



# What human sperm RNA-Seq tells us about the microbiome

Grace M. Swanson<sup>1</sup> · Sergey Moskvovtsev<sup>2</sup> · Clifford Librach<sup>2</sup> · J. Richard Pilsner<sup>3</sup> · Robert Goodrich<sup>1</sup> · Stephen A. Krawetz<sup>1</sup>

Received: 15 October 2019 / Accepted: 19 December 2019 / Published online: 4 January 2020  
© Springer Science+Business Media, LLC, part of Springer Nature 2020

## Abstract

**Purpose** The study was designed to assess the capacity of human sperm RNA-seq data to gauge the diversity of the associated microbiome within the ejaculate.

**Methods** Semen samples were collected, and semen parameters evaluated at time of collection. Sperm RNA was isolated and subjected to RNA-seq. Microbial composition was determined by aligning sequencing reads not mapped to the human genome to the NCBI RefSeq bacterial, viral and archaeal genomes following RNA-Seq. Analysis of microbial assignments utilized phyloseq and vegan.

**Results** Microbial composition within each sample was characterized as a function of microbial associated RNAs. Bacteria known to be associated with the male reproductive tract were present at similar levels in all samples representing 11 genera from four phyla with one exception, an outlier. Shannon diversity index ( $p < 0.001$ ) and beta diversity (unweighted UniFrac distances,  $p = 9.99e-4$ ; beta dispersion,  $p = 0.006$ ) indicated the outlier was significantly different from all other samples. The outlier sample exhibited a dramatic increase in *Streptococcus*. Multiple testing indicated two operational taxonomic units, *S. agalactiae* and *S. dysgalactiae* ( $p = 0.009$ ), were present.

**Conclusion** These results provide a first look at the microbiome as a component of human sperm RNA sequencing that has sufficient sensitivity to identify contamination or potential pathogenic bacterial colonization at least among the known contributors.

**Keywords** Human sperm RNA-seq · Microbiome · Bacterial identification · Microbial RNA

## Abbreviations

|      |                                  |
|------|----------------------------------|
| BH   | Benjamini-Hochberg               |
| CCLE | Cancer Cell Line Encyclopedia    |
| hNGS | Human sperm RNA-seq              |
| ICSI | Intracytoplasmic sperm injection |
| IUI  | Intrauterine Insemination        |

|           |   |
|-----------|---|
| IVF       | In vitro fertilization                                  |
| LB        | Live birth  |
| LB + NLB  | Combined LB and NLB group samples excluding the outlier |
| MS2       | Escherichia virus MS2                                   |
| NGS       | Next-generation sequencing                              |
| NLB       | No live birth   |
| NMDS      | Non-metric multidimensional scaling                     |
| OTU       | Operational taxonomic unit                              |
| PERMANOVA | Permutational multivariate analysis of variance test    |
| phiX      | Enterobacteria phage phiX174 sensu lato                 |
| rDNA      | DNA sequencing of rRNA                                  |
| rRNA      | Ribosomal RNA   |
| Seq       | 16S rDNA sequencing by NGS                              |
| TCGA      | The Cancer Genome Atlas                                 |
| TIC       | Timed intercourse                                       |
| TII       | Transcript Integrity Index                              |
| WHO       | World Health Organization                               |

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10815-019-01672-x>) contains supplementary material, which is available to authorized users.

✉ Stephen A. Krawetz  
steve@compbio.med.wayne.edu

<sup>1</sup> Department of Obstetrics and Gynecology, Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, 275 E. Hancock, Detroit, MI 48202, USA

<sup>2</sup> CReATe Fertility Centre Research Program, Toronto, Canada

<sup>3</sup> Department of Environmental Health Sciences, University of Massachusetts Amherst School of Public Health and Health Sciences, Amherst, MA 01003, USA

## Introduction

The current accepted standard for evaluating the male reproductive tract microbiome employs culturing techniques. These can be limiting [1, 2] as the majority of pathogens cannot be cultured [3]. To provide an overview, metagenomic/targeted DNA sequencing of 16S ribosomal RNA (rDNA) hypervariable regions and RNA sequencing of the 16S rRNA hypervariable regions using PCR or next-generation sequencing (NGS) has been performed on seminal fluid, semen, and urine [1, 2, 4–6]. These strategies are capable of genus and species identification [2], but are generally targeted to one or a group of bacteria [1]. Despite an increased sensitivity by 16S rRNA/rDNA sequencing, technical considerations and cost may be perceived as a limiting factor in some clinical laboratories [7]. Similarly, virus detection must be performed separately and relies on targeted viral culture or virus specific DNA/RNA sequencing strategies [8–10]. However, sequencing costs have decreased dramatically from the ~3.5 billion dollars for the first human genomes to less than 1000 dollars and are still declining with a non-targeted approach providing a more complete and picture.

Male genital tract bacteria are primarily introduced by direct contact to the source, e.g., sexual transmission [6]. *Escherichia coli* is the most frequently noted bacteria in males, associated with genital tract infection or urethral/post-urethral contamination upon semen collection [6, 11]. Comparative 16S rDNA sequencing between male and female partners indicates that 85% of the phylotypes are shared, primarily within the genera *Lactobacillus*, *Veillonella*, *Streptococcus*, *Porphyromonas*, and *Atopobium* [4]. Compared to vaginal bacterial communities, the male seminal fluid has a less concentrated, more diverse (alpha diversity) bacterial community [4].

We sought to determine if human sperm RNA-seq data could provide a sensitive method of detection. To accomplish this objective, the microbial composition among 85 individuals was compared to previously published reports of the male reproductive tract microbiome [1, 6, 12–15]. Here, we report that human sperm RNA-seq is a viable method for the evaluation of the seminal microbiome. By RNA-seq, we have identified a male that may present with *Streptococcus agalactiae* bacterial colonization.

