failed to amplify in many samples. This suggests my methods were not able to detect the full bacterial community present in the avian ejaculate. As the mock community is designed to present a challenge to PCR, and most bacteria found in the environmental samples in this study have intermediate levels of GC-content, biases might however not be as severe in the environmental samples. I found the between-sample variation within the fecal samples, where most sequences have intermediate GC-content, to be smaller than for the mock community, despite these samples having undergone separate DNA isolation in addition to the library preparation step. Thus, most bacteria in the environmental samples were likely not greatly affected by the GC-bias.

My DNA isolation method appears to shape the microbiota stronger than the library preparation protocol. The largest OTU in the standardized fecal sample was on average more than 16000 times more abundant in the present study than in samples from Muinck et al. (2017), suggesting that the PowerSoil 96 well DNA isolation kit utilized in that study was unable to effectively lyse the archaeon cell wall. As archaea feature divergent cell walls from bacteria, they can be resistant to lysing enzymes intended for bacteria, such as lysozyme (Gottlieb et al. 2016). However, pinpointing what caused the difference is not possible due to the contents of the PowerSoil 96 well DNA isolation kit being proprietary. Overall, I found gram-positive bacteria to be more abundant in the present study than in Muinck et al. (2017), likely due to the inclusion of the enzymatic lytic cocktail used the DNA isolation protocol. Increased relative abundance of gram-positives is in line with the original description of the protocol, suggesting that it increases lysis efficiency of gram-positive bacteria, thereby increasing the accuracy of the representation of the true bacterial community (Yuan et al. 2012). Despite this, I found gram-negative phyla to dominate the avian ejaculate microbiota, suggesting that these were in fact truly most abundant.

As contaminant DNA is ubiquitous in laboratory reagents (Salter et al. 2014), and it is possible that the amount of bacterial DNA in sparrow ejaculate is consistent with the contaminant-prone category of "low-biomass" samples, it is plausible that several OTUs found in this study were contaminants. I identified several suspects, but no particular OTU can be ascertained as contaminant with good confidence. Thus, I did not remove any OTUs from my dataset. Conveniently, removal of the suspected contaminants would not have significantly changed the results. I do however suggest caution in the interpretation of my results. Some of the bacteria shared between this and previous avian microbiota studies, for instance *Acinetobacter*, *Streptococcus, Pseudomonas,* and *Micrococcus*, are also known lab contaminants (Laurence et al. 2014; Salter et al. 2014). It is possible that different species or strains are associated with birds and laboratories, but the primers utilized by this and most other microbiota studies do not allow classification to this level.

All samples should be expected to have contained approximately equal amounts of contaminant template DNA, as reagents were homogenized before use. Thus, any reagent contaminants should in theory be ubiquitous in the sequenced samples. Ejaculate samples contain bacteria from the 'true microbiota' in addition to any reagent contaminants, and these should thus also be expected to be less abundant ejaculate samples than controls. While a few OTUs follow these patterns, the vast majority of potential contaminants were only present in one or two controls, and often only in one or a few ejaculate samples. Often, the contaminant candidate is also more abundant in the ejaculate samples than in any control. Low repeatability due to PCR bias and minute template amounts might explain some of this pattern, but as this study featured in total 9 controls, contaminants of relevance should have been evident. Some taxa present in controls, for instance unculturable Gracilibacteria and Parcubacteria (Brown et al. 2015), are improbable lab reagent colonizers. It is possible that these bacteria were introduced to controls from pipetting spillage from other samples, or contamination from the air during the library preparation. Handling of amplified samples without barcodes might have caused trace contamination into neighbouring wells, thereby introducing sample DNA to the controls.

The 3 most abundant OTUs (*Flavobacterium* and *Acinetobacter*) stand out as consistently being more abundant in controls than samples. Thus, they appear as more plausible contaminants. Both *Flavobacterium* and *Acinetobacter* have been identified as lab contaminants by several studies (Laurence et al. 2014; Salter et al. 2014; Glassing et al. 2016), but *Acinetobacter* is known also from avian ejaculate (Poiani 2010), while

Flavobacterium has been found in avian intestines (Vancanneyt et al. 1994), and is associated with the human ejaculate (Mändar et al. 2015). However, a separate study on fairy-wren (*Malurus*) ejaculate (Rowe, unpublished data) using the same methods, overlapping reagent batches, and same sequencing run as this study, found *Flavobacterium* and *Acinetobacter* only at low abundance in a few samples, and not in controls. Thus, they appear unlikely to be contaminants. Regardless, the most important analysis steps were rerun with the 3 most abundant OTUs removed from the dataset, but this did not cause significant changes to the results. *Flavobacterium* presence in sparrow ejaculate will be confirmed with culturing before journal publication. Regardless of whether the OTUs found in controls are genuine bacteria from the ejaculate or lab-contaminants, it is evident that they have somehow been spread to samples where no DNA should have been present. This warrants scrutiny of distribution patterns observed, in particular the status of *Flavobacterium* as being ubiquitous.

5 Concluding remarks and further work

I found the avian ejaculate microbiota to be highly diverse, with a community structure dominated by large between-individual variation. I found neither species nor location to be strong drivers on the avian ejaculate microbiota, although locations were marginally different. Thus, the drivers shaping the avian ejaculate microbiota remains unknown. My findings confirm the presence of several ejaculate-associated taxa known from previous culture-based studies, but discover a far wider range of taxa previously unknown to occur in the avian ejaculate. Notably among novel findings, I detect a remarkably high diversity of OTUs classified the candidate phyla Parcubacteria, of which virtually nothing is known other than its existence. Moreover, I find known avian pathogens *Borrelia* and *Helicobacter* to be highly abundant, and I suggest these as novel avian putative STDs. In view of my findings, I suggest *Flavobacterium*, *Campylobacter, Helicobacter, Borrelia* and Parcubacteria as priorities for research on the avian ejaculate microbiota.

The inclusion of bead beating and a lytic enzyme cocktail in my DNA isolation protocol likely increased my ability to detect the full bacterial community, and to more accurately represent relative abundances of community members. Conversely, my PCR protocol caused significant bias obfuscating the true community structure, and cannot be recommended for use without further optimization. In future studies, replicate PCR reactions for each sample could improve repeatability. It should also be considered to use a different primer set, as the 16S V4 region provides poor resolution in Enterobacteriaceae, which contains a range of putative STDs. Together with several possible contaminants being identified, the PCR bias means my results should be interpreted with caution. Importantly, the inclusion or exclusion of possible contaminants does not significantly alter my results or conclusions. Nonetheless, it is possible that the dominant genus found in this study is a contaminant. It is unfortunately uncommon in the published literature to include information on repeatability or negative controls, and thus caution is warranted for most other microbiota studies as well.

To further advance knowledge on the avian ejaculate microbiota in the future, several approaches should be taken. Knowing the habitat of each community members is the first step to understanding their ecological roles. While I detect a wide range of bacteria in the ejaculate, I do not know if they originate from the male cloaca or reproductive tract, or if they colonize the ejaculate itself. Thus, inclusion of a small number of cloacal flush samples in this study could have increased the knowledge acquired considerably. In a more thorough approach, the gastrointestinal tract, testes, and seminal glomera should be sampled separately and tested for bacterial load and community composition. In order to investigate the role of intracellular bacteria such *Borrelia*, *Chlamydia*, and *Campylobacter*, sperm cells could be investigated separately from the ejaculate. This would however greatly increase any potential contaminant-related problems caused by low sample biomass. Metatranscriptomic analyses are needed to assess the function of the microbiota and its interaction with the host.

The ecology of the male reproductive tract can be further investigated by excising reproductive tract mucosal surface and examining it with electron microscopy, to see if a community of bacteria adheres to it. If this is the case in most birds, i.e. if the host immune system in healthy birds does not repel the bacteria, it follows that a community of commensal bacteria occur, and that not every microbe in the reproductive tract can be classified as an infection. Flavobacterium should be a particular focus of this line of investigation. To investigate the trend of increased and converging levels of diversity through the season observed in this study, one population should be sampled continuously through the full mating season.

