University of Padua
School of Medicine and Surgery
Department of Molecular Medicine
Master's Degree Programme in Medical Biotechnologies
Director: Prof. Stefano Piccolo, PhD

DESIGN AND EVALUATION OF A METABARCODING APPROACH FOR THE DETECTION OF EUKARYOTIC INTESTINAL PARASITES

Supervisor: Prof. Stefano Toppo, PhD
University of Padua - Department of Molecular Medicine

Co-supervisor: Prof. Giorgio Valle, PhD
BMR Genomics S.r.l.

Candidate: Giovanni Rosson

Academic Year 2022/2023

INDEX

1. ABSTRACT

The incidence of parasitic intestinal infections is increasing worldwide, becoming of clinical concern for the human health. Nevertheless, clinicians continue to diagnose them with poorly sensitive and non-specific procedures, like the microscopic examination of faecal samples. Although several molecular methods have been developed in the last decade, allowing increasing the diagnostic precision, they still have some drawbacks that limit their application in diagnostics. With these premises, in this thesis we describe a feasible DNA metabarcoding approach to detect eukaryotic intestinal parasites in human faecal samples. To achieve this objective, we firstly tested *in silico* the amplification capability of several literature-available primer pairs targeting the 18S rRNA gene, and selected those able to identify the broadest spectrum of human intestinal parasites. Then, we tuned the experimental conditions to profitably extract total DNA from formalin-fixed stool and to increase the amplification efficiency. We used the selected primer pairs to amplify the metabarcodes of interest, and subsequently we sequenced the obtained samples on the Illumina MiSeq platform. Ultimately, we developed a general bioinformatics pipeline to analyse sequenced amplicons and to perform taxonomical identification of micro-eukaryotes. Overall, this work could be a promising starting point for developing a clinically relevant DNA metabarcoding diagnostic technique.

Keywords: Intestinal infectious diseases; 18S rRNA sequencing; Bioinformatics tools.

1. SOMMARIO

L'incidenza di infezioni da parassiti intestinali sta aumentando in tutto il mondo, richiedendo una maggiore attenzione a livello clinico. Nonostante ciò, queste infezioni vengono ancora diagnosticate tramite metodi a bassa sensibilità e specificità, come l'analisi al microscopio di campioni fecali. Negli ultimi dieci anni sono stati sviluppati metodi molecolari che permettono di aumentare la precisione diagnostica, ma essi possiedono ancora degli svantaggi che ne limitano l'applicazione in diagnosi.

Viste queste premesse, in questa tesi descriviamo un metodo di DNA metabarcoding per l'identificazione di parassiti intestinali eucariotici in campioni di feci umane. Inizialmente, abbiamo testato *in silico* le capacità di amplificazione di diverse coppie di primer specifiche per il gene 18S rRNA, e abbiamo selezionato quelle capaci di identificare il maggior numero di parassiti dell'intestino umano. Successivamente, abbiamo modificato alcune condizioni sperimentali per garantire un'efficiente estrazione del DNA da campioni fecali fissati in formalina e per aumentare l'efficienza di amplificazione. Abbiamo poi utilizzato le coppie di primer selezionate per amplificare i metabarcodes di interesse, e abbiamo sequenziato i campioni ottenuti nella piattaforma Illumina MiSeq. Infine, abbiamo sviluppato una pipeline bioinformatica per analizzare gli ampliconi sequenziati ed effettuare l'identificazione tassonomica dei micro-eucarioti. Questo progetto potrebbe consistere in un primo passo verso lo sviluppo di una tecnica diagnostica basata sul DNA metabarcoding.

Parole chiave: Malattie da infezione intestinale; Sequenziamento del gene 18S rRNA; Strumenti bioinformatici.

2. INTRODUCTION

2.1. The Human gut Microbiota

Microorganisms belonging to the Bacteria, Archaea and Eukarya domains, and the relative viruses, populate every surface of the human body in contact with the external environment. Overall, it is estimated that at least 10^{14} microbial cells and 10^{15} viral particles colonise almost all human tissues: they constitute the "human-associated microbiota", and their genomes compose the so called "host microbiome" (Clemente et al., 2012).

Humans and microorganisms have evolved symbiotic relations, and through co-evolution (i.e., when two or more species reciprocally affect each other's evolution through the process of natural selection), co-adaptation (i.e., when two or more species develop and maintain advantageous genetic traits, so that their mutual relations can persist) and codependency (i.e., when two or more species manifest a mutual dependence, therefore neither can function independently) human hosts have developed a mutualistic relationship with their microbiota, meaning that both receive benefits (Iebba et al., 2016). As a matter of fact, while microbes get nutrients, proliferate and survive in protected niches, their presence is fundamental for the host's health, because the collective functions they perform directly affect the (patho)physiology of humans (Lozupone et al., 2012).

The human gastrointestinal tract is composed of multiple organs (i.e., mouth, oesophagus, stomach, small intestine, large intestine and anus) and, being in contact with the external environment, is populated by a large diversity of microorganisms and viruses that collectively constitute the "human gut microbiota". While both stomach and small intestine are poorly colonised due to the presence of acids, digestive enzymes and antimicrobials that create an inhospitable environment, the large intestine is devoid of these compounds and shows the highest diversity and density of microbes in the human body (e.g., 10^{12} bacteria/g) (Piewngam et al., 2020).

The human gut-associated microscopic population can be divided in three macro groups: the intestinal bacteriome, the intestinal virome and the intestinal "eukaryome". Bacteria account for 70% of organisms of the gut microbiota, which comprises over 1,800 bacterial genera and over 35,000 bacterial species, 99% of which belong to the *Bacteroidetes* and *Firmicutes* phyla (other important, but less represented bacterial phyla are *Actinobacteria* and *Proteobacteria*) (Iebba et al., 2016). DNA bacteriophages are the most abundant viruses of the intestinal virome (10^8-10^9 Viral-Like Particles in the intestine), but this collection includes also RNA bacteriophages, DNA and RNA archaeal viruses and DNA and RNA eukaryotic viruses. The intestinal "eukaryome" is composed of Protists (i.e., unicellular

eukaryotes), helminths (i.e., parasitic worms of humans) and fungi (Lukeš et al., 2015); the latters, which compose the intestinal mycobiome, belong to the Ascomycota, Basidiomycota, Microsporidia and Zygomycota phyla, and show a lower diversity than the intestinal bacteriome or virome (Piewngam et al., 2020).
The taxonomic composition of the gut microbiota can be modulated by several factors, such as the gut morphology and structure, age, diet, hygiene, stress, drugs and antibiotic use; therefore, it is highly variable among individuals (interpersonal variability) and can change in time within the same person (intrapersonal variability). The interpersonal variability is demonstrated by the fact that different individuals are endowed with different microbial species and show diverse amounts of microbes belonging to the same genera, even if the microbial phyla present in the human intestine are similar across people. The intrapersonal variability depends mainly on the pathophysiological state of the host, who can thus manifest different gut microbiota compositions in various moments of their life according to their changing health conditions and lifestyle (Lozupone et al., 2012) and (Iebba et al., 2016).

As mentioned before, humans have evolved a mutualistic relationship with their gut microbiota, which performs several functions that are fundamental for the maintenance of the host health: (I) extraction of nutrients and energy from the diet (Iebba et al., 2016), (II) production of essential amino acids and vitamins, (III) regulation of the gut homeostasis (Nathan et al., 2021), (IV) modulation of the immune development and of host immune responses (Piewngam et al., 2020), (V) protection against enteropathogens ("colonisation resistance") (Robinson et al., 2010), and (VI) modulation of other organs' functionality (Iebba et al., 2016).
Regardless of the interpersonal variation in taxonomic composition at lower phylogenetic levels (i.e., genus and species), all gut microbiota functions are conserved among individuals. Indeed, host and host-associated microbial community have co-evolved to promote a functional redundancy, meaning that a single function is exerted by multiple, evolutionarily distant microbial species, and thus it is preserved in people with different microbiota compositions (as long as they possess the functional genes). Overall, individuals keep a core microbiome, but not a core microbiota. However, the small variances in both microbiota and microbiome must not be overlooked, because they might clarify the interindividual differences in metabolism of dietary substrates and drugs, hence help explain why people exhibit different responses to diets and medical treatments (Lozupone et al., 2012).

If the gut microbiota composition is excessively modified by the factors cited above (i.e., pathophysiological state of the host, diet, and drugs) and/or by pathogenic infections, thus the physiological balance among microbes and between microbes and host (defined "eubiosis") is disrupted, the outcome is the "intestinal dysbiosis" and the possible manifestation of infectious and non-infectious diseases (Piewngam et al., 2020). In the eubiotic state, the gut microbiota is rich in diverse species and is characterised by stability, meaning the ability to resist perturbations (resistance) and to return to the equilibrium state following perturbations (resilience) (Robinson et al., 2010). In the dysbiotic state, the microbial diversity is strongly reduced due to the overgrowth and/or loss of some members of the community, and so are stability, resistance and resilience, so the host's health is impaired.

Dysbiosis-related infections frequently occur after antibiotic treatments; as a matter of fact, these drugs cause a disruption of the commensal ecosystem and a reduction of "colonisation resistance", allowing the proliferation of opportunistic (e.g., *Clostridium difficile*) and exogenous pathogens (Iebba et al., 2016). On what concerns dysbiosis-related non-infectious diseases, they are usually complex and multifactorial in terms of pathogenesis and complications: acute diarrhoea, malnutrition, inflammatory bowel diseases (IBDs, e.g., Crohn's disease and ulcerative colitis), irritable bowel syndrome (IBS), obesity, type I diabetes, allergic and autoimmune diseases (e.g., rheumatoid arthritis), neurodegenerative diseases (e.g., multiple sclerosis and Parkinson's disease), neurological disorders (e.g., autism spectrum disorders and Rett syndrome), and cancer. Anyway, it is still debated whether the intestinal dysbiosis is a cause or a consequence of these health conditions (Robinson et al., 2010), (Clemente et al., 2012) and (Piewngam et al., 2020).

Interestingly, individuals of Western, industrialised populations are more frequently affected by these diseases than people of non-industrialized nations. One possible explanation might be the higher hygiene, consumption of "Western diet" foods and use of antibiotics of the firsts, habits leading to a lower diversity of their gut microbiota that, in some cases, may even resemble a dysbiotic state (Chabé et al., 2017). The fact that these pathological states can be treated by transplanting an exogenous healthy gut microbiota, which can restore the physiological microbial community, gives further proof that this host-associated microscopic population is fundamental for the maintenance of the human health (Clemente et al., 2012).

2.2. "Eukaryome" and eukaryotic parasites

Historically, any fungus, Protist and helminth detected in the human gastrointestinal tract was thought to be a pathogenic parasite, etiological agent of intestinal or even systemic diseases associated with parasite migration and induction of lesions and inflammation. Nowadays, the "eukaryome" is considered an essential component of the human gut microbiota: organisms belonging to more than 140 fungal genera and 15 protistan genera (Hamad et al., 2016) and some helminths stably colonise the human intestine, are well tolerated by the host immune system, and play their role in promoting the host's health. Therefore, they can be considered mutualists (Lukeš et al., 2015). As for the intestinal bacteriome, the intestinal "eukaryome" shows a low interpersonal variation if observed at the phylum level (fewer than 10 phylotypes are found in each individual), and a higher variety at the species (and strain) level (Parfrey et al., 2011). The Western, industrialised populations' lifestyle negatively affects also the "eukaryome" composition, causing the quite complete disappearance of gut eukaryotes and intestinal helminths (Chabé et al., 2017). As a consequence, Western individuals manifest a higher incidence of inflammatory and autoimmune diseases compared to non-industrialized people (Lukeš et al., 2015). Commensal eukaryotes are beneficial since, as demonstrated in several studies (Berrilli et al., 2012), (Chabé et al., 2017), (Leung et al., 2018) and (Dubik et al., 2022), they increase the gut microbiota species' richness and provide immunomodulatory cues to the host. Given these observations, it is not infrequent the administration of helminths to individuals affected by inflammatory, autoimmune and allergic diseases: this strategy, termed helminth therapy, has often been successful at preventing and treating these pathological conditions (Lukeš et al., 2015).

All eukaryotic microorganisms that can be found in the human gastrointestinal tract, being either commensal, opportunistic or parasites, fall into four Eukaryota supergroups: Amoebozoa, Excavates, Opisthokonta (of the Obazoa clade) and SAR (Stramenopiles-Alveolates-Rhizaria). Amoebozoa includes commensal and pathogenic amebae (e.g., *Entamoeba* genus); Excavates includes intestinal flagellates (e.g., *Dientamoeba*, *Giardia* and *Trichomonas* genera); Opisthokonta includes fungi and helminths, which can be divided into the four groups of acanthocephalans, cestodes (e.g., *Diphyllobothrium*, *Hymenolepis* and *Taenia* genera), nematodes (e.g., *Ancylostoma*, *Anisakis*, *Ascaris*, *Enterobius*, *Necator*, *Strongyloides*, *Trichinella* and *Trichuris* genera) and trematodes (e.g., *Fasciolopsis* and *Schistosoma* genera); SAR includes commensal, opportunistic and parasitic protozoa (i.e., animal-like Protists) (e.g., *Blastocystis*, *Cyclospora*, *Cystoisospora*, *Cryptosporidium*, *Neobalantidium* and *Toxoplasma* genera) (Hamad et al., 2016) and (Mathison and Sapp, 2021).

Protozoa belonging to the *Blastocystis*, *Entamoeba* and *Trichomonas* genera, and fungi, are the most common commensal eukaryotes detected in the gut of healthy individuals, who manifest an eubiotic state (Parfrey et al., 2014). However, if the health conditions of the host are impaired (e.g., malnutrition, dysbiosis, compromised immunity), some of those microorganisms, in addition to other less frequently found micro-eukaryotes (e.g., *Cryptosporidium parvum*), can behave as opportunistic pathogens and cause severe diarrhoea (Parfrey et al., 2011). In terms of medical importance, the most relevant protozoan and helminthic pathogenic parasites that cause the majority of human intestinal infections are *Blastocystis* spp. (in particular *Blastocystis hominis* ST3 and ST7), *Cyclospora cayetanensis*, *Cryptosporidium* spp., *Entamoeba histolytica* (i.e., the only pathogenic species of the *Entamoeba* genus), *Giardia intestinalis*, and members of the above-cited cestodes, nematodes and trematodes genera (Haque, 2007).

Intestinal parasites infect billions of people worldwide: the greatest incidence is in the tropical and subtropical developing countries with poor sanitation and quality of food and water, which favours food-borne and water-borne outbreaks of parasitic diseases; however, an increasing frequency of infection is visible in economically developed populations due to the transmission of pathogens from infected returning travellers and immigrant groups, and also due to the growing habit of raw food consumption (Jong, 2002).

Humans become infected with protozoan parasites via the faecal-oral route; indeed, by ingesting faecally contaminated food or water, individuals enter in contact with oocysts (i.e., the environmentally stable form of the parasite) that, once in the intestine, transform into the metabolically active, proliferating and pathogenic trophozoites (Hamad et al., 2016). If they succeed in overcoming the "colonisation resistance" promoted by the healthy gut microbiota and invade the intestinal epithelium (Partida-Rodríguez et al., 2017) and (Leung et al., 2018), there is a strong induction of mucosal inflammation (Dubik et al., 2022) that leads to the development of parasitic diseases, which usually manifest with diarrhoea, abdominal pain, malabsorption and weight loss. *Blastocystis* spp. colonises billions of people, but only few, immunocompromised patients manifest blastocytosis (i.e., the *Blastocystis*-related disease). *Cryptosporidium* spp. symptomatic infection (i.e., cryptosporidiosis) is one of the main causes of infants' death in developing countries and the first water-borne diarrhoeal disease in the United States (Guérin and Striepen, 2020). *Entamoeba histolytica* infects 50 million people worldwide, but only 10% of infections result in severe dysentery (i.e., diarrhoea containing blood) that causes 40,000-100,000 deaths annually (Haque, 2007). *Giardia intestinalis* is the most prevalent human intestinal parasite, and the related giardiasis is the first diarrhoeal disease in the developed nations. An important characteristic

of this parasite is its non-invasiveness (it simply attaches to the small intestine mucosa), thus it has been suggested that its pathogenicity depends on interactions between *Giardia* secreted factors (e.g., cysteine proteases) and the host's gut microbiota and immune system (Adam, 2021).
Helminths populate the gut of 3-5 billion people, 450 million of whom become ill (Hamad et al., 2016) due to a high parasite load and the invasion of extra-intestinal tissues by ectopic eggs, migrating larvae and adult worms. Individuals can be infected via various routes: (I) ingestion of undercooked or raw meat and fish (deriving from intermediate animal hosts of the helminths) contaminated with cysts and larvae of cestodes, some nematodes (i.e., *Anisakis* sp. and *Trichinella* spp.) and some trematodes (i.e., *Fasciolopsis buski*); (II) ingestion or inhalation of soil particles containing mature eggs of some nematodes (i.e., *Ascaris lumbricoides* and *Enterobius vermicularis*); (III) direct contact of the bare skin with humid soil and freshwater containing infective larvae of some nematodes (i.e., *Ancylostoma duodenale*, *Necator americanus*, *Strongyloides* spp. and *Trichuris trichiura*) and some trematodes (i.e., *Schistosoma* spp.) (Jong, 2002). Symptomatic helminthic infections are not lethal for the human host, but are usually associated with chronic inflammation, diarrhoea, malnutrition, weight loss, and delayed physical and cognitive growth of children (Haque, 2007).

Generally, parasitic infections are treated with efficacious and broad-spectrum antibiotic and anti-parasitic drugs, which allow reducing morbidity, mortality and transmission of pathogens. In addition, encouraging therapeutic strategies based on the administration of probiotics and prebiotics are emerging: probiotics consist in preparations of living microorganisms that confer health benefits to the host by performing the same protective functions of the resident gut microbiota; prebiotics promote the growth of beneficial commensals to re-establish an eubiotic state, such that the host-associated microbes can effectively contrast the infection (Leung et al., 2018). It is important to remember that interactions between microbiota and parasites can be different, indeed some pathogens compete with commensals to establish a successful infection while others rely on microbial cues to become more virulent (i.e., "feeding-dependent activation" (Berrilli et al., 2012)), hence decisions about the therapeutic intervention to adopt must be taken according to the pathogens' characteristics.

2.3. Diagnosis of intestinal parasitic infections

Clinicians may suspect the presence of intestinal pathogenic parasites in travellers affected by abdominal pain and diarrhoea, immunocompromised patients with chronic diarrhoea, migrants coming from countries with poor sanitation, individuals with history of skin contact with humid soil and freshwater in parasites-endemic areas, and people with exotic tastes in food (e.g., undercooked or raw meat and fish) (Jong, 2002). To confirm their suspicions, clinicians must perform diagnostic tests that reveal the infecting microorganism(s): with accurate results, they can administer the proper therapeutic treatment to eradicate the parasite(s), cure the patient, and prevent the pathogen transmission. Diagnosis can be performed via (I) microscopic examination of faeces, (II) cultivation, (III) serological methods, (IV) antigen-detection tests, and (V) nucleic acid-based analyses. Pros and cons of each method are summarised in **Table 1**.