## Materials and methods

### Study population, sperm RNA isolation, sequencing, and alignment

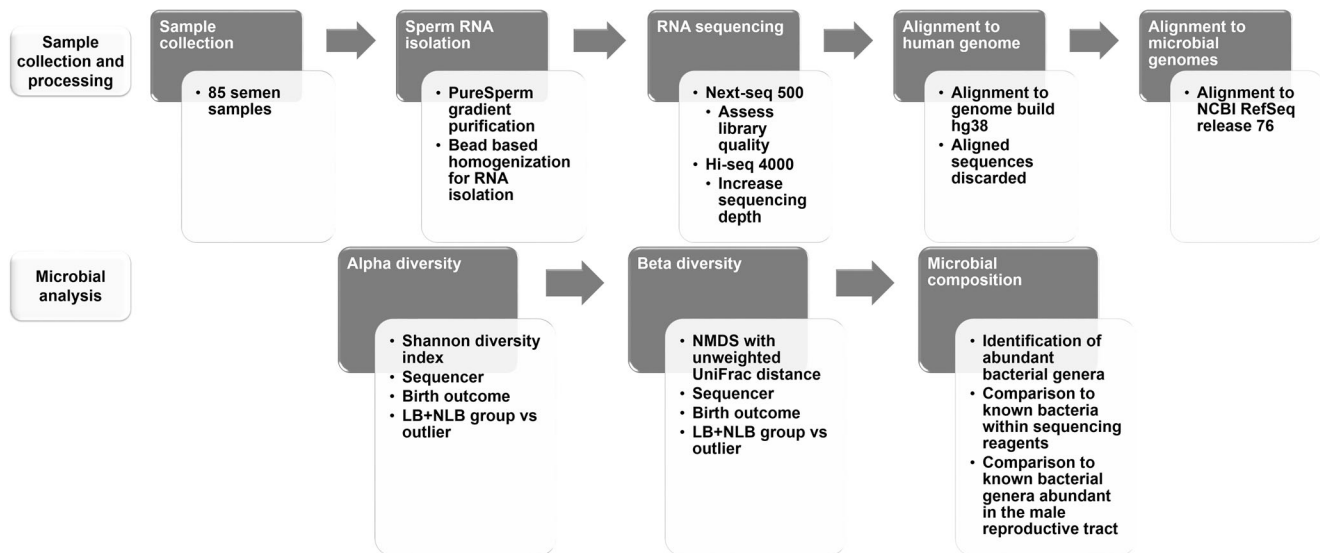
A total of 85 informed consented deidentified biobanked semen samples from idiopathic infertile couples were obtained from two independent groups of sites within the USA and a

single site in Canada. Males ranged in age from 24 to 44 years old (mean age = 34.73). Each couple had undergone at least one cycle of timed intercourse (TIC), intrauterine insemination (IUI), intracytoplasmic sperm injection (ICSI) or in vitro fertilization (IVF). By study completion, 48 couples achieved live birth (LB) and 37 couples failed to achieve a birth (NLB).

The corresponding deidentified frozen semen samples were processed and analyzed as summarized in Fig. 1 in accord with IRB protocol 095701MP2E(5R). Sperm RNA was isolated as previously described [16–19] and sequenced on an Illumina Next-seq 500 to assess library quality and then an Illumina Hi-seq 4000 to increase sequencing depth. Sequences aligned to the human genome [20] were removed from further analysis. Sequences not mapped to the human genome by HISAT2 [21] were aligned to the bacterial/viral/archaeal genomes (NCBI RefSeq release 95—all bacterial genomes, as well as NCBI RefSeq release 76—all bacterial, viral, archaeal genomes) with the 98% similarity taxonomic profiling method of Kraken v1 [22]. Using the Kraken-filter command, the threshold for calling a sequence was set at 0.15. The percentage of microbial sequencing reads was determined as a function of total sequence reads. For each sample, the percentage of microbial assigned reads from the total sequencing reads was calculated for total microbial assignments after removing internal controls, Enterobacteria phage phiX174 sensu lato (phiX) and Escherichia virus MS2 (MS2). Within each sample, the percentage of species-specific microbial sequences was calculated for each sequencing run. One sample presented an abnormally high number of microbial-assigned sequencing reads (outlier) and was considered separately from the remaining group of 84 samples which provided a background distribution. A subset of these samples including the outlier were also assessed using the Greengenes v13\_8 reference database aligning sequences based on a 99% percent similarity to the 16S rRNA.

### Estimated richness—alpha and beta diversity

The Kraken report was used to compute OTU tables by unique NCBI taxonomy ID and phylogeny tables, for analysis by phyloseq that was written in R [23, 24]. Taxonomic lineage and rank from the Kraken v1 [22] designated NCBI taxonomy IDs were assigned using Taxonkit [25]. Duplicate NCBI IDs were merged, and NCBI taxonomy IDs for phiX and MS2 controls were removed. Missing values were re-assigned a raw abundance of 0. The remaining NCBI taxonomy IDs were re-named beginning with OTU1 and abundance normalized to total library size using reads per million (RPM). Two OTU tables were generated (one with raw abundance and one with normalized abundance) and included the OTU ID and abundance value. The phylogeny tables were generated for each remaining OTU ID and associated taxonomic lineage (Kingdom, Phylum, Class, Order, Family, Genus and



**Fig. 1** Study design. Sample collection and processing is outlined. Sperm RNA from 85 semen samples was isolated and sequenced using a Next-seq 500 and Hi-Seq 4000 sequencer. Sequences were aligned to the human genome then aligned sequences discarded. The unmapped sequences were retained then aligned to the bacterial, viral, and archaeal

genomes. Estimated richness of each sample (alpha diversity) employed the Shannon diversity index. Samples were then evaluated as a function of beta diversity (between group richness). With the exception of a single outlier, 84 samples showed the same statistical distribution. Microbial composition was then determined

Species). Alpha diversity (within sample microbial diversity) was calculated using the functions, `estimate_richness()`, followed by `plot_richness()` [24] for visualization. The Shannon diversity index, a common measure of alpha diversity, was performed for all 4426 OTUs prior to RPM normalization and visualized based on sequencing run (Next-seq vs Hi-seq), birth outcome (LB vs NLB), and sample group (LB + NLB vs outlier). The R Mann-Whitney test [23] was used to evaluate group differences in alpha diversity.

A phyloseq object was then created using the normalized OTU abundance table and merged at the species level by the `tax_glom()` function. This function combines abundance and lineage for all OTUs assigned to the same species. For example, only 1 OTU will represent the 49 OTUs originally assigned to *Escherichia coli*. All OTUs with no species assigned were relabeled as NA at the species level but remain unmerged. Rarefaction of the species merged phyloseq object was performed so that each sample library was of equal size. Briefly, the `rarefy_even_depth()` function was performed with 711 random sub-samplings (“set.seed(711)”) without OTU replacement and OTU trimming if no longer observed in any sample. This resulted in the removal of 581 OTUs from further analysis of the microbial diversity between species (beta diversity). Beta diversity of the remaining 3645 OTUs was determined. This employed non-metric multidimensional scaling (NMDS) with unweighted UniFrac distances computed and visualized using the `plot_ordination()` function from phyloseq. Statistical significance of the unweighted UniFrac distances based on sequencing run, birth outcome, and sample group employed the R version of `vegan` [23, 26] using a permutational multivariate analysis of variance test (PERMANOVA). Differences between

sequencer, birth outcome, and sample group microbiome variance were further evaluated in `vegan` by testing beta dispersion using the `betadisp()` and `permutest()` functions in `vegan`. Both functions were calculated using 1000 permutations.