No bacteria are confirmed as avian STDs to date, and experimental studies are needed to confirm STD status. Sexual transmission can be confirmed by inoculating males with marked bacterial strains and testing for the presence of these in the partner, while in parallel controlling for its transmission in pairs where mating is prevented. Confirming pathogenicity can be done by observing detrimental effects such as decreased body mass and reproduction, or increased mortality, in inoculated birds.

This study presents a significant advance in the knowledge on the avian ejaculate microbiota, and lays the groundwork for future research. While my results do not suggest the ejaculate microbiota to have played a particularly large role in the evolutionary history relating to speciation in this hybrid system, STDs and other bacteria might still induce strong selective pressures shared by the three species.

6 References

Able, D. J. (1996) The contagion indicator hypothesis for parasite-mediated sexual selection. *Proceedings of the National Academy of Sciences* **93**, 2229-2233.

Ait Belkacem, A., Gast, O., Stuckas, H., Canal, D., LoValvo, M., Giacalone, G., and Päckert, M. (2016) North African hybrid sparrows (*Passer domesticus*, *P. hispaniolensis*) back from oblivion–ecological segregation and asymmetric mitochondrial introgression between parental species. *Ecology and Evolution* **6**, 5190-5206.

Alemka, A., Whelan, S., Gough, R., Clyne, M., Gallagher, M. E., Carrington, S. D., and Bourke, B. (2010) Purified chicken intestinal mucin attenuates *Campylobacter jejuni* pathogenicity in vitro. *Journal of Medical Microbiology* **59**, 898-903.

Amato, K. R., Martinez-Mota, R., Righini, N., Raguet-Schofield, M., Corcione, F. P., Marini, E., Humphrey, G., Gogul, G., Gaffney, J., Lovelace, E., Williams, L., Luong, A., Dominguez-Bello, M. G., Stumpf, R. M., White, B., Nelson, K. E., Knight, R., and Leigh, S. R. (2016) Phylogenetic and ecological factors impact the gut microbiota of two Neotropical primate species. *Oecologia* **180**, 717-733.

Andere, C. I., Monteavaro, C., Palacio, M. A., Catena, M., Rodríguez, E. M., and Collins, A. M. (2011) *Apis mellifera* semen: bacterial contamination and susceptibility to antibiotics. *Apidologie* **42**, 551-559.

Anders, S., and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology* **11**, R106.

Anderson, J. F. (1988) Mammalian and avian reservoirs for *Borrelia burgdorferi*. *Annals of the New York Academy of Sciences* **539**, 180-191.

Aumüller, G., and Riva, A. (1992) Morphology and functions of the human seminal vesicle. *Andrologia* **24**, 183-196.

Banks, J. C., Cary, S. C., and Hogg, I. D. (2009) The phylogeography of Adelie penguin faecal flora. *Environmental Microbiology* **11**, 577-588.

Belden, L. K., and Harris, R. N. (2007) Infectious diseases in wildlife: the community ecology context. *Frontiers in Ecology and the Environment* **5**, 533-539.

Belkaid, Y., and Naik, S. (2013) Compartmentalized and systemic control of tissue immunity by commensals. *Nature Immunology* **14**.

Benskin, C. M., Rhodes, G., Pickup, R. W., Wilson, K., and Hartley, I. R. (2010) Diversity and temporal stability of bacterial communities in a model passerine bird, the zebra finch. *Molecular Ecology* **19**, 5531-5544.

Birkhead, T., Veiga, J., and Moller, A. (1994) Male sperm reserves and copulation behaviour in the house sparrow, *Passer domesticus*. *Proceedings of the Royal Society of London B: Biological Sciences* **256**, 247-251.

Birkhead, T. R., and Møller, A. P. 1992. *Sperm competition in birds: evolutionary causes and consequences*. Academic Press, London, UK.

Black, R. E., Levine, M. M., Clements, M. L., Hughes, T. P., and Blaser, M. J. (1988) Experimental *Campylobacter jejuni* infection in humans. *Journal of Infectious Diseases* **157**, 472-479.

Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Brown, C. R., and Brown, M. B. (2003) Testis size increases with colony size in cliff swallows. *Behavioral Ecology* **14**, 569-575.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., and Banfield, J. F. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208-211.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., and Bauer, M. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal* **6**, 1621-1624.

Coon, K. L., Brown, M. R., and Strand, M. R. (2016) Mosquitoes host communities of bacteria that are essential for development but vary greatly between local habitats. *Molecular Ecology* **25**, 5806-5826.

Dallo, S., and Baseman, J. (2000) Intracellular DNA replication and long-term survival of pathogenic mycoplasmas. *Microbial Pathogenesis* **29**, 301-309.

Diemer, T., Huwe, P., Ludwig, M., Hauck, E., and Weidner, W. (2003) Urogenital infection and sperm motility. *Andrologia* **35**, 283-287.

Elgvin, T. O., Hermansen, J. S., Fijarczyk, A., Bonnet, T., Borge, T., Saether, S. A., Voje, K. L., and Sætre, G. P. (2011) Hybrid speciation in sparrows II: a role for sex chromosomes? *Molecular Ecology* **20**, 3823-3837.

Forstmeier, W., Nakagawa, S., Griffith, S. C., and Kempenaers, B. (2014) Female extra-pair mating: adaptation or genetic constraint? *Trends in Ecology & Evolution* **29**, 456-464.

Fox, J. (2005) Getting started with the R commander: a basic-statistics graphical user interface to R. *Journal of Statistical Software* **14**, 1-42.

Foxman, B. (2010) The epidemiology of urinary tract infection. *Nature Reviews Urology* **7**, 653-660.

Franzenburg, S., Walter, J., Künzel, S., Wang, J., Baines, J. F., Bosch, T. C., and Fraune, S. (2013) Distinct antimicrobial peptide expression determines host species-specific bacterial associations. *Proceedings of the National Academy of Sciences* **110**, E3730-E3738.

Fung, M., Scott, K. C., Kent, C. K., and Klausner, J. D. (2007) Chlamydial and gonococcal reinfection among men: a systematic review of data to evaluate the need for retesting. *Sexually Transmitted Infections* **83**, 304-309.

Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B., and Chiodini, R. J. (2016) Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut pathogens* **8**, 24.

Gohl, D. M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T. J., Clayton, J. B., Johnson, T. J., and Hunter, R. (2016) Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology* **34**, 942-949.

Gong, J., Qing, Y., Guo, X., and Warren, A. (2014) "*Candidatus* Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). *Systematic and Applied Microbiology* **37**, 35-41.

Gottlieb, K., Wacher, V., Sliman, J., and Pimentel, M. (2016) Review article: inhibition of methanogenic archaea by statins as a targeted management strategy for constipation and related disorders. *Alimentary pharmacology & therapeutics* **43**, 197-212.

Griffith, S. C., Owens, I. P., and Thuman, K. A. (2002) Extra pair paternity in birds: a review of interspecific variation and adaptive function. *Molecular Ecology* **11**, 2195-2212.

Haines, M., Parker, H., McDaniel, C., and Kiess, A. (2013) Impact of 6 different intestinal bacteria on broiler breeder sperm motility in vitro. *Poultry science* **92**, 2174-2181.

Hamilton, W. D. (1990) Mate choice near or far. *American Zoologist* **30**, 341-352.

Hamilton, W. D., and Zuk, M. (1982) Heritable true fitness and bright birds: a role for parasites? *Science* **218**, 384-387.

Handelsman, J. (2004) Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685.

Hermansen, J. S., Saether, S. A., Elgvin, T. O., Borge, T., Hjelle, E., and Sætre, G. P. (2011) Hybrid speciation in sparrows I: phenotypic intermediacy, genetic admixture and barriers to gene flow. *Molecular Ecology* **20**, 3812-3822.