Since the traditional diagnostic techniques have many pitfalls, in particular poor sensitivity and specificity, modern and better performing molecular procedures can be executed alongside them to assure performing diagnoses that are more reliable. In particular, shotgun metagenomics (MG) and polymerase chain reaction (PCR) -based test are two of the most promising molecular approaches, based on DNA sequencing and bioinformatics, able to detect pathogenic protozoa or helminths. Briefly, shotgun MG consists in sequencing every DNA molecule contained in a faecal sample, allowing the identification of the whole gut microbial community (i.e., bacteriome, virome and "eukaryome" collections) and the detection of any potentially infecting intestinal parasite. This technique has a high sensitivity and a high level of protocol standardisation: indeed, it can be used with no restrictions on any sample to detect even complex populations. However, it is highly laborious and expensive, requires specialised personnel, and delivers massive genetic information that must be extensively processed to obtain the final results (Lokmer et al., 2019). On the other hand, the PCR-based test consists in sequencing only specific DNA fragments, termed amplicons, that are the result of selective amplification of target genes by using properly designed oligonucleotide primer pairs. The amplified regions must be taxonomically informative, meaning that each sequence can be used to undoubtedly determine the presence of a certain organism. By applying this approach on faecal samples, it is possible to diagnose the occurrence of protozoan or helminthic infections, but the knowledge of the patient's clinical-epidemiological background is fundamental to properly decide which target genes should be analysed to confirm the assumed parasite's presence. The PCR-based test was applied for the first time in the parasitological field to discriminate between the morphologically indistinguishable *E. dispar* (non-pathogenic) and *E. histolytica* (pathogenic) by sequencing the small subunit ribosomal RNA (SSU rRNA) gene,

DIAGNOSTIC TECHNIQUE		DESCRIPTION	PROS	CONS	DIAGNOSIS	REFERENCES
TRADITIONAL METHODS	MICROSCOPIC ANALYSIS	Identification of eggs and adult parasites in fresh or fixed faeces	Gold standard approach	- Poor sensitivity - Does not distinguish organisms with same morphology - Complex sample preparation - Needs multisampling - Operator-dependent accuracy	Every parasite via specific tests, some diagnoses are very difficult (e.g., *Entamoeba histolytica*, *Strongyloides* spp. and *Trichinella* spp.)	(Hamad et al., 2016) (AMCLI-CoSP, 2022)
	CULTIVATION	*In vitro* growth of parasites isolated from fresh faeces	Test therapeutic efficacy of new drugs	- Poor sensitivity - Most parasites do not grow *in vitro*	- *Ancylostoma duodenale* - *Blastocystis* spp. - *Dientamoeba fragilis* - *Entamoeba* spp. - *Necator americanus* - *Strongyloides stercoralis* - Intestinal fungi	(Hamad et al., 2016) (AMCLI-CoSP, 2022)
	SEROLOGY-BASED APPROACH	Identification of parasite-specific Abs in patient's serum		- Variable sensitivity and specificity - Does not distinguish between present or past infection	Invasive pathogens (e.g., *Entamoeba histolytica*) and persistent infections (e.g., *Schistosoma* spp.)	(AMCLI-CoSP, 2022)
	ANTIGEN-DETECTION TEST	Identification of parasitic Ags in faeces	90% accuracy	Cross-reactions with non-target Ags and other Abs	- *Cryptosporidium* spp. - *Entamoeba histolytica* - *Giardia intestinalis* - *Schistosoma* spp. - *Strongyloides* spp. - *Taenia solium* - *Trichinella* spp.	(Haque, 2007) (Hamad et al., 2016)
NUCLEIC-ACID BASED ANALYSES	SHOTGUN MG	Sequencing of every DNA molecule in faecal sample	- High sensitivity - Standardised method - Identify whole, complex community	- Very laborious and expensive - Needs specialised personnel	Every parasite	(Lokmer et al., 2019)
	PCR-BASED TEST	Sequencing of taxonomically informative amplicons	- High sensitivity and specificity - Fast results	- Needs knowledge of patient's clinical-epidemiological background - Variable accuracy	- *Blastocystis* spp. - *Cryptosporidium* spp. - *Dientamoeba fragilis* - *Entamoeba histolytica* - *Giardia intestinalis* - *Schistosoma* spp. - *Taenia* spp.	(van Lieshout and Verweij, 2010) (Hamad et al., 2016)

Table 1. Description of the techniques exploited to diagnose intestinal parasitic infections.

and was subsequently extended to the identification of other parasites (e.g., *Blastocystis* STs, *Cryptosporidium* spp. and *Giardia intestinalis*) by analysing other targets. Due to its unmatched sensitivity and specificity, this technique can identify even low protozoa or helminth burdens and provides fast results by analysing a single faecal sample. Nevertheless, it allows detecting only a reduced amount of eukaryotic parasites and has a diagnostic accuracy that depends on several factors: the DNA extraction efficacy, the correct choice of the amplified target genes, the absence of primer biases (e.g., preferential amplification of more abundant DNA molecules that causes an underestimation of the frequency of poorly abundant genotypes and subtypes) and the lack of contamination (van Lieshout and Verweij, 2010) and (Hamad et al., 2016).

Even though the molecular diagnostic approaches are very consistent, nowadays they can only support the traditional microscopic examination of faeces, which is still the gold standard technique that must be performed in any case to provide clinically relevant diagnoses. Anyway, by developing new reagents and more standardised diagnostic protocols, in the future the modern methods might be exploited in place of the traditional ones, in particular when samples are collected from patients with a clear clinical history (e.g., travellers in parasites-endemic regions) and hence can be analysed by using specific diagnostic panels.

2.4. DNA metabarcoding as a diagnostic tool

Despite their limitations, molecular approaches are increasingly exploited in parasitology to take advantage of their high sensitivity and specificity; in particular, an evolution of the PCR-based tests, named "metabarcoding", is becoming more and more successful as it can economically characterise whole populations, even at species level, without knowing the clinical-epidemiological background of patients.

DNA metabarcoding is primarily exploited in ecological studies to investigate microbial, vegetal and metazoan (i.e., animal) communities, and consists in a PCR-based analysis of environmental DNA (eDNA; i.e., mixture of extracellular and intracellular genetic material from many different organisms) extracted from environmental samples, such as filtered air, water, sediment, soil, faeces, etc.. One of the main aims of this approach is to describe the biodiversity of a given ecosystem by amplifying and sequencing a single target gene (in this case termed "metabarcode") selected from the eDNA, thus the chosen gene must be taxonomically informative for a very broad spectrum of organisms (e.g., all prokaryotes, all plants or all Metazoa). The ideal metabarcode is a short DNA fragment composed by a highly variable sequence flanked by two conserved

regions: while the central sequence is unique for each species, the flanking regions are the same in every species of the target taxonomic group, and constitute the annealing sites for the DNA metabarcoding universal primers (Taberlet et al., 2018). As a matter of fact, after eDNA extraction the metabarcode variable sequence is amplified with a primer pair properly designed to bind to the flanking conserved regions, and the generated amplicons are sequenced using NGS platforms; then, by exploiting bioinformatics tools that compare the produced sequences with reference databases containing genetic data of the target taxa, each amplicon is associated to a unique species. Overall, DNA metabarcoding determines which components of the target taxonomic group are populating the sampled environment by analysing the variability of the amplified metabarcode central sequence.

The development of the DNA metabarcoding technique that is employed in the present days dates back to the 2006, when Sogin and colleagues provided a global description of the microbial diversity in the sea by amplifying and sequencing the V6 hypervariable region of the prokaryotic SSU rRNA gene, also known as the 16S rRNA gene (Sogin et al., 2006). This gene had been exploited for the taxonomic identification of bacteria and archaea since the 1980s (Robinson et al., 2010), as it is composed of 9 variable regions (V1 to V9), highly different among the various prokaryotes, interspersed between conserved regions, which sequences are identical across the Bacteria and Archaea domains. Therefore, it contains metabarcodes that can be exploited in DNA metabarcoding experiments. Similarly, the eukaryotic counterpart of the 16S rRNA gene, named 18S rRNA, is composed of an alternation of conserved and variable sequences (V1 to V9 as for the 16S rRNA nomenclature): the V4 and V9 regions show the highest degree of variability, while the V6 region is quite conserved and thus scarcely informative (Hadziavdic et al., 2014). Only in the most recent years, researchers have begun to exploit DNA metabarcoding to study eukaryotes, prevalently the microbial ones (i.e., Protists), by sequencing the V4 region; indeed, it offers the highest phylogenetic resolution and allows a more precise taxonomic assignment of the generated amplicons due to its optimal length and its good coverage in reference databases (Vaulot et al., 2022). Thus, the V4 region of the 18S rRNA gene can be considered the gold standard metabarcode for eukaryotic classification.

Nowadays, several DNA metabarcoding studies focus on the microbial eukaryotes' diversity in different environments, and high relevance is given to the examination of faecal samples to complete the characterization of the "eukaryome" collection in the human host, both in healthy and pathological conditions. Therefore, it could be possible to implement this approach to the clinical detection of intestinal pathogenic

protozoa and helminths, providing cost-effective diagnoses with high sensitivity (since it is a molecular technique) and without requiring a previous knowledge about the clinical-epidemiological background of the patient (the main limitation of the classic PCR-based tests). In addition, it would allow the simultaneous evaluation of how the gut microbiota might be affected by the pathogenic infection, providing a comprehensive picture of the patient's pathophysiological status based on their host-associated microbial communities.

2.5. Sequencing Technologies

All the above-cited molecular methods rely on sequencing the target nucleic acids and exploit bioinformatics tools to determine their organism(s) of origin, which populate the sampled environment.

The DNA sequencing history started almost 50 years ago, when Sanger and colleagues published their work based on chain-terminating and labelled dideoxynucleotides, thus developing the so-called "Sanger" or "First-generation" sequencing technique. This approach is still used in the present days, and for more than 30 years it was the only available sequencing method. However, a new technology able to produce a very large amount of data (e.g., many human genomes) in a single sequencing run was developed in 2006 and revolutionised the market. Indeed, compared to the previous method, "Second-generation" sequencing (better known as "Next Generation Sequencing" (NGS)) approaches allowed a huge decrease in operational time and costs through the parallel production of millions of high-quality short (250-800 bp) sequences, called reads. Despite the large number of analyses in which NGS is applied (e.g., metabarcoding, metagenomics, evaluation of gene expression, discovery of genetic variants, and many more), it is not the best choice for those applications requiring longer sequences, such as genome assembly or detection of alternative splicing isoforms. Hence, to overcome the limitations of NGS mainly derived from short reads production, in the last decade a couple of companies (i.e., Oxford Nanopore and PacBio) have developed new machineries capable of producing up to 100 kb long reads. These "Third-generation" sequencing approaches usually show a simpler sample preparation and a lower sequencing runtime than "Second-generation" methods, but they have lower throughput and accuracy, and a higher per-base cost. Therefore, to exploit the advantages of NGS and "Third-generation" techniques, researchers have developed hybrid methods that combine both kinds of reads; an example of their application is genome assembly, where long reads are used to build a scaffold on which high-quality short reads are mapped (Hu et al., 2021).

2.5.1. Next Generation Sequencing

Although NGS methods have already been described in the previous paragraph, the Illumina sequencing workflow is here briefly presented, as it constituted the NGS technology used in the experimental protocol proposed in this manuscript.

After obtaining the target DNA molecules for sequencing, they must be processed to prepare a sequencing library that is compatible with the used platform. Hence, adapters are ligated to both ends of each template: they include a binding region, a primer region and, if the library consists in a pool of DNA fragments belonging to multiple samples, a unique barcode sequence that allows identifying each sample. When the denatured library is loaded into the sequencer's flow cell, each single strand molecule hybridises to a surface-bound oligonucleotide which sequence is complementary to the adapter's binding region. Hybridised fragments are then clonally amplified through bridge amplification to create clusters, so that the fluorescent signals emitted during the sequencing step are more intense and more easily detected. Forward and reverse strands of each DNA molecule are sequenced in two separate phases using the "sequencing-by-synthesis" chemistry: the sequencing primer anneals to the adapter's primer region, and the hybridised fragment acts as a template for the synthesis of the complementary strand, which sequence is defined base-per-base while fluorescently labelled nucleotides (reversible dye-terminators) are incorporated in it. Finally, the formed reads are assigned to each sample of the library thanks to the unique barcode they possess, and the output genetic data can be analysed by bioinformatics (Bentley et al., 2008).

2.6. Bioinformatics Analysis

Bioinformatics "can be defined as the application of computational tools to organize, analyze, understand, visualise and store information associated with biological macromolecules" (Diniz and Canduri, 2017). Indeed, it integrates biology, biochemistry, mathematics and statistics into pipelines (i.e., defined sequences of data-processing algorithms) that allow the interpretation of extremely vast amounts of complex "omic" data resulting from molecular biology experiments (e.g., genomics, transcriptomics and proteomics). Bioinformatics tools can be exploited to analyse DNA sequences (e.g., identification of mutations, taxonomical classification of sequenced molecules, etc.), RNA sequences (e.g., evaluation of gene expression, identification of splicing isoforms and non-coding RNAs, etc.) and proteins (e.g., prediction of protein structure and function, evaluation of protein-protein interactions, etc.) (Diniz and Canduri, 2017).

As previously mentioned, NGS and the associated bioinformatics analysis revolutionised, among others, the investigation of microbial communities, allowing the development of shotgun MG and DNA metabarcoding -based approaches that are far more sensitive and specific than the traditional microscopy and culture -based methods. In particular, to characterise the species present in a specimen under analysis, reads generated by the high throughput sequencing of the target DNA molecules are usually processed with bioinformatics tools for performing 2 major steps: reads pre-processing and taxonomical classification. The first step involves processing the reads (e.g., quality filtering, trimming and denoising) to produce high-quality sequences representing the target DNA molecules. The second step consists in aligning the high-quality sequences to reference databases of genetic data so that it is possible to assign each sequenced DNA fragment to a defined organism.

2.6.1. 18S Databases

Since the volume of experimental data constantly increases, databases are fundamental to store and share the produced information, which can be exploited by the scientific community to interpret new experimental results and contribute to the growing of the scientific knowledge (Diniz and Canduri, 2017).

In the context of a DNA metabarcoding-based analysis of microbial eukaryotes, the reads produced by sequencing the gold standard metabarcode (i.e., the V4 region of the 18S rRNA gene) can be taxonomically classified by aligning them to the SILVA database (Quast et al., 2013) or the PR^2 database (Guillou et al., 2013). Between them, the PR^2 database should be preferred since it was specifically constructed for the analysis of the eukaryotic SSU rRNA gene, and overcomes several issues observed with SILVA and other reference databases to guarantee its ease-of-use and obtaining precise results. Briefly, it contains over 220,000 sequences accurately annotated using nine taxonomic levels (from domain to species): they are mainly 18S rRNA gene sequences of Protists, but also SSU rRNA gene sequences of Fungi, plants, Metazoa and a limited set of Bacteria and Archaea are included (Guillou et al., 2013).

For the above-cited reasons, the PR^2 database was exploited in the present thesis.

3. OBJECTIVES

DNA metabarcoding is a widespread technique useful for the characterisation of microbial communities, mainly bacteria (by analysing the 16S rRNA gene) and fungi (by analysing the ITS regions). However, its application for eukaryotic parasites' detection is still limited, as many of the metabarcodes designed on the 18S rRNA gene are not able to properly cover some taxa of interest. In addition, the length of the resulting amplicons usually exceeds 550 bp, preventing the usage of the bioinformatics pipelines that are commonly exploited for 16S and ITS analyses.

With these premises, this thesis aimed at developing a DNA metabarcoding protocol for the detection of protozoa and helminths parasitizing the human gastrointestinal tract, achieved by analysing the DNA extracted from human faecal samples.
The first part of this project included some bioinformatics analyses for the identification of suitable DNA metabarcoding universal primer pairs that would allow the amplification of the eukaryotic SSU rRNA V4 region. The second part of this project consisted in evaluating the feasibility of the developed DNA metabarcoding protocol in the detection of protozoa and helminths contaminating human faecal samples.

More in details, during the first part of the thesis we:

- Compared the properties of several literature-available universal primer pairs that amplify the 18S rRNA V4 region, in order to select the most suitable ones for developing the protocol.
- Tested *in silico* the taxonomic coverage of the selected primer pairs, in order to identify the ones that better amplify the target human eukaryotic intestinal parasites.

The objectives of the second part of the thesis were:

- To optimise the DNA metabarcoding protocol for a correct DNA extraction and target metabarcode amplification.
- To develop a bioinformatics pipeline for analysing the reads produced by sequencing the generated amplicons with an Illumina platform.

4. MATERIALS AND METHODS

4.1. Primer pairs selection

29 universal primer pairs targeting the hypervariable V4 and V9 regions of the eukaryotic 18S rRNA gene were obtained from the available literature (Bates et al., 2012), (Hadziavdic et al., 2014), (Hugerth et al., 2014), (Parfrey et al., 2014), (Bradley et al., 2016), (Moreno et al., 2018), (Del Campo et al., 2019), (Kounosu et al., 2019), (Choi and Park, 2020) and (Vaulot et al., 2022). They were filtered to discard those predicted to frequently amplify bacterial and archaeal genes (off-targets in this analysis) or to produce amplicons with mean length incompatible with the Illumina MiSeq 300PE sequencing strategy (i.e., shorter than 280 bp or longer than 550 bp) (see chapter 4.11). The remaining primer pairs were tested by *in silico* PCR to evaluate their coverage of bacterial, archaeal and eukaryotic SSU rRNA gene sequences using the PR^2 primer database (v. 2.0.0) (Vaulot et al., 2022), under the condition "Max mismatches = 1" for both forward and reverse primer sequences.

The PR^2 primer database (v. 2.0.0) contains "179 primers and 76 primer pairs that have been used for eukaryotic 18S rRNA metabarcoding" (Vaulot et al., 2022) and target the V4 and V9 regions. It offers an R-based web application that, among other options (e.g., download the pr2-primers database), allows testing by *in silico* PCR any user-defined primer set against the PR^2 database (v. 5.0.0) and the SILVA database (release 132). As the test is completed, it is possible to download the results as a .tsv file: for each sequence of the PR^2 database, indicated with its unique PR^2 accession number and a nine-level taxonomy, it is reported the potential annealing position of forward and reverse primers and, if both are predicted to bind to the sequence, the length of the produced amplicon.

The results obtained from the *in silico* PCR tests were further processed to focus on the primer pairs' ability to amplify the target metabarcode of the most common human eukaryotic intestinal parasites (i.e., *Ancylostoma duodenale*, *Anisakis sp.*, *Ascaris lumbricoides*, *Blastocystis spp.*, *Cryptosporidium spp.*, *Cyclospora cayetanensis*, *Cystoisospora belli*, *Dientamoeba fragilis*, *Diphyllobothrium dendriticum*, *Diphyllobothrium latum*, *Diphyllobothrium nihonkaiense*, *Diphyllobothrium pacificum*, *Entamoeba histolytica*, *Enterobius vermicularis*, *Fasciolopsis buski*, *Giardia intestinalis*, *Hymenolepis diminuta*, *Hymenolepis microstoma*, *Hymenolepis nana*, *Necator americanus*, *Neobalantidium coli*, *Schistosoma intercalatum*, *Schistosoma japonicum*, *Schistosoma mansoni*, *Schistosoma mekongi*, *Strongyloides fuelleborni*, *Strongyloides stercoralis*, *Taenia asiatica*, *Taenia saginata*, *Taenia solium*, *Toxoplasma gondii*, *Trichinella spp.* and *Trichuris trichiura*).

Total sequences of the PR^2 database were filtered to identify those of the parasites of interest; then, we evaluated the coverage of each primer pair and the lengths of the produced amplicons. The 3 best performing primer pairs were selected for the DNA metabarcoding protocol.

4.2. 5'-modification of the selected primer pairs

Since the DNA metabarcoding protocol required sequencing the target metabarcodes on the Illumina MiSeq platform (see chapter 4.11), we appended Illumina Adapter Sequences to the 5'-end of the selected forward and reverse primers to allow the following preparation of the sequencing libraries. The Read 1 Nextera Transposase Adapter (5' TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG) (Illumina) was appended to the forward primers, while the Read 2 Nextera Transposase Adapter (5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG) (Illumina) was appended to the reverse primers.

4.3. Standard PCR amplification

5 µL of total DNA were added to 20 µL of master mix (MM) containing 11.5 µL of AccuStart II PCR SuperMix (Quantabio), 6.2 µL of H_2O, 1.15 µL of 10 µM forward primer solution and 1.15 µL of 10 µM reverse primer solution. The PCR amplification consisted in an initial denaturation at 94°C for 3 min, followed by 8 cycles (94°C for 20 s, 46°C for 20 s and 72°C for 40 s) and 27 cycles (94°C for 20 s, 54°C for 20 s and 72°C for 40 s). The products were visualised using gel electrophoresis (1.5% agarose gel).

4.4. Primer pairs amplification test

The selected and 5'-modified primer pairs were bought from biomers.net GmbH.
To check their amplification efficiency, they were used to amplify positive controls composed of purified eukaryotic genomes already available in the laboratory. We followed the standard PCR amplification protocol.

4.5. Human faecal samples

We analysed 5 formalin-fixed human faecal samples (sample A-E) that were obtained from UKNEQAS Parasitology; previous microscopic examinations demonstrated their contamination by protozoa and helminths (**Table 2**).

SAMPLE	CONTAMINATION	PRIMER PAIR 1	PRIMER PAIR 2	PRIMER PAIR 3
A	*Taenia* and *Giardia*	1460764	1460769	1460774
B	*Trichomonas* and *Strongyloides*	1460765	1460770	1460775
C	*Diphyllobothrium*	1460766	1460771	1460776
D	*Fasciolopsis*	1460767	1460772	1460777
E	*Schistosoma*	1460768	1460773	1460778

Table 2. For each sample, it is reported the contamination declared by UKNEQAS Parasitology, along with the ID number we gave them during our analyses with the selected primer pairs.

4.6. Standard DNA extraction

300-500 µL of formalin-fixed faecal sample were aliquoted in a 2 mL Microcentrifuge tube (VWR International) containing 100-200 µL of Zirconia/Silica beads 0.1 mm of diameter (BioSpec). 800 µL of Solution CD1 (QIAGEN) were added, and the sample was incubated at 75°C for 15 min for thermal lysis. The lysed sample was then homogenised by mechanical shaking with the TissueLyser II (QIAGEN) at 25 Hz for 10 min, and centrifuged at 15,000 g for 1 min. Subsequently, 550 µL of supernatant were transferred to an S-block (QIAGEN), added with 250 µL of Solution CD2 (QIAGEN) to precipitate non-DNA organic and inorganic material, mixed by pipetting, and centrifuged at 4,500 g for 6 min. At last, 550 µL of supernatant were transferred to a new S-block, and DNA extraction was carried out automatically by the QIAcube HT instrument following the DNeasy 96 PowerSoil Pro protocol (QIAGEN).

4.7. DNA purification

126 µL of AMPure XP Beads (Beckman Coulter) were added to 70 µL of total DNA and mixed by pipetting; after 5 min, the sample was placed on a magnet for 5 min and then the supernatant was discarded. The beads were washed 2 times with 200 µL of EtOH 80% and air-dried, before adding 15 µL of H_2O; after 2 min, the sample was placed on a magnet for 2 min and then the DNA-enriched supernatant was collected.

4.8. Samples pre-treatment

Formalin-fixed faecal samples were subjected to a pre-treatment to remove the formalin fixative. 500 µL of stool were aliquoted in a 2 mL tube, washed with 400 µL of 100% EtOH, and centrifuged at 16,000 g for 10 min. The supernatant was discarded and the sample was air-dried to remove the EtOH residue, before adding 400 µL of 70% EtOH and repeating centrifugation (16,000 g for 10 min), supernatant discarding and air-drying.

500 µL of 0.5 M EDTA were added to inhibit DNase activity, and the sample was incubated at 55°C for 1 hr; then, 100 µL of proteinase K were added to remove proteins, the sample was incubated at 55°C for 2 hr, and the supernatant was discarded (Lee et al., 2019).