## Determining microbial background

Microbial background within human sperm RNA-seq data was determined as a function of the 84 LB + NLB group sample set that excluded the outlier from each sequencing run using normalized, unmerged, and unrarefied data. Sequencing runs were visualized as separate samples as the alpha, but not beta, diversities were statistically different. This enabled visualization of the microbial abundance between sequencing runs so that taxa were present in at least one sequencing run for all samples and in > 60% of the samples at an RPM > 1 in both Next-seq and Hi-seq sequencing runs. Bacteria were prevalent, followed by viral and archaeal assigned OTUs. Based on the above-threshold criteria, bacterial taxa displaying a relatively similar abundance in both sequencing runs were identified, beginning at phylum and moving towards genus taxonomic level. These were compared to bacterial phyla and genera reported as discernible in the male reproductive tract by culture or rRNA sequencing [1, 6, 12–15]. In addition, bacterial species reported to be contained within some reagents or from unknown sources that can be accentuated with low biomass samples like the sperm were also considered. They are characterized by water or soil bacteria, often associated with nitrogen fixation, including some, but not all *Streptococcus* species [27–29]. Background genera were then assessed as sample group mean

for the comparison to outlier genera. The Wilcoxon Mann-Whitney  $U$  test in R [23] was used to identify significant differences between sample sets LB + NLB vs outlier and respective OTUs for their contribution. Group means are reported along with the confidence interval set at 95%. Multiple testing (F-test with 100 permutations and Benjamini-Hochberg (BH) adjustment of  $p$  values) in phyloseq was used to evaluate OTUs assigned along the taxonomic lineage for genera exhibiting the most variability between the LB + NLB group and the outlier for their influence over alpha and beta diversity among sample groups.

## Data availability

The microbial sequence data is available at SRA Bioproject PRJNA595621 (<https://www.ncbi.nlm.nih.gov/sra/PRJNA595621>).

## Results

### Microbial sequences in human sperm RNA-Seq

The 85 idiopathic infertile sample cohorts were comprised of the 84 male LB + NLB group and the NLB outlier. Semen parameter group means were within the healthy range (Table 1), with 84% (71 of 85) of the males fulfilling the World Health Organization (WHO) fertile-male characteristics for all four parameters [30] and 96% (82 of 85) for at least three of four parameters. By study completion, 48 couples achieved a successful LB, and 37 did not achieve a live birth NLB. The total number of microbial assigned sequencing reads obtained from the sequencer were 1.0% (Next-seq 500) and 1.4% (Hi-seq 4000) on average. The microbial species-specific sequencing read average ranged from 0.41% (Next-seq 500) to 0.84% (Hi-seq 4000) of the total number of sequencing reads are in accord with that observed by others [31–34]. In comparison, the outlier sample far exceeded the average presenting a total microbial assigned sequencing reads of 32.1% (Next-seq 500) and 41.2% (Hi-seq 4000). This corresponded to 12.99% (Next-seq 500) and 15.78% (Hi-seq 4000) species-specific microbial sequences assigned respectively.

**Table 1** Mean standard semen parameters for the 85 semen samples separated by sample group as compared to World Health Organization (WHO) minimum guidelines: 84 LB + NLB samples and 1 NLB outlier

|                       | Volume (mL)         | Concentration (M/ml)  | Motility (%)         | % normal forms       |
|-----------------------|---------------------|-----------------------|----------------------|----------------------|
| WHO standards         | 1.5                 | 15                    | 32                   | 4                    |
| LB + NLB ( $n = 84$ ) | 3.14 ( $\pm 0.25$ ) | 66.41 ( $\pm 10.24$ ) | 55.71 ( $\pm 2.74$ ) | 22.35 ( $\pm 4.44$ ) |
| outlier ( $n = 1$ )   | 2.70                | 68.00                 | 38.00                | 40.00                |

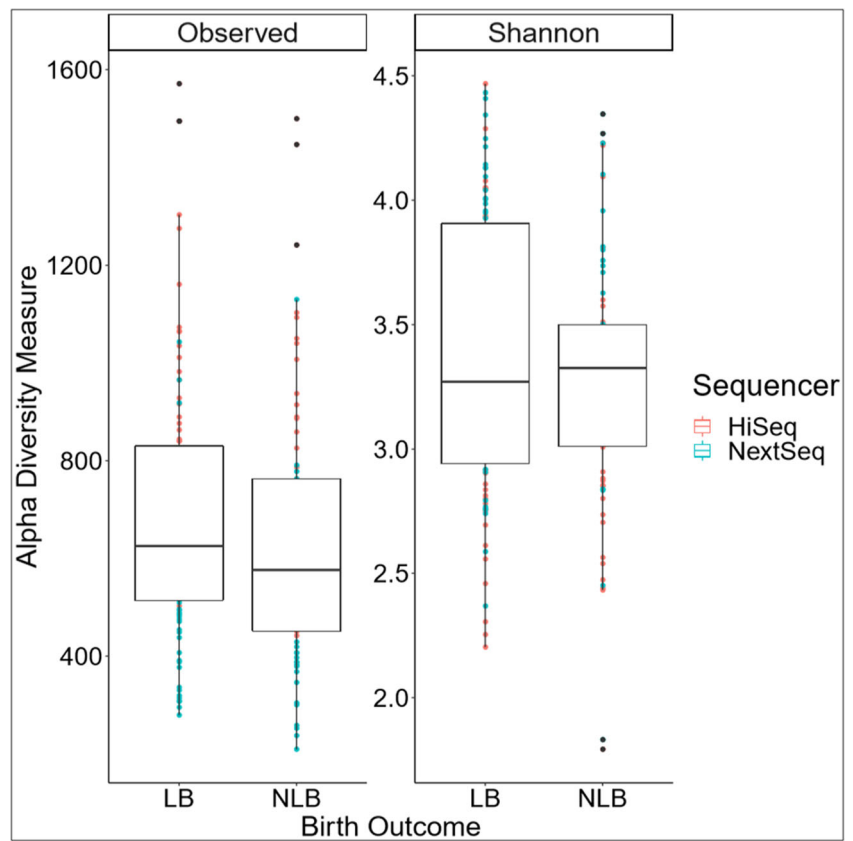
### Estimated microbial richness

Microbial diversity (alpha and beta) within each sample was examined to identify shared microbial profiles between the samples. Microbial richness within each sample (alpha diversity) was evaluated using the Shannon diversity index for each of the samples to determine the number of different microbes (or OTUs) within each sample. Differences between the two sequencers and birth outcome (LB vs NLB) were examined. A statistically significant different variance by sequencer ( $p$  value  $\leq 0.001$ , Online Resource 1) but not birth outcome ( $p$  value = 0.596, Fig. 2) was noted. For the LB group, the mean observed alpha diversity was 689.98 ( $\pm 53.87$ ) with a mean Shannon diversity index of 3.37 ( $\pm 0.11$ ) while the NLB group presented a mean observed alpha diversity of 642.20 ( $\pm 61.25$ ) and Shannon diversity index of 3.28 ( $\pm 0.12$ ). Differences in microbial diversity between samples (beta diversity) were then evaluated for both sequencer and birth outcome. The rarified data was plotted in two dimensions as determined by a stress level of 0.18 (stress type 1 or weak stress) using NMDS and unweighted UniFrac distances. Beta diversity was not statistically different between sequencer. However, the unweighted UniFrac distances ( $p = 9.99e-4$ ) and beta dispersion showed statistical significance (BH adjusted  $p$  value = 0.001) using the R vegan package when birth outcome was considered. To determine whether this difference was due to the presence of the outlier sample within the NLB group, the analysis was repeated following removal of the outlier. This yielded no significant difference between the remaining LB and NLB samples, indicating the difference observed in the original analysis was due to the outlier sample. As such, the 48 LB and 37 NLB samples were combined into a single group, LB + NLB.

