Genomics-Based Identification of Microorganisms in Human Ocular Body Fluid

Philipp Kirstahler[1], Søren Solborg Bjerrum[2], Alice Friis-Møller[3], Morten la Cour[2], Frank M. Aarestrup[1], Henrik Westh[3,4], and Sünje Johanna Pamp[1,*]

[1]Research Group for Genomic Epidemiology, Technical University of Denmark, Kgs. Lyngby, Denmark
[2]Department of Ophthalmology, Rigshospitalet, Glostrup, Denmark
[3]Department of Clinical Microbiology, Hvidovre Hospital, Copenhagen, Denmark
[4]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

*corresponding author sjpa@food.dtu.dk

ABSTRACT

Advances in genomics have the potential to revolutionize clinical diagnostics and management of infectious diseases. Here, we examine the microbiome of vitreous (intraocular body fluid) from patients who developed endophthalmitis following cataract surgery or intravitreal injection. Endophthalmitis is an inflammation of the intraocular cavity, usually due to microbial infection and can lead to the loss of vision. As controls, we included vitreous from endophthalmitis-negative patients, basal salt solution used during vitrectomy, and DNA extraction blanks. We compared two DNA isolation procedures and found that an ultraclean production of reagents appeared to reduce background DNA in these low microbial biomass samples. We designed a metagenomics workflow with filtering steps to reduce DNA sequences originating from: i) human hosts, ii) ambiguousness/contaminants in public microbial reference genomes, and iii) the environment. Our metagenomic read classification revealed in nearly all cases the same microorganism than was determined in cultivation- and mass spectrometry-based analyses. For some patients, we identified the sequence type of the microorganism and antibiotic resistance genes through analyses of whole genome sequence (WGS) assemblies of isolates, and metagenomic assemblies. Together, we conclude that genomics-based analyses of human ocular body fluid specimens can provide actionable information relevant to infectious disease management.

Introduction

Genomics-based analyses of patient specimens have the potential to provide actionable information that could facilitate faster and possibly more precise clinical diagnoses and guide treatment strategies in infectious diseases. A medical condition where a faster and more precise diagnosis could make a difference in clinical outcomes is endophthalmitis. Endophthalmitis is an acute intraocular inflammation that can lead to a permanent loss of vision. It often develops in response to microorganisms (usually bacteria and fungi) that enter the eye following eye surgery such as cataract surgery and intravitreal injection. The treatment strategy as well as visual outcome depends in part on the identity of the causative agents. For example, endophthalmitis cases involving coagulase-negative staphylococci have a better prognosis than cases involving enterococci or streptococci [1]. Often, the involving bacteria appear to originate from the patients' own microbiota, but may also be introduced through contaminated solutions or instruments used during eye surgery [2,3]. Rapid determination of the identity of the causing agents and their antimicrobial resistance profiles would facilitate the application of precise treatments and reduce blindness.

Cataract is a condition in which the lens of the eye becomes progressively opaque and is one of the major causes of reversible visual loss. It is estimated that every year 10 million cataract surgeries are performed around the world [4]. The risk of endophthalmitis after cataract surgery is 1.4-4 per 10,000 cataract surgeries in the US and Denmark, and can be higher in other countries [1,5,6]. About 1/3 of the eyes with endophthalmitis in cataract patients remain blind after treatment [7].

Intravitreal injection with anti-vascular endothelial growth factor (anti-VEGF) has revolutionized the treatment of wet age-related macular degeneration, as well as diabetic maculopathy, and retinal vein occlusions during the last decade. It is the fastest growing procedure in ophthalmology and it was estimated that the number of intravitreal injections in the US would reach nearly 6 million in 2016 [8]. The risk of endophthalmitis after intravitreal injection is approximately 4.9 per 10,000 intravitreal injections [9].

The diagnosis and treatment of endophthalmitis is performed by vitrectomy surgery. The vitreous body of the eye, which is the immobile gel-like fluid that occupies the space between the lens and retina, is aspirated and replaced by basal salt solution. Antibiotics are being injected into the body of the eye to limit the occurrence of infections. The vitreous is often examined for infectious agents in the clinical laboratory using cultivation-based techniques.

In the clinical setting it is challenging to distinguish between infectious endophthalmitis and non-infectious ("sterile") endophthalmitis. Studies have shown that the proportion of culture-positive cases can be as low as 39% after cataract surgery and 52% after intravitreal injection [9,10]. Polymerase chain reaction can increase the rate of identifying the microorganisms by 20% [11], but in many endophthalmitis cases a causative agent cannot be identified. It is also unclear, whether the vitreous in endophthalmitis may contain multiple microorganisms that are not all being detected with the current methods. Furthermore, from a clinical perspective it is of importance to have a method that facilitates the identification of the cases of non-infectious endophthalmitis. Non-infectious endophthalmitis can present as a variant of TASS (toxic anterior segment syndrome), and these patients may benefit from steroid instead of antibiotic treatment to obtain a better visual outcome [12].

Genomics approaches have the potential to revolutionize clinical diagnostic and therapeutic approaches in particular in the area of infectious diseases. Using shotgun metagenomic sequencing, a range of microorganisms and possible causing agents (e.g. bacteria, archaea, fungi, protozoa, viruses) can be identified [13,14]. In addition, upon cultivation-based isolation of microorganisms from the patient specimen, these can be subjected to whole genome sequencing (WGS) and *in silico*-determination of their taxonomic affiliation, phylogenetic relationships, potential antibiotic resistance genes, and virulence-associated genes [15,16].

Here, we perform metagenomic sequencing on vitreous specimens obtained from patients with endophthalmitis, and a range of control samples. We evaluate two DNA isolation procedures for vitreous, and describe a bioinformatics workflow for data analysis and identification of potential infectious agents. The workflow includes *in silico* filtering steps for the removal of human DNA sequences, ambiguous and contaminant sequences in reference genomes from public repositories, and background DNA detected in control samples. We compare the metagenomics-based results with the results from the routine clinical cultivation- and mass spectrometry-based analysis, as well as to WGS-based identification of isolates obtained from the vitreous. Our findings suggest that metagenomics analysis together with WGS-based analysis is suitable for the identification of the potential infectious agents from human ocular body fluid, and could guide therapeutic strategies including targeted antimicrobial therapy and the choice of steroids.

Results

Study design and metagenomic sequencing

To evaluate the use of shotgun metagenomic sequencing for the identification of potential disease-causing agents in postoperative endophthalmitis, we collected vitreous during vitrectomy from 14 patients with endophthalmitis (7 post cataract surgery, 7 post intravitreal injection) (Figure 1, Supplementary Table S1). As control, we obtained vitreous from 7 patients without endophthalmitis during macula hole surgery. Additional controls included 6 balanced salt solution (BSS) aliquots, of which 3 originated from individual bottles (BSS-B), and 3 from the vitrectomy BSS infusion lines (to be inserted into the eye) after the bottle had been connected to the vitrectomy system (BSS-S) (Figure 1). As there exist no standard procedure for the isolation of DNA from vitreous, we examined two procedures using the QIAamp DNA Mini Kit (QIA) and QIAamp UCP Pathogen Mini kit (UCP), and 4 extraction (blank) controls were included per kit (Figure 1).

The 62 samples were sequenced using Illumina MiSeq sequencing technology and a total of 90.6 million raw read-pairs were obtained. The average number of read-pairs after quality control for the endophthalmitis patients were 2.1/2.3 million read-pairs (QIA/UCP), and for the endophthalmitis-negative vitreous samples 1.0/0.6 million read-pairs (QIA/UCP). The average number of read-pairs for the BSS samples were 52,899/6,067 (QIA/UCP), and for the DNA extraction controls 20,931/3,134 (QIA/UCP). Overall, more read-pairs were obtained on average for the control samples when extracted with the QIA kit, while more read-pairs were obtained for the vitreous from the endophthalmitis patients when extracted with the UCP kit (Supplementary Fig. S1, Supplementary Table S2).

Identification of human-affiliated DNA sequences

In a first-pass analysis, in which we mapped the reads against a set of reference genomes, we detected a high number of reads affiliated with human DNA sequences, which was anticipated in particular in the endophthalmitis cases that can experience an infiltration of immune cells into the

131 vitreous chamber. Hence, we implemented a 2-step filtering process to remove the reads affiliated
132 with human genome sequences (Figure 2). In the first step we removed the reads that mapped to the
133 human reference genome (GRCh8.p10). Due to the genetic individuality of humans some reads
134 might not map to this reference genome, and therefore we removed in a second step all reads that
135 aligned to any human DNA sequence entry in the NCBI nt database (Supplementary Fig. S2,
136 Supplementary Table S2).
137
138 **Identification of ambiguous and contaminant DNA sequences in genomes from public**
139 **repositories**
140 In the initial first-pass analysis involving mapping of reads against reference genomes, we observed
141 that some genomes recruited particular high numbers of reads. These included *Hammondia*
142 *hammondi* strain H.H.34, *Alcanivorax hongdengensis* Strain A-11-3, *Toxoplasma gondii* ME49,
143 and *Arthrobacter* sp. Soil736. Upon inspection of these genomes we found that the reads mapped
144 only to specific genome sequence fragments such as short contigs and scaffolds (Supplementary
145 Fig. S3). To examine why specific contigs and scaffolds recruited high numbers of reads, we
146 aligned these against the nucleotide collection nt (NCBI). We found that the Top10 matches for
147 most of these contigs and scaffolds included several human DNA sequence entries that are not part
148 of the human reference genome GRCh8.p10 (Supplementary Table S3). While a few scaffolds of
149 *Hammondia hammondi* strain H.H.34 aligned with human DNA sequences (e.g. scaffold
150 NW_008644893.1), many aligned to *Bradyrhizobium* spp. genomes in the nt database
151 (Supplementary Table S3).
152
153 **Construction of a curated microbial genome database**
154 Our analysis suggested that some microbial reference genomes contain ambiguous/contaminant
155 sequences and we aimed at constructing a curated microbial genome database, devoid of these
156 sequences. Removing these sequences could reduce the number of false positive hits that are the
157 result of either contaminant sequences in the (incomplete) genome assemblies, or because highly
158 similar sequence regions naturally exist across genera that result in the classification of reads to a
159 different genus. We examined 5754 microbial reference and representative genomes (archaea,
160 bacteria, fungi, protozoa) (Supplementary Table S4) and aligned all sequences ≤10 kb against the
161 nucleotide collection nt (for a detailed description, see Supplementary Methods). A total of 70,478
162 ambiguous sequences (contigs and scaffolds) were identified, of which the majority were detected
163 in incomplete microbial genomes. A total of 62% of all incomplete microbial genomes had
164 sequencs flagged as ambiguous (range: 1 - 10,590; average: 28 sequence fragments). Ambiguous
165 sequences were identified in 43% of all bacterial and 72% of all protozoan genomes, and on
166 average comprised 0.36% and 0.84% of the total genome sequence, respectively (Table 1, and
167 https://figshare.com/s/045b1252bd7555b50ef0, https://figshare.com/s/c42158cdee23f25489cd). The
168 ambiguous sequences were removed and the resulting reference microbial genome database
169 contained a total of 5,755 genomes with 34 Tb (including 3.1 Tb for the human genome).
170
171 **Identification of contaminant (environmental background) DNA sequences in samples**
172 From the sequencing of DNA extraction (blank) control samples we obtained sequencing data,
173 albeit at a lower frequency compared to the patient specimens (Supplementary Fig. S1). The *in*
174 *silico* identification and removal of background DNA sequences are of critical importance,
175 particularly from specimens where the potential infectious agent may be present in low abundance.
176 We carefully examined the eight DNA extraction control samples and devise a list of the most
177 abundant and frequent environmental contaminant taxa in these samples (Supplementary Table S5,
178 Supplementary Fig. S4). We did not include taxa in the list that were occasionally observed in