4.9. Manual DNA extraction

500 µL of pre-treated faecal sample were aliquoted in a 2 mL Microcentrifuge tube containing 100-200 µL of Zirconia/Silica beads 0.1 mm of diameter. 800 µL of Solution CD1 were added, and the sample was incubated at 75°C for 15 min for thermal lysis. The lysed sample was then homogenised by mechanical shaking with the TissueLyser II at 25 Hz for 10 min, and centrifuged at 15,000 g for 1 min. Successively, 550 µL of supernatant were added with 250 µL of Solution CD2, mixed by pipetting, and centrifuged at 4,500 g for 6 min. 550 µL of supernatant were transferred in a new tube with 450 µL of Solution CD3 (QIAGEN) and mixed by pipetting; then, the 1 mL solution was transferred in a spin column and centrifuged at 14,000 g for 1 min. The filtered solution was removed, before adding 400 µL of Solution AW1 (QIAGEN) to the spin column and centrifuging at 14,000 g for 1 min. This step was repeated three more times, the first one with 600 µL of Solution AW1, the second one with 600 µL of Solution AW2 (QIAGEN), and the third one with 400 µL of EtOH 96%. Then, the filtered solution was removed and the sample was centrifuged at 14,000 g for 3 min. 60 µL of Solution C6 (QIAGEN) were added to the spin column, which was placed in a new tube and centrifuged at 14,000 g for 1 min; subsequently, the filtered solution was collected and poured in the same spin column, which was centrifuged again at 14,000 g for 1 min to allow the collection of 60 µL of total DNA.

4.10. Optimised PCR amplification

11.2 µL of total DNA were added to 13.8 µL of master mix (MM*) containing 11.5 µL of AccuStart II PCR SuperMix (Quantabio), 1.15 µL of 10 µM forward primer solution and 1.15 µL of 10 µM reverse primer solution. The PCR amplification consisted in an initial denaturation at 94°C for 3 min, followed by 8 cycles (94°C for 20 s, 46°C for 20 s and 72°C for 40 s) and 32 cycles (94°C for 20 s, 54°C for 20 s and 72°C for 40 s). The products were visualised using gel electrophoresis (1.5% agarose gel).

4.11. Sequencing library preparation and metabarcode sequencing

PCR products were subjected to an exonuclease treatment to remove single-strand DNA molecules (e.g., non-annealed primers): 5 µL of amplicons were added to 5 µL of exonuclease solution, consisting in 1 µL of NEBuffer 3.1 (New England Biolabs), 0.5 µL of Exonuclease I (New England Biolabs) and 3.5 µL of H_2O; the sample was then heated at 37°C for 4 min and at 80°C for 1 min.

Subsequently, we created the sequencing library by adding Illumina sequencing adapters and dual-index barcodes to the purified amplicons using the Nextera XT Index Kit v2 Set A (Illumina). The required PCR reaction, performed on 2.5 µL of amplicons, included 22.5 µL of master mix containing 12.5 µL of KAPA HiFi HotStart ReadyMix (Roche), 5 µL of H_2O, 2.5 µL of Nextera XT Index 1 Primers (N7XX) and 2.5 µL of Nextera XT Index 2 Primers (S5XX). The PCR condition consisted in an initial denaturation at 95°C for 3 min, followed by 9 cycles (98°C for 20 s, 55°C for 30 s and 72°C for 40 s) and a final elongation at 72°C for 5 min. The products were visualised using gel electrophoresis (1.5% agarose gel).

Barcoded amplicons were then normalised by means SequalPrep^{TM} Normalisation Plate Kit 96-well (Thermo Fisher Scientific), 5 µL of each sample were pooled in a unique vial, and the formed pool was concentrated in a 30 µL volume by pouring it in a spin column that was centrifuged at 16,000 g for 20 min.

Subsequently, we purified the pool DNA: 30 µL of AMPure XP Beads (Beckman Coulter) were added and mixed by pipetting; after 7 min, the sample was placed on a magnet for 2 min and then the supernatant was discarded; the beads were washed 2 times with 200 µL of EtOH 80% and air-dried, before adding 30 µL of H_2O; after 7 min, the sample was placed on a magnet for 2 min and then the DNA-enriched supernatant was collected.

Quality control of the concentrated and purified barcoded amplicons was performed by capillary electrophoresis with the Agilent 2100 Bioanalyzer instrument (Agilent Technologies) following the manufacturers' instructions.

Successively, 2 µL of the pool were diluted 1:20 with 38 µL of 1X Tris-EDTA pH 8.0 in order to be quantified with the Qubit 2.0 Fluorometer (Thermo Fisher Scientific). 10 µL of the dilution were mixed with 190 µL of Qubit working solution (Thermo Fisher Scientific) and quantified following the manufacturers' instructions.

According to the quality control and quantification results, we prepared the library for sequencing on the Illumina MiSeq platform: 5 µL of 0.2 N NaOH were added to 5 µL of 4 nM pool DNA for chemical denaturation of the amplicons; after centrifugation at 280 g for 1 min and incubation at room temperature for 5 min, 990 µL of HT1 (Illumina) were added. The created 20 pM denatured library was further diluted with HT1 to form 600 µL of 7.5 pM denatured library, from which 60 µL

were discarded and replaced by 60 µL of denatured 20 pM PhiX control library (10% PhiX control spike-in).
In the end, the combined library was heat denatured at 95°C for 2 min, and then placed on ice for 5 minutes or until it was loaded in the pre-filled, ready-to-use reagent cartridge of the MiSeq Reagent Kit v3 (600-cycle) (Illumina). Thus, the library was loaded in the Illumina MiSeq platform and sequenced with the 300PE strategy.

The Illumina MiSeq 300PE strategy allows sequencing DNA fragments with a length between 280 bp and 550 bp. As a matter of fact, the system is designed to sequence 300 bp at both ends of the DNA fragment, creating two reads that should overlap at their 3'-end. Then, they can be merged into a single contig that represents with high accuracy the original DNA fragment sequence. If the fragment is shorter than 280 bp, both 300-bases-reads will have random nucleotides at their 3'-end due to the NGS platform sequencing features; if the fragment is longer than 550 bp, the forward and reverse reads will not overlap.

4.12. Bioinformatics analysis of <550 bp amplicons

NGS-produced raw reads, available as FASTQ format files, were downloaded from BaseSpace (Illumina) and analysed using FastQC (v. 0.11.9) (Andrews, 2010) for a preliminary quality control of the sequencing run. QIIME 2 (v. 2023.5) (Bolyen et al., 2019) was used for further analyses. In particular, raw reads were imported into QIIME 2 environment exploiting the import command of the q2-tools plugin with `SampleData[PairedEndSequencesWithQuality]` as type and `PairedEndFastqManifestPhred33V2` as input format. Then, forward and reverse primer sequences and residual adapters were removed with q2-cutadapt (Martin, 2011). We exploited the q2-dada2 plugin (Callahan et al., 2016) to perform trimmed reads filtering (i.e., truncation of the reads at a specified position and removal of shorter reads), dereplication, denoising (i.e., removal of reads with sequencing errors) and merging (i.e., creation of Amplicon Sequence Variants - ASVs - by merging overlapping forward and reverse reads), and to remove chimeric sequences. Subsequently, the RESCRIPt plugin (Robeson et al., 2021) was used to train a naive Bayes taxonomic classifier based on the PR^2 database (v. 5.0.0) (downloaded on 01/05/2023 from https://github.com/pr2database/pr2database/releases). Taxonomy was assigned to the ASVs by exploiting the `classify-sklearn` method of the q2-feature-classifier plugin (Bokulich et al., 2018).
Workflow parameters slightly changed depending on the selected primers pair and are summarised in **Table 3**.

QIIME 2 PLUGIN	PARAMETER	PRIMER PAIR		
		515FY - NSR951	E572f - 897r	616*f - 1132r
q2-cutadapt	forward_adapter_and _primer_sequence	TCGTCGGCAGCGTCAGAT GTGTATAAGAGACAGGTG YCAGCMGCCGCGGTA	TCGTCGGCAGCGTCAGAT GTGTATAAGAGACAGCYG CGGTAATTCCAGCTC	TCGTCGGCAGCGTCAGAT GTGTATAAGAGACAGTTA AARVGYTCGTAGTYG
	reverse_adapter_and _primer_sequence	GTCTCGTGGGCTCGGAGA TGTGTATAAGAGACAGTT GGYRAATGCTTTCGC	GTCTCGTGGGCTCGGAGA TGTGTATAAGAGACAGTC YDAGAATTYCACCTCT	GTCTCGTGGGCTCGGAGA TGTGTATAAGAGACAGCC GTCAATTHCTTYAART
q2-dada2	read1_truncation_position	280	280	280
	read2_truncation_position	230	230	230
q2-feature-classifier	forward_primer_sequence	GTGYCAGCMGCCGCGGTA	CYGCGGTAATTCCAGCTC	TTAAARVGYTCGTAGTYG
	reverse_primer_sequence	TTGGYRAATGCTTTCGC	TCYDAGAATTYCACCTCT	CCGTCAATTHCTTYAART
	primer_pair_name	515FY-NSR951	E572f-897r	616f-1132r
RESCRIPt	primer_pair_name	515FY-NSR951	E572f-897r	616f-1132r

Table 3. Parameters exploited when working with QIIME 2.

4.13. Bioinformatics analysis of amplicons of any length

The selected primer pairs could also produce amplicons longer than 550 bp, thus the corresponding 300-bases forward and reverse reads would not overlap. Since QIIME 2 requires overlapping reads to work properly, we exploited a different approach to further analyse the sequencing results.

Raw reads were pre-processed with fastp (v. 0.23.2) (Chen et al., 2018): low-quality reads were filtered out and the remaining were trimmed to remove primer sequences and residual adapters. Differently from the previous pre-processing method, filtered reads were not merged. Unmerged, pre-processed reads were then aligned in paired-end mode to the sequences of the PR^2 database (v. 5.0.0) with Bowtie 2 (v. 2.5.1) (Langmead and Salzberg, 2012), and finally, the number of paired-end reads aligned to each sequence of the PR^2 database was assessed with featureCounts (v. 2.0.1) (Liao et al., 2014). Workflow parameters are summarised in **Table 4**.

BIOINFORMATICS TOOL	PARAMETER	PRIMER PAIR		
		515FY - NSR951	E572f - 897r	616*f - 1132r
fastp	phred_quality	20	20	20
	unqualified_limit	30	30	30
	average_phred_score	25	25	25
	read1_trim_front_position	18	18	18
	read1_trim_tail_position	20	20	20
	read2_trim_front_position	17	18	18
	read2_trim_tail_position	50	50	50
Bowtie 2	alignment_option	--very-sensitive	--very-sensitive	--very-sensitive
	min_fragment_length	400	400	400
	max_fragment_length	2000	2000	2000
featureCounts	min_fragment_length	400	400	400
	max_fragment_length	2000	2000	2000

Table 4. Parameters exploited when working with fastp + Bowtie 2 + featureCounts.

To focus our analysis on the human eukaryotic intestinal parasites of interest (see chapter 4.1), we repeated the paired-end alignment of unmerged, pre-processed reads against a user-produced database containing only their 18S rRNA gene sequences. This was created working on the PR^2 database (v. 5.0.0) web application (https://app.pr2-database.org/pr2-database/), where the target taxa were selected on the available filter and the filtered sequences were downloaded as a FASTA format file (on 08/05/2023).

4.14. Analysis of the results

We exploited matplotlib-venn python library (v. 0.11.9) to create several Venn diagrams that quantitatively describe the results.

To compare the results of the analyses performed with QIIME 2 (v. 2023.5) and with fastp (v. 0.23.2) + Bowtie 2 (v. 2.5.1) + featureCounts (v. 2.0.1), we first collapsed the feature table (output of the q2-dada2 plugin) to the species level (level-9) using the command `collapse` of the q2-taxa plugin, and then we generated a custom script for evaluating which taxonomies were identically assigned by both pipelines. Indeed, the script (see "`final_comparison.py`" in chapter 9.1.5) takes as inputs the level-9 collapsed feature table and the output of featureCounts, and returns a file reporting the shared taxonomies.

The exploited bioinformatics pipelines and scripts are described in **9. SUPPLEMENTARY INFORMATION**. Directories (in light blue) are indicated with a general term: each user who wants to follow the pipelines should name them as they please. Parameters written in green are reported in **Table 3** and **Table 4**.

5. RESULTS

5.1. Primer pairs selection

In this thesis, we focused on the detection of human eukaryotic intestinal parasites in faecal samples and we decided to use the Illumina MiSeq 300PE strategy to sequence the amplified target metabarcodes. Therefore, literature-available information about each of the 29 selected universal primer pairs (**Table 5**) was exploited to perform a preliminary filtering and remove those pairs that would be inadequate for the following analyses. Primer pairs #1 (Uni18Sf - Uni18Sr (Zhan et al., 2013)) and #2 (3NDf - V4 euk R1 (Bråte et al., 2010)) were discarded because the research groups that firstly exploited them had reported frequent amplification of prokaryotic sequences; primer pairs #27 (1380f - 1510r (Amaral-Zettler et al., 2009)), #28 (1389f - 1510r (Amaral-Zettler et al., 2009)) and #29 (1391f - EukB (Stoeck et al., 2010)) were castoff because produce amplicons with mean length shorter than 280 bp; primer pairs #5 (515f - 1119r (Parfrey et al., 2014)), #6 (563f - 1132r (Hugerth et al., 2014)) and #23 (18S-EUK581f - 18S-EUK1134r (Carnegie et al., 2003)) were removed because they produce amplicons with mean length longer than 550 bp. Even if primer pair #20 (574*f - 1132r (Hugerth et al., 2014)) produces 580 bp-long amplicons on average, it was not discarded in the first filtering because Hugerth and colleagues had demonstrated its high efficiency in amplifying the eukaryotic SSU rRNA gene; thus, we planned to further analyse it.

In silico PCR tests performed on the PR^2 primer database with the remaining 21 primer pairs allowed evaluating their performance in amplifying the V4 or V9 region of the 18S rRNA gene, while checking for possible off-targets relative to the amplification of the prokaryotic 16S rRNA gene. For each tested primer pair, we generated a graph describing its coverage on different taxonomic groups of the PR^2 database (i.e., Archaea, Bacteria, Amoebozoa, Archaeplastida, CRuMs, Cryptista, Eukaryota_X, Excavata, Haptista, Obazoa, Provora, TSAR and Eukaryota:plas) (**Figures 1A-1U**). Primer pair #3 (3NDf - V4 euk R2 (Bråte et al., 2010)) amplified at least 79.19% of sequences of the considered eukaryotic taxa, with the exception of Excavata (20.53%) and Obazoa (55.15%); the off-target probability was very low (0.00% for Archaea, 1.99% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1A**). Primer pair #4 (515FY - NSR951 (Lambert et al., 2019)) amplified at least 77.43% of sequences of the considered eukaryotic taxa, with the exception of Amoebozoa (55.12%) and Excavata (8.77%); the off-target probability was very low (0.00% for Archaea, 2.10% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1B**). Primer pair #7 (V4 1f - TAReukREV3 (Bass et al., 2016)) amplified at least 75.70% of sequences of the considered eukaryotic taxa, with the exception of Excavata (42.55%); the off-target probability was very low (0.00% for Archaea,

#	PRIMER PAIR	18S rRNA REGION	MEAN AMPLICON LENGTH	FORWARD PRIMER SEQUENCE	REVERSE PRIMER SEQUENCE	REFERENCE
1	Uni18Sf - Uni18Sr	V4	480 bp	AGGGCAAKYCTGGTGCCAGC	GRCGGTATCTRATCGYCTT	Zhan et al., 2013
2	3NDf - V4 euk R1	V4	450 bp	GGCAAGTCTGGTGCCAG	GACTACGACGGTATCTRATCRTCTTCG	Bråte et al., 2010
3	3NDf - V4 euk R2	V4	450 bp	GGCAAGTCTGGTGCCAG	ACGGTATCTRATCRTCTTCG	Bråte et al., 2010
4	515FY - NSR951	V4	400 bp	GTGYCAGCMGCCGCGGTA	TTGGYRAATGCTTTCGC	Lambert et al., 2019
5	515f - 1119r	V4	600 bp	GTGCCAGCMGCCGCGGTAA	GGTGCCCTTCCGTCA	Parfrey et al., 2014
6	563f - 1132r	V4	570 bp	GCCAGCAVCYGCGGTAAY	CCGTCAATTHCTTYAART	Hugerth et al., 2014
7	V4 1f - TAReukREV3	V4	420 bp	CCAGCASCYGCGGTAATWCC	ACTTTCGTTCTTGATYRA	Bass et al., 2016
8	TAReuk454FWD1 - TAReukREV3	V4	400 bp	CCAGCASCYGCGGTAATTCC	ACTTTCGTTCTTGATYRA	Stoeck et al., 2010
9	TAReuk454FWD1 Choi - TAReukREV3 Choi	V4	420 bp	CCAGCAGCCGCGGTAATTCC	ACTTTCGTTCTTGATTAA	Choi and Park, 2020
10	TAReuk454FWD1 - V4 18S Next.Rev	V4	440 bp	CCAGCASCYGCGGTAATTCC	ACTTTCGTTCTTGATYRATGA	Piredda et al., 2017
11	TAReuk454FWD1 - V4RB	V4	440 bp	CCAGCASCYGCGGTAATTCC	ACTTTCGTTCTTGATYRR	Balzano et al., 2015
12	V4f - V4r	V4	420 bp	CCAGCASCYGCGGTAATTCC	ACTTTCGTTCTTGAT	Bradley et al., 2016
13	E572f - 897r	V4	350 bp	CYGCGGTAATTCCAGCTC	TCYDAGAATTYCACCTCT	Enberg et al., 2018
14	EUKAf - EUKAr	V4	440 bp	GCCGCGGTAATTCCAGCTC	CYTTCGYYCTTGATTRA	Moreno et al., 2018
15	E572f - E1009r	V4	440 bp	CYGCGGTAATTCCAGCTC	AYGGTATCTRATCRTCTTYG	Comeau et al., 2011
16	Nex_0597_F - Nex_0964_R	V4	440 bp	CCGCGGTAATTCCAGCTC	GATCCCYYAACTTTCGTTCTTGA	Kim et al., 2016
17	NSF563 - NSR951	V4	380 bp	CGCGGTAATTCCAGCTCCA	TTGGYRAATGCTTTCGC	Mangot et al., 2013
18	528iF - 964iR	V4	400 bp	GCGGTAATTCCAGCTCCAA	ACTTTCGTTCTTGATYRR	Fadeev et al., 2018
19	A-528f - B-706r	V4	350 bp	GCGGTAATTCCAGCTCCAA	AATCCRAGAATTTCACCTCT	Xu et al., 2020
20	574*f - 1132r	V4	580 bp	CGGTAAYTCCAGCTCYV	CCGTCAATTHCTTYAART	Hugerth et al., 2014
21	EK-565f - 18S-EUK1134r	V4	550 bp	GCAGTTAAAAAGCTCGTAGT	TTTAAGTTTCAGCCTTGCG	Simon et al., 2015
22	616*f - 1132r	V4	520 bp	TTAAARVGYTCGTAGTYG	CCGTCAATTHCTTYAART	Hugerth et al., 2014
23	18S-EUK581f - 18S-EUK1134r	V4	600 bp	GTGCCAGCAGCCGCG	TTTAAGTTTCAGCCTTGCG	Carnegie et al., 2003
24	1183f - 1631r	V9	450 bp	AATTTGACTCAACACGGG	TACAAAGGGCAGGGACGTAAT	Hadziavdic et al., 2014
25	V8f - V9r	V9	370 bp	ATAACAGGTCTGTGATGCCCT	CCTTCYGCAGGTTCACCTAC	Bradley et al., 2016
26	Nex_18S_1434_F - Nex_18S_1757_R	V9	370 bp	GAGGCAATAACAGGTCTGTGATG	CAGGTTCACCTACGGAAACCT	Kim et al., 2016
27	1380f - 1510r	V9	180 bp	CCCTGCCHTTTGTACACAC	CCTTCYGCAGGTTCACCTAC	Amaral-Zettler et al., 2009
28	1389f - 1510r	V9	130 bp	TTGTACACACCGCCC	CCTTCYGCAGGTTCACCTAC	Amaral-Zettler et al., 2009
29	1391f - EukB	V9	145 bp	GTACACACCGCCCGTC	TGATCCTTCTGCAGGTTCACCTAC	Stoeck et al., 2010

Table 5. 29 universal primer pairs were selected from the available literature: 23 target the V4 region of the 18S rRNA gene, 6 target the V9 region of the 18S rRNA gene. For each pair, it is reported the mean length of the produced amplicons, the forward and reverse primers sequences, and the reference of the published article from which information for the preliminary filtering of the pairs was obtained.