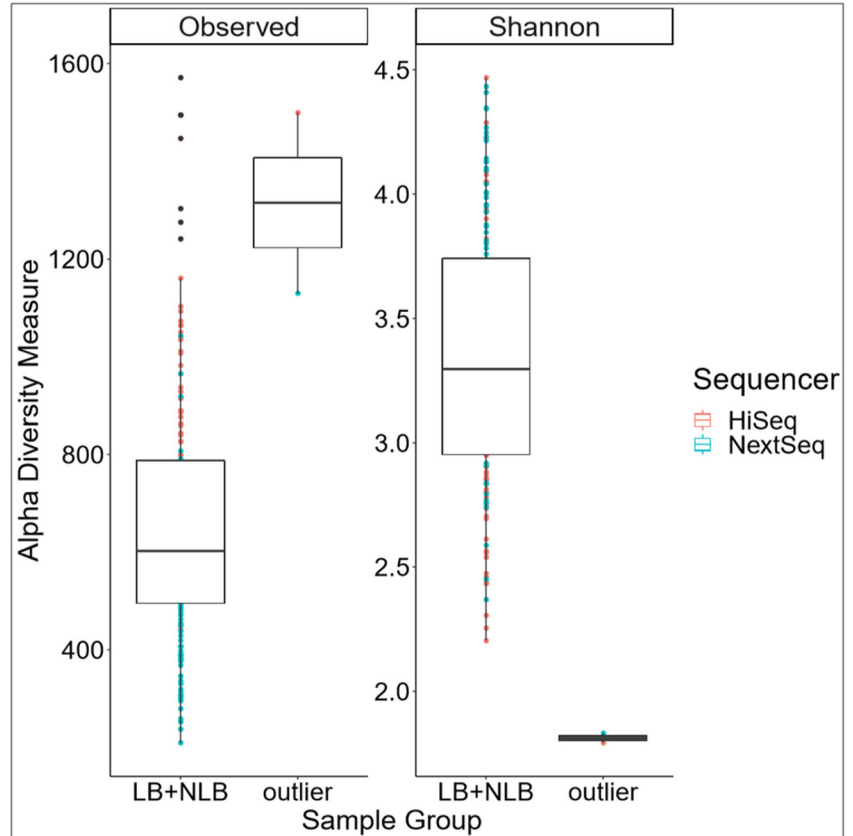
Re-examining the samples based on sample group (LB + NLB) versus NLB outlier showed a dramatic difference in the Shannon diversity index ( $p$  value = 0.015, Fig. 3). This corresponded to major difference in alpha diversity in the LB + NLB group (mean observed = 661.49 ( $\pm 39.42$ ), mean Shannon diversity index = 3.35 ( $\pm 0.08$ ) compared to the outlier (mean observed = 1315.00 ( $\pm 362.59$ ), mean Shannon diversity index = 1.18 ( $\pm 0.04$ )). Significance of the unweighted UniFrac distances ( $p$  value = 9.99e-4) and beta dispersion, microbiome variance ( $p$  value = 9.99e-4), for the sample group was significant. Similar to alpha diversity, this difference in beta diversity remained independent of taxonomic

sample with an over-abundance of microbial sequencing reads. The 95% confidence interval is reported for group LB + NLB only

**Fig. 2** Estimated sample richness (alpha diversity) by birth outcome; live birth (LB) versus no live birth (NLB). The observed microbial richness (Observed) and Shannon diversity index (Shannon) based on birth outcome is not significantly different ( $p = 0.596$ ). Both sequencing runs were considered as separate samples so that each sample has two reported values. The color of the dots indicates the sequencing run, while the NLB black dots indicate an outlier



**Fig. 3** Estimated sample richness (alpha diversity) by sample groups. Group (LB + NLB;  $n = 84$ ) was compared to the NLB outlier ( $n = 1$ ). The observed microbial richness (Observed) and Shannon diversity index (Shannon) of the outlier sample compared to LB + NLB group is significantly different ( $p$  value = 0.015). Both sequencing runs were considered as separate samples so that each sample has two reported values. The color of the dots indicates the sequencing run, while black dots indicate the sample replicate was an outlier



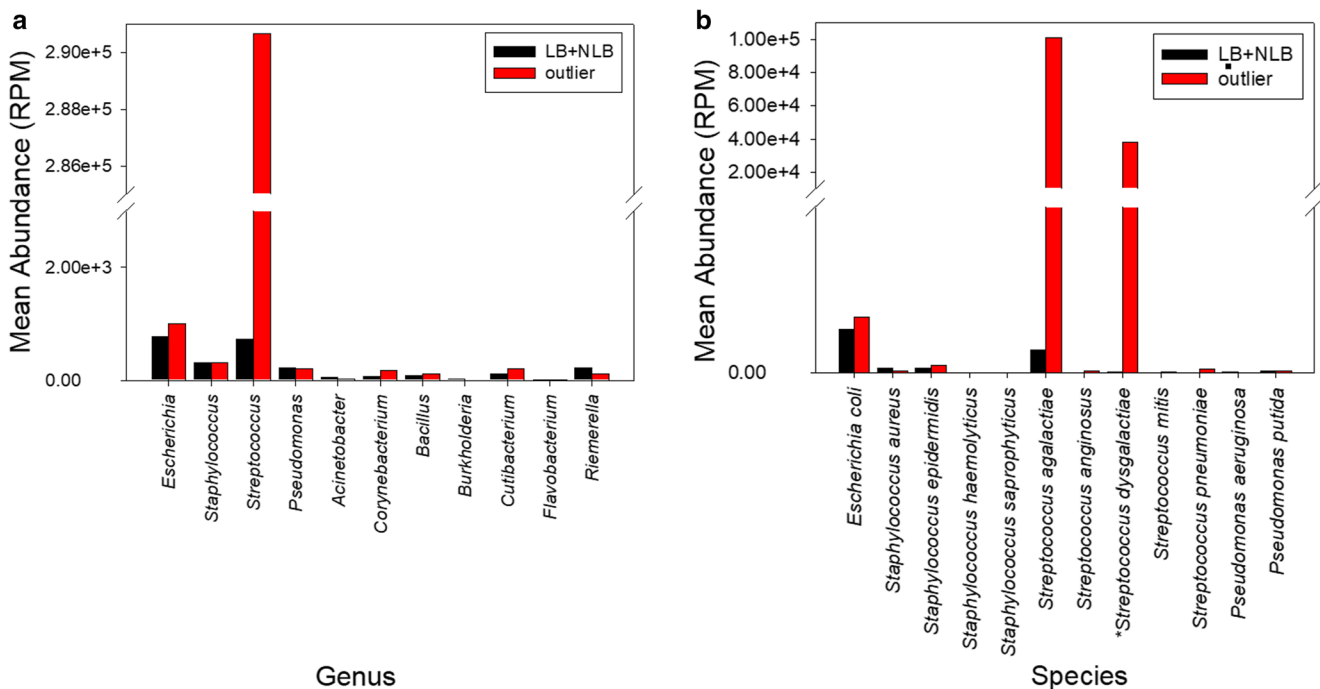
database (Greengenes, NCBI RefSeq) used. The results were consistent with the above Kraken v1 assignments. This indicated that the sample in question was truly an outlier.