179 endophthalmitis-positive patients and that were detected at a higher abundance in these samples
180 than in the respective DNA extraction controls. These non-contaminant taxa include *Enterococcus*
181 *faecalis*, *Escherichia coli*, *Micrococcus luteus*, *Staphylococcus aureus*, and *Staphylococcus*
182 *epidermidis* (Figure 3). The microbial composition patterns in the DNA extraction control samples
183 appeared to be influenced by the choice of DNA isolation kit, the day of DNA extraction, and
184 sequencing run (Supplementary Fig. S4). The contaminant taxa (Supplementary Table S5) were
185 removed from the datasets of all endophthalmitis patients.
186
187 **The microbial composition in endophthalmitis-negative and basal salt solution samples is**
188 **similar to DNA extraction controls**
189 The contaminant taxa that were identified in the DNA extraction controls were often present at
190 similar abundances in the endophthalmitis-negative (vitreous control) and basal salt solution
191 samples (Figure 3, Supplementary Fig. S5). We found certain taxa to be specific for the DNA
192 isolation method (QIA or UCP) in round C of DNA extractions (Supplementary Fig. S5,
193 Supplementary Table S2). Samples processed using the QIA method contained *Pseudomonas* spp.,
194 *Acinetobacter* spp., and *Janthinobacterium* spp. among others, and samples processed with the UCP
195 method included mainly *Bradyrhizobium* spp. Other organisms appeared to be present across all
196 samples (Supplementary Fig. S5). For example, *Cutibacterium acnes* and *Propionibacterium*
197 *humerusii* were detected in most samples and they might represent environmental bacteria
198 originating from the staff handling the samples or fomites such as the laboratory equipment and
199 supplies.
200
201 **Microorganisms in endophthalmitis-positive patients as determined by metagenomics**
202 For 13 out of 14 endophthalmitis patients a dominant microorganism was identified in the vitreous
203 (for all UCP-extracted, and most QIA-extracted specimens) using the read classification approach
204 (Figure 4). These organisms included *Staphylococcus epidermidis* (six patients), *Enterococcus*
205 *faecalis* (two patients), *Serratia marcescens* (one patient), *Paenibacillus* spp. (one patient), and
206 *Staphylococcus hominis* (one patient). For two additional patients, *Commamonas testosteronii* and
207 *Escherichia coli*, or *Caulobacter* spp. were identified as the most dominant organisms respectively
208 (C1, I7), however, these were only represented by <25 reads. In one patient (C5), a number of
209 different organisms were identified, most dominantly *E. coli* in the UCP-extracted specimen (>3000
210 reads), and *Moraxella catarrhalis* (11 reads) in the QIA-extracted specimen (Figure 4, and
211 https://figshare.com/s/5feabfad1d8c495bf7a3). In the seven patients that contracted endophthalmitis
212 following cataract surgery, the most frequent bacteria were *Enterococcus faecalis* (two patients),
213 *Staphylococcus epidermidis* (two patients), and *Serratia marcescens* (one patient). In the seven
214 patients with endophthalmitis following intravitreal injection, the most frequent bacteria were
215 *Staphylococcus epidermidis* (four patients), *Paenibacillus* spp. (one patient), and *Staphylococcus*
216 *hominis* (one patient) (Figure 4). Overall, potential causing agents where identified with 58 reads
217 (*Paenibacillus* spp.) as a lower bound in patient I2, and 2,999,838 reads as the highest detected read
218 count (*Staphylococcus epidermidis*) in patient I4.
219
220 In addition to the read classification approach, we also constructed metagenomic assemblies for the
221 individual samples and characterized these according to a number of taxonomic and functional
222 categories, including bacterial species affiliation, sequence type, genomic MLST, resistance genes,
223 virulence-associated genes, and plasmids. For the three patients for whom we obtained high
224 numbers of classified reads using the read classification approach described above (C6, C7, I4)
225 (Figure 4), we obtained information in nearly all categories using the metagenomics assembly
226 approach (Supplementary Table S6). The taxonomic information we obtained using the

227 metagenomic read classification compared to the metagenomic assembly approach was in
228 agreement in all cases. Furthermore, using metagenomics assembly analysis we detected a number
229 of resistance genes in the specimens for which we also obtained taxonomic information. In addition,
230 we detected a streptogramin B resistance gene (vat(B)) in sample C3_UCP, an aminoglycoside
231 resistance gene (aadD) in sample C7_QIA, and a Col plasmid origin of replication in sample
232 C5_UCP. Of note, five out of the seven total samples, for which we obtained information using the
233 metagenomics assembly approach, were ocular body fluid samples that had been processed using
234 the UCP DNA isolation protocol (Supplementary Table S6).
235
236 **Bacterial isolates from endophthalmitis patients have in most cases the same identity as**
237 **the most abundant organism determined by metagenomics**
238 At the hospital microbiology laboratory, bacteria could be isolated from the vitreous right after
239 vitrectomy for 12 out of 14 patients. The identity of the isolates was determined by MALDI-TOF
240 mass spectrometry (MS), and in nine cases the same agent was identified as in the metagenomic
241 analysis (Figure 4). In addition, *Micrococcus luteus* was isolated from patient C5, and while 9/45
242 (QIA/UCP) reads classified as *Micrococcus luteus* were also identified using metagenomics, this
243 organism was not the most abundant one identified using the metagenomics sequencing-based
244 method (Figure 4). Using the cultivation-based method a *Bacillus* sp. (Order: Bacillales) was
245 determined for patient I2, and reads classified as *Paenibacillus* spp. (Order: Bacillales) were
246 identified using the metagenomics analysis. A *Staphylococcus aureus* culture was obtained in the
247 hospital for patient I3, and *S. aureus* was also represented with 22/18 (QIA/UCP) reads in the
248 metagenomics analysis in this patient, even though *S. epidermidis* was the most abundant organisms
249 identified using this approach (150(QIA)/162(UCP) reads). In the two cases for which the culture-
250 based approach was negative (C1, I7), only fewer than 25 reads were classified using the
251 metagenomics approach.
252
253 At the research laboratory, additional isolates were recultured from the frozen vitreous samples for
254 six patients. Different colony morphotypes on the agar plates were obtained and analysed using
255 whole genome sequencing (WGS) and MALDI-TOF mass spectrometry (MS). These isolates had
256 the same species affiliation according to WGS and MS analyses as the isolates obtained at the
257 hospital for the same patient, and as identified in the vitreous using metagenomics analysis (Figure
258 4). Using WGS we found that for multiple morphotypes the same organism and sequence type was
259 identified, with one exception. For patient I6, we obtained three *Staphylococcus epidermidis*
260 isolates and of which two belonged to sequence type ST-88 and one to ST-487. The *Staphylococcus*
261 *epidermidis* isolates obtained from other patients (C3, C7, I4) belonged to different sequence types
262 (Figure 4, Supplementary Table S7), suggesting that they have different origins. Each
263 *Staphylococcus epidermidis* sequence type exhibited its own set of antibiotic resistance genes,
264 including genes facilitating resistance to macrolides, β-lactams, aminoglycosides, and tetracyclines.
265 The *Enterococcus faecalis* from two patients (C2, C6) belonged to different sequence types, and
266 both sequence types shared a gene facilitating resistance towards macrolides (Figure 4,
267 Supplementary Table S7). Several resistance genes that were identified in the sequenced isolates
268 were also identified in the metagenomic assembly analysis (Figure 4, Supplementary Table S6).
269 Some of the resistance genes and their predicted functions identified using the genomics approaches
270 were also in alignment with results from the phenotypic antibiotic susceptibility testing of the
271 isolates obtained during the 1st culturing at the hospital
272 (https://figshare.com/s/e579abea97dfc8c77a6a).
273
274 **Detection of bacteriophages and human DNA viruses**

As we did not identify a dominant microorganism in two endophthalmitis patients (C1, I7) we examined whether these or any of the other specimens contained DNA viruses not represented in our microbial genome database. We added an additional 7,175 virus genome sequences to the 39 virus genomes (https://figshare.com/s/b040289827b79d3a60df) in our database and classified our metagenomic sequencing data using kraken. We identified several *Enterococcus*, *Staphylococcus* and *Propionibacterium* bacteriophages among others in specimens that also were identified to contain the respective bacterial host (https://figshare.com/s/ff0527509828d1529ad9).

To evaluate whether our metagenomics approach (Figure 2) would facilitate the identification of human DNA viruses, we analysed metagenomic data obtained from patients with uveitis in which human DNA viruses had been detected previously [17]. We obtained similar results as described by Doan and colleagues, including the identification of herpes simplex virus 1 (HSV-1) in subject 1, and rubella virus in subject 6 (Supplementary Table S8). In the previous study *Hammondia hammondi* was identified in subject 3 as the second most abundant organism after *Toxoplasma gondii*. We also detected *Toxoplasma gondii* as the most abundant organism in this specimen (represented by 4,410 reads), but did only detect 4 reads for *Hammondia hammondi*; most likely because we had removed DNA sequences from the *Hammondia hammondi* genome in the database that were flagged as ambiguous. In addition, in subject 5 we detected *Ochrobactrum anthropi*, an agent that had been identified previously in eye infections such as endophthalmitis and keratitis [18,19]. However, we detected *Ochrobactrum anthropi* in high abundance in the water control sample, and therefore it may here rather represent an environmental contaminant.

295 ## Discussion

296

297 Metagenomic sequencing-based analyses of complex patient specimens and whole genome
298 sequencing (WGS) of microbial isolates will advance clinical diagnostics and treatment strategies in
299 infectious diseases [20-22]. One example, for which this strategy may be advantageous is postoperative
300 endophthalmitis as currently a causing microbial agent can only be identified in a fraction of these
301 cases [1]. Immediate diagnosis and treatment of endophthalmitis is required to prevent vision loss of
302 the affected eye. Challenges in clinical metagenomics remain at several levels, from specimen
303 collection and processing to the generation of actionable information. We examine here vitreous
304 samples from endophthalmitis patients together with a variety of control samples, evaluate two
305 DNA isolation procedures, create a curated microbial reference database, and present a pipeline for
306 sequencing data analysis. We compare the results from metagenomic read analysis to WGS and
307 MALDI-TOF mass spectrometry identification of isolates obtained for several patients, as well as to
308 results from metagenomic assembly analysis.

309

310 Vitreous samples were collected from 14 patients with endophthalmitis. Seven patients developed
311 endophthalmitis post cataract surgery, in which the natural intraocular lens was exchanged with an
312 artificial one, without the introduction of surgical instruments into the vitreous body. Another seven
313 patients developed endophthalmitis post intravitreal injection, a procedure in which drugs were
314 introduced into the vitreous body using surgical instruments to treat retinal diseases such as age-
315 related macular degeneration. As controls, we included i) vitreous samples from seven
316 endophthalmitis-negative patients, ii) basal salt solution used during vitrectomy from both,
317 individual bottles and the vitrectomy system after the solution had passed the vitrectomy infusion
318 lines, and iii) DNA extraction (blank) controls (Figure 1).

319

320 We investigated two DNA isolation procedures, QIAamp DNA Mini Kit (QIA) and QIAamp UCP
321 Pathogen Mini kit (UCP), for metagenomics analysis to determine possible infectious agents in the
322 vitreous fluid. We obtained more reads (total and classified) on average for endophthalmitis-
323 positive specimens when vitreous fluid was extracted with the UCP kit compared to the QIA
324 procedure (Supplementary Figure S1 and Figure 4). In contrast, lower numbers of reads on average
325 were obtained from the three types of control samples when they were extracted with the UCP kit
326 compared to the QIA procedure (Supplementary Figure S1). Our analysis revealed that UCP-
327 extracted control samples harboured a lower microbial diversity compared to QIA-extracted ones
328 (Supplementary Figures S4 and S5). Even though we identified distinct QIA and UCP kit
329 "fingerprints", bacteria such as *Cutibacterium acnes* and *Propionibacterium humerusii* were present
330 as background DNA across samples, independent of the DNA isolation kit. These bacteria likely
331 originated from the staff handling the samples and/or additional laboratory supplies that were used
332 during sample handling and processing [23]. Contaminant background DNA has been identified
333 previously in other DNA isolation kits [24-26], and our analysis suggests that an ultraclean production
334 of reagents and consumables reduced the amount of background DNA in the UCP DNA isolation
335 kit reagents and/or supplies. Contaminant viral DNA has been identified in previous sequencing-
336 based studies as well such as hybrid parvovirus-like virus NIH-CQV/PHV DNA from silica
337 column-based nucleic acid isolation kits [27,28], and which we detected in our samples, too
338 (https://figshare.com/s/ff0527509828d1529ad9). We did not, however, detect torque teno virus
339 DNA, as previously described for some endophthalmitis cases [29]. Overall, the UCP kit appeared to
340 be suited for the isolation of DNA from vitreous, and may potentially be useful for other human
341 body fluids and biological specimens that are assumed to have a low microbial biomass.

342
343 Our metagenomics data analysis workflow included three filtering steps (Figure 2) to reduce i)
344 human host DNA sequences, ii) false positive hits due to ambiguous and contaminant DNA
345 sequences in reference genomes, and iii) environmental background DNA sequences introduced by
346 kit reagents, potentially other laboratory supplies, as well as laboratory staff. We particularly found
347 that ambiguous/contaminant sequences in public genomes, which serve as reference in many
348 metagenomic studies, could lead to the false positive identification of microorganisms. Our initial
349 read classification, in which we used the original reference genomes, revealed *Toxoplasma gondii*
350 (false positive) across samples, even after filtering reads that mapped to the human reference
351 genome. Some microbial reference genomes appeared to harbour human DNA sequences not
352 present in the human reference genome, thus making it challenging to detect these sequences in the
353 initial human DNA sequence filtering step. This effect is especially critical when analysing clinical
354 specimens, since the patient's DNA is expected to be found in such samples. In addition, we noticed
355 that certain microbial genomes contained sequences that had a high similarity to other
356 microorganisms belonging to a different genus. These can be correct naturally occurring DNA
357 sequence regions that have a high similarity across a range of microbial taxa (including regions
358 acquired via horizontal gene transfer). In other cases they can be contaminant contigs or scaffolds in
359 primarily incomplete genome sequence assemblies. In either case, the read classification can lead to
360 a false positive identification of microorganisms. Contaminant DNA sequences in published
361 genomes have been previously found, particularly in human and animal genome assemblies [30-32].
362 Hence, we systematically examined 5,754 microbial reference and representative genomes (archaea,
363 bacteria, fungi, protozoa), and in 62% of all incomplete microbial genomes sequences were flagged
364 as ambiguous. We removed 70,478 ambiguous sequences (including human and microbial
365 contaminants), reflecting 0.35% of the bases in total from the microbial reference genomes
366 (https://figshare.com/s/045b1252bd7555b50ef0). The majority of the removed sequence fragments
367 are a correct part of the respective genome. However, for the metagenomics data analysis it is
368 advantageous to remove such redundant information, since shared sequences between multiple
369 organisms have a low information value on species level and can change a taxonomy assignment
370 based on a few kmers. According to the insight gained from this study we recommend using curated
371 microbial reference genomes in microbiome studies and particularly for the analysis of clinical
372 samples with an assumed low microbial biomass. We provide a script that facilitates the generation
373 of curated databases, including the one used in this study
374 (https://github.com/philDTU/endoPublication).
375
376 Our analysis further demonstrated the benefit of including a variety of control samples. In fact, the
377 number of control samples in this study exceeded the number of the main samples under
378 investigation by a factor of 1.2. By analysing vitreous from endophthalmitis-negative patients,
379 aliquots of basal salt solution (from bottle and vitrectomy infusion line), and blank DNA extraction
380 controls, we determined the background levels of organisms in the respective environments. All
381 control sample types had a similar microbiome pattern, characterized by organisms found in the
382 corresponding DNA extraction (blank) controls (QIA and UCP) and typical skin inhabitants. We
383 did not identify specific microorganisms for the endophthalmitis-negative patients, similar to
384 previous cultivation-based assessments, suggesting that vitreous fluid is a sterile body part or only
385 contains few microbial cells in individuals without eye infections [33,34]. We also did not identify any
386 specific organisms residing in basal salt solution that was infused into the patient's eye, in addition
387 to the ones identified in DNA extraction controls. In all cases, we cannot however exclude that
388 DNA sequences from other microorganisms would have been found if a deeper DNA sequencing
389 had been performed, or RNA had been isolated and analysed by deep sequencing. In our analysis of