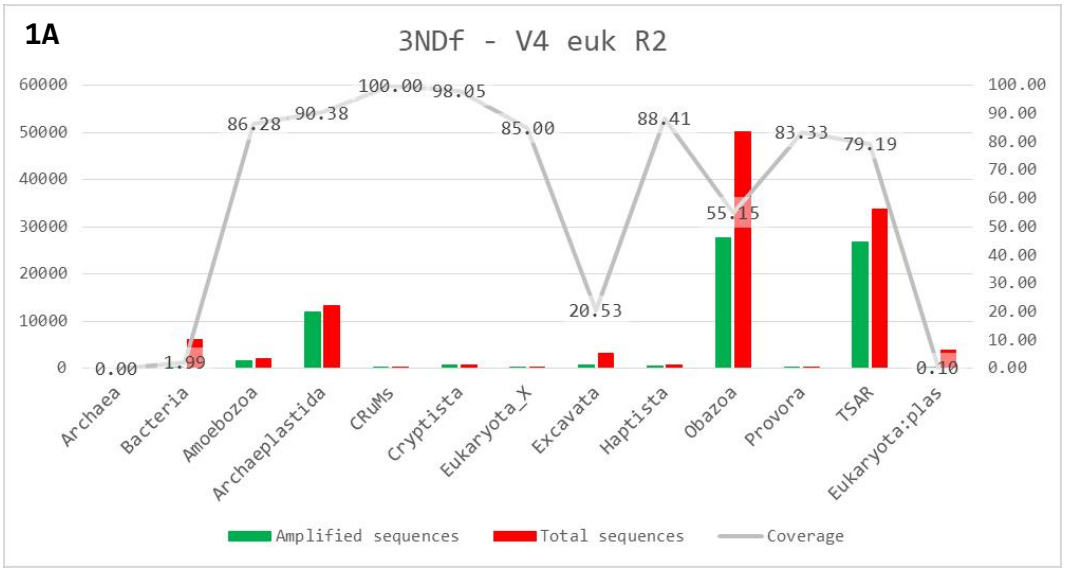
2.00% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1C**). Primer pair #8 (TAReuk454FWD1 - TAReukREV3 (Stoeck et al., 2010)) amplified at least 75.55% of sequences of the considered eukaryotic taxa, with the exception of Excavata (42.36%); the off-target probability was very low (0.00% for Archaea, 2.00% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1D**). Primer pair #9 (TAReuk454FWD1 Choi - TAReukREV3 Choi (Choi and Park, 2020)) amplified at least 71.39% of sequences of the considered eukaryotic taxa, with the exception of Excavata (9.47%), Haptista (21.49%) and Provora (66.67%); the off-target probability was very low (0.00% for Archaea, 2.00% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1E**). Primer pair #10 (TAReuk454FWD1 - V4 18S Next.Rev (Piredda et al., 2017)) amplified at least 71.78% of sequences of the considered eukaryotic taxa, with the exception of Amoebozoa (66.90%) and Excavata (13.27%); the off-target probability was very low (0.00% for Archaea, 2.00% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1F**). Primer pair #11 (TAReuk454FWD1 - V4RB (Balzano et al., 2015)) amplified at least 75.71% of sequences of the considered eukaryotic taxa, with the exception of Excavata (43.24%); the off-target probability was very low (0.00% for Archaea, 2.00% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1G**). Primer pair #12 (V4f - V4r (Bradley et al., 2016)) amplified at least 76.10% of sequences of the considered eukaryotic taxa, with the exception of Excavata (44.03%); the off-target probability was very low (0.00% for Archaea, 2.02% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1H**). Primer pair #13 (E572f - 897r (Enberg et al., 2018)) amplified at least 80.00% of sequences of the considered eukaryotic taxa, with the exception of Amoebozoa (53.14%) and Excavata (59.31%); the off-target probability was very low (0.00% for Archaea, 2.12% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1I**). Primer pair #14 (EUKAf - EUKAr (Moreno et al., 2018)) amplified at least 76.08% of sequences of the considered eukaryotic taxa, with the exception of Excavata (39.87%) and Haptista (21.19%); the off-target probability was very low (0.00% for Archaea, 2.05% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1J**). Primer pair #15 (E572f - E1009r (Comeau et al., 2011)) amplified at least 80.29% of sequences of the considered eukaryotic taxa, with the exception of Excavata (61.57%) and Obazoa (58.44%); the off-target probability was very low (0.00% for Archaea, 2.00% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1K**). Primer pair #16 (Nex_0597_F - Nex_0964_R (Kim et al., 2016)) amplified at least 72.61% of sequences of the considered eukaryotic taxa, with the exception of Excavata (6.54%) and Obazoa (25.07%); the off-target probability was extremely low (0.00% for Archaea, 1.31% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1L**). Primer pair #17 (NSF563 - NSR951 (Mangot et al., 2013)) amplified at least 76.76% of sequences of the considered eukaryotic taxa, with the exception of Amoebozoa (53.75%) and Excavata (10.94%); the off-target probability was very low (0.00%

for Archaea, 2.12% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1M**). Primer pair #18 (528iF - 964iR (Fadeev et al., 2018)) amplified at least 71.60% of sequences of the considered eukaryotic taxa, with the exception of Excavata (46.98%); the off-target probability was very low (0.00% for Archaea, 2.03% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1N**). Primer pair #19 (A-528f - B-706r (Xu et al., 2020)) amplified at least 75.47% of sequences of the considered eukaryotic taxa, with the exception of Amoebozoa (38.09%), Excavata (4.43%), Haptista (21.49%) and Obazoa (63.84%); the off-target probability was extremely low (0.00% for Archaea, 1.41% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1O**). Primer pair #20 (574*f - 1132r (Hugerth et al., 2014)) amplified at least 82.50% of sequences of the considered eukaryotic taxa; the off-target probability was very low (2.88% for Archaea, 1.91% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1P**). Primer pair #21 (EK-565f - 18S-EUK1134r (Simon et al., 2015)) amplified at least 82.16% of sequences of the considered eukaryotic taxa, with the exception of Amoebozoa (63.59%), Eukaryota_X (65.00%), Excavata (14.91%) and Obazoa (36.25%); the off-target probability was extremely low (0.00% for Archaea, 1.20% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1Q**). Primer pair #22 (616*f - 1132r (Hugerth et al., 2014)) amplified at least 75.00% of sequences of the considered eukaryotic taxa, with the exception of Excavata (59.72%); the off-target probability was low (10.58% for Archaea, 1.99% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1R**). Primer pair #24 (1183f - 1631r (Hadziavdic et al., 2014)) amplified at least 78.96% of sequences of the considered eukaryotic taxa, with the exception of Amoebozoa (45.59%), CRuMs (66.67%) and Excavata (52.26%); the off-target probability was extremely low (0.00% for Archaea, 1.15% for Bacteria, 0.10% for Eukaryota:plas) (**Figure 1S**). Primer pairs #25 (V8f - V9r (Bradley et al., 2016)) and #26 (Nex_18S_1434_F - Nex_18S_1757_R (Kim et al., 2016)) amplified less than 50.00% of sequences of all considered eukaryotic taxa; their off-target probability was extremely low (0.00% for Archaea, less than 0.57% for Bacteria, less than 0.05% for Eukaryota:plas) (**Figures 1T-1U**).

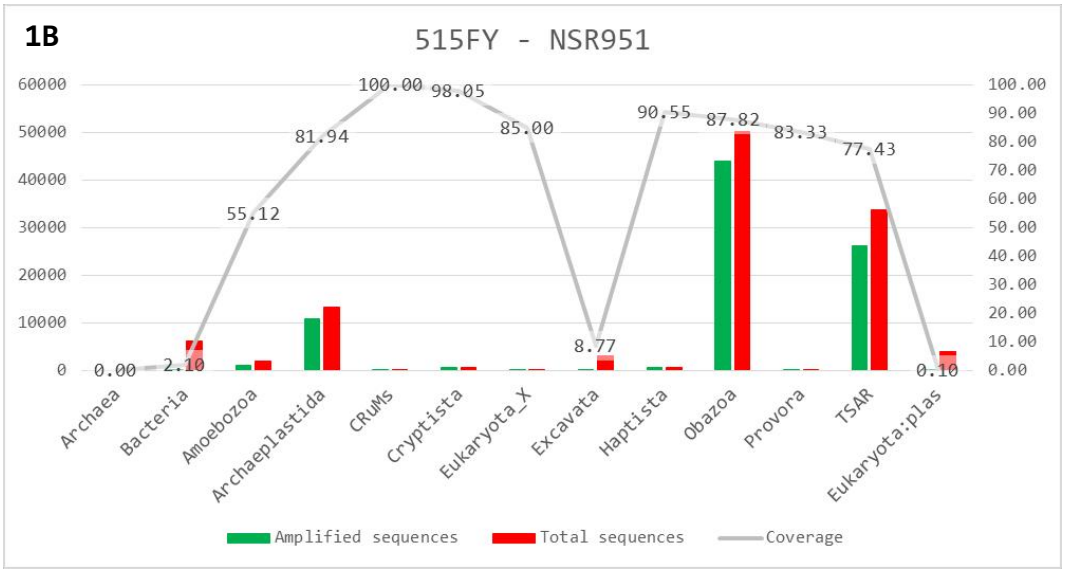
Overall, the tested pairs showed an extremely low rate of off-targets, thus all of them could be feasible for an unbiased detection of eukaryotic organisms in bacteria-rich faecal samples. Archaeal sequences were never amplified, except for primer pairs #20 (2.88%) and #22 (10.58%) (**Figure 2A**); less than 2.12% of bacterial sequences and less than 0.10% of Eukaryota:plas sequences were predicted to be amplified by any primer pair (**Figures 2B and 2M**).

Figures 1A-1U. Each figure represents the result of an *in silico* PCR test performed with a primer pair using the PR^2 primer database (under the condition "Max mismatches = 1").
Red columns indicate the number of sequences available in the PR^2 database 5.0.0 for each taxon, while green columns show how many of these sequences are predicted to be amplified by the tested primer pair. The value above each pair of columns corresponds to the percentage of amplification (i.e., the tested primer pair's coverage of the considered taxonomic group).

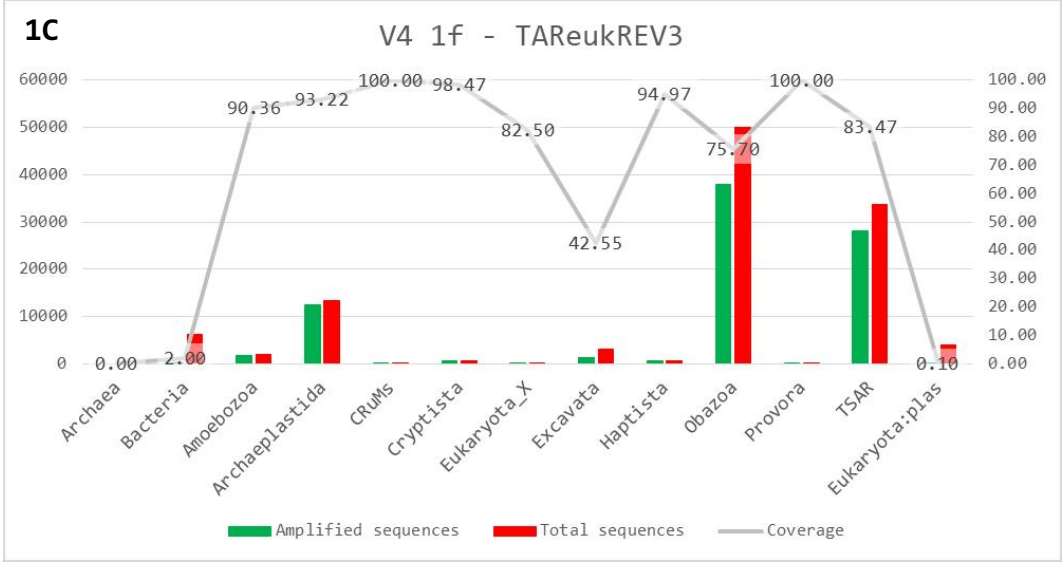
1A


3NDf - V4 euk R2
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
1.99
86.28
90.38
100.00
98.05
85.00
20.53
88.41
55.15
83.33
79.19
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

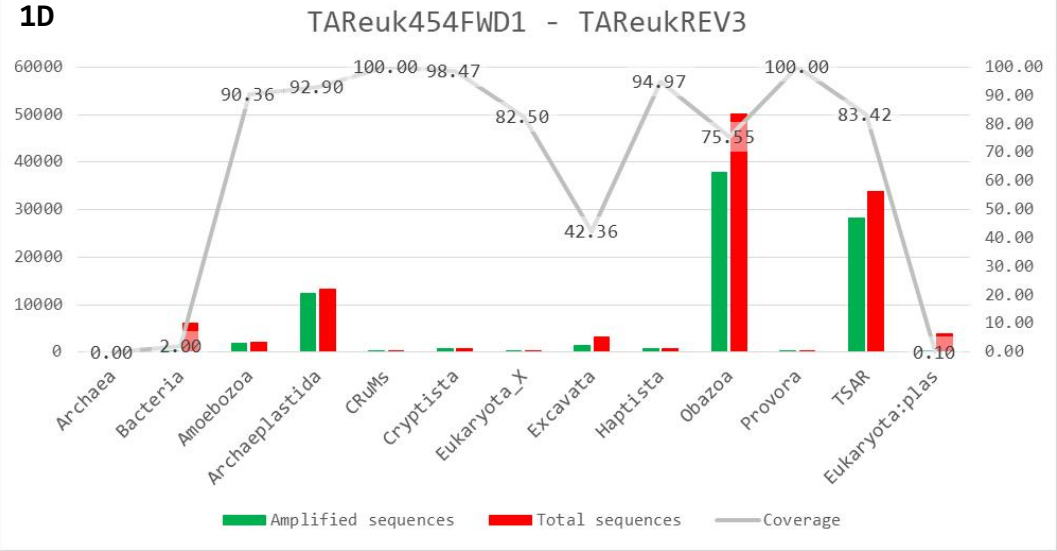
1B


515FY - NSR951
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.10
55.12
81.94
100.00
98.05
85.00
8.77
90.55
87.82
83.33
77.43
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

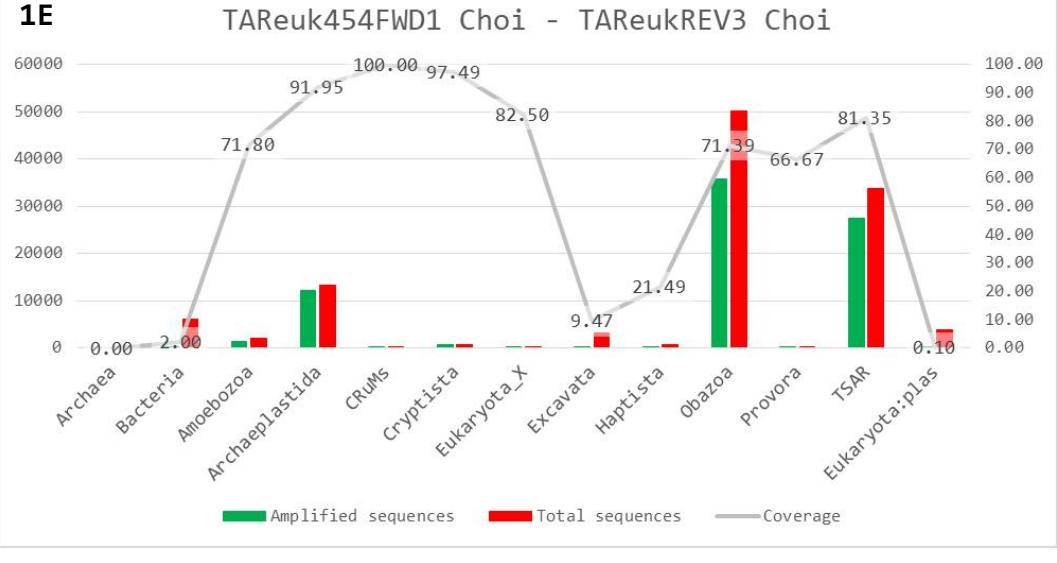
1C


V4 1f - TAReukREV3
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.00
90.36
93.22
100.00
98.47
82.50
42.55
94.97
75.70
100.00
83.47
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

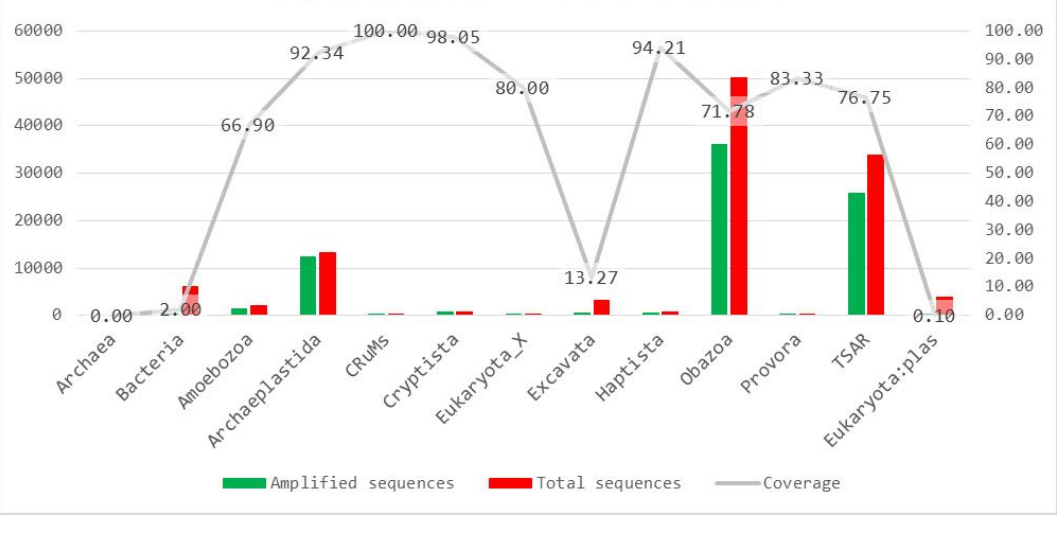
1D


TAReuk454FWD1 - TAReukREV3
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.00
90.36
92.90
100.00
98.47
82.50
42.36
94.97
75.55
100.00
83.42
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

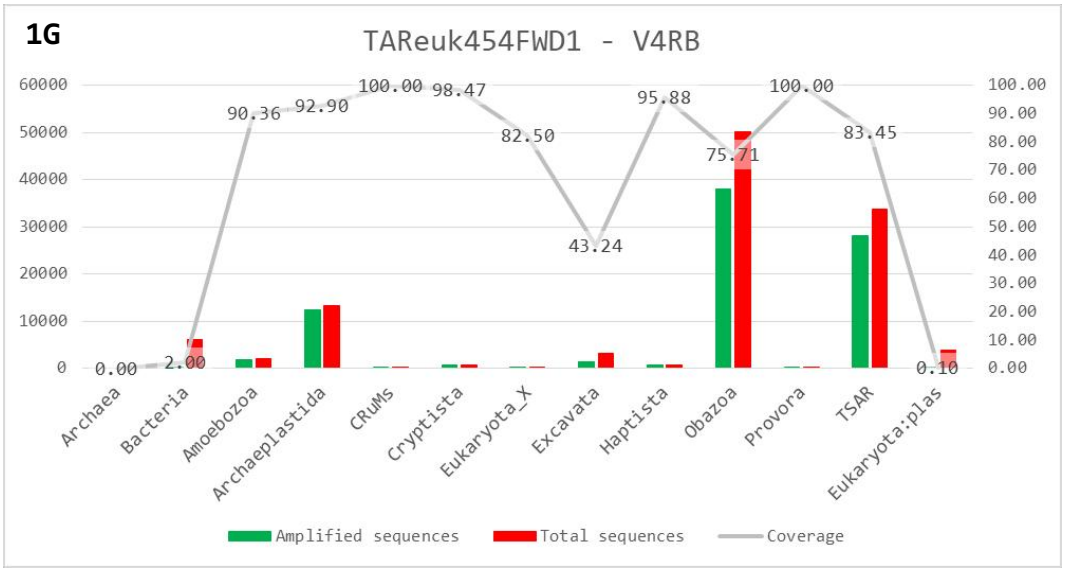
1E


TAReuk454FWD1 Choi - TAReukREV3 Choi
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.00
71.80
91.95
100.00
97.49
82.50
9.47
21.49
71.39
66.67
81.35
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

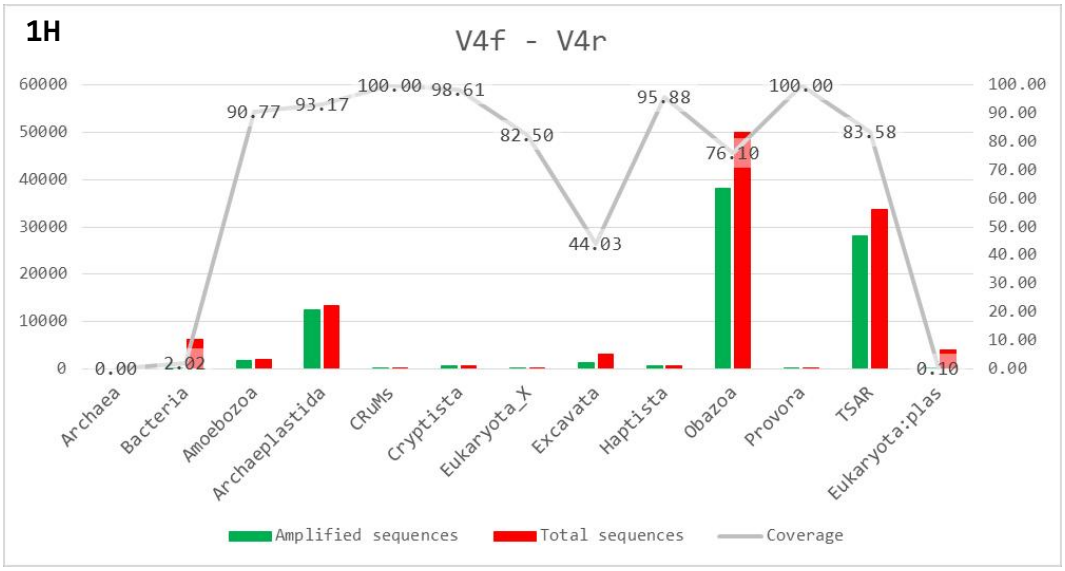
1F


TAReuk454FWD1 - V4 18S Next.Rev
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.00
66.90
92.34
100.00
98.05
80.00
13.27
94.21
71.78
83.33
76.75
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