## Sperm microbiota

Birth outcome did not impact the alpha diversity. Accordingly, to define the microbiota background, the outlier sample was excluded and the LB + NLB samples were combined as a group ( $n = 84$ ) and replicates used to define the microbiota background (Fig. 4a). Four bacterial phyla: *Firmicutes*, *Proteobacteria*, *Bacteroidetes* and *Actinobacteria* were consistently highlighted in all sperm samples. OTUs from these phyla were assigned to the 11 genera previously identified in semen by PCR and/or 16S rDNA sequencing (Seq) and/or hNGS (human sperm RNA-seq) and/or bacterial culture (Table 2) [1, 6, 12–15]. The genera *Escherichia*, *Staphylococcus*, and *Streptococcus* were observed independent of assay used (culture-PCR-Seq-hNGS) while the genera *Corynebacterium* and *Pseudomonas* were observed using

PCR-Seq-hNGS but not by bacterial culture. In comparison, the genus *Bacillus* was only observed by Seq and hNGS, and the genus *Acinetobacter* was identified by PCR and hNGS. Finally, the genera *Burkholderia*, *Cutibacterium*, *Flavobacterium*, and *Riemerella* were only observed using hNGS (Table 2). The background level was established across the genera observed in all assays within the LB + NLB group. *Corynebacterium*, *Staphylococcus*, and *Streptococcus* were identified as significantly different between the LB + NLB group and the NLB outlier (Fig. 4a; Table 3).

Seventeen bacterial species within the 11 genera were identified in previous reports as present in the male reproductive tract [1, 6, 15]. Of these, 12 species were sequenced and are visualized in Fig. 4b. As shown in Fig. 4b, the NLB outlier OTUs were directly assigned to *S. agalactiae* (9.21 to 11.03% total sequencing reads) and *S. dysgalactiae* (3.37 to 4.25% total sequencing reads) which comprise the majority of sequences assigned to the genus *Streptococcus*. In total, these account for 48.13% of all *Streptococcus* assigned sequencing reads (Fig. 4). The majority of these reads corresponded to OTU3813, assigned to



**Fig. 4** Average group abundance of assigned bacteria. **a** Genera are ordered based on assay identification; *Escherichia*, *Streptococcus*, and *Staphylococcus* (PCR-Seq-hNGS-culture); *Acinetobacter*, *Corynebacterium*, and *Pseudomonas* (PCR-Seq-hNGS); *Bacillus* (Seq-hNGS); *Burkholderia*, *Cutibacterium*, *Flavobacterium*, and *Riemerella* (hNGS). **b** Representative bacterial species are ordered by assay; PCR-Seq-hNGS-culture for *Escherichia coli* (49 OTUs); the *Staphylococcus* species *S. aureus* (31 OTUs), *S. epidermidis* (3 OTUs), *S. haemolyticus* (1 OTU), *S. saprophyticus* (1 OTU); and the *Streptococcus* species *S. agalactiae* (8 OTUs), *S. anginosus* (3 OTUs), *S. dysgalactiae* (95 OTUs), *S. mitis* (1 OTU), *S. pneumoniae* (22 OTUs), and PCR-Seq-hNGS for *Pseudomonas aeruginosa* (14 OTUs) and *Pseudomonas putida* (11 OTUs) known to be discernible within the seminal

microbiome. *S. dysgalactiae* infection of the male reproductive tract has recently been reported and is abundant in the outlier sample only. Background genera were determined using the 84 LB + NLB sample group in which summed genera level OTUs had an RPM > 1 in both sequencing runs for > 60% of samples. A total of 545 OTUs directly assigned a taxonomic rank (to the genera level) or indirectly (to the species level within the given genera) were used for determination of genera abundance (a). Only OTUs assigned directly to the representative species were used to determine species abundance (b). Both sequencing runs were considered as separate samples. Bar color indicates sample group. PCR, 16S rDNA sequencing by PCR; NGS, next-generation sequencing; Seq, 16S rDNA sequencing by NGS; hNGS, human sperm RNA-seq

**Table 2** Background genera identified among sperm samples. The 11 background genera identified by human sperm RNA-seq (hNGS) were compared with those reported as present by 16S rDNA sequencing (PCR, Seq; represents > 1% of sequences) and/or bacterial culture (cultured or not cultured). The data is summarized as present (+) and absent (–) by assay type. Background genera were determined using the LB +

NLB group samples. Both sequencing runs were considered as separate samples. Summed OTUs for each genus had an RPM > 1 in both sequencing runs for > 60% of samples. PCR 16S rDNA sequencing by PCR; NGS next generation sequencing; Seq 16S rDNA sequencing by NGS