390 the vitreous fluid from endophthalmitis-positive patients we removed the background contaminant
391 organisms *in silico* that were detected in the respective DNA extraction controls and were not
392 present in higher abundance in the endophthalmitis-positive patients. To trace the origin of detected
393 organisms, including infectious agents, additional controls in future studies could include samples
394 from: i) the patients skin, eye lid, conjunctiva, or other body sites that are in proximity to the
395 surgical site, ii) the surgical instruments, iii) blank tubes and/or devices used for the collection of
396 the patient specimen, as well as iv) reference mock communities with known composition. Careful
397 analysis of control samples may assist in the design of harmonized standards and guidelines for the
398 sequencing-based analysis of clinical samples and other biological specimens.
399
400 Through our metagenomics read classification data analysis workflow we identified a single
401 potential causing microorganism in 11 out of 12 culture-positive cases, and which in most cases
402 agreed with the cultivation-based analyses (Figure 4). For patient C5 we did not identify a single
403 potential causing agent and instead obtained different microbiome patterns for the two sequenced
404 aliquots (QIA and UCP). In both samples we detected *Micrococcus luteus*, in alignment with the
405 cultivation-based analysis. However, *Micrococcus luteus* was not the most abundant organism in
406 the metagenomic analysis. *Escherichia coli* was the most abundant organism in the UCP-extracted
407 sample, but which may also be a contamination introduced during sample handling. For patient I2
408 we revealed *Paenibacillus* spp. as a possible causing agent, whereas in the cultivation-dependent
409 analysis the isolate was identified as a related *Bacillus* sp using MALDI-TOF. For patient I3, our
410 metagenomic analysis suggests a potential infection by *Staphylococcus epidermidis* together with
411 *Staphylococcus aureus*. In the cultivation-based analysis *Staphylococcus aureus* was identified as
412 the potential causing agent. Only a few metagenomic reads were classified for patients C1 and I7,
413 and which were regarded as contaminants. For these two patients no microorganisms could be
414 isolated by cultivation-based methods, neither at the hospital nor the research laboratory. Therefore,
415 these two patients are assumed to have a non-infectious (sterile) endophthalmitis.
416
417 Both, the analysis of metagenome assemblies as well as whole genome sequences of isolates can
418 reveal the presence of antibiotic resistance genes that could potentially guide therapeutic treatment
419 strategies. In addition, the particular sequence type for infectious agents, such as *Enterococcus*
420 *faecalis*, *Staphylococcus epidermidis*, and other organisms, can be identified and assist in the source
421 tracking and epidemiology of the particular agent. Detailed evolutionary relationships between
422 isolates can be revealed if sufficient genome sequence information has been obtained. Our analysis
423 of the identified *Enterococcus faecalis* and *Staphylococcus epidermidis* suggests that they originate
424 from the individuals involved in the surgery or the immediate environment, as different bacterial
425 sequence types and resistant profiles were identified across patients. The metagenomics analysis did
426 not reveal these bacteria to be present in the basal salt solution samples, further pointing towards an
427 acquisition from another source.
428
429 Our previous clinical microbiology research on urinary tract infections and diarrhoeal diseases [35,36]
430 had some limitations for assessing clinical metagenomics as a technology, and whose analysis could
431 now be improved by new insight gained from this study. For example, the examined urine samples
432 were pre-processed by using a centrifugation step to remove human cells [35]. In this step, also
433 microbial cells may have been removed that have a similar density than the human cells and/or
434 were attached to these. While this step can be advantageous to limit human contaminant sequences,
435 it could be of interest to examine samples with and without the sedimentation step and using the
436 DNA isolation procedure and/or data analysis pipeline described in the present study. Furthermore,
437 the presence of potential contaminant DNA sequences was not examined in the previous study. In

438 the study concerning diarrhoeal diseases one challenge was to differentiate between natural
439 intestinal inhabitants, possible infectious agents, and potential contaminants [36]. Careful
440 bioinformatics filtering steps and inclusion of control samples, as used in the present study, might
441 allow for more robust identifications in the future. To facilitate a more standardized workflow for
442 sample analysis, we have created a list of recommendations for the design and execution of
443 metagenomic sequencing projects (https://figshare.com/s/2a0709b1f0c5e18754df), in addition to
444 specific details described in this study.
445
446 In summary, we find that metagenomics analysis, supported by WGS of isolates, may be a
447 promising strategy for the identification and characterization of infectious agents from human
448 ocular body fluid. Nucleic acid extraction from patient specimens, followed by high-throughput
449 sequencing may ultimately provide more rapid insight in regard to the identity of the causing
450 agent(s) than cultivation-based techniques, in particular in light of recent developments in long-read
451 nanopore sequencing and real-time analysis [37-39]. In cases where the metagenomic sequencing depth
452 of coverage of the microorganism is sufficiently high, valuable functional information such as
453 antibiotic resistance and virulence-associated genes can be revealed. Prerequisites for a robust data
454 analysis are suitable procedures that facilitate the isolation of nucleic acid from microorganisms
455 residing in complex samples, the analysis of relevant control samples, as well as high-quality
456 genome sequence reference databases for data analysis, as exemplified in this study. A video
457 summary of this study is available from figshare at https://figshare.com/s/38fe043f6a8ef1710444.
458
459

460 # Methods

461 ### Vitreous Samples

462 A total of 21 vitreous samples from 21 individual patients were examined in this study. From April
463 2012 to November 2013, vitreous samples from 14 eyes with postoperative endophthalmitis
464 following cataract surgery (n=7) and intravitreal injection (n=7) were collected. As control, vitreous
465 was collected from 7 patients without endophthalmitis during macula hole surgery. Approximately
466 1-2 ml of vitreous body fluid was aspirated from each eye. It was at the discretion of the
467 vitreoretinal surgeon whether to aspirate the vitreous sample before or after balanced salt solution
468 installation. About half of each collected sample was cultured in the acute clinical setting at the
469 Department of Microbiology, Hvidovre Hospital, Denmark, and the remaining material was stored
470 at -80°C.
471

472 ### Balanced salt solution samples

473 During vitrectomy, balanced salt solution (BSS PLUS, Alcon) is infused into the eye in order to
474 keep the appropriate tension of the eye. BSS PLUS is a sterile physiological saltwater solution
475 containing bicarbonate, dextrose and glutathione. Subsequently, 2.25 mg ceftazidime and 1 mg
476 vancomycin dissolved in 0.1 ml sterile salt solution are injected into the vitreous chamber. We
477 examined 3 paired sets of samples, i.e. 6 BSS samples in total. Aliquots were taken directly from
478 separate BSS PLUS bottles before vitrectomy at different time points during the study period.
479 Subsequently, BSS was collected from the vitrectomy surgical system after the BSS bottle had been
480 connected and BSS had passed through the vitrectomy infusion line. The aliquot obtained from the
481 vitrectomy system represents the fluid that is infused into the eye of the patient. The BSS samples
482 were stored at -80°C.
483

484 ### Isolation of DNA from complex samples

485 DNA was isolated from the vitreous fluid and balanced salt solution samples using two different
486 DNA isolation procedures, i) the QIAamp DNA Mini Kit (51304, Qiagen), and ii) the QIAamp
487 UCP Pathogen Mini Kit (50214, Qiagen). For each round of DNA isolation, one extraction control
488 (blank) was included. For details, see Supplementary Methods.
489

490 ### Metagenomic sequencing

491 The DNA was prepared and sequenced according to the Nextera XT DNA Library Preparation
492 Guide, Part # 15031942 Rev. D. Sequencing was performed on an Illumina MiSeq sequencer using
493 paired-end sequencing with v3 chemistry and 2×250 cycles. A total of 90,599,659 read pairs were
494 obtained from the samples. The number of read pairs was in a range of 711,886 – 4,633,576 for the
495 samples from patients with endophthalmitis, with the exception of sample I6_QIA for which only
496 106 read-pairs were obtained (Supplementary Table S2).
497

498 ### Metagenomic sequencing data analysis

499 The metagenomics analysis was carried out in five steps. 1) Adapter and quality trimming as well as
500 low complexity filtering of raw reads was performed using BBDuk of BBMap version 35.82
501 (http://jgi.doe.gov/data-and-tools/bbtools/). 2) Removal of human-affiliated reads from samples in a
502 2-step approach: i) reads that mapped against the reference genome GRCh38.p10
503 (GCF_000001405.36), and ii) reads that aligned to human sequences in the non-redundant
504 nucleotide collection (nt) database from NCBI. 3) Detection of ambiguous sequences in public
505 reference genomes and creation of curated microbial genome database that was composed of 5755
506 different genomes: archaea (251), bacteria (5166), fungi (225), protozoa (73), viruses (39) and the
507 human reference GRCh38.p7 (Supplementary Table S4). 4) Classification of reads in samples using

508 Kraken followed by bayesian reestimation of abundance with Bracken [40,41], and 5) Classification of
509 reads using BLASTn of BLAST version 2.6.0 [42]. For details, see Supplementary Methods.
510
511 **Cultivation and mass spectrometry (Clinical Microbiology lab)**
512 Aliquots from the vitreous specimens were cultivated for 12 days on 5% horse blood agar,
513 chocolate agar, brain heart infusion broth under anaerobic conditions, and on at anaerobic plates
514 (SSI Diagnostica, Denmark) under anaerobic conditions at 35°C according to the standard operating
515 procedure at the Department of Clinical Microbiology, Hvidovre Hospital. Species identification
516 was performed using MALDI-TOF mass spectrometry analysis (MALDI Biotyper 3.1, Bruker
517 Daltonics Microflex LT, database MBT DB-5627) from colony material. Antimicrobial
518 susceptibility was tested towards a range of compounds and the results were interpreted in
519 accordance to EUCAST breakpoints (http://www.eucast.org/clinical_breakpoints/).
520
521 **Cultivation (Research lab).**
522 To isolate bacteria and fungi from the vitreous body and balanced salt solution samples, 100 µl
523 aliquots were distributed on chocolate agar (SSI Diagnostica, Denmark) and Sabouraud agar with
524 Chloramphenicol (Fischer Scientific). The chocolate agar was incubated for 2 days at 37°C.
525 Colonies from the chocolate agar plates were harvested and stored in Protect Multipurpose TS80
526 preservation tubes (Technical Service Consultants Ltd, UK) at -80°C. One representative colony
527 morphotype per sample was selected for whole genome sequencing. No growth after incubation for
528 5 days was observed on the Sabouraud agar plates.
529
530 **Whole genome sequence analysis**
531 Isolates were sequenced (2x150 bp paired-end) on a MiSeq system (Illumina, San Diego, CA, USA)
532 as previously described [43]. Reads were adapter trimmed and filtered for phiX reads using BBduk.
533 The high-quality reads where assembled using the SPAdes assembler [44], and the genome sequence
534 assemblies analysed using the Bacterial Analysis Pipeline [45]. For details, see Supplementary
535 Methods.
536
537 **Ethics**
538 This study was performed in accordance with the Declaration of Helsinki. It was approved by the
539 Danish Data Protection Agency (journal number: 2011-41-5881) and by the local ethics committee
540 (journal number: H-2-2011-004), and took place at public clinics in the capital region of Denmark
541 (Region Hovedstaden).
542
543 **Data availability**
544 The data generated or analyzed during this study are available in this published article (including its
545 Supplementary Information files), and from DDBJ/ENA/GenBank under the umbrella project
546 PRJEB21503, Figshare (https://figshare.com/account/home#/projects/21038, The umbrella project
547 directory is made public post peer review. The individual links to the figure and tables in this
548 manuscript are already accessible), and GitHub (https://github.com/philDTU/endoPublication).