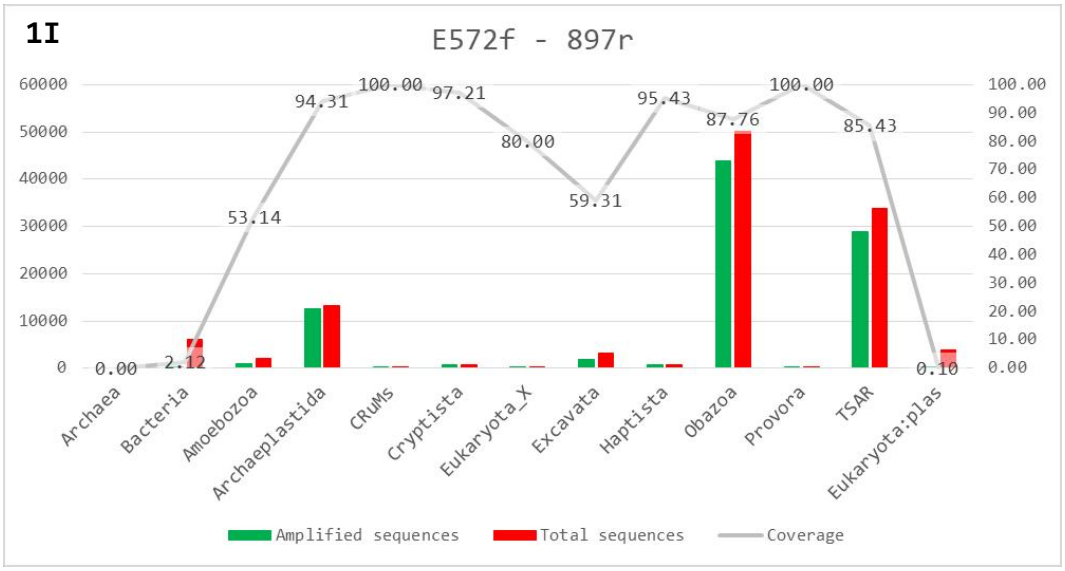
1G


TAReuk454FWD1 - V4RB
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.00
90.36
92.90
100.00
98.47
82.50
43.24
95.88
75.71
100.00
83.45
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

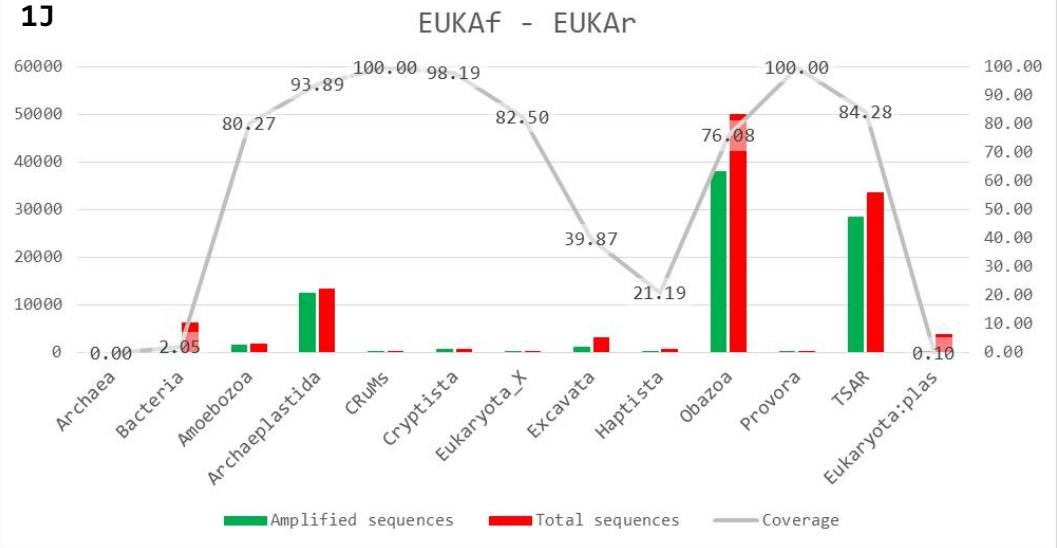
1H


V4f - V4r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.02
90.77
93.17
100.00
98.61
82.50
44.03
95.88
76.10
100.00
83.58
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

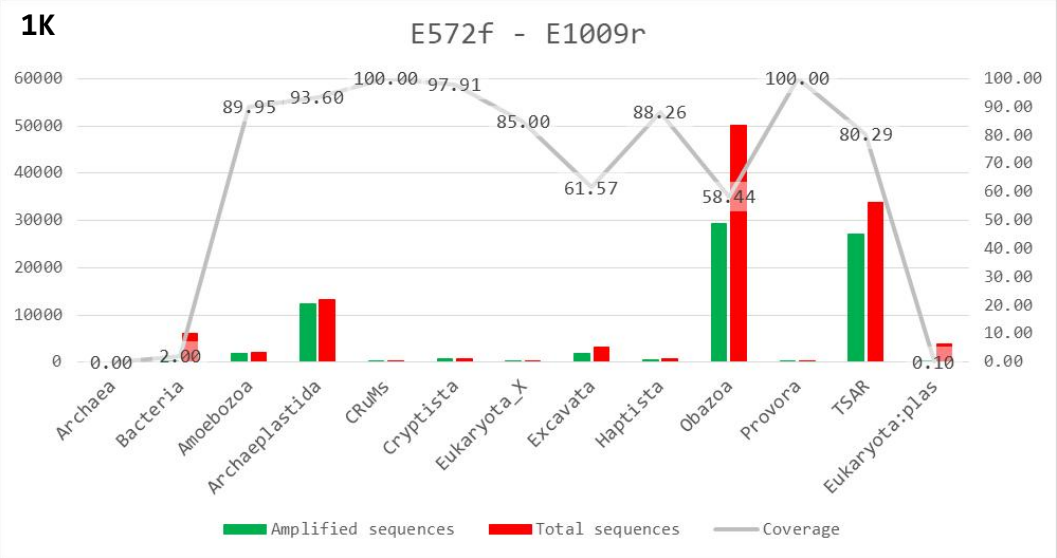
1I


E572f - 897r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.12
53.14
94.31
100.00
97.21
80.00
59.31
95.43
87.76
100.00
85.43
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

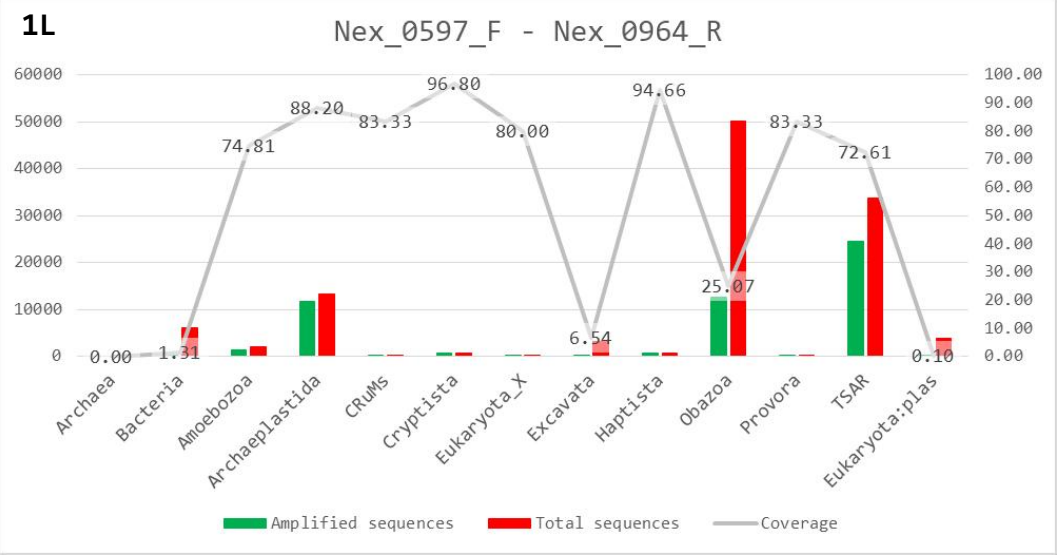
1J


EUKAf - EUKAr
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.05
80.27
93.89
100.00
98.19
82.50
39.87
21.19
76.08
100.00
84.28
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

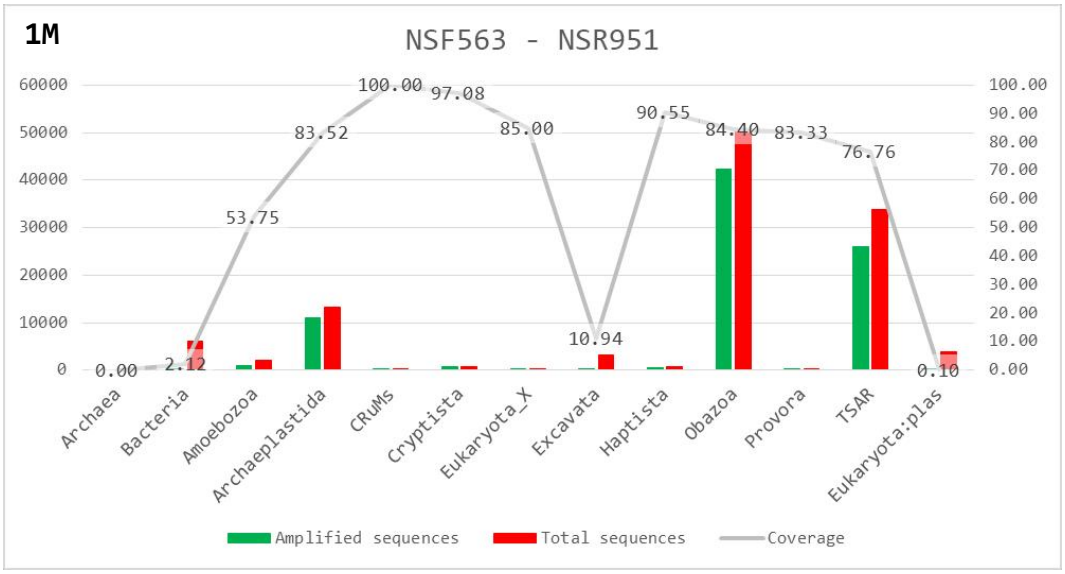
1K


E572f - E1009r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.00
89.95
93.60
100.00
97.91
85.00
61.57
88.26
58.44
100.00
80.29
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

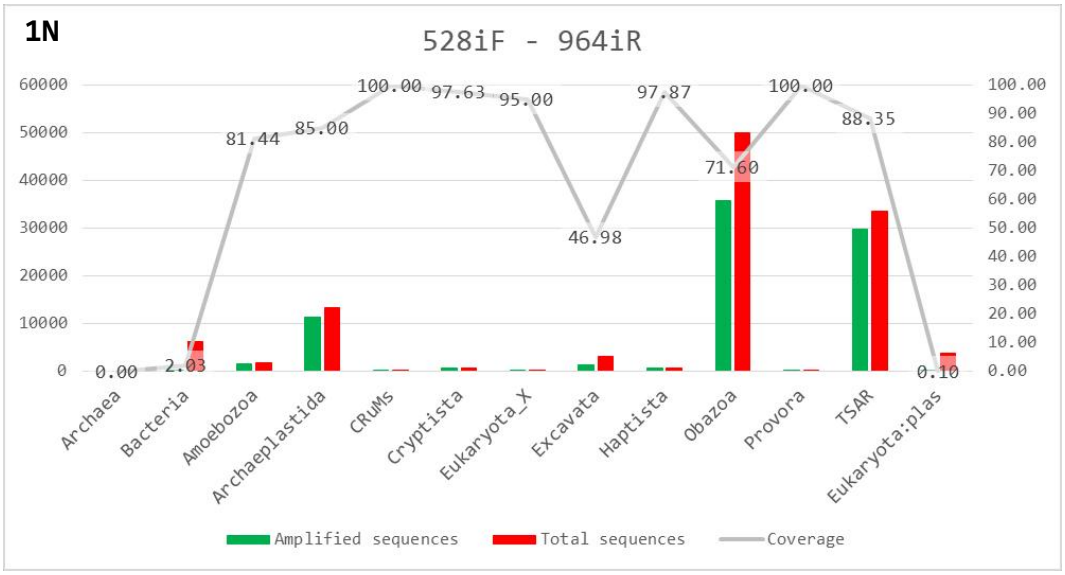
1L


Nex_0597_F - Nex_0964_R
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
1.31
74.81
88.20
83.33
96.80
80.00
6.54
94.66
25.07
83.33
72.61
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

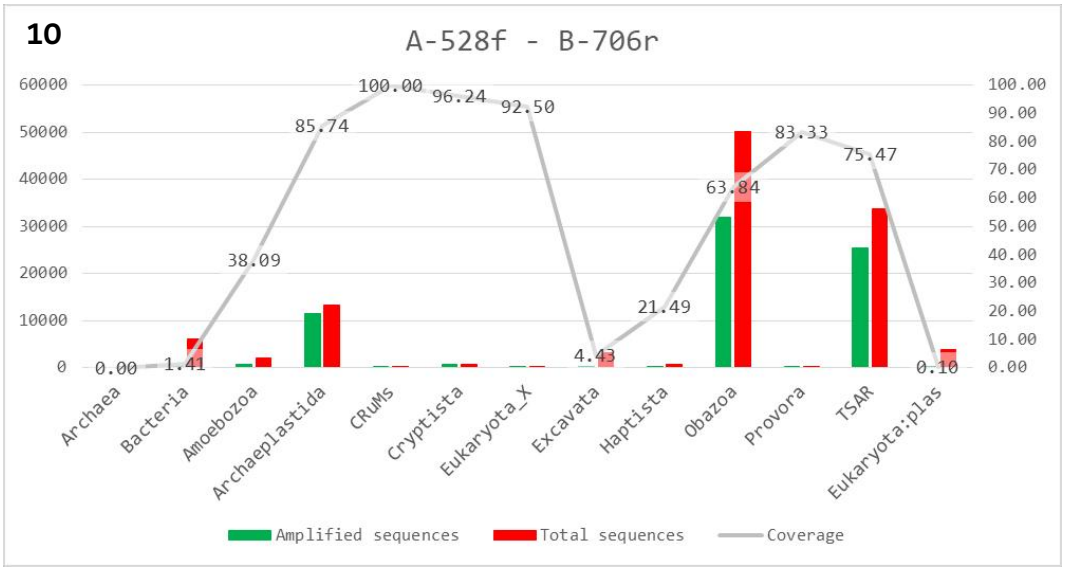
1M


NSF563 - NSR951
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.12
53.75
83.52
100.00
97.08
85.00
10.94
90.55
84.40
83.33
76.76
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

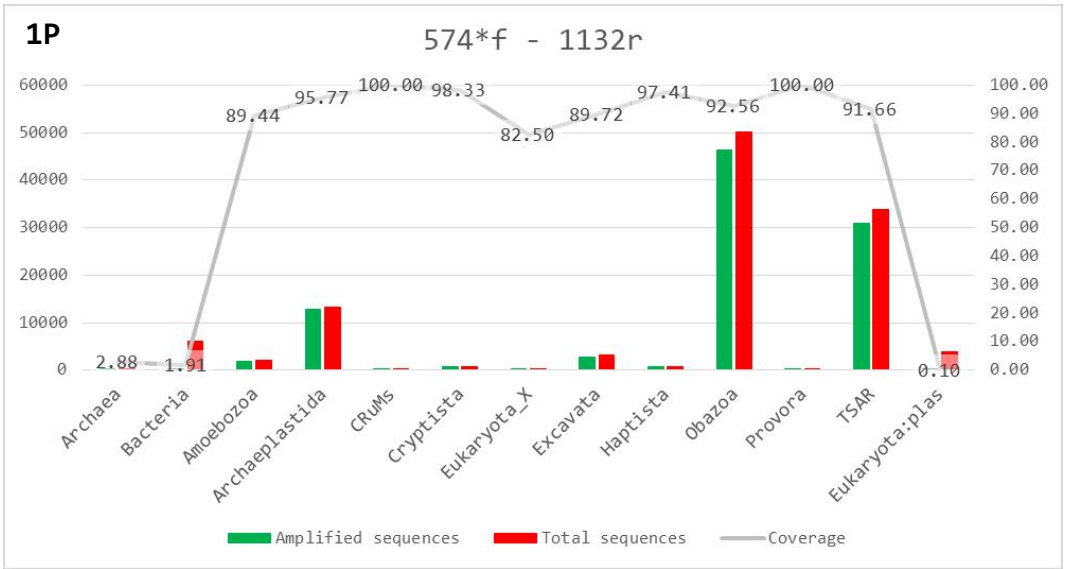
1N


528iF - 964iR
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
2.03
81.44
85.00
100.00
97.63
95.00
46.98
97.87
71.60
100.00
88.35
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

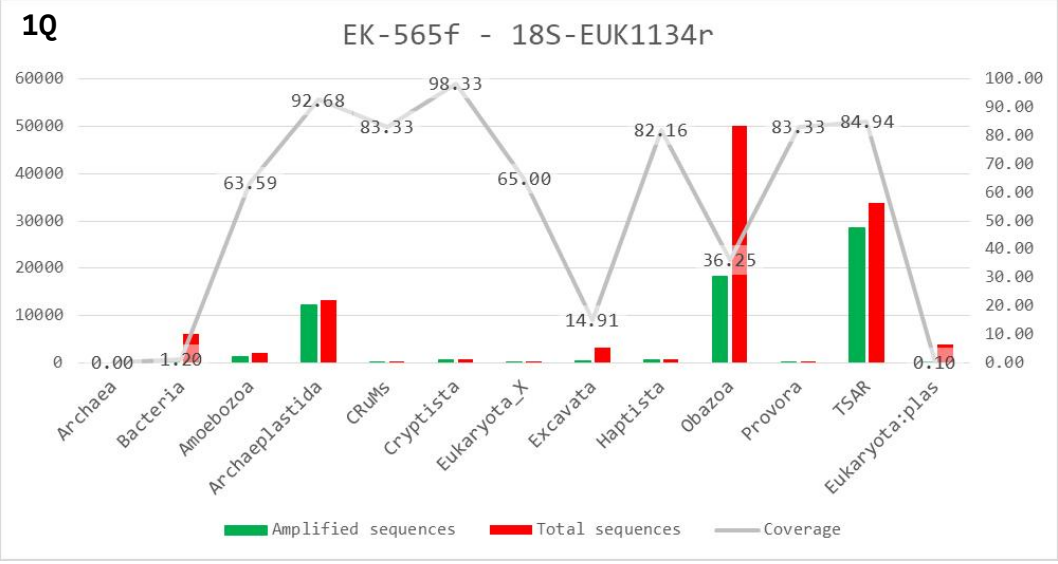
1O


A-528f - B-706r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
1.41
38.09
85.74
100.00
96.24
92.50
4.43
21.49
63.84
83.33
75.47
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

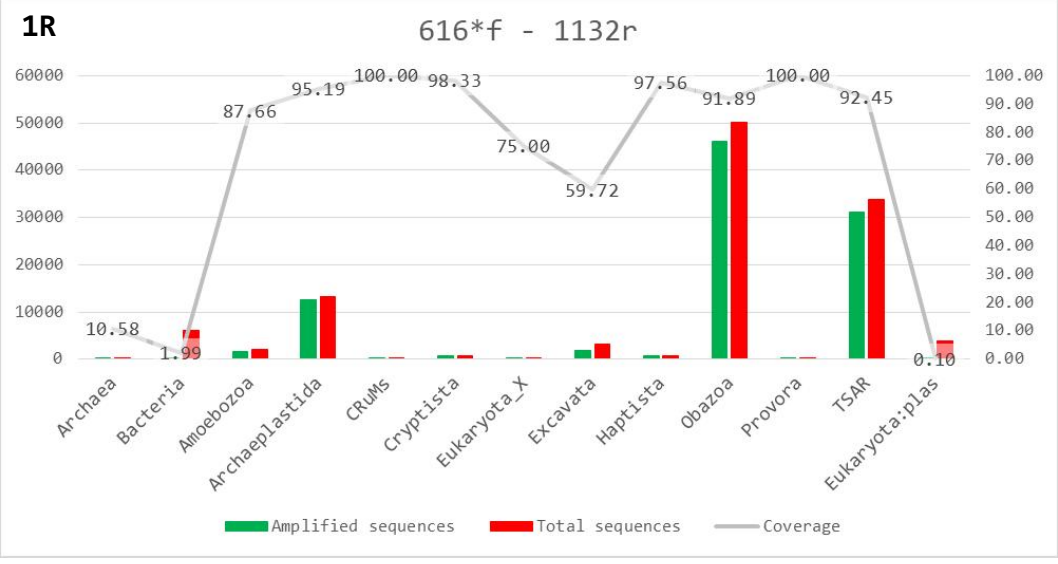
1P

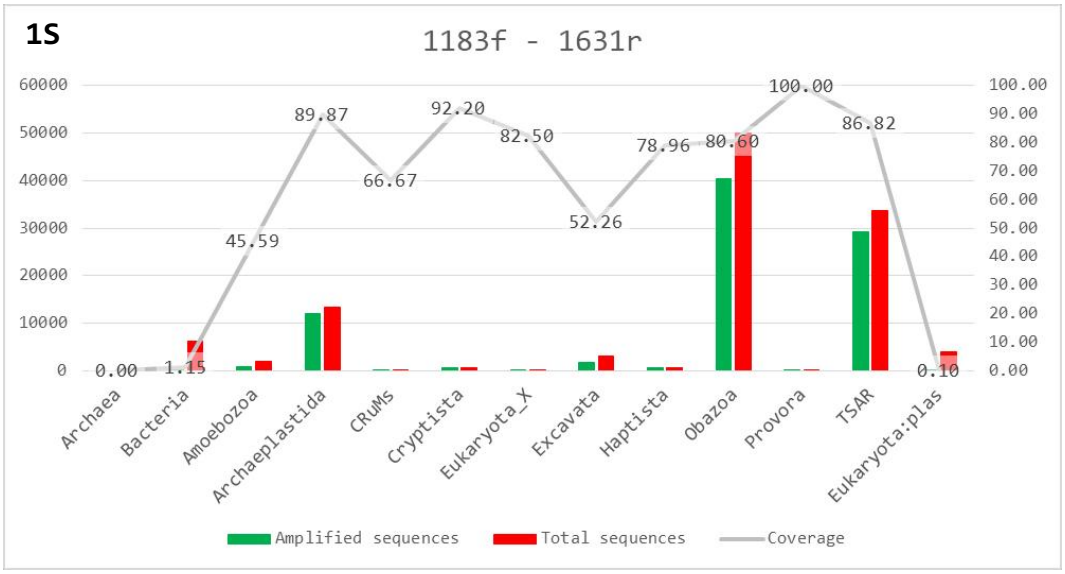

574*f - 1132r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
2.88
1.91
89.44
95.77
100.00
98.33
82.50
89.72
97.41
92.56
100.00
91.66
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