| Genera                       | 16S rDNA by PCR <sup>a, b</sup><br>(greater than 1% total reads) | 16S rDNA by NGS (Seq) <sup>c</sup><br>(greater than 1% total reads) | Culture <sup>d, e, f</sup><br>(presence/absence) | Human sperm RNA-seq<br>(hNGS) background |
|------------------------------|--|---|--|--|
| <b>Phylum Firmicutes</b>     |  |   |  |  |
| <i>Bacillus</i>              | –  | +   | –  | +  |
| <i>Staphylococcus</i>        | +  | +   | + <sup>g</sup>                                   | +  |
| <i>Streptococcus</i>         | +  | +   | + <sup>g</sup>                                   | +  |
| <i>Lactobacillus</i>         | + <sup>g</sup>   | +   | –  | –  |
| <i>Finegoldia</i>            | + <sup>g</sup>   | +   | –  | –  |
| <i>Veillonella</i>           | + <sup>g</sup>   | –   | –  | –  |
| <i>Anaerococcus</i>          | + <sup>g</sup>   | –   | –  | –  |
| <i>Peptoniphilus</i>         | + <sup>g</sup>   | –   | –  | –  |
| <i>Aerococcus</i>            | + <sup>g</sup>   | –   | –  | –  |
| <i>Gemella</i>               | + <sup>g</sup>   | –   | –  | –  |
| <i>Granulicatella</i>        | + <sup>g</sup>   | –   | –  | –  |
| <i>Clostridium</i>           | –  | +   | –  | –  |
| <i>Brevibacillus</i>         | –  | +   | –  | –  |
| <i>Enterococcus</i>          | –  | –   | + <sup>g</sup>                                   | –  |
| <b>Phylum Actinobacteria</b> |  |   |  |  |
| <i>Corynebacterium</i>       | +  | +   | –  | +  |
| <i>Gardnerella</i>           | –  | +   | –  | –  |
| <i>Cutibacterium</i>         | –  | –   | –  | +  |
| <i>Atopobium</i>             | + <sup>g</sup>   | –   | –  | –  |
| <b>Phylum Proteobacteria</b> |  |   |  |  |
| <i>Acinetobacter</i>         | + <sup>g</sup>   | –   | –  | +  |
| <i>Pseudomonas</i>           | + <sup>g</sup>   | +   | –  | +  |
| <i>Escherichia</i>           | + <sup>g</sup>   | +   | +  | +  |
| <i>Rhodanobacter</i>         | –  | +   | –  | –  |
| <i>Campylobacter</i>         | –  | +   | –  | –  |
| <i>Haemophilus</i>           | –  | +   | –  | –  |
| <i>Ralstonia</i>             | + <sup>g</sup>   | –   | –  | –  |
| <i>Burkholderia</i>          | –  | –   | –  | +  |
| <i>Pelomonas</i>             | + <sup>g</sup>   | –   | –  | –  |
| <i>Stenotrophomonas</i>      | –  | +   | –  | –  |
| <i>Acidovorax</i>            | + <sup>g</sup>   | –   | –  | –  |
| <i>Bradyrhizobium</i>        | + <sup>g</sup>   | –   | –  | –  |
| <i>Novosphingobium</i>       | + <sup>g</sup>   | –   | –  | –  |
| <i>Diaphorobacter</i>        | + <sup>g</sup>   | –   | –  | –  |
| <i>Morganella</i>            | –  | –   | + <sup>g</sup>                                   | –  |
| <b>Phylum Bacteroidetes</b>  |  |   |  |  |
| <i>Prevotella</i>            | + <sup>g</sup>   | +   | –  | –  |
| <i>Riemerella</i>            | –  | –   | –  | +  |
| <i>Porphyromonas</i>         | + <sup>g</sup>   | –   | –  | –  |
| <i>Flavobacterium</i>        | –  | –   | –  | +  |
| <i>Cloacibacterium</i>       | + <sup>g</sup>   | –   | –  | –  |
| <b>Other phyla</b>           |  |   |  |  |
| <i>Chlamydia</i>             | –  | –   | + <sup>g</sup>                                   | –  |
| <i>Ureaplasma</i>            | + <sup>g</sup>   | –   | + <sup>g</sup>                                   | –  |

<sup>a</sup> Keissling et al. [1]

<sup>b</sup> Hou et al. [12]

<sup>c</sup> Weng et al. [15]

<sup>d</sup> Vetrosky and White [14]

<sup>e</sup> Weidner et al. [13]

<sup>f</sup> Moretti et al. [6]

<sup>g</sup> Genera is described in one but not all corresponding publications

**Table 3** Background genera significantly different in the outlier. The proportion of operational taxonomic units (OTUs) contributing to this difference are reported. Background genera were determined using the LB + NLB group samples. Both sequencing runs were considered as

separate samples. Summed OTUs for each genus had an RPM > 1 in both sequencing runs for >60% of samples. *P* values were determined using the Wilcoxon Mann-Whitney *U* test

| Genus                  | <i>p</i> value | Number of significant OTUs | Total number of OTUs | Proportion of OTUs |
|------------------------|----------------|----------------------------|----------------------|--------------------|
| <i>Corynebacterium</i> | 0.039          | 11                         | 29                   | 0.379              |
| <i>Staphylococcus</i>  | 0.039          | 26                         | 46                   | 0.565              |
| <i>Streptococcus</i>   | 0.015          | 110                        | 220                  | 0.500              |

*S. agalactiae* and OTU3848, assigned to *S. dysgalactiae*, reflective of the alpha and beta diversity difference exhibited by the outlier following multiple testing in phyloseq (Fig. 3). These observations were independent of the taxonomic database assignment method used (Greengenes, NCBI RefSeq).

## Discussion

Human RNA-seq data includes sequencing reads that do not align and map back to the human genome [31]. For example, bacterial, viral, fungal, and archaeal sequences have now been observed in several large datasets including the Cancer Cell Line Encyclopedia (CCLE) [34], 1000 Genome Project [33], SRA database [31], and The Cancer Genome Atlas (TCGA) [32]. As shown above, species-specific microbial-associated sequencing reads associated with sperm are present at frequency of 0.41 to 0.84% of the total. This is in accord with the average microbial load obtained by merging the sequencing data irrespective of sequencing depth or sample type (e.g., intestine, brain, skin) [31–34].

Samples were considered within the context of the LB + NLB samples versus the NLB outlier sample following at least one cycle of TIC, IUI, ICSI, or IVF. No difference in alpha diversity was determined based on birth outcome (Fig. 2); however, one sample was identified as an outlier in the NLB group and was separately tested from the remaining samples, group LB + NLB versus NLB outlier. On one hand, the presence of bacterial components can be traced to their introduction to reagents during manufacture [27–29]. On the other hand, this could reflect the sample itself. This poses a particular challenge for low microbial biomass samples such as semen, as the small amount of starting material may be overpowered by these reagent bacteria [29]. Computationally, this noise can be removed from analysis. Again, as illustrated in Fig. 3, the Shannon diversity index suggested the outlier's microbiome was vastly different from all other samples.

The most dramatic difference between the outlier and the remaining 84 samples was *Streptococcus* (Fig. 4a). This microbial difference was identified irrespective of microbial reference

genome (Greengenes v13\_8 99%—16S rRNA, NCBI RefSeq release 95—all bacterial genomes, NCBI RefSeq release 76—all bacterial, viral, archaeal genomes) or the specific algorithm used for taxonomic assignment. Multiple testing of the contribution of the OTU towards differences in microbial composition between all 85 samples was performed for the taxonomic lineages of each of the 11 genera. The OTUs yielding significance were for *S. agalactiae* (OTU3813) and *S. dysgalactiae* (OTU3848). When sub-setting along the taxonomic lineage within the phylum *Firmicutes* (BH adjusted *p* value = 0.044), class *Bacilli* (BH adjusted *p* values = 0.033), order *Lactobacillales* (BH adjusted *p* values = 0.015), family *Streptococcaceae* (BH adjusted *p* values = 0.011) and genus *Streptococcus* (BH adjusted *p* values = 0.009) were evident. Considering abundance and contribution towards alpha and beta diversity, the outlier was likely colonized with *S. agalactiae* and potentially with *S. dysgalactiae*.