549 ## References

550

551 1. Durand, M. L. Bacterial and Fungal Endophthalmitis. *Clinical Microbiology Reviews* **30,** 597–
552 613 (2017).
553 2. Bannerman, T. L., Rhoden, D. L., McAllister, S. K., Miller, J. M. & Wilson, L. A. The source of
554 coagulase-negative staphylococci in the Endophthalmitis Vitrectomy Study. A comparison of
555 eyelid and intraocular isolates using pulsed-field gel electrophoresis. *Arch Ophthalmol* **115,**
556 357–361 (1997).
557 3. Buchta, V. *et al.* Outbreak of Fungal Endophthalmitis Due to Fusarium oxysporum Following
558 Cataract Surgery. *Mycopathologia* **177,** 115–121 (2014).
559 4. Foster, A. Cataract and 'Vision 2020—the right to sight' initiative. *British Journal of*
560 *Ophthalmology* **85,** 635–639 (2001).
561 5. Solborg Bjerrum, S., Kiilgaard, J. F., Mikkelsen, K. L. & la Cour, M. Outsourced cataract
562 surgery and postoperative endophthalmitis. *Acta Ophthalmol* **91,** 701–708 (2013).
563 6. Shorstein, N. H., Winthrop, K. L. & Herrinton, L. J. Decreased postoperative endophthalmitis
564 rate after institution of intracameral antibiotics in a Northern California eye department.
565 *Journal of Cartaract & Refractive Surgery* **39,** 8–14 (2013).
566 7. Gower, E. W. *et al.* Characteristics of Endophthalmitis after Cataract Surgery in the United
567 States Medicare Population. *Ophthalmology* **122,** 1625–1632 (2015).
568 8. Avery, R. L. *et al.* Intravitreal injection technique and monitoring: updated guidelines of an
569 expert panel. *Retina* **34 Suppl 12,** S1–S18 (2014).
570 9. McCannel, C. A. Meta-analysis of endophthalmitis after intravitreal injection of anti-vascular
571 endothelial growth factor agents: causative organisms and possible prevention strategies.
572 *Retina* **31,** 654–661 (2011).
573 10. Yao, K. *et al.* The incidence of postoperative endophthalmitis after cataract surgery in China:
574 a multicenter investigation of 2006–2011. *British Journal of Ophthalmology* **97,** 1312–1317
575 (2013).
576 11. Seal, D. *et al.* Laboratory diagnosis of endophthalmitis: Comparison of microbiology and
577 molecular methods in the European Society of Cataract & Refractive Surgeons multicenter
578 study and susceptibility testing. *Journal of Cataract & Refractive Surgery* **34,** 1439–1450
579 (2008).
580 12. Barry, P., Cordovés, L. & Gardner, S. *ESCRS guidelines for prevention and treatment of*
581 *endophthalmitis following cataract surgery: data, dilemmas and conclusions*. (ESCRS, 2013).
582 13. Wilson, M. R. *et al.* Actionable Diagnosis of Neuroleptospirosis by Next-Generation
583 Sequencing. *New England Journal of Medicine* **370,** 2408–2417 (2014).
584 14. Oh, J. *et al.* Temporal Stability of the Human Skin Microbiome. *Cell* **165,** 854–866 (2016).
585 15. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *Journal of*
586 *Antimicrobial Chemotherapy* **67,** 2640–2644 (2012).
587 16. Joensen, K. G. *et al.* Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance,
588 and Outbreak Detection of Verotoxigenic Escherichia coli. *Journal of Clinical Microbiology*
589 **52,** 1501–1510 (2014).
590 17. Doan, T. *et al.* Illuminating uveitis: metagenomic deep sequencing identifies common and
591 rare pathogens. *Genome Medicine* 1–9 (2016). doi:10.1186/s13073-016-0344-6
592 18. Mattos, F. B., Saraiva, F. P., Angotti-Neto, H. & Passos, A. F. Outbreak of Ochrobactrum
593 anthropi endophthalmitis following cataract surgery. *Journal of Hospital Infection* **83,** 337–

594 340 (2013).
595 19. Venkateswaran, N., Wozniak, R. A. F. & Hindman, H. B. Ochrobactrum anthropi Keratitis
596 with Focal Descemet's Membrane Detachment and Intracorneal Hypopyon. *Case Reports in*
597 *Ophthalmological Medicine* **2016,** 1–4 (2016).
598 20. Relman, D. A. Actionable Sequence Data on Infectious Diseases in the Clinical Workplace.
599 *Clinical Chemistry* **61,** 38–40 (2014).
600 21. Pallen, M. J. Diagnostic metagenomics: potential applications to bacterial, viral and parasitic
601 infections. *Parasitology* **141,** 1856–1862 (2014).
602 22. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A. & Crook, D. W. Transforming clinical
603 microbiology with bacterial genome sequencing. *Nature Reviews Genetics* **13,** 601–612
604 (2012).
605 23. Mollerup, S. *et al.* Propionibacterium acnes: Disease-Causing Agent or Common
606 Contaminant? Detection in Diverse Patient Samples by Next-Generation Sequencing.
607 *Journal of Clinical Microbiology* **54,** 980–987 (2016).
608 24. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-
609 based microbiome analyses. *BMC Biology* 1–12 (2014). doi:10.1186/s12915-014-0087-z
610 25. Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. & Chiodini, R. J. Inherent bacterial DNA
611 contaminationof extraction and sequencing reagents may affect interpretation of
612 microbiota in low bacterial biomass samples. *Gut Pathogens* 1–12 (2016).
613 doi:10.1186/s13099-016-0103-7
614 26. Tanner, M. A., Goebel, B. M., Dojka, M. A. & Pace, N. R. Specific ribosomal DNA sequences
615 from diverse environmental settings correlate with experimental contaminants. *Applied and*
616 *Environmental Microbiology* **64,** 3110–3113 (1998).
617 27. Naccache, S. N. *et al.* The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like
618 Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. *Journal of Virology* **87,**
619 11966–11977 (2013).
620 28. Smuts, H., Kew, M., Khan, A. & Korsman, S. Novel hybrid parvovirus-like virus, NIH-
621 CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. *Journal of*
622 *Virology* **110,** 10264–10269 (2014).
623 29. Lee, A. Y., Akileswaran, L., Tibbetts, M. D., Garg, S. J. & Van Gelder, R. N. Identification of
624 Torque Teno Virus in Culture-Negative Endophthalmitis by Representational Deep DNA
625 Sequencing. *Ophthalmology* **122,** 524–530 (2015).
626 30. Merchant, S., Wood, D. E. & Salzberg, S. L. Unexpected cross-species contamination in
627 genome sequencing projects. *PeerJ* **2,** e675 (2014).
628 31. Kryukov, K. & Imanishi, T. Human Contamination in Public Genome Assemblies. *PLoS ONE*
629 **11,** e0162424 (2016).
630 32. Longo, M. S., O'Neill, M. J. & O'Neill, R. J. Abundant Human DNA Contamination Identified in
631 Non-Primate Genome Databases. *PLoS ONE* **6,** e16410 (2011).
632 33. Harper, D. R. A comparative study of the microbiological contamination of postmortem
633 blood and vitreous humour samples taken for ethanol determination. *Forensic Sci Int* **43,**
634 37–44 (1989).
635 34. Egger, S. F. *et al.* Bacterial growth in human vitreous humor. *Exp Eye Res* **65,** 791–795
636 (1997).
637 35. Hasman, H. *et al.* Rapid Whole-Genome Sequencing for Detection and Characterization of
638 Microorganisms Directly from Clinical Samples. *Journal of Clinical Microbiology* **52,** 139–146

639 (2014).
640 36. Joensen, K. G. *et al.* Evaluating next-generation sequencing for direct clinical diagnostics in
641 diarrhoeal disease. *European Journal of Clinical Microbiology & Infectious Diseases* **36,**
642 1325–1338 (2017).
643 37. Greninger, A. L. *et al.* Rapid metagenomic identification of viral pathogens in clinical
644 samples by real-time nanopore sequencing analysis. *Genome Medicine* 1–13 (2015).
645 doi:10.1186/s13073-015-0220-9
646 38. Cao, M. D. *et al.* Streaming algorithms for identification of pathogens and antibiotic
647 resistance potential from real-time MinIONTM sequencing. *GigaScience* **5,** 32 (2016).
648 39. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other
649 virus genomes directly from clinical samples. *Nat Protoc* **12,** 1261–1276 (2017).
650 40. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using
651 exact alignments. *Genome Biology* **15,** 1–12 (2014).
652 41. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance
653 in metagenomics data. *PeerJ Computer Science* **3,** e104 (2017).
654 42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search
655 tool. *J Mol Biol* **215,** 403–410 (1990).
656 43. Bartels, M. D. *et al.* Comparing Whole-Genome Sequencing with Sanger Sequencing for spa
657 Typing of Methicillin-Resistant Staphylococcus aureus. *Journal of Clinical Microbiology* **52,**
658 4305–4308 (2014).
659 44. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to
660 Single-Cell Sequencing. *Journal of Computational Biology* **19,** 455–477 (2012).
661 45. Thomsen, M. C. F. *et al.* A Bacterial Analysis Platform: An Integrated System for Analysing
662 Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance. *PLoS*
663 *ONE* **11,** e0157718 (2016).

Acknowledgements

We thank Jacob Dyring Jensen (Technical University of Denmark) for technical assistance related to DNA sequencing. Thuy Doan (University of California San Francisco) is acknowledged for providing clarifying information regarding publication PMID: 27562436. This work was supported by the European Union's Framework Programme for Research and Innovation, Horizon2020 (643476). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions statement

P.K., S.S.B., H.W., and S.J.P. designed the research; P.K., S.S.B., A.F.M, H.W. and S.J.P. performed the research; P.K., F.M.A., H.W., and S.J.P. contributed analytic tools; P.K., S.S.B., A.F.M., H.W., and S.J.P. analysed the data; P.K. and S.J.P. wrote the manuscript; and S.S.B, A.F.M, F.M.A., and H.W. edited the manuscript. All authors have read and approved the manuscript as submitted.

Additional information

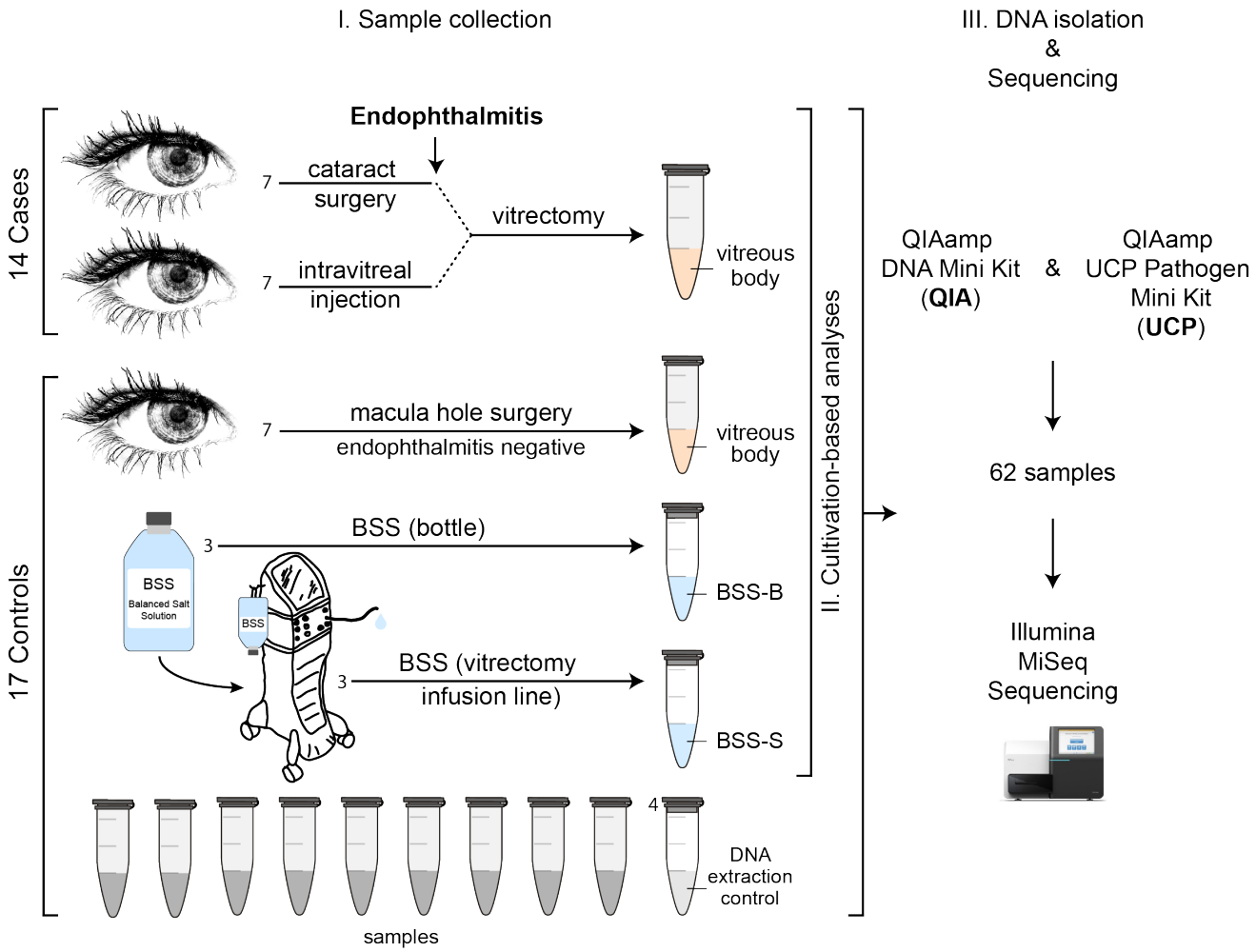
Accession numbers

The genomic sequencing data for this project are available from DDBJ/ENA/GenBank under the umbrella project PRJEB21503, including metagenomics shotgun reads (ERS1830261-ERS1830322), WGS reads (ERS1827480-ERS1827489), and WGS assemblies (ERZ468526-ERZ468535).