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

Hupton, G., Portocarrero, S., Newman, M., and Westneat, D. F. (2003) Bacteria in the reproductive tracts of red-winged blackbirds. *The Condor* **105**, 453-464.

Javurek, A. B., Spollen, W. G., Ali, A. M. M., Johnson, S. A., Lubahn, D. B., Bivens, N. J., Bromert, K. H., Ellersieck, M. R., Givan, S. A., and Rosenfeld, C. S. (2016) Discovery of a Novel Seminal Fluid Microbiome and Influence of Estrogen Receptor Alpha Genetic Status. *Scientific reports* **6**, 23027.

Javurek, A. B., Spollen, W. G., Johnson, S. A., Bivens, N. J., Bromert, K. H., Givan, S. A., and Rosenfeld, C. S. (in press) Consumption of a high-fat diet alters the seminal fluid and gut microbiomes in male mice. *Reproduction, Fertility and Development*.

Kennedy, K., Hall, M. W., Lynch, M. D., Moreno-Hagelsieb, G., and Neufeld, J. D. (2014) Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. *Applied and Environmental Microbiology* **80**, 5717-5722.

Kreisinger, J., Cizkova, D., Kropackova, L., and Albrecht, T. (2015) Cloacal microbiome structure in a long-distance migratory bird assessed using deep 16sRNA pyrosequencing. *PLoS One* **10**, e0137401.

Kropáčková, L., Těšický, M., Albrecht, T., Kubovčiak, J., Čížková, D., Tomášek, O., Martin, J. F., Bobek, L., Králová, T., and Procházka, P. (in press) Co-diversification of gastrointestinal microbiota and phylogeny in passerines is not explained by ecological divergence. *Molecular Ecology*.

Kueneman, J. G., Parfrey, L. W., Woodhams, D. C., Archer, H. M., Knight, R., and McKenzie, V. J. (2014) The amphibian skin-associated microbiome across species, space and life history stages. *Molecular Ecology* **23**, 1238-1250.

Kulkarni, S., and Heeb, P. (2007) Social and sexual behaviours aid transmission of bacteria in birds. *Behavioural Processes* **74**, 88-92.

Laurence, M., Hatzis, C., and Brash, D. E. (2014) Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* **9**, e97876.

Lewis, W. B., Moore, F. R., and Wang, S. (2016) Characterization of the gut microbiota of migratory passerines during stopover along the northern coast of the Gulf of Mexico. *Journal of Avian Biology* **47**, 659-668.

Li, J., Nasidze, I., Quinque, D., Li, M., Horz, H.-P., André, C., Garriga, R. M., Halbwax, M., Fischer, A., and Stoneking, M. (2013) The saliva microbiome of *Pan* and *Homo*. *BMC Microbiology* **13**, 204.

Lockhart, A. B., Thrall, P. H., and Antonovics, J. (1996) Sexually transmitted diseases in animals: ecological and evolutionary implications. *Biological Reviews* **71**, 415-471.

Lombardo, M. P. (1998) On the evolution of sexually transmitted diseases in birds. *Journal of Avian Biology*, 314-321.

Lombardo, M. P., and Thorpe, P. A. (2000) Microbes in tree swallow semen. *Journal of Wildlife Diseases* **36**, 460-468.

Lombardo, M. P., Thorpe, P. A., Cichewicz, R., Henshaw, M., Millard, C., Steen, C., and Zeller, T. (1996) Communities of cloacal bacteria in tree swallow families. *The Condor* **98**, 167-172.

Lombardo, M. P., Thorpe, P. A., and Power, H. W. (1999) The beneficial sexually transmitted microbe hypothesis of avian copulation. *Behavioral Ecology* **10**, 333-337.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012) Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220-230.

Lung, O., Kuo, L., and Wolfner, M. (2001) *Drosophila* males transfer antibacterial proteins from their accessory gland and ejaculatory duct to their mates. *Journal of Insect Physiology* **47**, 617-622.

Ma, Y., Sturrock, A., and Weis, J. J. (1991) Intracellular localization of *Borrelia burgdorferi* within human endothelial cells. *Infection and Immunity* **59**, 671-678.

Macpherson, A. J., and Uhr, T. (2004) Compartmentalization of the mucosal immune responses to commensal intestinal bacteria. *Annals of the New York Academy of Sciences* **1029**, 36-43.

Magoč, T., and Salzberg, S. L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963.

Marius-Jestin, V., Menec, M. L., Thibault, E., Moisan, J., L'Hospitalier, R., Alexandre, M., and Coiffard, J. M. (1987) Normal phallus flora of the gander. *Zoonoses and Public Health* **34**, 67-78.

Martín, L. O. M., Muñoz, E. C., De Cupere, F., Van Driessche, E., Echemendia-Blanco, D., Rodríguez, J. M. M., and Beeckmans, S. (2010) Bacterial contamination of boar semen affects the litter size. *Animal Reproduction Science* **120**, 95-104.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12.

McCord, A. I., Chapman, C. A., Weny, G., Tumukunde, A., Hyeroba, D., Klotz, K., Koblings, A. S., Mbora, D. N., Cregger, M., White, B. A., Leigh, S. R., and Goldberg, T. L. (2014) Fecal microbiomes of non-human primates in Western Uganda reveal species-specific communities largely resistant to habitat perturbation. *American Journal of Primatology* **76**, 347-354.

McKenzie, V. J., Bowers, R. M., Fierer, N., Knight, R., and Lauber, C. L. (2012) Co-habiting amphibian species harbor unique skin bacterial communities in wild populations. *The ISME journal* **6**, 588-596.

McMurdie, P. J., and Holmes, S. (2013) phyloseq: an R Package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217.

Middelveen, M. J., Burke, J., Sapi, E., Bandoski, C., Filush, K. R., Wang, Y., Franco, A., Timmaraju, A., Schlinger, H. A., and Mayne, P. J. (2015) Culture and identification of *Borrelia* spirochetes in human vaginal and seminal secretions. *F1000Research* **3**, 309.

Mirón, L., Mira, A., Rocha-Ramírez, V., Belda-Ferre, P., Cabrera-Rubio, R., Folch-Mallol, J., Cardénas-Vázquez, R., DeLuna, A., Hernández, A. L., and Schondube, J. (2014) Gut bacterial diversity of the house sparrow (*Passer domesticus*) inferred by 16S rRNA sequence analysis. *Metagenomics* **3**, 1-11.

Moeller, A. H., Peeters, M., Ndjango, J. B., Li, Y., Hahn, B. H., and Ochman, H. (2013) Sympatric chimpanzees and gorillas harbor convergent gut microbial communities. *Genome Research* **23**, 1715-1720.

Moller, A. P. (1991) Sperm competition, sperm depletion, paternal care, and relative testis size in birds. *The American Naturalist* **137**, 882-906.

Muinck, E. J. d., Trosvik, P., Gilfillan, G. D., and Sundaram, A. Y. (2017) A novel ultra high-throughput 16S rRNA amplicon sequencing library preparation method on the Illumina HiSeq platform. *bioRxiv*.

Mukhopadhya, I., Hansen, R., El-Omar, E. M., and Hold, G. L. (2012) IBD—what role do Proteobacteria play? *Nature Reviews Gastroenterology and Hepatology* **9**, 219-230.

Mändar, R., Punab, M., Borovkova, N., Lapp, E., Kiiker, R., Korrovits, P., Metspalu, A., Krjutškov, K., Nolvak, H., and Preem, J.-K. (2015) Complementary seminovaginal microbiome in couples. *Research in Microbiology* **166**, 440-447.

Ncbi Resource Coordinators. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **44**, D7-D19.

Nelson, W. C., and Stegen, J. C. (2015) The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Frontiers in microbiology* **6**, 713.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2017) *vegan: Community Ecology Package*. R package version 2.4-2. https://CRAN.R-project.org/package=vegan.