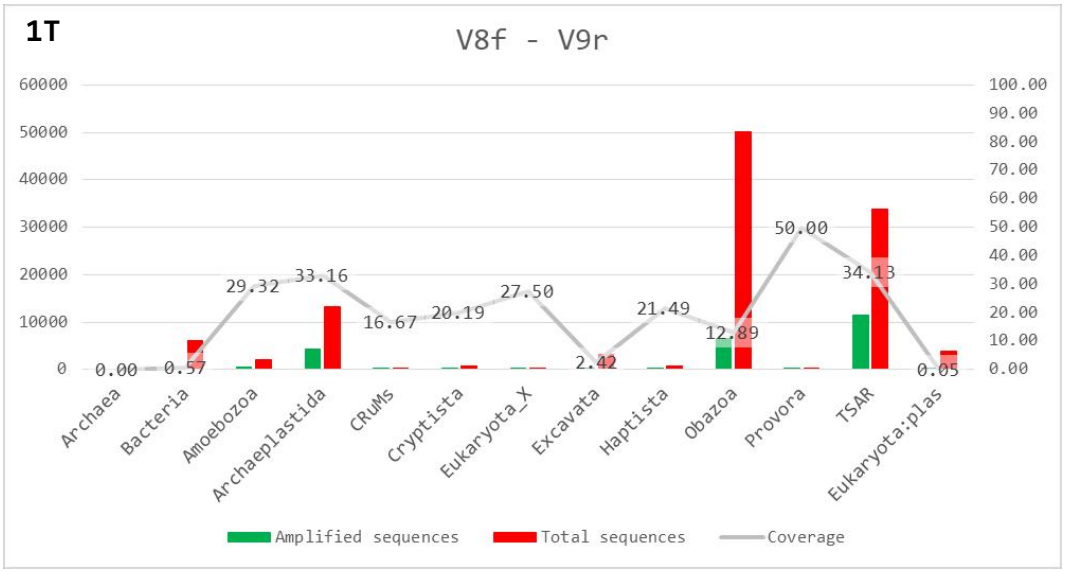
1Q

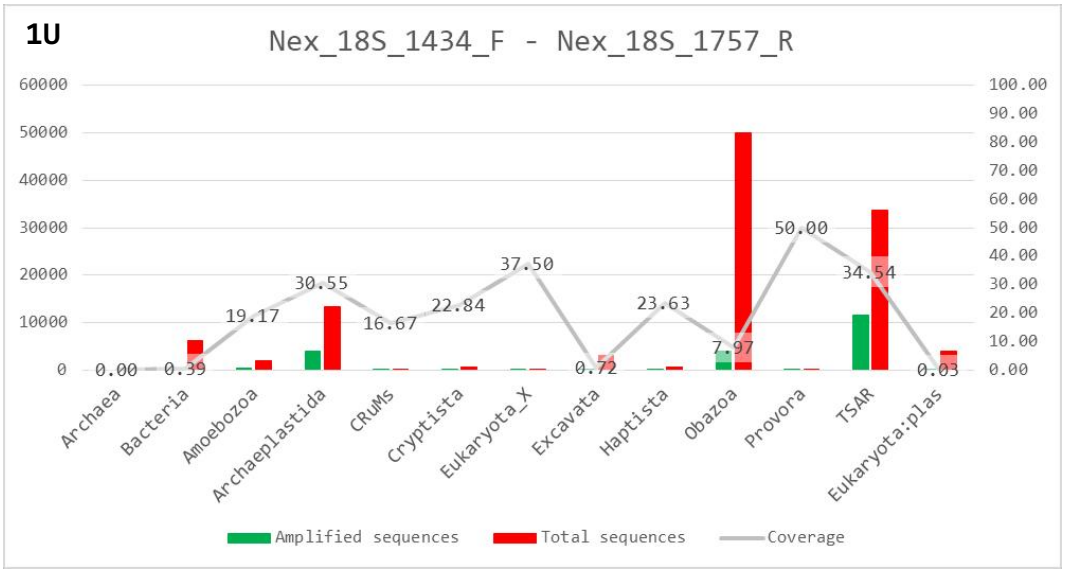

EK-565f - 18S-EUK1134r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
1.20
63.59
92.68
83.33
98.33
65.00
14.91
82.16
36.25
83.33
84.94
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

1R


616*f - 1132r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
10.58
1.99
87.66
95.19
100.00
98.33
75.00
59.72
97.56
91.89
100.00
92.45
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage


1S
1183f - 1631r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
1.15
45.59
89.87
66.67
92.20
82.50
52.26
78.96
80.60
100.00
86.82
0.10
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage


1T
V8f - V9r
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.57
29.32
33.16
16.67
20.19
27.50
2.42
21.49
12.89
50.00
34.13
0.05
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

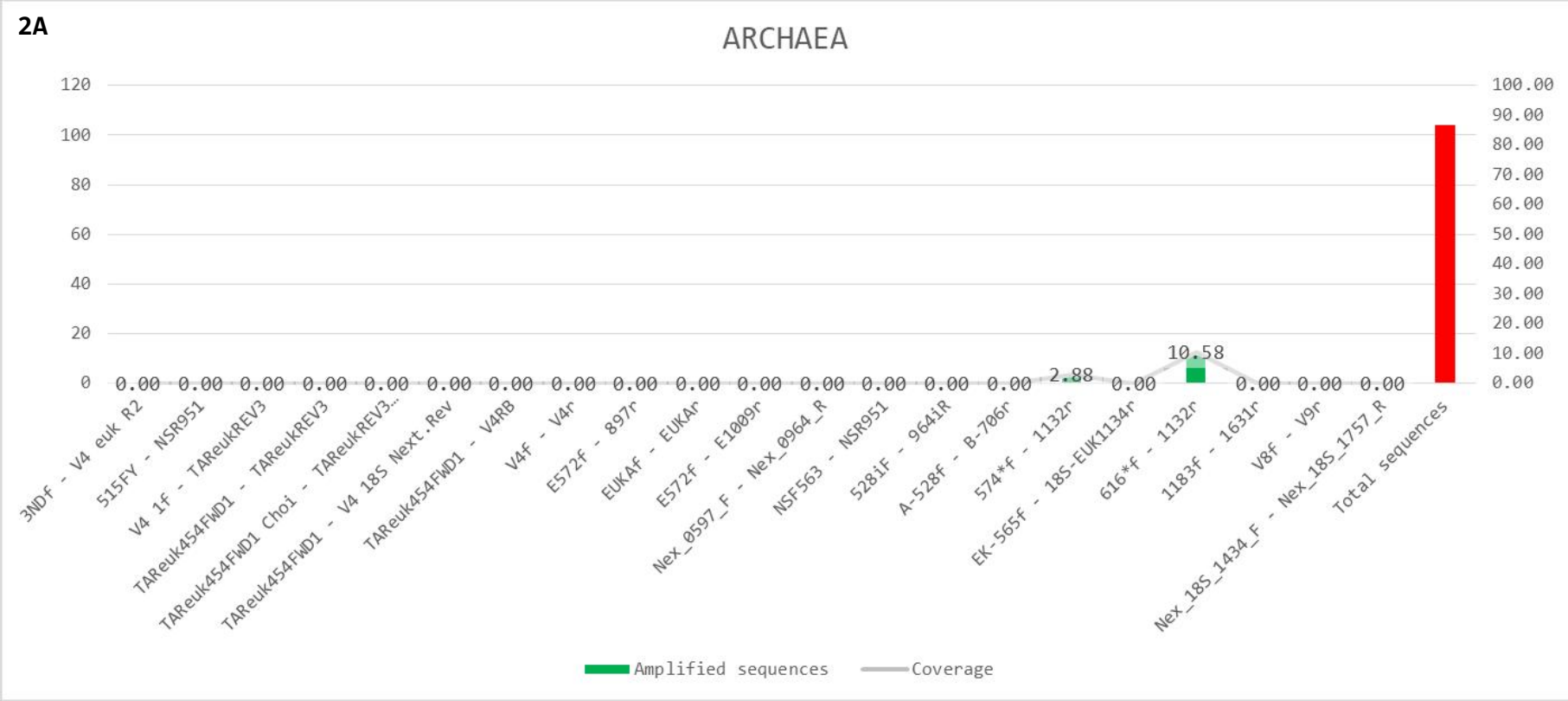

1U
Nex_18S_1434_F - Nex_18S_1757_R
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00
0.39
19.17
30.55
16.67
22.84
37.50
0.72
23.63
7.97
50.00
34.54
0.03
Archaea
Bacteria
Amoebozoa
Archaeplastida
CRuMs
Cryptista
Eukaryota_X
Excavata
Haptista
Obazoa
Provora
TSAR
Eukaryota:plas
Amplified sequences
Total sequences
Coverage

All pairs, except for #25 and #26, performed very well with sequences of Archaeplastida (>81.94%, **Figure 2D**), CRuMs (>83.33%, except for primer pair #24 (66.67%), **Figure 2E**), Cryptista (>92.20%, **Figure 2F**), Eukaryota_X (>75.00%, except for primer pair #21 (65.00%), **Figure 2G**), Haptista (>78.96%, except for primer pairs #9, #14 and #19 (21.49%), **Figure 2I**), Provora (>83.33%, except for primer pair #9 (66.67%), **Figure 2K**) and TSAR (>72.61%, **Figure 2L**). On the other hand, highly variable amplification efficiency was manifested with Amoebozoa (between 38.09% and 90.77%, **Figure 2C**), Excavata (between 4.43% and 89.72%, **Figure 2H**) and Obazoa (between 25.07% and 92.56%, **Figure 2J**) sequences.
Considering the total amount of sequences in the PR^2 database, primer pairs with the highest amplification efficiency were #13 (E572f - 897r, 78.84%), #20 (574*f - 1132r, 84.45%) and #22 (616*f - 1132r, 83.46%); those with the poorest coverage were #16 (Nex_0597_F - Nex_0964_R, 45.58%), #25 (V8f - V9r, 20.52%) and #26 (Nex_18S_1434_F - Nex_18S_1757_R, 17.97%) (**Figure 2N**).

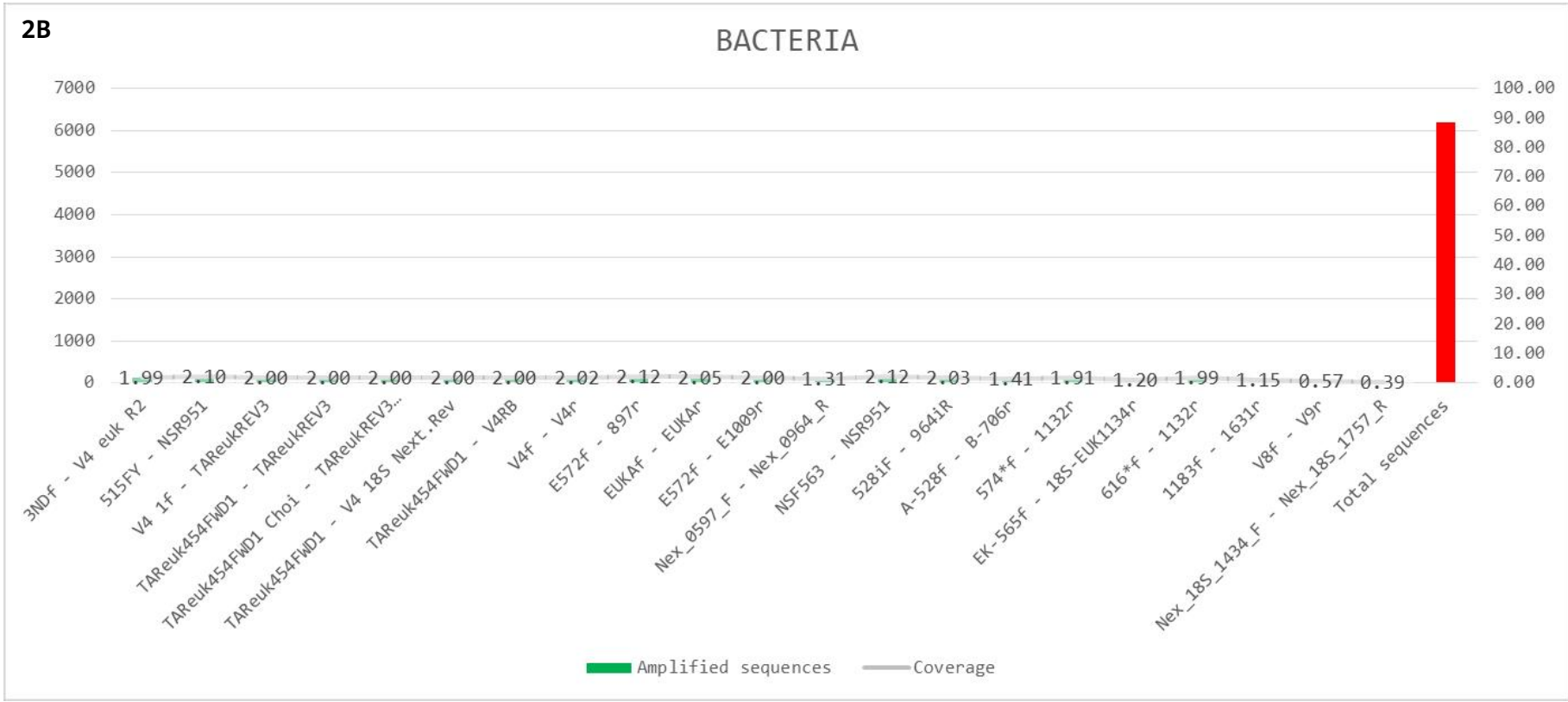
While the *in silico* PCR results provided a comprehensive description of the universal primer pairs performances, we needed to evaluate the pairs' coverage of the human eukaryotic intestinal parasites of interest to choose those more suitable for the DNA metabarcoding protocol. Hence, for each primer pair, we checked how many parasitic sequences were amplified *in silico*, and how long the potentially produced amplicons were predicted to be (**Table 6**).
Ancylostoma duodenale was amplified by all primer pairs but #3, #15, #16, #19, #21, #25 and #26; only #20 produced long amplicons (550 bp < length < 600 bp). *Anisakis sp.*, *Ascaris lumbriocoides* and *Enterobius vermicularis* were not amplified by pairs #3, #15, #16, #17, #18, #19, #21, #25 and #26; only #20 produced long amplicons. *Blastocystis spp.* was amplified by all pairs but #3 and #15; only #20 produced long amplicons. *Cryptosporidium spp.* and *Toxoplasma gondii* were amplified by all primer pairs, but #20 produced long amplicons for the latter. *Cyclospora cayetanensis* and *Cystoisospora belli* were not amplified by pair #16 and pair #25, respectively; both were amplified by #20 producing long amplicons. *Dientamoeba fragilis* was amplified only by primer pairs #20, #22 and #23. All *Diphyllobothrium*, *Hymenolepis*, *Schistosoma* and *Taenia* species considered in our analysis were amplified by pairs #3, #4, #13, #15, #17, #20, #22 and #24; however, almost all produced amplicons were longer than 550 bp, except for *Diphyllobotrium*

Figures 2A-2N. Each figure represents coverage information of all the tested primer pairs on the considered taxonomic group (A-M) or on the whole PR^2 database (N). Red columns indicate the number of sequences available in the PR^2 database 5.0.0 for the specified taxon, while green columns show how many of these sequences are predicted to be amplified by each tested primer pair. The value above each column corresponds to the percentage of amplification (i.e., the tested primer pair's coverage of the considered taxonomic group).

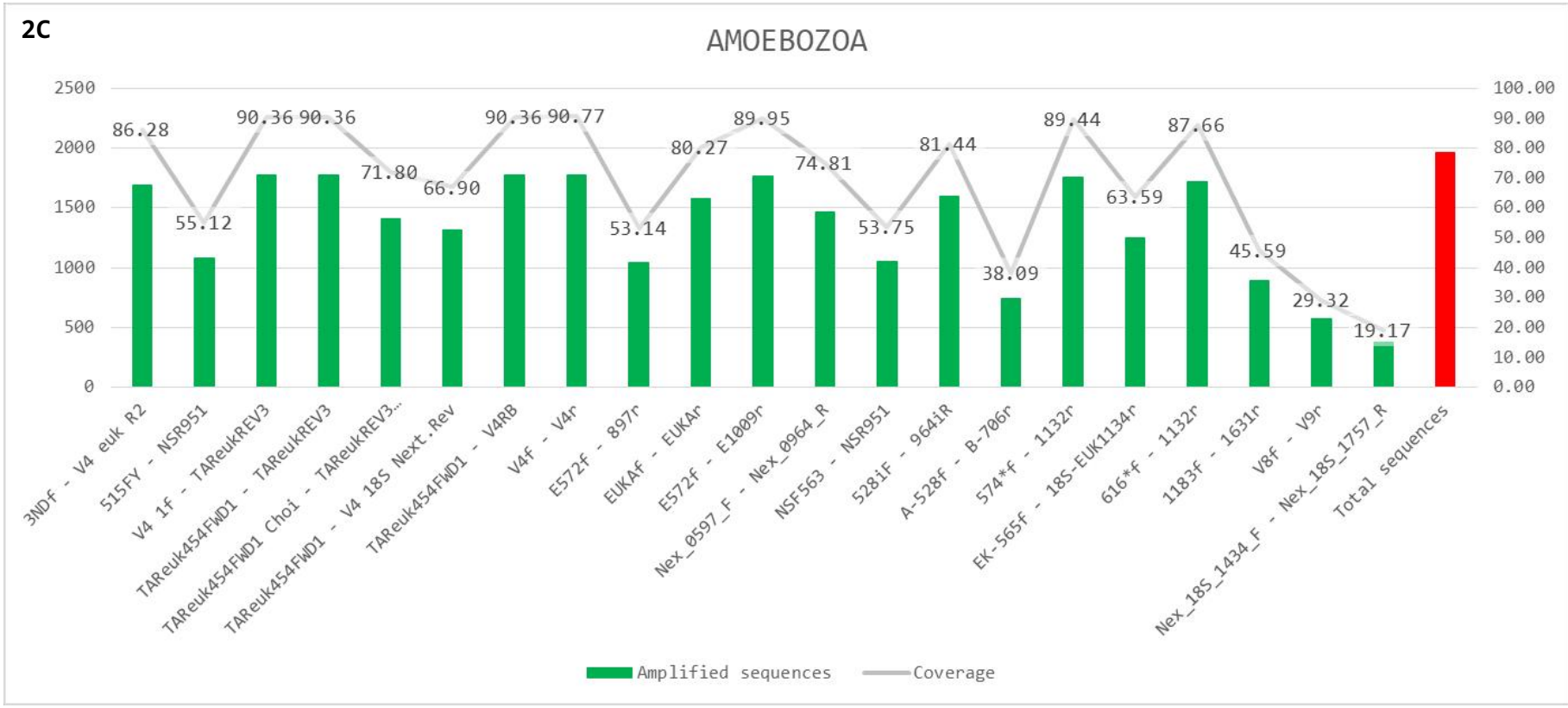
2A


ARCHAEA
120
100
80
60
40
20
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 2.88 0.00 10.58 0.00 0.00 0.00
3NDf - V4 euk R2
515FV - NSR951
V4 1f - TAReuKREV3
TAReuk454FWD1 - TAReuKREV3
TAReuk454FWD1 Choi - TAReuKREV3…
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