Competing financial interests

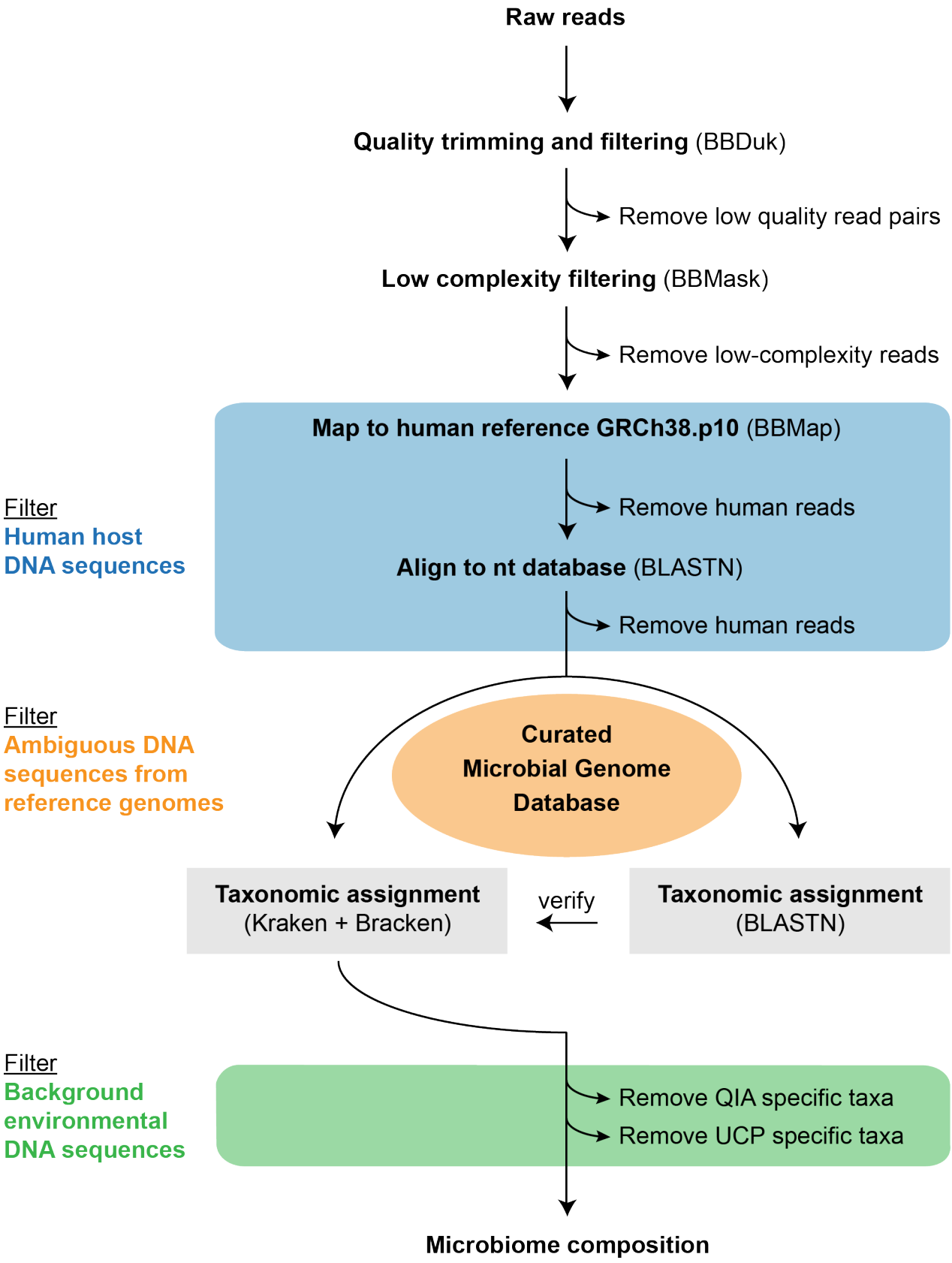
The authors declare that they have no competing interests.

688 **Figures**
689
690
691



692
693
694 **Figure 1: Sample collection, DNA isolation, and shotgun metagenomic sequencing.**
695 I. Sample collection: Vitreous body was collected through vitrectomy from 14 patients
696 with endophthalmitis following cataract surgery (n=7) and intravitreal injection (n=7). As
697 control, vitreous was collected from 7 patients without postoperative endophthalmitis
698 during macula hole surgery. Six aliquots (3 sample pairs) were obtained from balanced
699 salt solution (BSS) that is infused into the eye during vitrectomy. Three aliquots were
700 collected from separate BSS bottles (BSS-B), and the second set of aliquots was
701 collected from the vitrectomy surgical system (BSS-S) after it had passed through the
702 vitrectomy infusion line, respectively. II. Cultivation-based analyses: Aliquots of the
703 vitreous body fluid and basal salt solution samples were subjected to cultivation-based
704 analyses at the hospital and research laboratory. Obtained isolates were analyzed using
705 mass spectrometry and whole genome sequencing. III. DNA isolation & Metagenomic
706 shotgun sequencing: Samples were extracted using two DNA isolation procedures:
707 QIAamp DNA Mini Kit (QIA), and QIAamp UCP Pathogen Mini kit (UCP). A DNA extraction
708 (blank) control was included at each round of DNA isolation, i.e. one DNA extraction

709 control for 12-14 samples in total per extraction round (more vitreous samples were
710 extracted than analyzed in this study). A total of 62 samples were sequenced using
711 Illumina MiSeq sequencing technology.


Raw reads
Quality trimming and filtering (BBDuk)
Remove low quality read pairs
Low complexity filtering (BBMask)
Remove low-complexity reads
Map to human reference GRCh38.p10 (BBMap)
Remove human reads
Align to nt database (BLASTN)
Remove human reads
Filter
Human host
DNA sequences
Filter
Ambiguous DNA
sequences from
reference genomes
Curated
Microbial Genome
Database
Taxonomic assignment
(Kraken + Bracken)
verify
Taxonomic assignment
(BLASTN)
Filter
Background
environmental
DNA sequences
Remove QIA specific taxa
Remove UCP specific taxa
Microbiome composition

712
713

714 **Figure 2: Workflow for metagenomic data analysis.** In a first step, sequencing
715 adapters, low quality bases, and reads with low complexity were removed. Subsequently,
716 reads that mapped against the human reference genome sequence, or aligned with
717 human sequences in the nt database were removed. The taxonomic classification of the
718 reads was performed with Kraken together with Bracken using a curated microbial
719 genome database containing 5754 microbial and 1 human reference genome sequence
720 (for details, see Supplementary Methods). Additional reads that in this step were
721 classified as human were removed. To verify the classification results, the reads were also
722 aligned to the reference genomes using BLASTn. Organisms specific for the DNA
723 extraction (blank) controls were filtered from the patient samples.

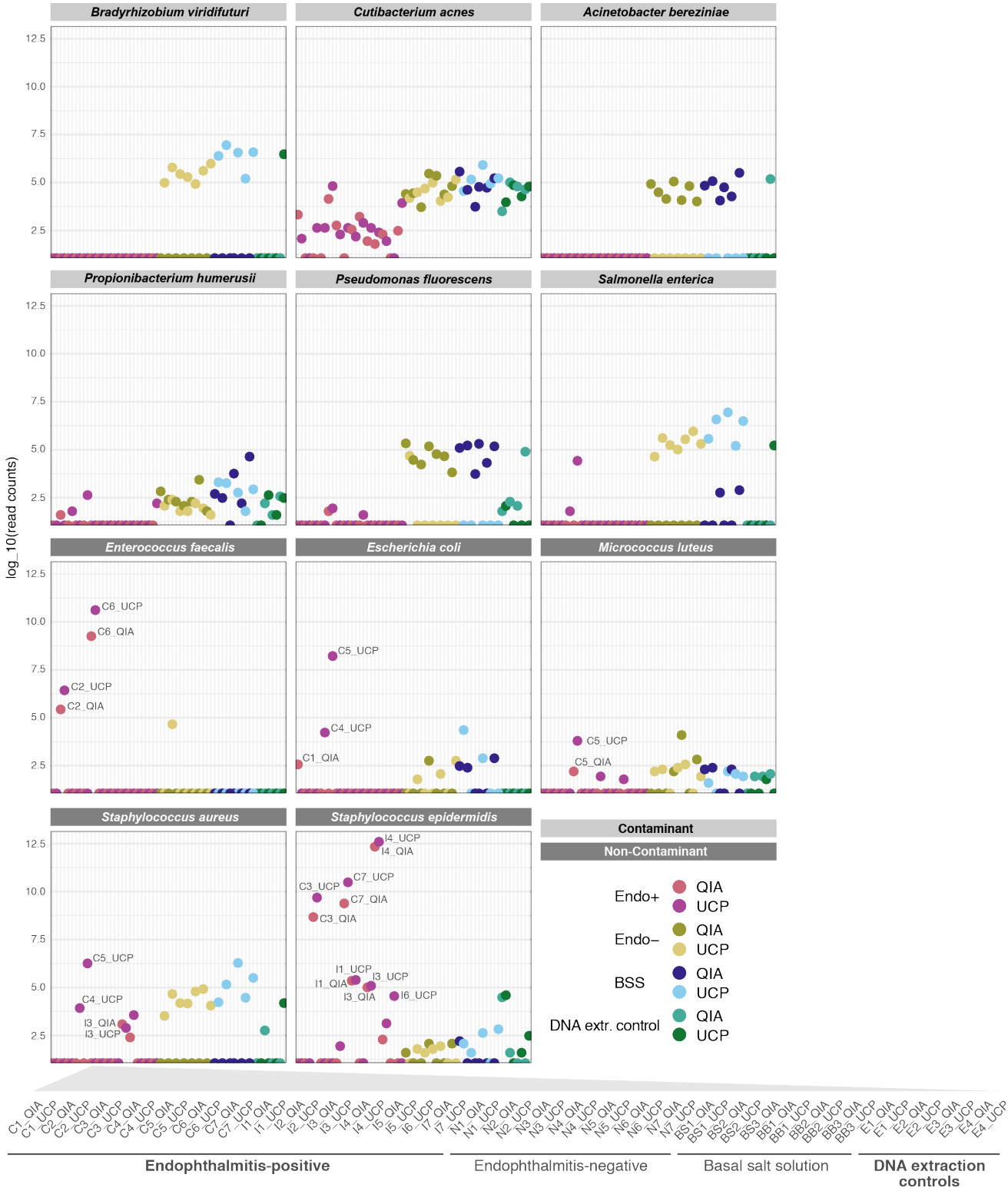


Figure 3: Selected contaminant and non-contaminant organisms based on evaluation of DNA extraction control samples. Contaminant organisms (light grey) were present in higher abundance in DNA extraction controls (green) compared to the endophthalmitis-positive samples (red). The contaminant organisms were detected in similar abundance in the endophthalmitis-negative (yellow) and/or basal salt solution

731 samples (blue) as in the DNA extraction control samples. Organisms that were detected in
732 higher abundance in patient samples (dark grey), compared to their respective DNA
733 extraction control samples, were not regarded as sample contaminants. Read counts are
734 presented as counts per million in relation to the total non-human read counts per
735 sample, respectively. An interactive version of this figure that includes individual sample
736 information, including read counts, is available from
737 https://figshare.com/s/a4fd9d84260e8456ab72. For a detailed list of contaminant
738 organisms, see Supplementary Table S5.

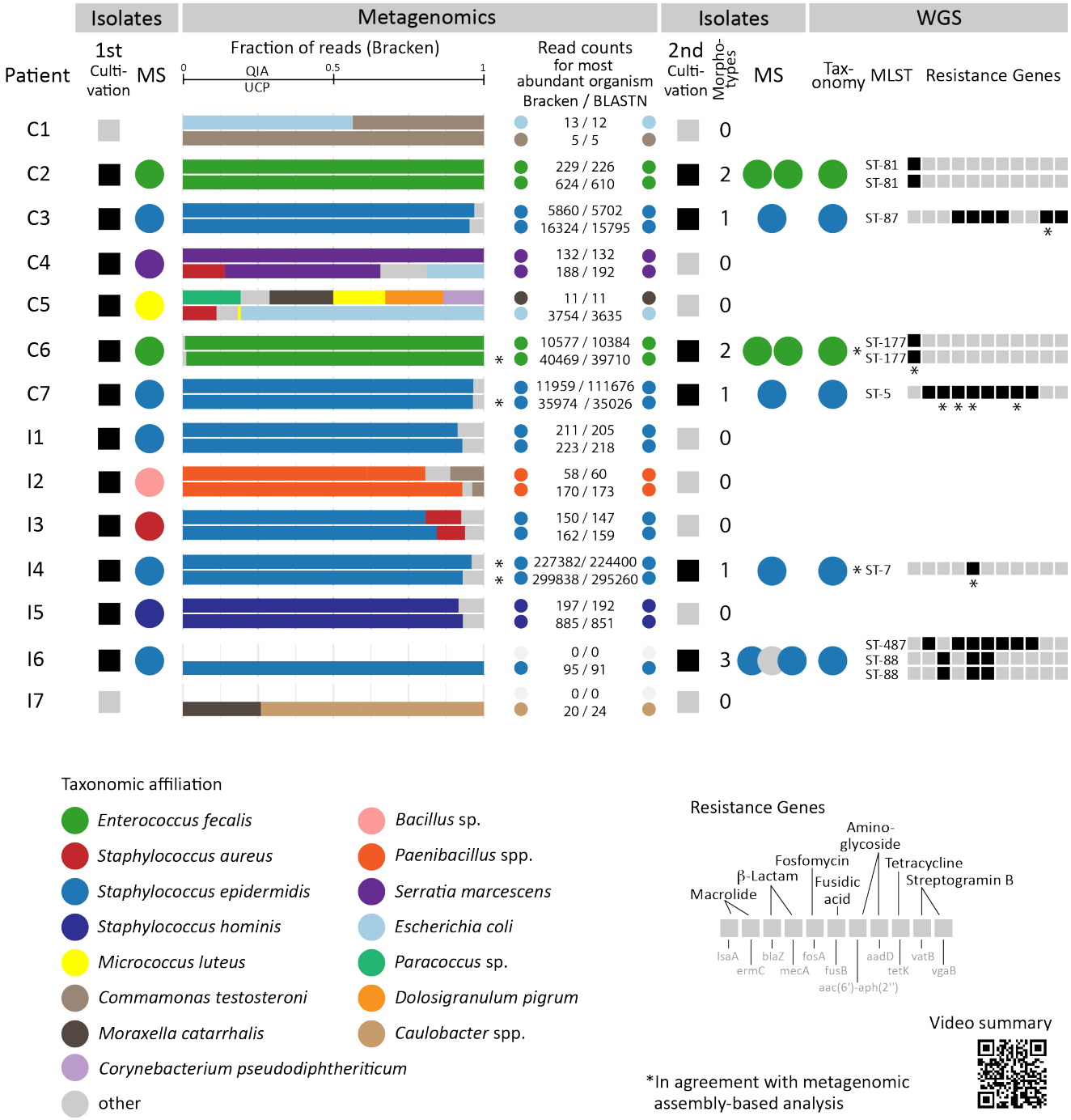


Figure 4: Summary of cultivation-based, metagenomics, and whole genome sequence analyses.