Male genital tract infection by *S. dysgalactiae*, but not colonization, has been reported [35, 36]. However, the frequency of *S. agalactiae* transient colonization in college-aged males of 20%, as determined in urine (first of the day), swabs from the bottom of the rectum, mouth, and interdigital spaces is half that observed in females [37–39]. Sexual partners share the same isolates of *S. agalactiae* when colonized indicating these bacteria are sexually transmitted [37, 40]. As partners will share isolates of *S. agalactiae*, the female partner of the outlier may have been colonized. *S. agalactiae* infection in women can resolve as both stillbirth or pregnancy loss [41]. This may reflect that *S. agalactiae* actively forms biofilms and increases inflammation via leukocyte number [39, 42, 43]. Interestingly, this couple did not achieve a pregnancy after three ICSI cycles. While *S. agalactiae* colonization has been detected in sperm from couples seeking reproductive care [6], there was no difference in standard semen parameters (Table 1), suggesting that the presence of bacteria (specific or in groups) did not impact semen parameters [2, 12]. This remains to be reconciled with the data of others that suggest that certain bacterial infection(s) alter semen parameters including sperm count and the degree of necrosis [1, 5, 6, 42] when larger studies are undertaken.



## Conclusions

Un-biased sequencing of microorganisms (bacteria, viruses, archaea) has the potential to provide enhanced sensitivity and specificity as compared to targeted-culturing, or -rDNA/rRNA sequencing. This information can be recovered from the data typically cast aside as part of RNA-seq unmapped reads. Using this in hand resource provided the opportunity to identify a sample presenting a noticeable level of *S. agalactiae* RNA-seq reads. While it cannot be excluded, it is unlikely that the presence of *S. agalactiae* is due to contamination introduced during processing considering the mode of transmission and use of personal protection equipment as part of standard clinical practice. Given the recent increase and severity of *Streptococcus (agalactiae)* infection as well as others in adults, neonates, and newborns [40], in addition to providing fertility status [16], non-targeted human sperm RNA-seq data may also prove useful as a diagnostic for microbial status. Perhaps it may become part of routine sample quality assessment as we move towards personalized care.

**Acknowledgments** The use of samples from the Eunice Kennedy Silver National Institute of Child Health and Human Development (Assessment of Multiple Intrauterine Gestations from Ovarian Stimulation (AMIGOS) study are gratefully acknowledged. Support from the Postdoctoral Recruiting Fellowship from Wayne State University to GMS and from the Charlotte B. Failing Professorship to SAK and the Wayne State University OPR Grants Boost award to SAK is gratefully acknowledged. The authors would like to thank Dr. Kevin Theis, Department of Biochemistry, Microbiology and Immunology and Obstetrics and Gynecology, Wayne State University School of Medicine for his thoughtful review of the manuscript. Merck KGaA Darmstadt, Germany reviewed the manuscript for medical accuracy only before journal submission. The authors are fully responsible for the content of this manuscript, and the views and opinions described in the publication reflect solely those of the authors. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NICHD or NIH.

**Author's contributions** SAK and GMS were responsible for study design. Sample acquisition was by SM, CL, and JRP. RNA isolation and sequencing were performed by RG. Sequence alignment, analysis, and manuscript preparation was performed by GMS. All the authors contributed to the editing of the manuscript.

**Funding information** This study was funded by a 2016 Grant for Fertility Innovation (25RJY1) from Merck KGaA Darmstadt, Germany and a National Institute of Health (NIH)/National Institute of Environmental Health Sciences (NIEHS) Grant (R01-ES028298).

## Compliance with ethical standards

**Conflict of interest** Stephen Krawetz has received grants from EMD Serono and GFI Fertility Innovation. Stephen Krawetz reports honoraria from Taylor and Francis and KINBRE.

## References

- Kiessling AA, Desmarais BM, Yin H-Z, Loverde J, Eyre RC. Detection and identification of bacterial DNA in semen. *Fertil Steril*. 2008;90(5):1744–56. <https://doi.org/10.1016/j.fertnstert.2007.08.083>.
- Franasiak JM, Scott RT. Reproductive tract microbiome in assisted reproductive technologies. *Fertil Steril*. 2015;104(6):1364–71. <https://doi.org/10.1016/j.fertnstert.2015.10.012>.
- Franasiak JM, Scott RT. Introduction: microbiome in human reproduction. *Fertil Steril*. 2015;104(6):1341–3. <https://doi.org/10.1016/j.fertnstert.2015.10.021>.
- Mändar R. Microbiota of male genital tract: impact on the health of man and his partner. *Pharmacol Res*. 2013;69(1):32–41. <https://doi.org/10.1016/j.phrs.2012.10.019>.
- Qing L, Song Q-X, Feng J-L, Li H-Y, Liu G, Jiang H-H. Prevalence of Chlamydia trachomatis, *Neisseria gonorrhoeae*, Mycoplasma genitalium and Ureaplasma urealyticum infections using a novel isothermal simultaneous RNA amplification testing method in infertile males. *Ann Clin Microbiol Antimicrob*. 2017;16(1):45. <https://doi.org/10.1186/s12941-017-0220-2>.
- Moretti E, Capitani S, Figura N, Pammolli A, Federico MG, Giannerini V, et al. The presence of bacteria species in semen and sperm quality. *J Assist Reprod Genet*. 2009;26(1):47–56. <https://doi.org/10.1007/s10815-008-9283-5>.
- Clarridge JE 3rd. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev*. 2004;17(4):840–62. <https://doi.org/10.1128/CMR.17.4.840-862.2004>.
- Chen M, Cai L-Y, Kanno N, Kato T, Lu J, Jin F, et al. Detection of human herpesviruses (HHVs) in semen of human male infertile patients. *J Reprod Dev*. 2013;59(5):457–62. <https://doi.org/10.1262/jrd.2013-023>.
- Cordeiro CN, Bano R, Washington Cross CI, Segars JH. Zika virus and assisted reproduction. *Curr Opin Obstet Gynecol*. 2017;29(3):175–9. <https://doi.org/10.1097/GCO.0000000000000366>.
- Lai YM, Yang F-P, Pao CC. Human papillomavirus deoxyribonucleic acid and ribonucleic acid in seminal plasma and sperm cells\*\*Supported by research grant NSC83-0412-B182-001 from National Science Council of Republic of China and by medical research grant CMRP-343 from Chang Gung College of Medicine and Technology and Memorial Hospital, Taipei, Taiwan, Republic of China, both awarded to C.C.P. *Fertil Steril*. 1996;65(5):1026–30. [https://doi.org/10.1016/S0015-0282\(16\)58281-8](https://doi.org/10.1016/S0015-0282(16)58281-8).
- Diemer T, Huwe P, Ludwig M, Schroeder-Printzen I, Michelmann HW, Schiefer HG, et al. Influence of autogenous leucocytes and Escherichia coli on sperm motility parameters in vitro. *Andrologia*. 2003;35(2):100–5. <https://doi.org/10.1046/j.1439-0272.2003.00523.x>.
- Hou D, Zhou X, Zhong X, Settles ML, Herring J, Wang L, et al. Microbiota of the seminal fluid from healthy and infertile men. *Fertil Steril*. 2013;100(5):1261–9.e3. <https://doi.org/10.1016/j.fertnstert.2013.07.1991>.
- Weidner W, Jantos C, Schiefer HG, Haidl G, Friedrich HJ. Semen parameters in men with and without proven chronic prostatitis. *Arch Androl*. 1991;26(3):173–83. <https://doi.org/10.3109/01485019108987640>.
- Vetrosky D, White GL Jr. Prostatitis. *Lippincotts Prim Care Pract*. 1997;1(4):437–41.
- Weng S-L, Chiu C-M, Lin F-M, Huang W-C, Liang C, Yang T, et al. Bacterial communities in semen from men of infertile couples: metagenomic sequencing reveals relationships of seminal