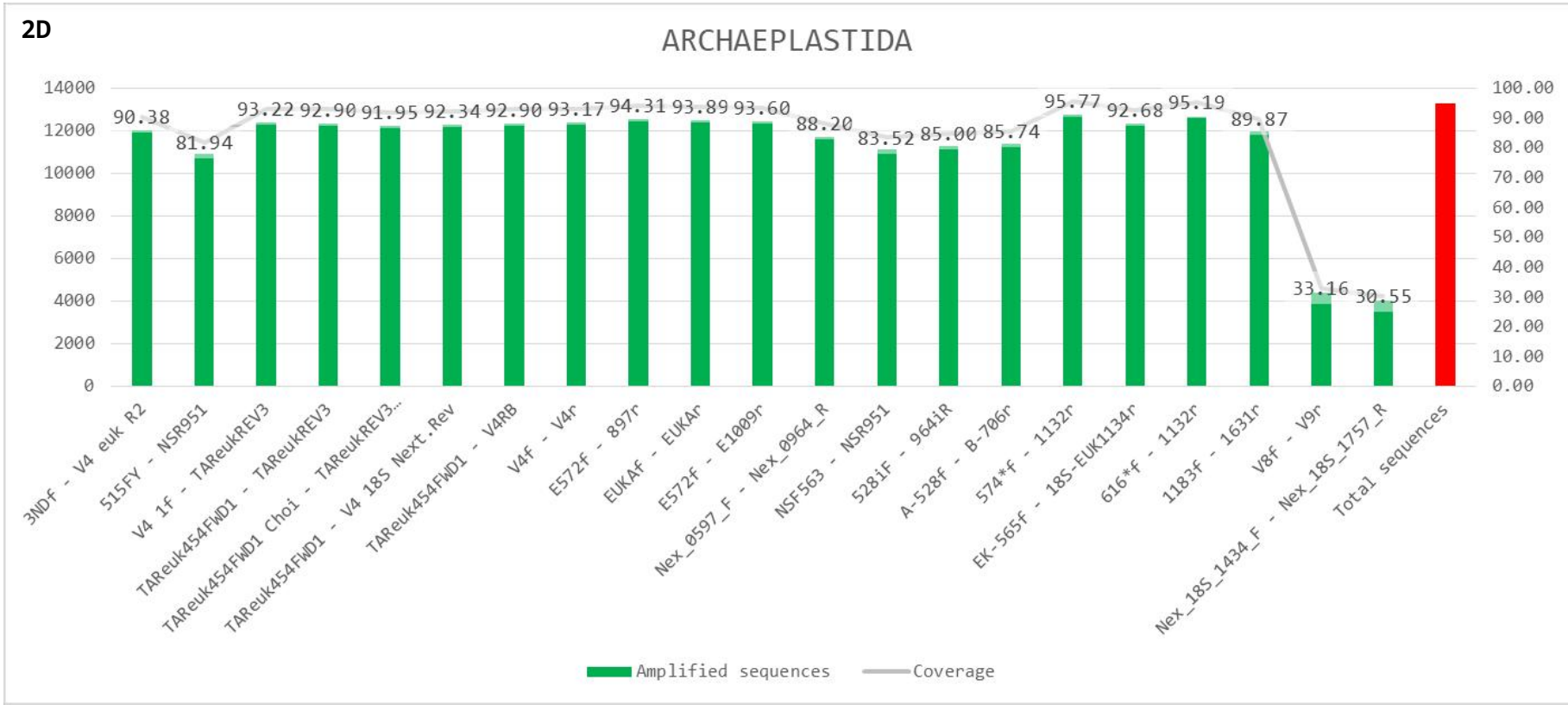
2B


BACTERIA
7000
6000
5000
4000
3000
2000
1000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1.99
2.10
2.00
2.00
2.00
2.00
2.00
2.02
2.12
2.05
2.00
1.31
2.12
2.03
1.41
1.91
1.20
1.99
1.15
0.57
0.39
3NDf - V4 euk R2
515FY - NSR951
V4 1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

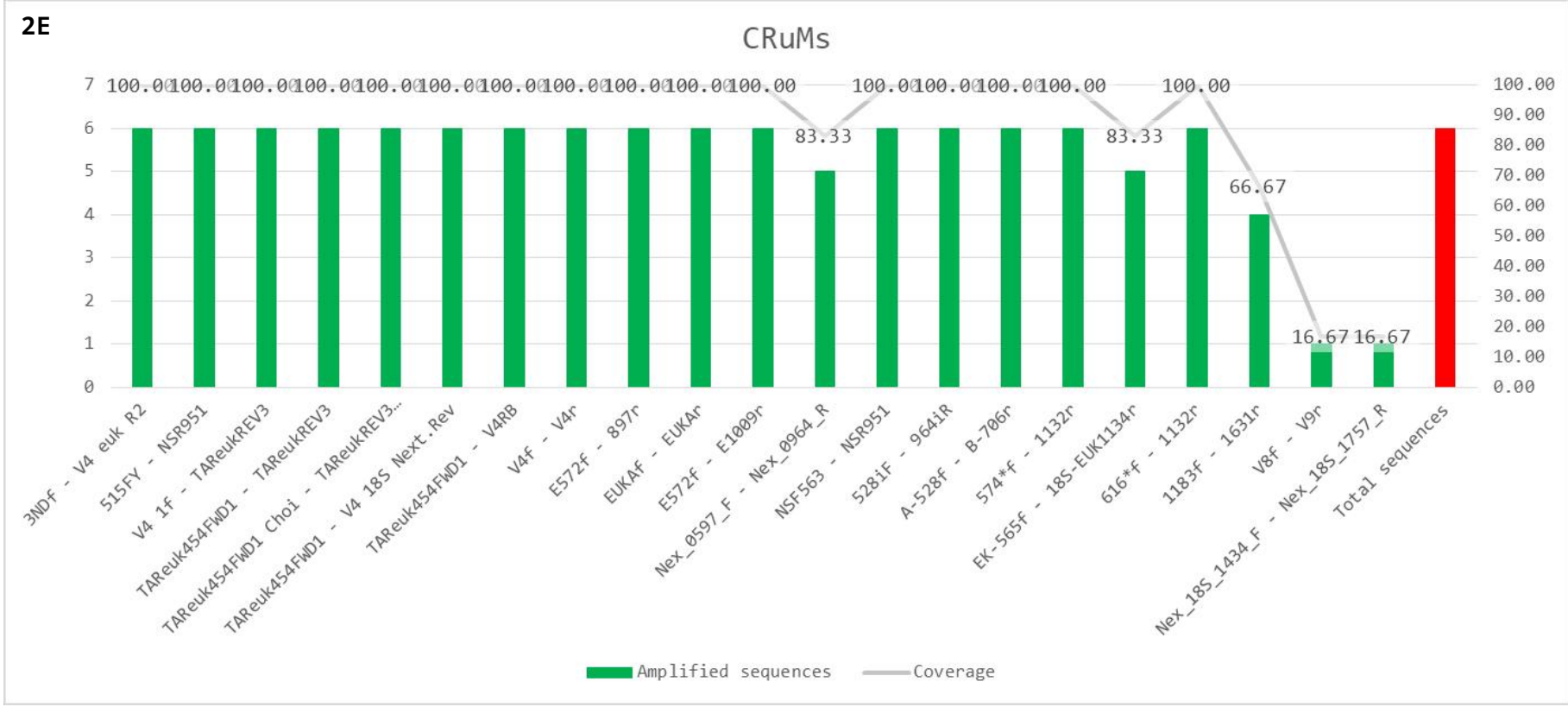
2C


AMOEBOZOA
2500
2000
1500
1000
500
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
86.28
55.12
90.36
90.36
71.80
66.90
90.36
90.77
53.14
80.27
89.95
74.81
53.75
81.44
38.09
89.44
63.59
87.66
45.59
29.32
19.17
3NDf - V4 euk R2
515FY - NSR951
V4 1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

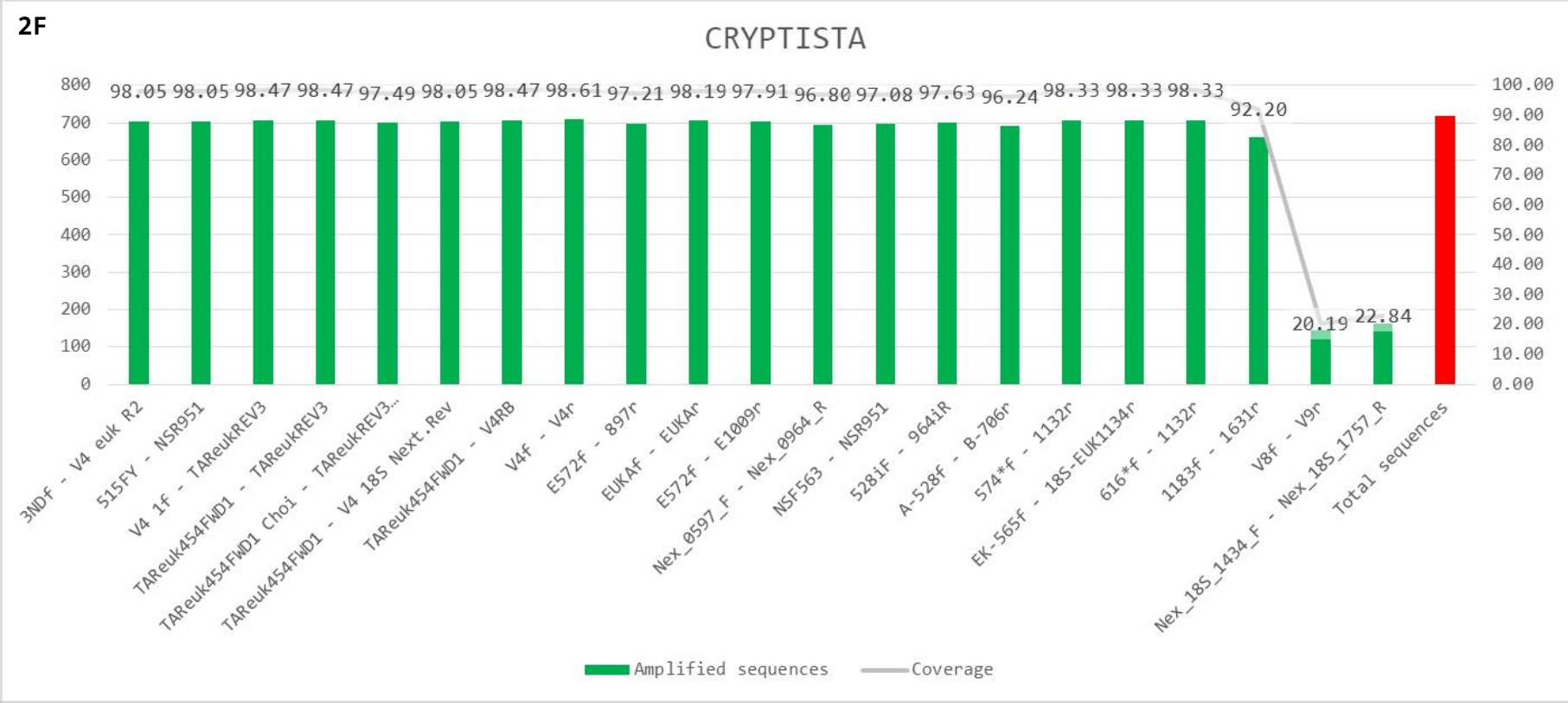
2D


ARCHAEPLASTIDA
14000
12000
10000
8000
6000
4000
2000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
90.38
81.94
93.22
92.90
91.95
92.34
92.90
93.17
94.31
93.89
93.60
88.20
83.52
85.00
85.74
95.77
92.68
95.19
89.87
33.16
30.55
3NDf - V4 euk R2
515FY - NSR951
V4 1f - TAReuKREV3
TAReuk454FWD1 - TAReuKREV3
TAReuk454FWD1 Choi - TAReuKREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

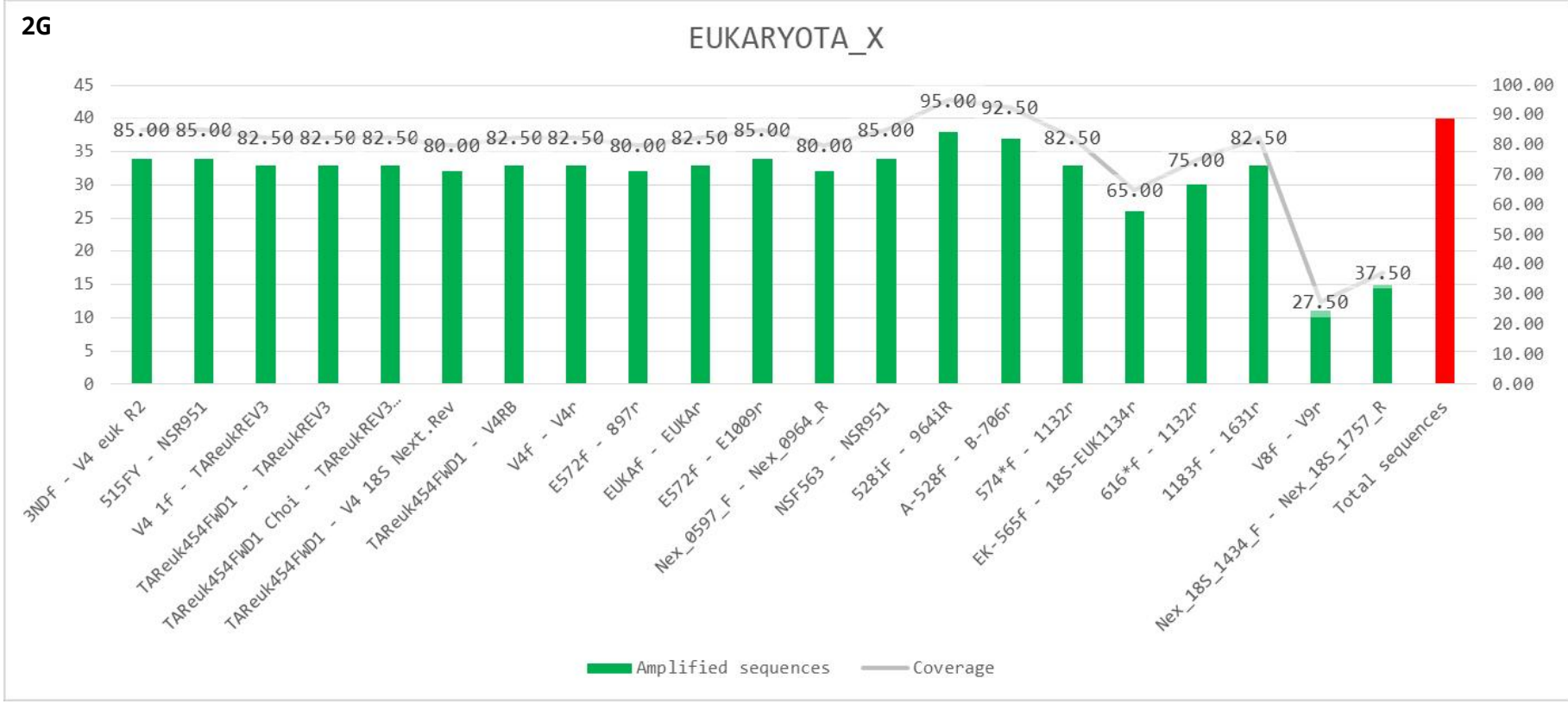
2E


CRuMs
100.00
83.33
66.67
16.67
16.67
3NDf - V4 euk R2
515FY - NSR951
V4 1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3..
TAReuk454FWD1 - V4 18S Next.Rev
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

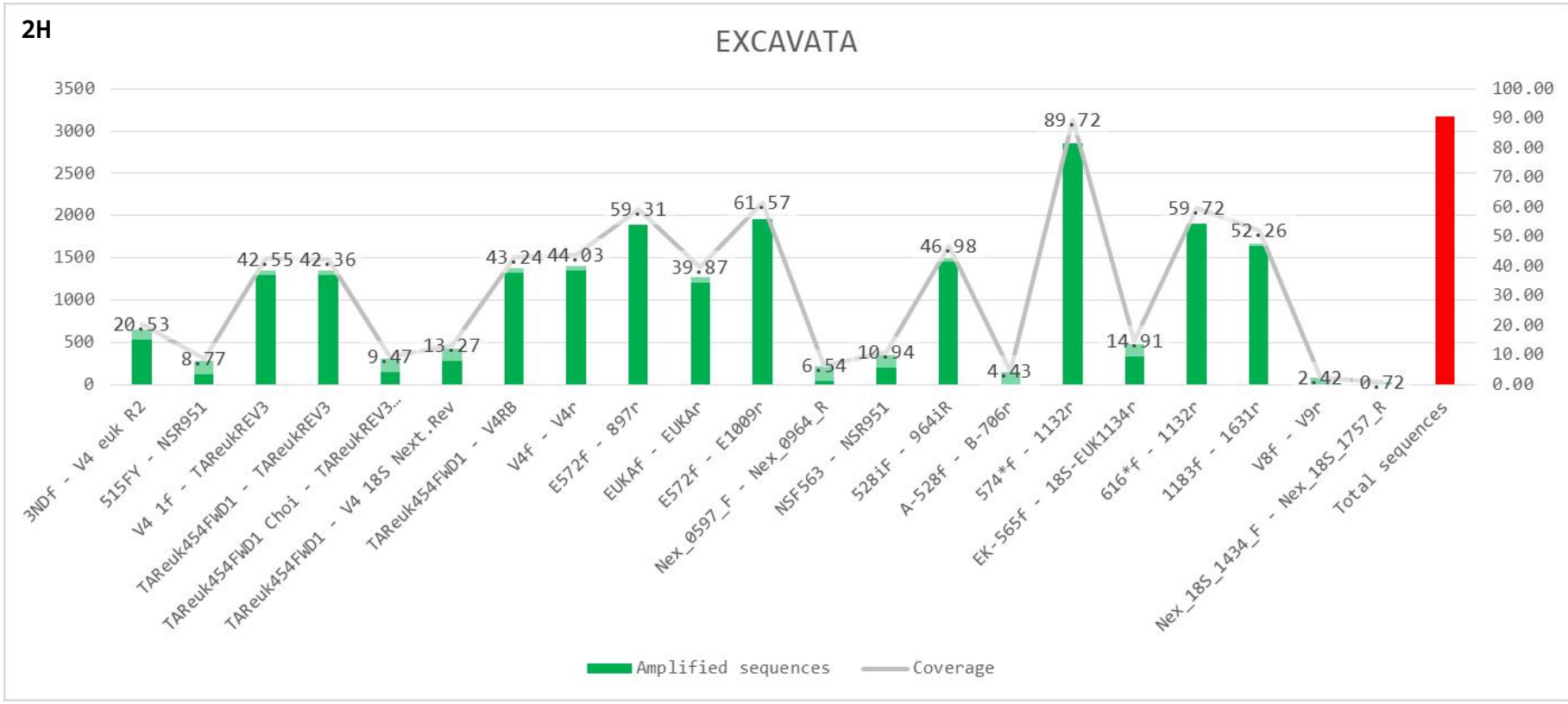
2F


CRYPTISTA
800
700
600
500
400
300
200
100
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
98.05 98.05 98.47 98.47 97.49 98.05 98.47 98.61 97.21 98.19 97.91 96.80 97.08 97.63 96.24 98.33 98.33 98.33 92.20 20.19 22.84
3NDf - V4 euk R2
515FY - NSR951
V4_1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1_Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

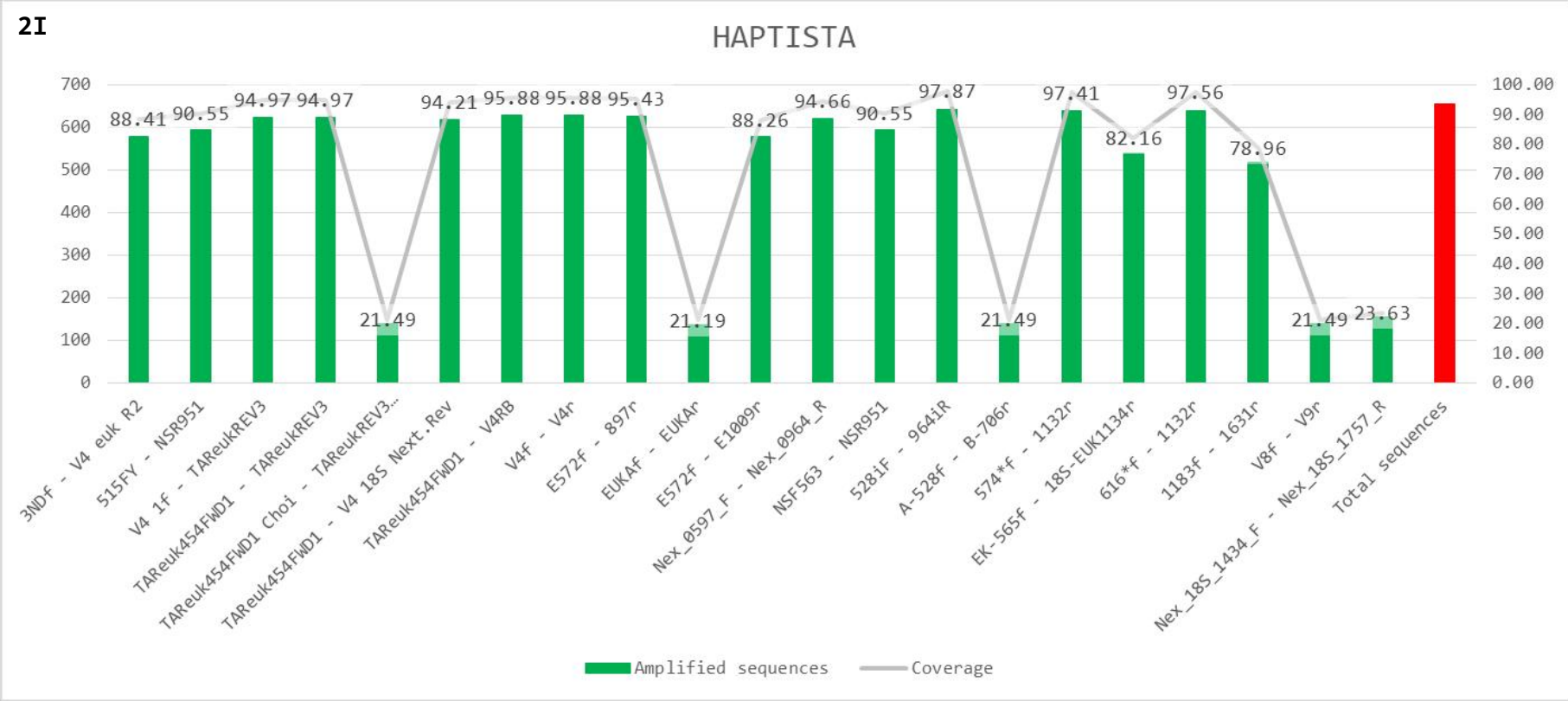
2G


EUKARYOTA_X
45
40
35
30
25
20
15
10
5
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
85.00
85.00
82.50
82.50
82.50
80.00
82.50
82.50
80.00
82.50
85.00
80.00
85.00
95.00
92.50
82.50
65.00
75.00
82.50
27.50
37.50
3NDf - V4 euk R2
515FY - NSR951
V4_1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1_Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