Bacterial isolates were obtained (1st cultivation) at the hospital laboratory from vitreous from endophthalmitis patients following cataract surgery (C1-7) and intravitreal injection (I1-7) and the taxonomic affiliation of the isolates were determined by MALDI-TOF mass spectrometry (MS). Vitreous was analyzed through metagenomics at the research laboratory using two DNA isolation methods (QIAamp DNA Mini Kit, QIA; QIAamp UCP Pathogen Mini kit, UCP) and the taxonomic affiliation of reads was determined. The relative fraction of reads and most abundant identified organisms following

750 Kraken+Bracken analysis is indicated for both DNA isolation methods to the left. The read
751 counts for the most abundant organism according to the Kraken+Bracken (all reads) and
752 BLASTN (forward read) analyses are indicated to the right. The read counts for the most
753 abundant organisms per sample as determined by Kraken, Bracken, and BLASTn
754 analyses are available through figshare at https://figshare.com/s/5feabfad1d8c495bf7a3.
755 Bacterial isolates were obtained in a second round of cultivation (2nd cultivation) at the
756 research laboratory, and one representative per colony morphotype per vitreous sample
757 were subjected to MS and whole genome sequencing (WGS). The taxonomic affiliation of
758 isolates was determined through classification of assembled genomes using a k-mer
759 based approach and genomic MLST, and antibiotic resistance genes were identified
760 using ResFinder. Furthermore, metagenomic assemblies were generated from the
761 shotgun metagenomic reads and analyzed with regards to taxonomic affiliation and
762 selected functional characteristics (Supplementary Table S6). A video summary is
763 available from figshare at https://figshare.com/s/38fe043f6a8ef1710444.

764 **Table 1. Ambiguous sequences in public microbial genomes**
765

	Total		Ambiguous sequences*		
	Genomes	Bases	Genomes	Bases	Bases (%)
Archaea	251	673,145,451	65	1,813,095	0.27
Bacteria	5166	20,854,687,300	2251	75,888,994	0.36
Fungi	225	6,486,874,847	126	6,642,500	0.10
Protozoa	73	2,930,167,033	53	26,447,579	0.84
Sum	5714	30,944,874,631	2922	110,858,157	0.35

766 *Genomic sequence regions ≤10 kb (incl. contigs and scaffolds) who had a match (e-value ≤ 1e-6; query
767 coverage ≥ 70%) belonging to a different genus than their stated genus definition when aligned against the
768 non-redundant nucleotide collection (nt) database from NCBI. For more details, see Supplementary Methods
769 and https://figshare.com/s/045b1252bd7555b50ef0, https://figshare.com/s/c42158cdee23f25489cd.

Supplementary Methods

Genomics-Based Identification of Microorganisms in Human Ocular Body Fluid

Philipp Kirstahler[1], Søren Solborg Bjerrum[2], Alice Friis-Møller[3], Morten la Cour[2], Frank M. Aarestrup[1], Henrik Westh[3,4], and Sünje Johanna Pamp[1,*]

[1]Research Group for Genomic Epidemiology, Technical University of Denmark, Kgs. Lyngby, Denmark
[2]Department of Ophthalmology, Rigshospitalet, Glostrup, Denmark
[3]Department of Clinical Microbiology, Hvidovre Hospital, Copenhagen, Denmark [4]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

*corresponding author sjpa@food.dtu.dk

Isolation of DNA from complex samples

As there was no standardized protocol for the extraction of DNA from vitreous prior to this study, we analysed each sample using two different DNA isolation procedures. To avoid freeze-thaw cycles, DNA was extracted on the same day as cultivation was performed. We isolated DNA from the vitreous fluid as well as balanced salt solution samples using two procedures, i) the QIAamp DNA Mini Kit (51304, Qiagen), and ii) the QIAamp UCP Pathogen Mini Kit (50214, Qiagen). In a pilot experiment, we extracted DNA from vitreous fluid using three procedures, i) the QIAamp DNA Mini Kit, ii) QIAamp UCP Pathogen Mini Kit, and iii) QIAamp UCP Pathogen Mini Kit + bead beating step. The microbiome profile resulting from procedure i) and iii) was most similar to each other, and hence we conducted the present study with the two complementing DNA isolation procedures i) and ii). For each round of DNA isolation, one extraction control (blank) was included that did not include any sample input, and was processed in the same way as the complex sample. One extraction control (blank) was included for 12-14 study samples (more samples were extracted than sequenced in the current study). i) DNA was isolated from 200 µl vitreous body using the QIAamp DNA Mini Kit according to the manufacturer's protocol "DNA Purification from Blood or Body Fluids" with minor modifications. The DNA was eluted in 100 µl AE buffer and stored at -20°C until further use. ii) DNA was isolated from 200 µl vitreous body using the QIAamp UCP Pathogen Mini Kit according to the manufacturer's protocol "Pretreatment of Microbial DNA from Biological Fluids or Cultures" and "Sample Prep Spin Protocol" with minor modifications. The sample pellet was resuspended in 400 µl ATL buffer, and the DNA was eluted in 50 µl AE buffer and stored at -20°C until further use. The DNA concentration was determined using Qubit® dsDNA High Sensitivity Assay Kit on a Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA).

Metagenomic sequencing data analysis

1. *Quality trimming and filtering*

Adapter sequences were filtered out and poor quality bases were removed from the end of the read using BBDuk of BBMap version 35.82 (http://jgi.doe.gov/data-and-tools/bbtools/). The minimal read length was set to 75 bases and the phread score to 20. In addition, reads with low complexity were removed using an entropy value of 0.7. Only intact pairs where kept for further analysis; i.e. both reads of a pair were removed also if only one did not fulfil the quality criteria.

2. *Removal of human reads*

We employed a 2-step filter approach for a thorough removal of human reads. In a first step, we mapped all high-quality reads against the reference genome GRCh38.p10 (GCF_000001405.36) using BBDuk with minimal identity set to 0.65, and through which the majority of human-affiliated reads was removed (Supplementary Figure S2). However, because of human genetic individuality not all human DNA sequences are represented in the single human reference genome sequence. In a second step, we therefore aligned all unmapped reads to a precompiled non-redundant nucleotide collection (nt) database from NCBI (downloaded 27.01.2017) using BLASTn of BLAST version 2.6.0 [1]. We removed all reads that had aligned to human DNA sequences with a minimal e-value of 1e-6.

3. Detection of ambiguous sequences & *creation of curated microbial genome databases*

Detection of ambiguous sequences: We discovered in our analysis of patient samples that certain microbial reference genomes accumulated reads (e.g. *Alcanivorax hongdengensis*, Hammondia hammondi). We did not expect these particular organisms to be present in the eye in high numbers, and we could not explain this phenomenon with human sequence contamination only. We noticed that the reads were mainly recruited to specific contigs, scaffolds, or particular genomic regions only in the particular organisms, and that mostly small sequence fragments (e.g. small contigs) were affected, or short regions within larger genomic fragments that were in close proximity to stretches of ambiguous base calls (N's). To get a better understanding, we split the microbial reference sequences (archaea, bacteria, fungi, protozoa from NCBI RefSeq database (20.12.2016)) at stretches of 10 Ns or more into separated contigs. After splitting, contigs with less than 100 bases where removed. A total of 239,009 sequences below 10 kb (these were original contigs and scaffolds, and contigs originating from the split at N's) where then aligned against the non-redundant nucleotide collection (nt) database from NCBI using BLASTn with a minimal e-value of 1e-6. [The 10 kb as threshold for short sequences was selected for computational reasons and because most sequences with a high kraken label score, when classifying the microbial genomes of the database against the human reference, were below 10 kb (https://figshare.com/s/5cc8f045347a93743739)]. Only sequences with a best hit outside of their own genus and that covered the query sequence to at least 70% where considered ambiguous. The best hit for a contig was determined based on the bitscore of the alignment. An overview and sequence list of the 35,809 ambiguous contigs is available from figshare at https://figshare.com/s/045b1252bd7555b50ef0 and https://figshare.com/s/c42158cdee23f25489cd. These ambiguous contigs were removed from the genomes before building the databases.

Creation of curated microbial genome databases: A custom Kraken database was build and contained 5755 different genomes (Supplementary Table S4). It included archaeal (251), bacterial (5166), fungal (225), protozoan (73), and viral (39) genomes from which ambiguous sequenced had been removed (see above), and the human reference genome GRCh38.p7. Low complexity regions in this library of genomes where masked using dustmasker from the C++ Toolkit version 12 [2] with standard settings. Masked bases where then converted to N. The database was build using Kraken (0.10.6-unreleased) with the standard parameter for k=31 and M=15 [3]. A script for generating the kraken database is available from github (https://github.com/philDTU/endoPublication). We also generated a BLAST database, which contained the same genomes as the kraken database and was build using makeblastdb.

Evaluation of a threshold for generating a curated microbial genome database: To build the database we used a conservative filtering of ambiguous sequence fragments (contigs/scaffolds). We deemed a sequence as ambiguous in the cases where we detected at least one BLAST hit that was affiliated with a different genus than the stated genus of the query sequence (under conditions: e-value ≤ 1e-6; query coverage ≥ 70%). To evaluate the impact of such a conservative filtering we analysed the BLAST hits in more detail.
For each sequence fragment we determined the ratio between the number of negative BLAST hits (query genus ≠ subject genus) and all BLAST hits. A low ration would indicate a higher chance for the sequence fragment to be ambiguous. We filtered the potentially ambiguous sequences by different ratio levels between 0.05 and 1.00. For the 0.05 threshold, a sequence fragment was flagged as ambiguous when the query genus

accounted for less than 5% of all BLAST hits. When we set the threshold to 1.00, we have the most conservative case with only one BLAST hit outside of the query genus being required to classify a sequence as ambiguous.
The impact of the threshold varied from microorganism to microorganism. For the organisms relevant to our study the impact was minimal. The bacterial genomes in the database are mostly closed and return no fragments below 10.000 bases with the exception of small plasmids. A decline over the threshold was observed for *Toxoplasma gondii*. Our method performed though well on the *Toxoplasma gondii* sample from the Doan *et al*. study (see main text), mainly because even in the most conservative case we removed only around 1% of the genome sequence. Additionally, it is to expect that most of the kmers in those ambiguous sequences were linked to a higher taxonomic level in kraken. The results from this analysis are summarized in a table available from figshare (https://figshare.com/s/b2db263f05db3b571aed).

4. *Classification of reads using Kraken and Bracken*
We employed Kraken to assign a read to its respective taxon. To classify a set of reads, Kraken splits each read into all possible *k-mers* and searches for perfect matches in its database [3]. The database consists of a hash-table matching each *k-mer* to all genomes in the database it is found in. One *k-mer* can be assigned to multiple organisms and is then assigned to the lowest common ancestor. For example, if a *k-mer* is found in two differed *Bacillus* species the *k-mer* is assigned to the genus instead of species level. The taxon label for the whole read is then evaluated based on the taxon labels for the individual *k-mers*. Kraken returns read counts for each label in the taxonomy tree, thus one will obtain counts on all taxonomic levels. Species abundance is then estimated using Bracken [4]. Read are re-distributed to the desired level by Bayesian probability estimation. Further only species with at least 5 reads assigned by kraken are considered for taking up new reads. This is to minimize the number of false positive species.

5. *Classification of reads using BLASTN*
In addition, we used BLASTN to align the reads to genomes in a BLAST database. The BLAST database contained the same genomes as the kraken database. Hits need at least a p-value of 1e-6 and a minimal query coverage of 80.

6. *Metagenomic assemblies*

Metagenomic shotgun reads were assembled, after human-affiliated reads had been removed, using SPAdes (3.10.1). The assembled contigs were submitted to the Bacterial Analysis Pipeline [5], which is described in the paragraph below.