Olsen, B., Gylfe, Å., and Bergström, S. (1996) Canary finches *(Serinus canaria*) as an avian infection model for Lyme borreliosis. *Microbial Pathogenesis* **20**, 319-324.

Otti, O., McTighe, A. P., and Reinhardt, K. (2013) In vitro antimicrobial sperm protection by an ejaculate-like substance. *Functional Ecology* **27**, 219-226.

Pace, N. R. (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**, 734-740.

Pan, D., and Yu, Z. (2014) Intestinal microbiome of poultry and its interaction with host and diet. *Gut Microbes* **5**, 108-119.

Paradis, E., Claude, J., and Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289-290.

Partecke, J., and Schwabl, H. (2008) Organizational effects of maternal testosterone on reproductive behavior of adult house sparrows. *Developmental neurobiology* **68**, 1538-1548.

Pellati, D., Mylonakis, I., Bertoloni, G., Fiore, C., Andrisani, A., Ambrosini, G., and Armanini, D. (2008) Genital tract infections and infertility. *European Journal of Obstetrics & Gynecology and Reproductive Biology* **140**, 3-11.

Phillips, C. D., Phelan, G., Dowd, S. E., McDonough, M. M., Ferguson, A. W., Delton Hanson, J., Siles, L., Ordonez-Garza, N., San Francisco, M., and Baker, R. J. (2012) Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Molecular Ecology* **21**, 2617-2627.

Pinto, A. J., and Raskin, L. (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One* **7**, e43093.

Poiani, A. (2006) Complexity of seminal fluid: a review. *Behavioral Ecology and Sociobiology* **60**, 289-310.

Poiani, A. (2010) Do cloacal pathogenic microbes behave as sexually transmitted parasites in birds? *The Open Ornithology Journal* **3**, 72-85.

Poiani, A., and Gwozdz, J. (2002) Cloacal microorganisms and mating systems of four Australian bird species. *Emu* **102**, 291-296.

Poiani, A., and Wilks, C. (2000) Sexually transmitted diseases: a possible cost of promiscuity in birds? *The Auk* **117**, 1061-1065.

Polz, M. F., and Cavanaugh, C. M. (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* **64**, 3724-3730.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590-D596.

R Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reiber, M., McInroy, J., and Conner, D. (1995) Enumeration and identification of bacteria in chicken semen. *Poultry science* **74**, 795-799.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584.

Rowe, M., Czirják, G. Á., Lifjeld, J. T., and Giraudeau, M. (2013) Lysozyme-associated bactericidal activity in the ejaculate of a wild passerine. *Biological Journal of the Linnean Society* **109**, 92-100.

Rowe, M., Czirják, G. Á., McGraw, K. J., and Giraudeau, M. (2011) Sexual ornamentation reflects antibacterial activity of ejaculates in mallards. *Biology Letters* **7**, 740-742.

Rowe, M., and Pruett-Jones, S. (2011) Sperm competition selects for sperm quantity and quality in the Australian Maluridae. *PLoS One* **6**, e15720.

Ryu, H., Grond, K., Verheijen, B., Elk, M., Buehler, D. M., and Santo Domingo, J. W. (2014) Intestinal microbiota and species diversity of *Campylobacter* and *Helicobacter* spp. in migrating shorebirds in Delaware Bay. *Applied and Environmental Microbiology* **80**, 1838-1847.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, 87.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., and Robinson, C. J. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537-7541.

Sheldon, B. C. (1993) Sexually transmitted disease in birds: occurrence and evolutionary significance. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **339**, 491-497.

Shreiner, A. B., Kao, J. Y., and Young, V. B. (2015) The gut microbiome in health and in disease. *Current opinion in gastroenterology* **31**, 69-75.

Skirrow, M. (1994) Diseases due to *Campylobacter*, *Helicobacter* and related bacteria. *Journal of comparative pathology* **111**, 113-149.

Stephens, R. S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R. L., and Zhao, Q. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**, 754-759.

Stewart, R., and Rambo, T. B. (2000) Cloacal microbes in house sparrows. *The Condor* **102**, 679-684.

Stipkovits, L., Varga, Z., Czifra, G., and Dobos-Kovács, M. (1986) Occurrence of mycoplasmas in geese affected with inflammation of the cloaca and phallus. *Avian Pathology* **15**, 289-299.

Summers-Smith, D. 1988. *The sparrows*. A&C Black, London, UK.

Sætre, G. P., Riyahi, S., Aliabadian, M., Hermansen, J. S., Hogner, S., Olsson, U., Gonzalez Rojas, M., Sæther, S., Trier, C., and Elgvin, T. (2012) Single origin of human commensalism in the house sparrow. *Journal of Evolutionary Biology* **25**, 788-796.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.

Tang, B., Li, N., Gu, J., Zhuang, Y., Li, Q., Wang, H.-G., Fang, Y., Yu, B., Zhang, J.-Y., and Xie, Q.-H. (2012) Compromised autophagy by MIR30B benefits the intracellular survival of *Helicobacter pylori*. *Autophagy* **8**, 1045-1057.

Trier, C. N., Hermansen, J. S., Sætre, G.-P., and Bailey, R. I. (2014) Evidence for mito-nuclear and sex-linked reproductive barriers between the hybrid Italian sparrow and its parent species. *PLoS Genetics* **10**, e1004075.

Triplett, M., Parker, H., McDaniel, C., and Kiess, A. (2016) Influence of 6 different intestinal bacteria on Beltsville Small White turkey semen. *Poultry science* **95**, 1918-1926.

Vancanneyt, M., Segers, P., Hauben, L., Hommez, J., Devriese, L., Hoste, B., Vandamme, P., and Kersters, K. (1994) *Flavobacterium meningosepticum*, a pathogen in birds. *Journal of Clinical Microbiology* **32**, 2398-2403.

Vences, M., Granzow, S., Künzel, S., Tebbe, C. C., Baines, J. F., and Dohrmann, A. B. (2015) Composition and variation of the skin microbiota in sympatric species of European newts (Salamandridae). *Amphibia-Reptilia* **36**, 5-12.

Videnska, P., Faldynova, M., Juricova, H., Babak, V., Sisak, F., Havlickova, H., and Rychlik, I. (2013) Chicken faecal microbiota and disturbances induced by single or repeated therapy with tetracycline and streptomycin. *BMC Veterinary Research* **9**, 30.

Waldenström, J., Axelsson-Olsson, D., Olsen, B., Hasselquist, D., Griekspoor, P., Jansson, L., Teneberg, S., Svensson, L., and Ellström, P. (2010) *Campylobacter jejuni* colonization in wild birds: results from an infection experiment. *PLoS One* **5**, e9082.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**, 5261-5267.

Watson, R. O., and Galán, J. E. (2008) *Campylobacter jejuni* survives within epithelial cells by avoiding delivery to lysosomes. *PLoS Pathogens* **4**, e14.

Weiss, S., Amir, A., Hyde, E. R., Metcalf, J. L., Song, S. J., and Knight, R. (2014) Tracking down the sources of experimental contamination in microbiome studies. *Genome biology* **15**, 564.

Weng, S.-L., Chiu, C.-M., Lin, F.-M., Huang, W.-C., Liang, C., Yang, T., Yang, T.-L., Liu, C.-Y., Wu, W.-Y., and Chang, Y.-A. (2014) Bacterial communities in semen from men of infertile couples: metagenomic sequencing reveals relationships of seminal microbiota to semen quality. *PLoS One* **9**, e110152.

Westcott, S. L., and Schloss, P. D. (2015) De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**, e1487.

Westneat, D. F., and Rambo, T. B. (2000) Copulation exposes female Red-winged Blackbirds to bacteria in male semen. *Journal of Avian Biology* **31**, 1-7.

White, J., Mirleau, P., Danchin, E., Mulard, H., Hatch, S. A., Heeb, P., and Wagner, R. H. (2010) Sexually transmitted bacteria affect female cloacal assemblages in a wild bird. *Ecology Letters* **13**, 1515-1524.