- microbiota to semen quality. PLoS One. 2014;9(10):e110152-e. <https://doi.org/10.1371/journal.pone.0110152>.
16. Jodar M, Sandler E, Moskovtsev SI, Librach CL, Goodrich R, Swanson S, et al. Absence of sperm RNA elements correlates with idiopathic male infertility. Sci Transl Med. 2015;7(295):295re6. <https://doi.org/10.1126/scitranslmed.aab1287>.
  17. Goodrich R, Johnson G, Krawetz SA. The preparation of human spermatozoal RNA for clinical analysis. Arch Androl. 2007;53(3):161–7. <https://doi.org/10.1080/01485010701216526>.
  18. Goodrich RJ, Anton E, Krawetz SA. Isolating mRNA and small noncoding RNAs from human sperm. Methods Mol Biol. 2013;927:385–96. [https://doi.org/10.1007/978-1-62703-038-0\\_33](https://doi.org/10.1007/978-1-62703-038-0_33).
  19. Mao S, Goodrich RJ, Hauser R, Schrader SM, Chen Z, Krawetz SA. Evaluation of the effectiveness of semen storage and sperm purification methods for spermatozoa transcript profiling. Syst Biol Reprod Med. 2013;59(5):287–95. <https://doi.org/10.3109/19396368.2013.817626>.
  20. Estill MS, Hauser R, Krawetz SA. RNA element discovery from germ cell to blastocyst. Nucleic Acids Res. 2018. <https://doi.org/10.1093/nar/gky1223>.
  21. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60. <https://doi.org/10.1038/nmeth.3317> <https://www.nature.com/articles/nmeth.3317#supplementary-information>.
  22. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. <https://doi.org/10.1186/gb-2014-15-3-r46>.
  23. Team RC. R: a language and environment for statistical computing. 2013.
  24. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One. 2013;8(4):e61217. <https://doi.org/10.1371/journal.pone.0061217>.
  25. Shen W, Xiong J. TaxonKit: a cross-platform and efficient NCBI taxonomy toolkit. bioRxiv.org. 2019;513523. <https://doi.org/10.1101/513523>.
  26. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2019.
  27. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5(1):52. <https://doi.org/10.1186/s40168-017-0267-5>.
  28. Barton H, Taylor N, Lubbers B, Pemberton A. DNA extraction from low-biomass carbonate rock: an improved method with reduced contamination and the low-biomass contaminant database. J Microbiol Methods. 2006;66(1):21–31.
  29. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12(1):87. <https://doi.org/10.1186/s12915-014-0087-z>.
  30. Cooper TG, Noonan E, von Eckardstein S, Auger J, Baker HW, Behre HM, et al. World Health Organization reference values for human semen characteristics. Hum Reprod Update. 2010;16(3):231–45. <https://doi.org/10.1093/humupd/dmp048>.
  31. Simon LM, Karg S, Westermann AJ, Engel M, Elbehery AHA, Hense B, et al. MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. Gigascience. 2018;7(6):giy070. <https://doi.org/10.1093/gigascience/gyi070>.
  32. Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, et al. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog. 2014;10(11):e1004437. <https://doi.org/10.1371/journal.ppat.1004437>.
  33. Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. Large scale comparison of non-human sequences in human sequencing data. Genomics. 2014;104(6 Pt B):453–8. <https://doi.org/10.1016/j.ygeno.2014.08.009>.
  34. Uphoff CC, Pommerenke C, Denkmann SA, Drexler HG. Screening human cell lines for viral infections applying RNA-Seq data analysis. PLoS One. 2019;14(1):e0210404-e. <https://doi.org/10.1371/journal.pone.0210404>.
  35. Anantha RV, Kasper KJ, Patterson KG, Zeppa JJ, Delpont J, McCormick JK. Fournier's gangrene of the penis caused by Streptococcus dysgalactiae subspecies equisimilis: case report and incidence study in a tertiary-care hospital. BMC Infect Dis. 2013;13(1):381. <https://doi.org/10.1186/1471-2334-13-381>.
  36. Takakura S, Gibo K, Takayama Y, Shiiki S, Narita M. Clinical characteristics of *Streptococcus pyogenes*, *Streptococcus agalactiae* and *Streptococcus dysgalactiae* subsp. *equisimilis* bacteremia in adults: a 15-year retrospective study at a major teaching hospital in Okinawa, Japan. Open Forum Infect Dis. 2017;4(suppl\_1):S559-S. <https://doi.org/10.1093/ofid/ofx163.1457>.
  37. Bliss SJ, Pearlman MD, Marrs CF, Tallman P, Manning SD, Foxman B, et al. Group B Streptococcus colonization in male and nonpregnant female university students: a cross-sectional prevalence study. Clin Infect Dis. 2002;34(2):184–90. <https://doi.org/10.1086/338258>.
  38. Foxman B, Gillespie BW, Manning SD, Marrs CF. Risk factors for group B streptococcal colonization: potential for different transmission systems by capsular type. Ann Epidemiol. 2007;17(11):854–62. <https://doi.org/10.1016/j.annepidem.2007.05.014>.
  39. D'Urzo N, Martinelli M, Pezzicoli A, De Cesare V, Pinto V, Margarit I, et al. Acidic pH strongly enhances in vitro biofilm formation by a subset of Hypervirulent ST-17 Streptococcus agalactiae strains. Appl Environ Microbiol. 2014;80(7):2176–85. <https://doi.org/10.1128/aem.03627-13>.
  40. Furfaro LL, Chang BJ, Payne MS. Perinatal Streptococcus agalactiae epidemiology and surveillance targets. Clin Microbiol Rev. 2018;31(4):e00049–18. <https://doi.org/10.1128/cmr.00049-18>.
  41. Parks T, Barrett L, Jones N. Invasive streptococcal disease: a review for clinicians. Br Med Bull. 2015;115(1):77–89. <https://doi.org/10.1093/bmb/ldv027>.
  42. Parida R, Samanta L. In silico analysis of candidate proteins sharing homology with Streptococcus agalactiae proteins and their role in male infertility. Syst Biol Reprod Med. 2017;63(1):15–28. <https://doi.org/10.1080/19396368.2016.1243741>.
  43. Rosini R, Margarit I. Biofilm formation by *Streptococcus agalactiae*: influence of environmental conditions and implicated virulence factors. Front Cell Infect Microbiol. 2015;5(6). <https://doi.org/10.3389/fcimb.2015.00006>.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.