Whole genome sequence analysis

Reads were adapter trimmed and filtered for phiX reads using BBduk. In addition, we removed reads mapping to the masked human reference genome hg19 using the BBmap suite The masked human reference was generated by the BBsuit developer (https://drive.google.com/file/d/0B3llHR93L14wd0pSSnFULUlhcUk/edit?usp=sharing). Masked regions are of low entropy, multiple repeat or conserved regions from other fungal and plants. Masking a total of around 1.4% from the reference.

The high-quality reads where assembled using the SPAdes assembler [6] in careful mode with k-mer sizes 21, 33, 55, 77, 99, and 127. For each assembly, contigs smaller than 5 kb where aligned to the non-redundant nucleotide collection (nt) database from NCBI using BLASTn. Contigs that were larger than 1kb and exhibited a strong association to the same organism where kept in the final assembly. A quality assessment of the assemblies was performed using QUAST [7] (Supplementary Table S7). The genome sequence assemblies were analysed using the Bacterial Analysis Pipeline [5]. To determine the taxonomic identity of isolate the pipeline utilizes a k-mer based approach [8]. Subsequently, a genomic MLST-typing is performed based on the allele and sequence profiles from PubMLST [9]. Antimicrobial resistance genes are detected using ResFinder with a minimal sequence identity of 90% and minimal resistance gene length coverage of 60% [10].
The average depth of coverage for the genome assemblies ranged between 88 times and 198 times. The number of contigs per assembly ranged in between 19 and 37 for *E. faecalis* and 44 to 74 for *S. epidermidis*. The total number of assembled bases and GC content was close to the median reference length and GC content reported at NCBI. 3 Mb and a GC content of 37.3% for *E. faecalis* and 2.4 Mb for *S. epidermidis* with a GC content of 31.9%. The N50s were between 300k and 350k for *E. faecalis* and between 94k and 156k for *S. epidermidis* (Supplementary Table S7).

Extended data figures and tables

The additional data figures and tables listed in this study are available through the figshare project https://figshare.com/account/home#/projects/21038 (The umbrella project directory is made public post peer review. The individual links to the figure and tables in this manuscript are already accessible).

References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–410 (1990).
2. Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13,** 1028–1040 (2006).
3. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15,** 1–12 (2014).
4. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* **3,** e104 (2017).
5. Thomsen, M. C. F. *et al.* A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance. *PLoS ONE* **11,** e0157718 (2016).
6. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19,** 455–477 (2012).
7. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29,** 1072–1075 (2013).
8. Hasman, H. *et al.* Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples. *Journal of Clinical Microbiology* **52,** 139–146 (2014).
9. Larsen, M. V. *et al.* Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria. *Journal of Clinical Microbiology* **50,** 1355–1361 (2012).
10. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy* **67,** 2640–2644 (2012).

Supplementary Figures

Genomics-Based Identification of Microbial Agents in Human Ocular Body Fluid

[1]Philipp Kirstahler, [2]Søren Solborg Bjerrum, [3]Alice Friis-Møller, [2]Morten la Cour, [1]Frank M. Aarestrup, [3,4]Henrik Westh, and [1]Sünje Johanna Pamp[1*]

[1]Research Group for Genomic Epidemiology, Technical University of Denmark, Kgs. Lyngby, Denmark; [2]Department of Ophthalmology, Rigshospitalet, Copenhagen, Denmark; [3]Department of Clinical Microbiology, Hvidovre Hospital, Copenhagen, Denmark; [4]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

*corresponding author sjpa@food.dtu.dk.

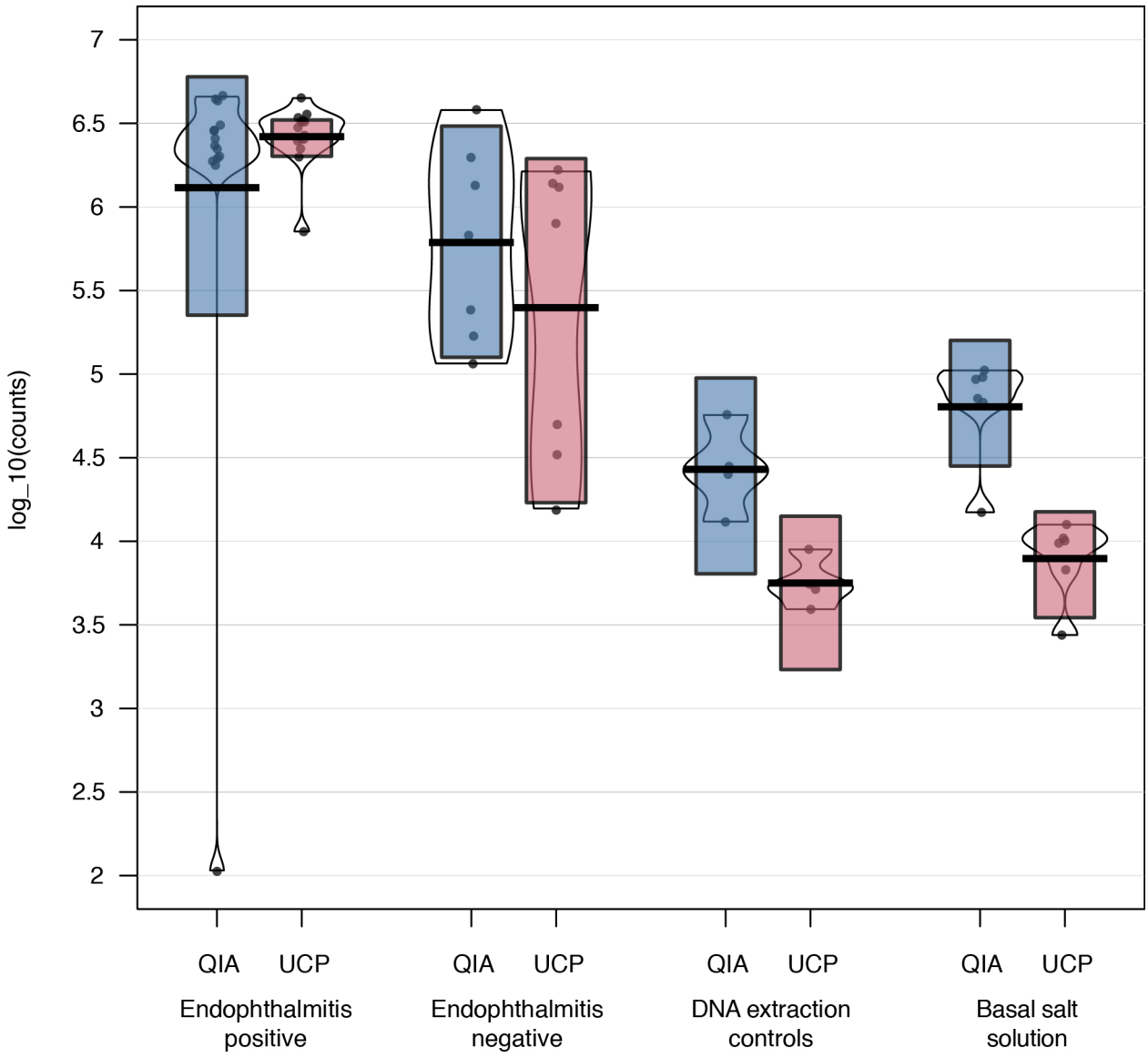


Figure S1: Read counts per sample group and DNA isolation method obtained using shotgun metagenomic sequencing. Number of reads obtained using metagenomic sequencing (MiSeq) that passed the quality trimming and filtering. The samples are listed according to sample group (endophthalmitis negative, endophthalmitis positive, DNA extraction controls, basal salt solution) and DNA extraction method (QIAamp DNA Mini Kit [QIA], QIAamp UCP Pathogen Mini kit [UCP]). The points in the plot represent the individual samples, the bean represents the smoothed reads count density based on the samples, the colored rectangle indicates the 95% highest density interval, and the horizontal bar indicates the central tendency.

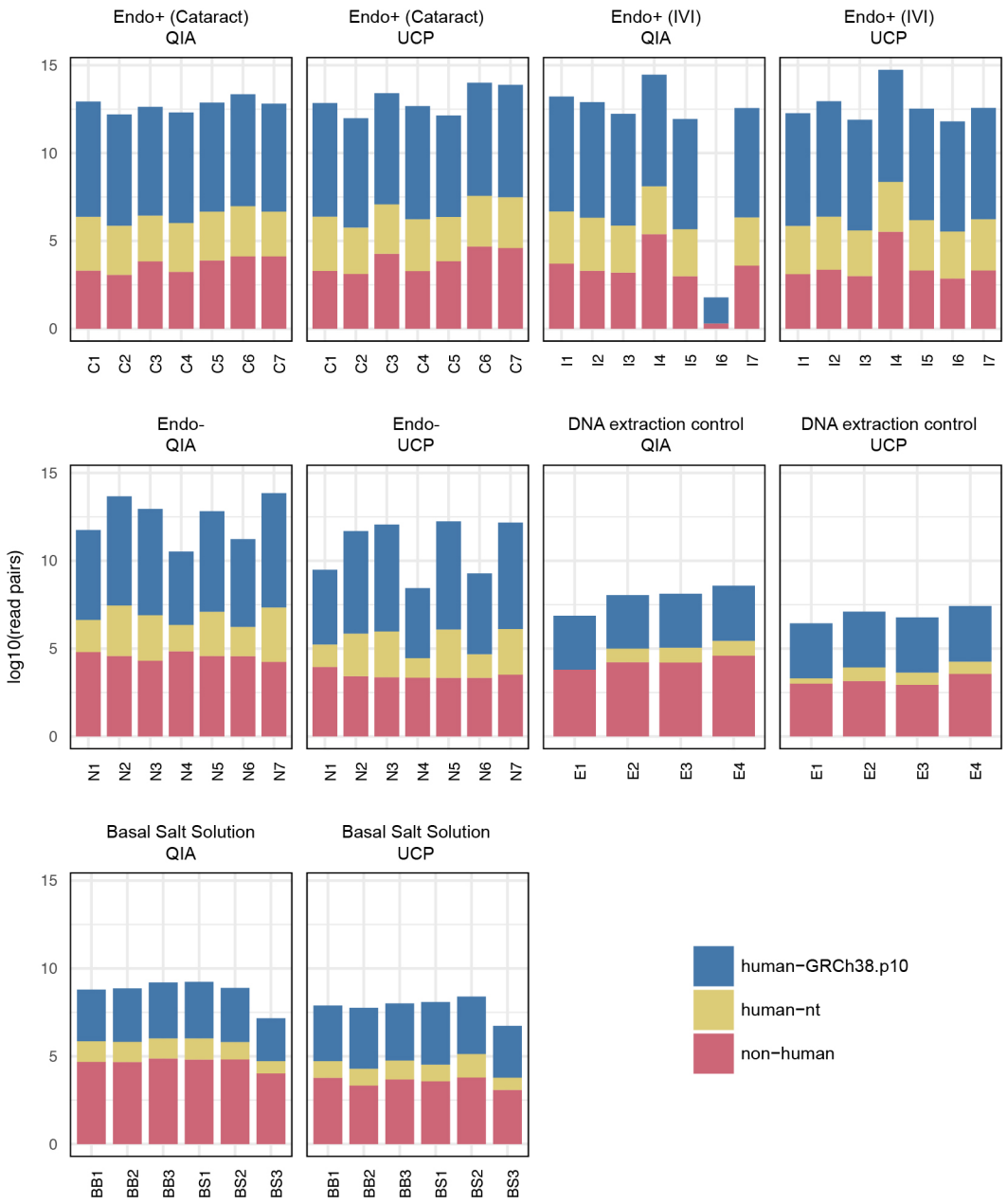


Figure S2: Human-affiliated reads in the metagenomic sequencing data. The reads from endophthalmitis-negative patients, endophthalmitis-positive patients (post cataract, and post intravitreal injection (IVI)), as well as the DNA extraction controls, and basal salt solution samples were mapped against the human genome assembly GRCh38.p10. The non-mapped reads were aligned using BLASTn to the non-redundant nucleotide collection nt database from NCBI. Reads mapped to GRCh38.p10 (blue), reads aligned to human genome sequences in NCBI nt database (yellow), and reads that were not affiliated with any human genome sequences in the two databases (red).