Wickham, H. 2017. *ggplot2: elegant graphics for data analysis*. Springer, New York, NY, USA.

Wolfson, A. (1952) The cloacal protuberance: a means for determining breeding condition in live male passerines. *Bird-banding* **23**, 159-165.

Xenoulis, P. G., Gray, P. L., Brightsmith, D., Palculict, B., Hoppes, S., Steiner, J. M., Tizard, I., and Suchodolski, J. S. (2010) Molecular characterization of the cloacal microbiota of wild and captive parrots. *Veterinary Microbiology* **146**, 320-325.

Yildirim, S., Yeoman, C. J., Janga, S. C., Thomas, S. M., Ho, M., Leigh, S. R., Primate Microbiome, C., White, B. A., Wilson, B. A., and Stumpf, R. M. (2014) Primate vaginal microbiomes exhibit species specificity without universal *Lactobacillus* dominance. *The ISME journal* **8**, 2431-2444.

Yildirim, S., Yeoman, C. J., Sipos, M., Torralba, M., Wilson, B. A., Goldberg, T. L., Stumpf, R. M., Leigh, S. R., White, B. A., and Nelson, K. E. (2010) Characterization of the fecal microbiome from non-human wild primates reveals species specific microbial communities. *PLoS One* **5**, e13963.

Young, K. T., Davis, L. M., and DiRita, V. J. (2007) *Campylobacter jejuni*: molecular biology and pathogenesis. *Nature Reviews Microbiology* **5**, 665-679.

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., and Forney, L. J. (2012) Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7**, e33865.

Appendix

Appendix 1: Detailed information on sampled populations

Location	Coordinates	Collection date	Species	Sympatry	Individuals (Post-rarefaction)
Oslo, Norway	59 55 1.14N 10 46 19.9E	05.05-22.06.2016	House	Allopatric house	14(9)
Montanari, Italy	41 54 36.8N 15 51 13.0E	21.05-22.05.2016	Italian	Allopatric Italian	16(16)
Badajoz, Spain	38 40 56.2N 7 11 19.8W	17.03-21.04.2016	House	Sympatric house/Spanish	18(15)
Badajoz, Spain	38 40 56.2N 7 11 19.8W	17.03-21.04.2016	Spanish	Sympatric house/Spanish	17(14)
Lago Salso, Italy	41 32 29.6N 15 53 24.2E	23.05-28.05.2016	Italian	Sympatric Spanish/Italian	12(12)
Lago Salso, Italy	41 32 29.6N 15 53 24.2E	23.05-28.05.2016	Spanish	Sympatric Spanish/Italian	30(28)

Total number of samples: 107(94)

Appendix 2: Methods development

DNA Isolation Protocol Development

There are a wide range of both 16S-based microbiota DNA isolation and amplification protocols, for instance via the Earth Microbiome Project (Gilbert et al. 2014) or the Human Microbiome Project (Human Microbiome Project Consortium 2012). However, none were obvious choices for our novel samples, and we did not feel confident that DNA isolation and amplification would succeed with any particular protocol. Thus, we performed preliminary work to test several protocols. We trialled four DNA isolation methods: (1) Phenol-chloroform (Yuan et al. 2012), (2) FastDNA-96 Soil Microbe DNA Kit (MP Biomedicals, Irvine, CA, USA), (3) DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA), and (4) DNeasy Blood & Tissue Kit with an added enzymatic lytic cocktail. The enzymatic lytic cocktail contained 25 µL lysozyme (10 mg/mL), 3 µL mutanolysin (6250 U/mL) and 1.5 µL lysostaphin (4000 U/mL). All methods included bead beating. For the FastDNA-96 Soil Microbe DNA Kit we used the provided beads, while we included one minute of bead beating at 4.0 m/s with 250 mg ≤106 µm acid-washed silica beads (Sigma-Aldrich, St. Louis, MO, USA) for the other methods. The addition of bead beating and the lytic cocktail has been shown to increase lysis efficiency, thereby producing isolated DNA that more accurately represents the microbial communities (Yuan et al. 2012). We evaluated the protocols based on DNA yield determined by Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA, USA) and NanoDrop ND-1000 (Thermo scientific, Waltham, MA, USA) values, gel electrophoresis band quality, and theoretical ability to produce isolated DNA as representative of the bacterial communities as possible. The main aim of this work was to ensure that at least one of the trialled methods allowed us to isolate sufficient amounts of DNA for PCR amplification.

All methods produced sufficient yields (1.2-14 µg per sample), and amplified with PCR. However, we found the DNeasy Blood & Tissue Kit with the added enzymatic lytic cocktail to perform marginally better than the other protocols. Moreover, the DNeasy Blood & Tissue Kit was more economical to use than the FastDNA-96 Soil Microbe DNA Kit, and safer to use than phenol-chloroform. Thus, we opted to use it.

PCR Protocol Development

While our DNA isolation trials produced sufficient yields, PCR amplification proved challenging. A possible reason for our problems could be that the measured DNA contains only a small amount of DNA originating from bacteria in proportion to that from the sparrow host. PCR inhibition was also a potential issue, but no inhibition was detected when performing an experiment spiking positive controls with increasing amounts of sample DNA. To attempt to find a working protocol, we tested 4 polymerases according to manufacturer's instructions: (1) 5Prime HotMaster Taq (Quantabio, Beverly, MA, USA), (2) AccuPrime Pfx DNA Polymerase (Thermofisher, Waltham, MA, USA), (3) Phusion High-Fidelity DNA Polymerase (New England Biolabs (NEB), Ipswich, MA, USA), and (4) Q5 High-Fidelity DNA Polymerase (NEB). (1) and (2) have previously been used successfully in several microbiota studies (e.g.: Star et al. 2013, Kueneman et al. 2014). In order to optimize the protocols, we tweaked PCR parameters annealing temperature, cycle number, template amount, primer concentration, and length of annealing, denaturation, and extension steps. Amplification success and quality was measured in terms of gel electrophoresis band quality.

Despite testing primers and PCR conditions proven to work in the past, and amplification of positive controls *(E. coli* DNA and sheep epithelium) being successful, our ejaculate samples proved challenging to amplify. 5Prime HotMaster Taq required too high cycle numbers (45) to be used reliably, while Phusion High-Fidelity DNA Polymerase was unable to amplify even at 45 cycles. AccuPrime Pfx DNA Polymerase amplified successfully, but also generated large amounts of non-specific amplicons. Q5 High-Fidelity DNA Polymerase produced good results at 35 cycles, and was thus found to be the best option. However, repeating the finalized method with indexed primers proved unsuccessful. A possible reason for this could be the index sequences preventing sufficiently strong alignment. To solve this potential issue, we attempted to find an optimal annealing temperature by performing and gradient PCR (for temperatures 42-58 ºC, 53 ºC was used previously), but without success. To work around the problem, we instead decided to use a three step PCR as outlined in the lab protocol section. This protocol involved in total 50 PCR cycles.

References:

Gilbert, J. A., Jansson, J. K., and Knight, R. (2014) The Earth Microbiome project: successes and aspirations. *BMC Biology* **12**, 69.

Human Microbiome Project, C. (2012) Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207-214.

Kueneman, J. G., Parfrey, L. W., Woodhams, D. C., Archer, H. M., Knight, R., and McKenzie, V. J. (2014) The amphibian skin-associated microbiome across species, space and life history stages. *Molecular Ecology* **23**, 1238-1250.

Star, B., Haverkamp, T. H., Jentoft, S., and Jakobsen, K. S. (2013) Next generation sequencing shows high variation of the intestinal microbial species composition in Atlantic cod caught at a single location. *BMC Microbiology* **13**, 248.

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., and Forney, L. J. (2012) Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7**, e33865.

Appendix 3: Read-merging batch file

```
mkdir ~/sparrow_ejaculate_final/merged_fastq_files

mkdir ~/sparrow_ejaculate_final/fasta_files


module load flash
#Merge plate 1 reads with FLASH (v.1.2.11)
cd ~/sparrow_ejaculate_final/fastq_files/combo_40_cutadapt_halfplate
ls *_cutadapt.fastq.gz | paste - - > filename_table
while read -r a b; do
        flash --output-prefix=${a%_R1_cutadapt.fastq.gz} --output-
directory=../../merged_fastq_files --min-overlap=90 --max-overlap=250 $a $b
done <filename_table

#Merge plate 2 reads with FLASH (v.1.2.11)
cd ~/sparrow_ejaculate_final/fastq_files/P2_40_cutadapt_fullplate
ls *_cutadapt.fastq.gz | paste - - > filename_table
while read -r a b; do
        flash --output-prefix=${a%_R1_cutadapt.fastq.gz} --output-
directory=../../merged_fastq_files --min-overlap=90 --max-overlap=250 $a $b
done <filename_table

#Change - to _ in filenames
cd ~/sparrow_ejaculate_final/merged_fastq_files
for filename in *.extendedFrags.fastq; do mv -v "$filename" $(echo "$filename" | tr '-' '_');
done

#Convert to FASTA format, remove primers
for file in *extendedFrags.fastq
do
SAMPLE=$(echo "$file" | cut -d . -f 1)
OUT="$SAMPLE.fas"
touch ../fasta_files/"$OUT"
cat "$file" | awk 'NR%4==1{printf ">%s\n", substr($0,2)}NR%4==2{print}' | sed '2~2
s/....................$//' | sed '2~2 s/^...................//' >> ../fasta_files/"$OUT"
done

#Add plate-specific tag to all sequence names (ensures that all sequence have unique names)
cd ~/sparrow_ejaculate_final/fasta_files

for file in P1*
do
  sed -i 's/>/>P1_/g' "$file"
done

for file in P2*
do
  sed -i 's/>/>P2_/g' "$file"
done
```

Appendix 4: Mothur OTU pipeline batch file

```
set.seed(seed=12345)
set.current(fasta=sparrows_and_controls.fasta, group=sparrows_and_controls.groups)

#Remove sequences with ambiguous bases or wrong length:
screen.seqs(fasta=current, group=current, maxambig=0, minlength=243, maxlength=263)

#Find unique sequences:
unique.seqs(fasta=current)

#Find and remove singleton sequences
split.abund(fasta=current, name=current, cutoff=1, accnos=true)
set.current(accnos=sparrows_and_controls.good.rare.accnos,
fasta=sparrows_and_controls.good.unique.fasta, group=sparrows_and_controls.good.groups,
name=sparrows_and_controls.good.names)
remove.seqs(accnos=current, fasta=current, group=current, name=current)

#Align the unique sequences to SILVA database
align.seqs(fasta=current, reference=~/nobackup/silva_v123/silva.v4_region.fasta)

#Remove sequences aligning outside of the appropriate range
screen.seqs(fasta=current, name=current, group=current, start=1, end=9582)
filter.seqs(fasta=current, vertical=T, trump=.)
unique.seqs(fasta=current, name=current)

#VSEARCH de novo chimera detection:
chimera.vsearch(fasta=current, name=current, group=current, reference=self, dereplicate=f)
remove.seqs(fasta=current, name=current, group=current, accnos=current, dups=t)

#Classify the sequences:
classify.seqs(fasta=current, name=current, group=current,
reference=~/nobackup/silva_v123/silva.v4_region.fasta,
taxonomy=~/nobackup/silva_v123/silva.nr_v123.tax, cutoff=80)

#Remove non-bacteria/archaea sequences:
remove.lineage(fasta=current, name=current, group=current, taxonomy=current,
taxon=Chloroplast-Mitochondria-unknown-Eukaryota)

#VSEARCH 97% clustering:
cluster(fasta=current, name=current, method=dgc, cutoff=0.03)
```

```
#Consensus taxonomy for each OTU:
classify.otu(list=current, name=current, taxonomy=current, cutoff=51, label=0.03)
rename.file(taxonomy=current, prefix=otu_taxonomy_table)

#Make OTU table:
make.shared(list=current, group=current, label=0.03)
rename.file(shared=current, prefix=otu_samples_table)


#Get representative (most abundant) sequence for each OTU:
get.oturep(list=current, name=current, fasta=sparrows_and_controls.good.unique.fasta,
method=abundance)
rename.file(fasta=current, prefix=otu_representative_seqs)
system(sed 's/.*\t/>/g' otu_representative_seqs.fasta | sed 's/|.*$//g' >
otu_representative_seqs_fixed.fasta)
set.current(fasta=otu_representative_seqs_fixed.fasta)

#Align sequences, make NJ tree:
align.seqs(fasta=current, reference=~/nobackup/silva_v123/silva.v4_region.fasta)
filter.seqs(fasta=current, vertical=T)
clearcut(fasta=current, DNA=T, neighbor=t)

#Make taxonomy table phyloseq-compatible:
system(awk '!($2="")'
sparrows_and_controls.good.unique.pick.good.filter.unique.pick.pick.dgc.0.03.cons.taxonom
y | sed 's/;/\t/g' | sed 's/Taxonomy/Domain\tPhylum\tClass\tOrder\tFamily\tGenus/g' | sed
's/OTU /OTU\t/g' | sed 's/ Archaea/\t"Archaea/g' | sed 's/ Bacteria/\t"Bacteria/g' | sed
's/[[:space:]]*$/"/' | sed 's/)\t/)"\t"/g' | sed 's/Genus"/Genus/g' | sed -e 's/([^()]*)//g' | sed 's/
\t/\t/g' > phyloseq_table.taxonomy)
```

Appendix 5: Reads removed by filtering

During filtering, 2% of the reads were discarded due to having the wrong length, 12.4% were detected as singletons, 8% were detected as chimeras, and 4% were classified as non-bacteria/archaea and discarded. To investigate the origin of these sequences I ran a separate pipeline with no filtering other than chimera detection, singleton sequence removal, and removal of all sequences classified as bacteria or archaea. OTUs were generated as in the main analysis, and they were classified using BLAST+ against the NCBI nucleotide collection database. The 126 OTUs with more than 50 reads contained in total 950604 reads. Several of the OTUs were likely to originate from the sparrow host (i.e. they had 100% hits to bird genomes), but none were identified as human. The low number of reads originating from the host was expected, as primers align poorly to the house sparrow mitochondrion (forward primer: 2 mismatches, reverse primer: 3 mismatches).

3 Spanish and 5 house sparrows from Badajoz had in total 67197 reads likely to originate Alveolata. The bird with the highest abundance had in total 48183 reads of putative alveolates, nearly as many as those classified as bacteria and archaea (61483). One of the putative alveolate OTUs had 91% identity hits to the apicoplast of both *Leucocytozoon* and *Plasmodium,* while another had 96% identity hits to *Naegleria* mitochondria. While the exact classification of these sequences is unclear, *Leucocytozoon* and *Plasmodium* are known avian parasites (Shurulinkov and Golemansky 2003), and it is probable that the sequences detected here originate from relatives with similar ecology. Thus, protists should be considered as possible STDs, and further investigated in future research. While no *Chlamydia* or Chlamydiaceae was detected in this study, Parachlamydiaceae, which is associated with infection of protists (Greub and Raoult 2002), was present at some abundance and diversity. It is possible that these interact with the protists described here, and do not detriment the avian host. The 515f-806r primers fit the *Naegleria* mitochondrion (GenBank: AY376153) fully on the forward primer, but with 2 mismatches on the reverse primer, while they fit the *Plasmodium* apicoplast (GenBank: AB649422) with 1 mismatch on each primer. This might suggest that the ejaculate contained a significant amount of Alveolates compared to bacteria, although it is possible that a GC-content of just 27% caused over-amplification of these sequences.