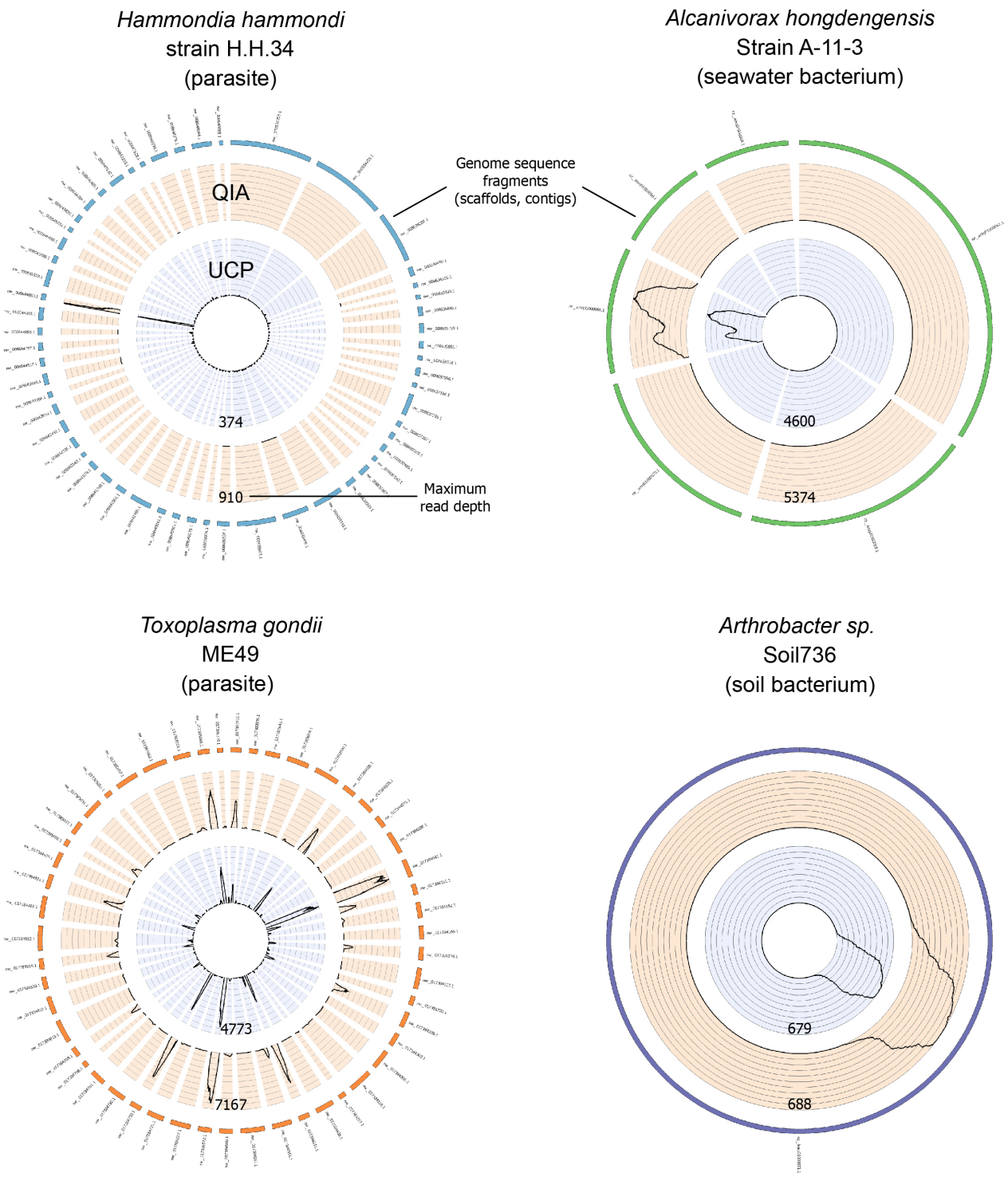


Figure S3: Selected genomes from public databases containing ambiguous (putative contaminant) DNA sequences. Reads of this study mapped to specific genome sequence fragments (scaffolds, contigs) of the genomes of *Hammondia hammondi* strain H.H.34 (GCF_000258005.1), *Alcanivorax hongdengensis* Strain A-11-3 (AMRJ00000000), *Toxoplasma gondii* ME49 (GCF_000006565.2), and *Arthrobacter* sp. Soil736 (GCF_001428005.1). The most outer circle displays genome sequence fragments (scaffolds, contigs), and the Top50 contigs with at least a minimum coverage of 5 per nucleotide position are shown. When these contigs were aligned to the nucleotide collection nt (NCBI), their Top10 hits included in most cases human genomic sequences and in other

cases bacterial sequences (see Supplementary Table S3). The orange and blue inner circles display the depth of mapped reads originating from the samples that were extracted with the QIA and UCP DNA extraction methods, respectively. The corresponding maximum read depth is indicated at the bottom of the ring. For a detailed list of scaffolds and contigs predicted to be ambiguous in the 5754 public microbial genomes examined in this study, see https://figshare.com/s/c42158cdee23f25489cd.

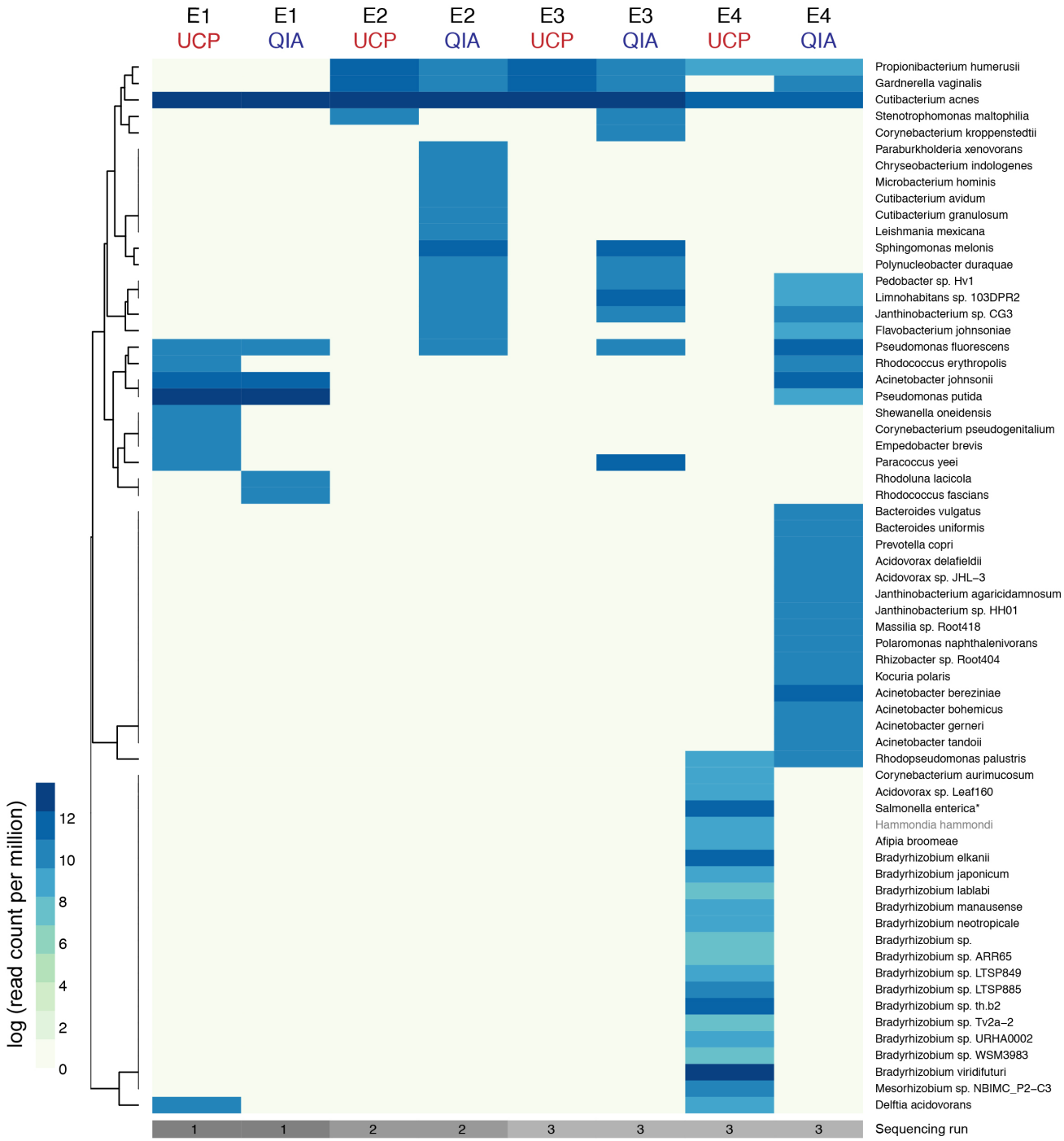


Figure S4: Abundance of different organisms in DNA extraction (blank) control samples. The Top20 most abundant organisms per sample for DNA extraction control E1 to E4, extracted at four different time points with the QIAamp DNA Mini Kit (QIA) and QIAamp UCP Pathogen Mini kit (UCP) respectively, are displayed. In some samples fewer than 20 different microorganisms were detected. Read counts are presented as counts per million. See Supplementary Table S5 for a complete list of contaminant taxa detected in the individual samples. *Salmonella enterica* (labeled with an asterisk) may originate from carry-

over contamination during sequencing as it was sequenced on the same Illumina MiSeq machine in a prior sequencing run. *Hammondia hammondi* (grey) may be a false positive, as its public reference genome sequence contained ambiguous (potential contaminant) contigs or scaffolds (see Supplementary Table S3).

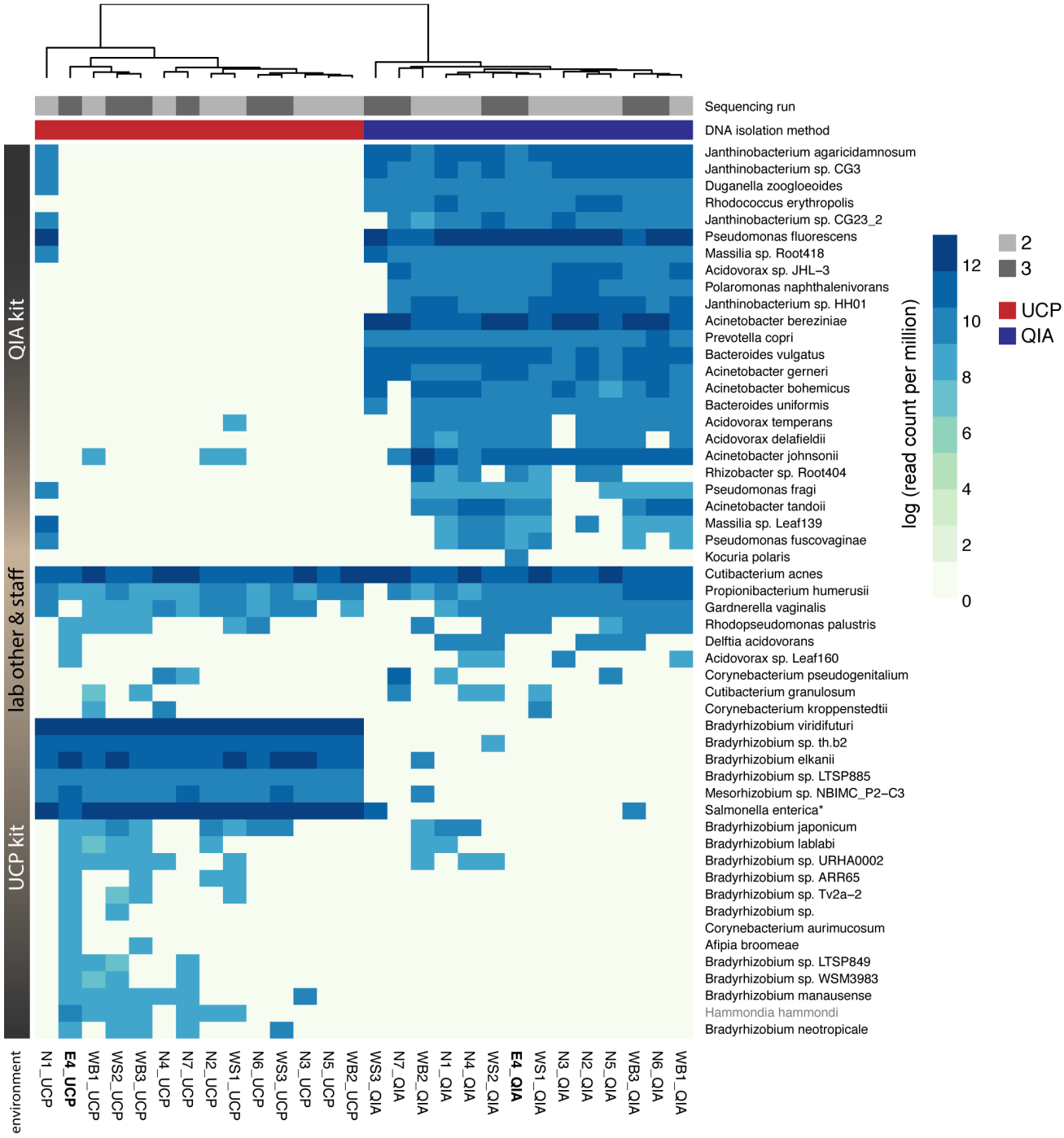


Figure S5: Abundance of taxa in endophthalmitis-negative, basal salt solution, and associated DNA extraction (blank) control samples. The 20 most abundant taxa per sample group are displayed, i.e. the Top20 organisms for E1_QIA, E1_UCP, endophthalmitis-negative QIA (N1-7_QIA), endophthalmitis-negative UCP (N1-7_UCP), basal salt solution QIA (BB1-3_QIA, BS1-3_QIA), and basal salt solution UCP (BB1-3_UCP, BS1-3_UCP), respectively. The abundance of taxa is displayed as counts per million. Three main groups of organisms emerge: organisms that are predominantly found in QIA-related samples, UCP-related samples, or both sample groups. The latter group may be environmental organisms originating from the laboratory equipment and supplies, and/or laboratory staff. Some taxa may also originate from carry-over contamination during sequencing as the following species were sequenced on the same Illumina MiSeq machine

in respective prior sequencing runs: *Salmonella enterica* (labeled with an asterisk), and *Plasmodium falciparum* and *Escherichia coli* (not displayed, as not among the Top20 organisms). *Hammondia hammondi* (grey) may be a false positive, as its public genome sequence contained ambiguous (potential contaminant) contigs or scaffolds (see Supplementary Table S3). It should be noted that fewer microorganisms were detected in the DNA extraction controls E1 – E3 for both kits (not displayed).