Both primers had 100% identity match to several chloroplast genomes, for instance *Lilium cernuum* (GenBank: KX354692), and 7 OTUs containing in total 420000 reads were classified with 98-100% identity BLAST hits as chloroplast. Most OTUs could only be identified to broad taxonomic levels, including monocots and eudicots and orders Laurales, Lamiales, and Poales. However, an OTU only present in 1 Italian sparrow from Lago Salso could be classified as *Medicago*, a genus that includes agricultural legumes that are likely present at the location and consumed by sparrows. Ejaculate samples containing plant reads had significantly higher alpha diversity than those without (figure 1, Simpson diversity, p = 0.001, Mann-Whitney U-test & $R^2 = 0.08$, p = 0.003, linear model with non-normalized plant-classified read numbers), suggesting that a greater range of gut-associated bacteria might have been incorporated into the ejaculate in these birds. Moreover, the variance in Simpson diversity appears to be considerably lower between most of these birds.

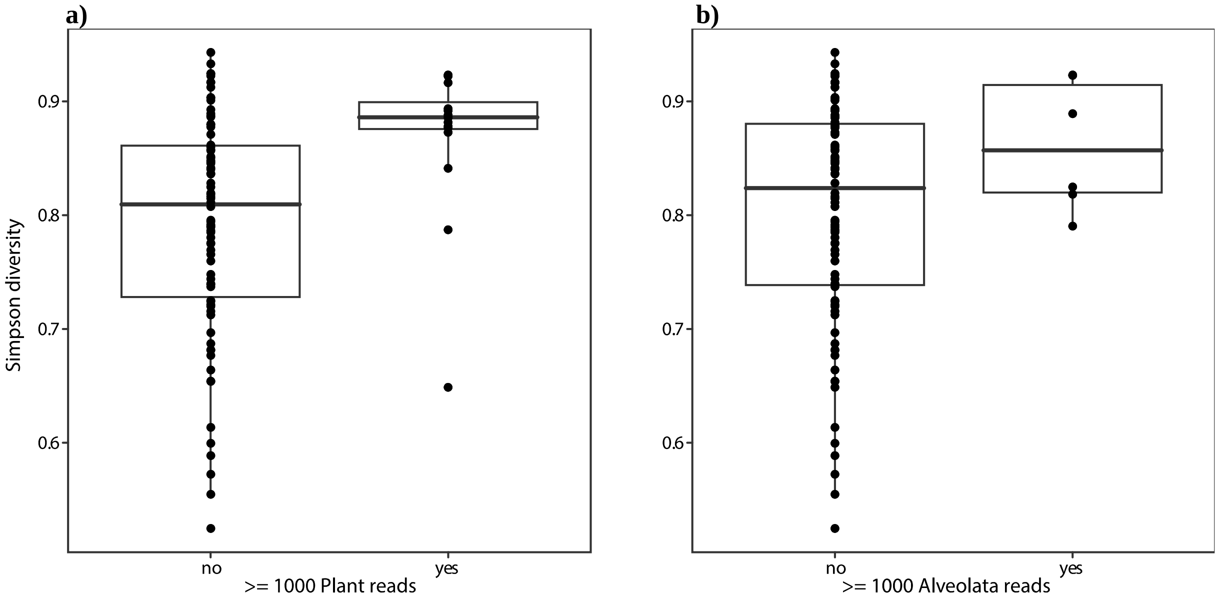

Figure 1 Alpha diversity of samples with alveolate or plant reads. Boxplots comparing Simpson diversity of birds with more than 1000 a) plant and b) alveolate reads. Samples with plant reads were significantly more diverse than the others, while those with alveolate reads were not (p = 0.13, Mann-Whitney U-test).

References:

Greub, G., and Raoult, D. (2002) Parachlamydiaceae: potential emerging pathogens. *Emerging Infectious Diseases* **8**, 625-630.

Shurulinkov, P., and Golemansky, V. (2003) *Plasmodium* and *Leucocytozoon* (Sporozoa: Haemosporida) of wild birds in Bulgaria. *Acta Protozoologica* **42**, 205-214.

Appendix 6: Vital, non-redundant, R code

```
set.seed(12345)
library(phyloseq)
```

Load data and make Phyloseq experiment-level object

```
#Import sample metadata and files from Mothur
otu_table_file <- read.table("input/otu_samples_table.an.shared", header=TRUE, sep="\t", dec=",", row.names = 2)
otu_table_file$label <- NULL
otu_table_file$numOtus <- NULL
otu_table <- t(otu_table_file)

otumat <- as.matrix(otu_table)
taxmat <- as.matrix(read.csv("input/phyloseq_table.taxonomy", header = TRUE, sep="\t", row.names = 1))

taxmat <- gsub("unclassified", "Unclassified", taxmat)
sam_dat <- read.table("input/sample_info_fixed_sample_names.txt", header = TRUE, dec = ".", row.names = 1)

library(phytools)
phylo_tree <- read.newick("otu_representative_seqs_fixed.filter.tre")

#Define phyloseq-classes
OTU <- otu_table(otumat, taxa_are_rows = TRUE, errorIfNULL = TRUE)
TAX <- tax_table(taxmat)
SAM_DAT <- sample_data(sam_dat, errorIfNULL = TRUE)
TREE <- read_tree(phylo_tree, errorIfNULL = TRUE)

#Combine into experiment-level object
phyloseq_explvl_object = phyloseq(OTU, TAX, SAM_DAT, TREE)
phyloseq_explvl_object_50 <- filter_taxa(phyloseq_explvl_object, function(x) sum(x) > 50, TRUE)
rarefied_explvl_object <- rarefy_even_depth(phyloseq_explvl_object_50, sample.size = 42049)
```

Alpha diversity

```
#Make data.frame containing alpha diversity measurements and sample info:
sampl <- as.data.frame(sample_data(rarefied_explvl_object), keep.rownames = TRUE)
sampl <- as.matrix(sampl)
sampl <- as.data.frame(sampl, row.names = NULL, optional = FALSE)

#Alpha diversity table calculated by phyloseq
alpha_diversity_table <- estimate_richness(rarefied_explvl_object, split = TRUE, measures = c("Observed", "Chao1", "ACE", "Shannon", "Simpson", "InvSimpson", "Fisher"))

#Merge alpha diversity and sample info data frames
```

```
alpha_diversity_table_sample_info <- merge(alpha_diversity_table, sampl, c("row.names"))

library(stats)
#Kruskal-Wallis one-way analysis of variance
kruskal.test(Observed ~ Population, data = alpha_diversity_table_sample_info)

#Mann-Whitney U test
badajoz_house <- subset(alpha_diversity_table_sample_info, Population ==
c("Badajoz_house"), select=c(Simpson))
badajoz_house <- as.vector(badajoz_house$Simpson)

badajoz_spanish <- subset(alpha_diversity_table_sample_info, Population ==
c("Badajoz_Spanish"), select=c(Simpson))
badajoz_spanish <- as.vector(badajoz_spanish$Simpson)

wilcox.test(badajoz_spanish, badajoz_house, paired = FALSE)
```

Collector’s curve

```
library(vegan)
OTU_vegan_all_populations <- otu_table(rarefied_explvl_object)

OTU_vegan_all_populations <- as.matrix(OTU_vegan_all_populations)
OTU_vegan_all_populations  <- t(OTU_vegan_all_populations)
specaccum_all_populations <- specaccum(OTU_vegan_all_populations, method = "random",
permutations = 10000)

plot(specaccum_all_populations)
```

Weighted UniFrac newick-tree of samples

```
#Create weighted UniFrac distance matrix
rarefied_no_tbd_or_repeats_wunifrac_matrix<- UniFrac(rarefied_explvl_object,
weighted=TRUE, normalized=TRUE, parallel=FALSE, fast=TRUE)

#Create tree with neighbour joining
library(ape)
rarefied_no_tbd_or_repeats_wunifrac_tree <-
nj(rarefied_no_tbd_or_repeats_wunifrac_matrix)

#Write newick-tree to file
write.tree(rarefied_no_tbd_or_repeats_wunifrac_tree, file =
"rarefied_no_tbd_or_repeats_wunifrac_tree.tre", append = FALSE,
        digits = 10, tree.names = FALSE)
```

Adonis and ANOSIM

```
library(vegan)
#Calculate Bray-Curtis distance matrix
bray_distance_matrix <- phyloseq::distance(rarefied_explvl_object, method = "bray")
```

```
#Make a data frame of sample data
sample_df <- data.frame(sample_data(rarefied_explvl_object))

#Adonis
adonis(bray_distance_matrix ~ Sparrow_species, strata = sample_df$Location, data = sample_df, permutations=10000)
adonis(bray_distance_matrix ~ Location, data = sample_df, permutations=10000, strata = sample_df$Sparrow_species)

#ANOSIM
anosim(bray_distance_matrix, sample_df$Sparrow_species, permutations = 10000, strata = sample_df$Location)
anosim(bray_distance_matrix, sample_df$Location, permutations = 10000, strata = sample_df$Sparrow_species)
```

Appendix 7: Full list of phylum-level classifications

This table show classifications as assigned from the SILVA database. Many are not formally described taxa, and thus do not have correct naming and taxonomy.

Domain	Phylum	Abundance	OTUs
Bacteria	Bacteroidetes	41.251 %	159
Bacteria	Proteobacteria	26.173 %	377
Bacteria	Firmicutes	11.464 %	157
Bacteria	Parcubacteria	6.081 %	173
Bacteria	Actinobacteria	2.750 %	85
Bacteria	Bacteria_Unclassified	2.729 %	82
Bacteria	Verrucomicrobia	1.951 %	43
Bacteria	Gracilibacteria	1.310 %	17
Bacteria	Spirochaetae	1.288 %	6
Bacteria	Microgenomates	1.111 %	13
Bacteria	Cyanobacteria	0.930 %	10
Bacteria	Planctomycetes	0.598 %	37
Bacteria	Candidate_division_SR1	0.537 %	11
Bacteria	Saccharibacteria	0.444 %	23
Bacteria	Fusobacteria	0.338 %	7
Bacteria	Acidobacteria	0.271 %	14
Bacteria	Chloroflexi	0.131 %	10
Bacteria	Deinococcus-Thermus	0.121 %	3
Bacteria	SM2F11	0.102 %	5
Bacteria	Chlamydiae	0.088 %	16
Archaea	Archaea_Unclassified	0.053 %	2
Bacteria	WCHB1-60	0.050 %	1
Bacteria	TM6	0.046 %	4
Archaea	Thaumarchaeota	0.046 %	1
Bacteria	Lentisphaerae	0.034 %	5
Bacteria	Tenericutes	0.032 %	4
Bacteria	Armatimonadetes	0.022 %	4
Archaea	Woesearchaeota	0.019 %	1
Archaea	Euryarchaeota	0.009 %	2
Bacteria	Elusimicrobia	0.007 %	5
Bacteria	Nitrospirae	0.003 %	1
Bacteria	OC31	0.002 %	1
Bacteria	Candidate_division_OP3	0.002 %	1
Bacteria	Omnitrophica	0.002 %	1
Bacteria	Fibrobacteres	0.001 %	1
Bacteria	Synergistetes	0.001 %	1
Bacteria	Caldiserica	0.001 %	1

Appendix 8: Abundant order- and genera-level classifications

These tables show classifications as assigned from the SILVA database. Many are not formally described taxa, and thus do not have correct naming and taxonomy. Only 72% of reads are classified to the genus level.

Phylum	Order	Abundance	OTUs
Bacteroidetes	Flavobacteriales	36.527 %	38
Parcubacteria	Parcubacteria_Unclassified	6.081 %	173
Proteobacteria	Pseudomonadales	5.493 %	17
Firmicutes	Lactobacillales	5.322 %	32
Proteobacteria	Campylobacterales	5.156 %	5
Proteobacteria	Burkholderiales	3.862 %	40
Firmicutes	Clostridiales	2.839 %	72
Firmicutes	Bacilli_Unclassified	2.729 %	82
Proteobacteria	B1-7BS	2.654 %	35
Bacteroidetes	Bacteroidales	2.470 %	33
Proteobacteria	Proteobacteria_Unclassified	1.999 %	26
Proteobacteria	Sphingomonadales	1.615 %	39
Gracilibacteria	Gracilibacteria_Unclassified	1.310 %	17
Spirochaetae	Spirochaetales	1.288 %	6
Proteobacteria	Enterobacteriales	1.253 %	5
Proteobacteria	Rhodobacterales	1.210 %	8
Microgenomates	Microgenomates_Unclassified	1.111 %	13
Proteobacteria	Oceanospirillales	0.982 %	4
Actinobacteria	Micrococcales	0.966 %	24
Bacteroidetes	Sphingobacteriales	0.939 %	49
Bacteroidetes	Cytophagales	0.879 %	16
Proteobacteria	Legionellales	0.845 %	35
Proteobacteria	Rhizobiales	0.830 %	47
Actinobacteria	Corynebacteriales	0.811 %	16
Cyanobacteria	Cyanobacteria_Unclassified	0.692 %	4
Verrucomicrobia	OPB35_soil_group_Unclassified	0.665 %	13
Verrucomicrobia	Verrucomicrobiales	0.658 %	19
Proteobacteria	Gammaproteobacteria_Unclassified	0.569 %	13
Candidate_division_SR1	Candidate_division_SR1_Unclassified	0.537 %	11
Actinobacteria	Frankiales	0.535 %	11
Firmicutes	Selenomonadales	0.482 %	9
Planctomycetes	Planctomycetales	0.478 %	25
Saccharibacteria	Saccharibacteria_Unclassified	0.444 %	23

Phylum	Genus	Abundance	OTUs
Bacteroidetes	Flavobacterium	34.692 %	23
Proteobacteria	Acinetobacter	4.332 %	6
Proteobacteria	Helicobacter	3.213 %	2
Firmicutes	Streptococcus	2.722 %	8
Bacteroidetes	Bacteroides	2.334 %	16
Proteobacteria	Campylobacter	1.941 %	2
Firmicutes	Staphylococcus	1.497 %	2
Spirochaetae	Borrelia	1.244 %	1
Bacteroidetes	Fluviicola	1.004 %	3
Proteobacteria	Halomonas	0.986 %	3
Proteobacteria	Yersinia	0.967 %	1
Firmicutes	Clostridium_sensu_stricto_1	0.884 %	5
Proteobacteria	Pseudomonas	0.862 %	6
Firmicutes	Lactococcus	0.793 %	3
Proteobacteria	Polynucleobacter	0.727 %	3
Firmicutes	Catellicoccus	0.640 %	2
Proteobacteria	Legionella	0.636 %	27
Bacteroidetes	Pseudarcicella	0.587 %	1
Firmicutes	Lactobacillus	0.585 %	8
Firmicutes	Lachnoclostridium	0.499 %	2
Bacteroidetes	Cloacibacterium	0.342 %	1
Firmicutes	Romboutsia	0.305 %	1
Actinobacteria	hgcI_clade	0.303 %	4
Actinobacteria	Corynebacterium_1	0.293 %	8
Firmicutes	Domibacillus	0.285 %	1
Proteobacteria	Sphingomonas	0.280 %	8
Firmicutes	Veillonella	0.273 %	2
Proteobacteria	Moraxella	0.265 %	2
Proteobacteria	Porphyrobacter	0.233 %	1
Firmicutes	Finegoldia	0.231 %	1
Proteobacteria	Brevundimonas	0.231 %	3
Firmicutes	Dolosigranulum	0.216 %	1
Actinobacteria	Fodinicola	0.212 %	1
Acidobacteria	Blastocatella	0.205 %	4