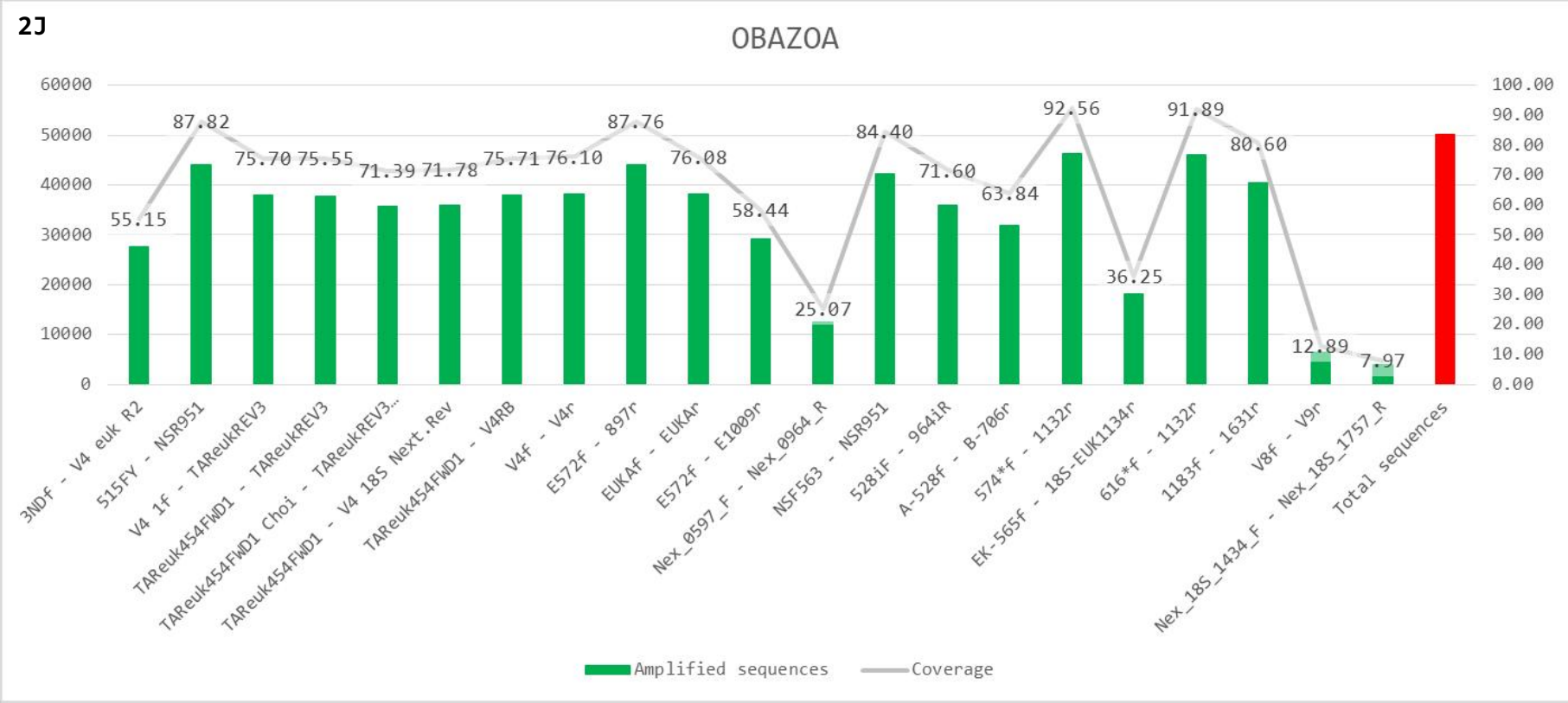
2H


EXCAVATA
3500
3000
2500
2000
1500
1000
500
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
20.53
8.77
42.55
42.36
9.47
13.27
43.24
44.03
59.31
39.87
61.57
6.54
10.94
46.98
4.43
89.72
14.91
59.72
52.26
2.42
0.72
3NDf - V4 euk R2
515FY - NSR951
V4_1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

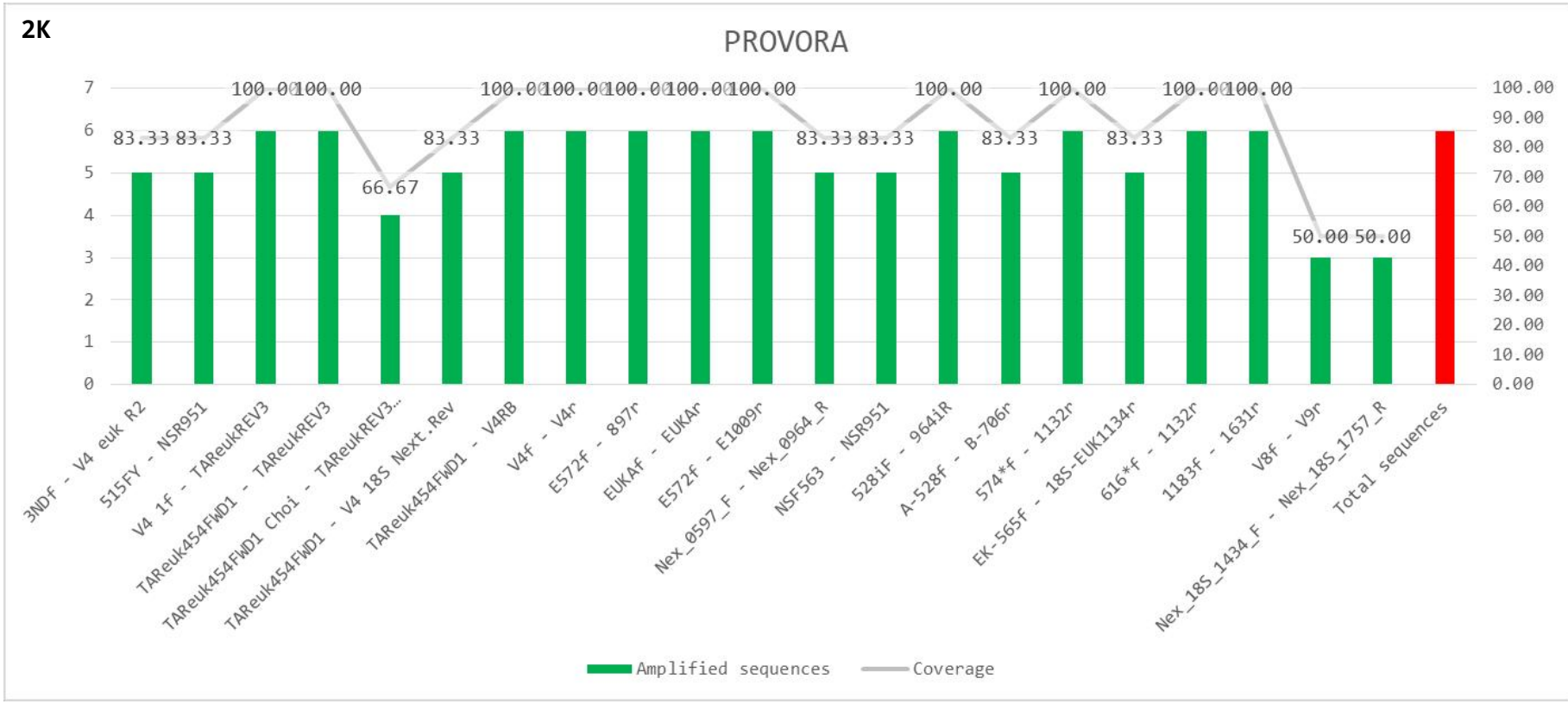
2I


HAPTISTA
700
600
500
400
300
200
100
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
88.41
90.55
94.97
94.97
21.49
94.21
95.88
95.88
95.43
21.19
88.26
94.66
90.55
97.87
21.49
97.41
82.16
97.56
78.96
21.49
23.63
3NDf - V4 euk R2
515FY - NSR951
V4 1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

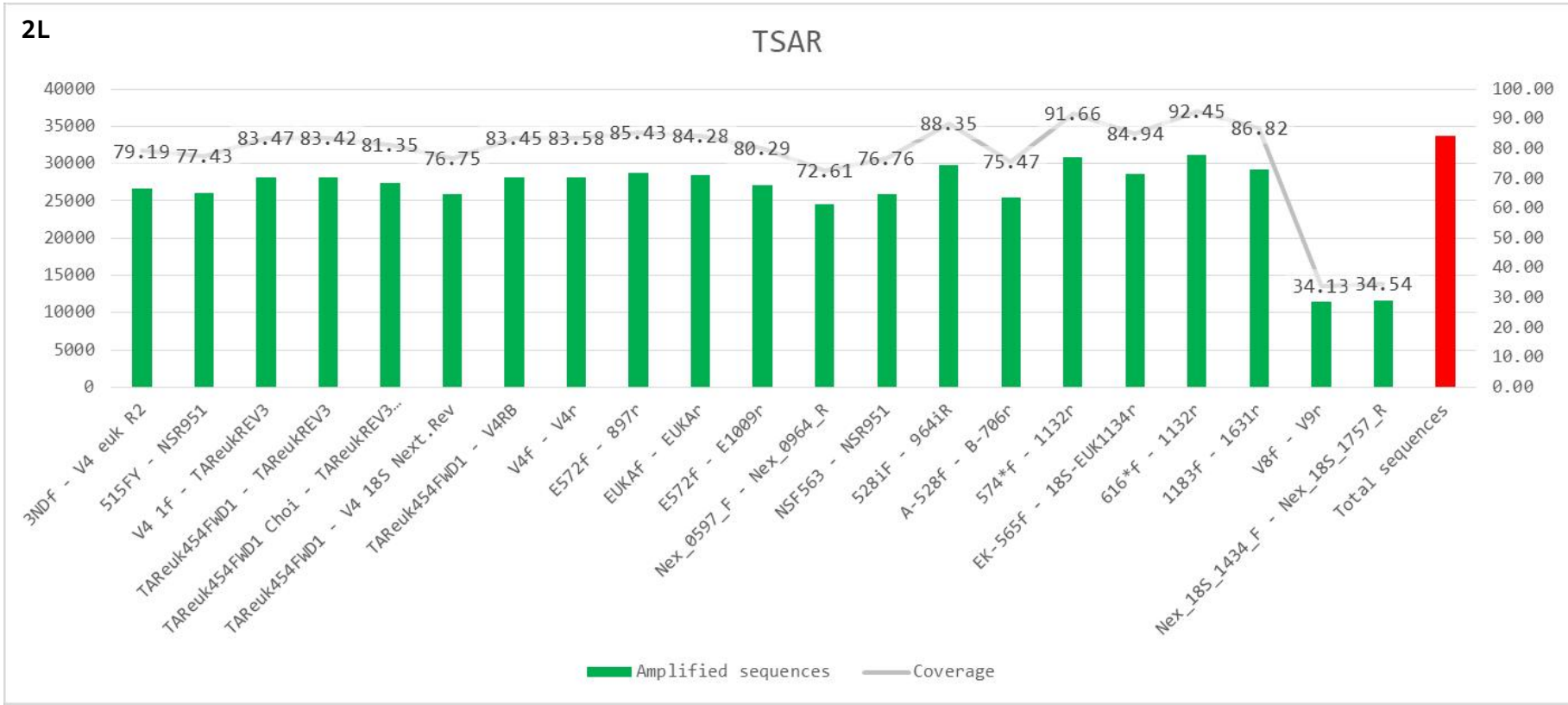
2J

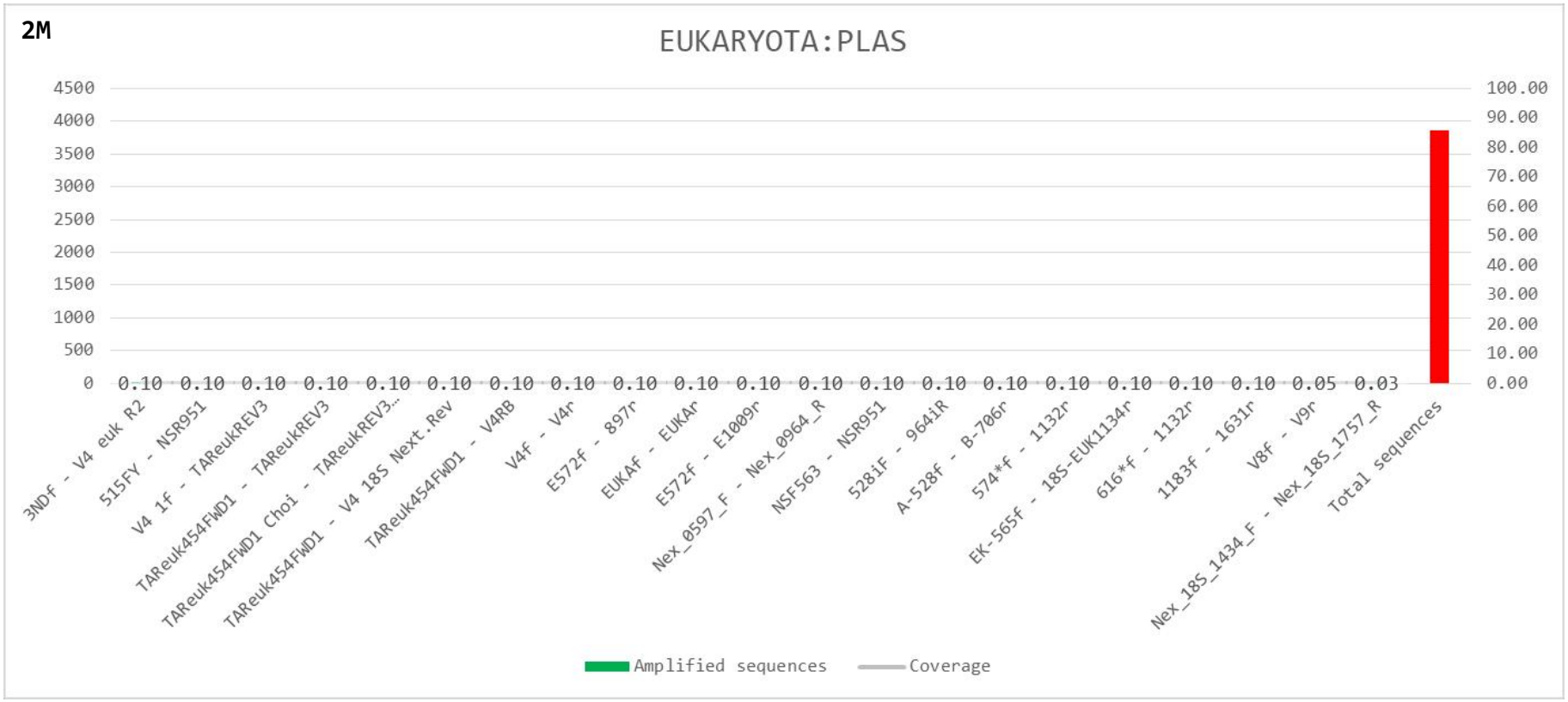

OBAZOA
60000
50000
40000
30000
20000
10000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
55.15
87.82
75.70
75.55
71.39
71.78
75.71
76.10
87.76
76.08
58.44
25.07
84.40
71.60
63.84
92.56
36.25
91.89
80.60
12.89
7.97
3NDf - V4 euk R2
515FY - NSR951
V4_1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1_Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

2K

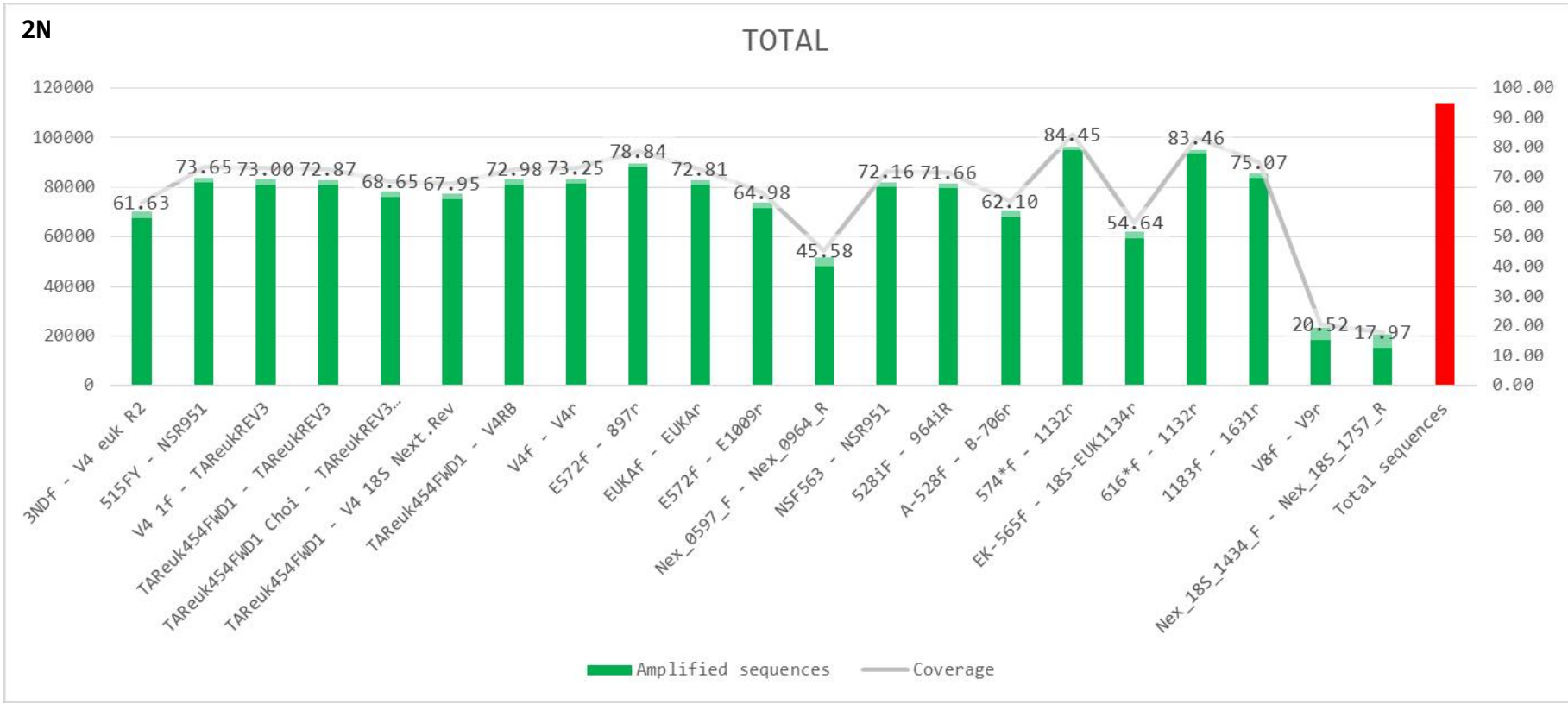

PROVORA
7
6
5
4
3
2
1
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
83.33
83.33
100.00
100.00
66.67
83.33
100.00
100.00
100.00
100.00
100.00
83.33
83.33
100.00
83.33
100.00
83.33
100.00
100.00
50.00
50.00
3NDf - V4 euk R2
515FY - NSR951
V4 1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

2L


TSAR
40000
35000
30000
25000
20000
15000
10000
5000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
79.19
77.43
83.47
83.42
81.35
76.75
83.45
83.58
85.43
84.28
80.29
72.61
76.76
88.35
75.47
91.66
84.94
92.45
86.82
34.13
34.54
3NDf - V4 euk R2
515FY - NSR951
V4_1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1_Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage


2M
EUKARYOTA:PLAS
4500
4000
3500
3000
2500
2000
1500
1000
500
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.05 0.03
3NDf - V4 euk R2
515FY - NSR951
V4 1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3…
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

2N


TOTAL
120000
100000
80000
60000
40000
20000
0
100.00
90.00
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
61.63
73.65
73.00
72.87
68.65
67.95
72.98
73.25
78.84
72.81
64.98
45.58
72.16
71.66
62.10
84.45
54.64
83.46
75.07
20.52
17.97
3NDf - V4 euk R2
515FV - NSR951
V4 1f - TAReukREV3
TAReuk454FWD1 - TAReukREV3
TAReuk454FWD1 Choi - TAReukREV3...
TAReuk454FWD1 - V4 18S Next.Rev
TAReuk454FWD1 - V4RB
V4f - V4r
E572f - 897r
EUKAf - EUKAr
E572f - E1009r
Nex_0597_F - Nex_0964_R
NSF563 - NSR951
528iF - 964iR
A-528f - B-706r
574*f - 1132r
EK-565f - 18S-EUK1134r
616*f - 1132r
1183f - 1631r
V8f - V9r
Nex_18S_1434_F - Nex_18S_1757_R
Total sequences
Amplified sequences
Coverage

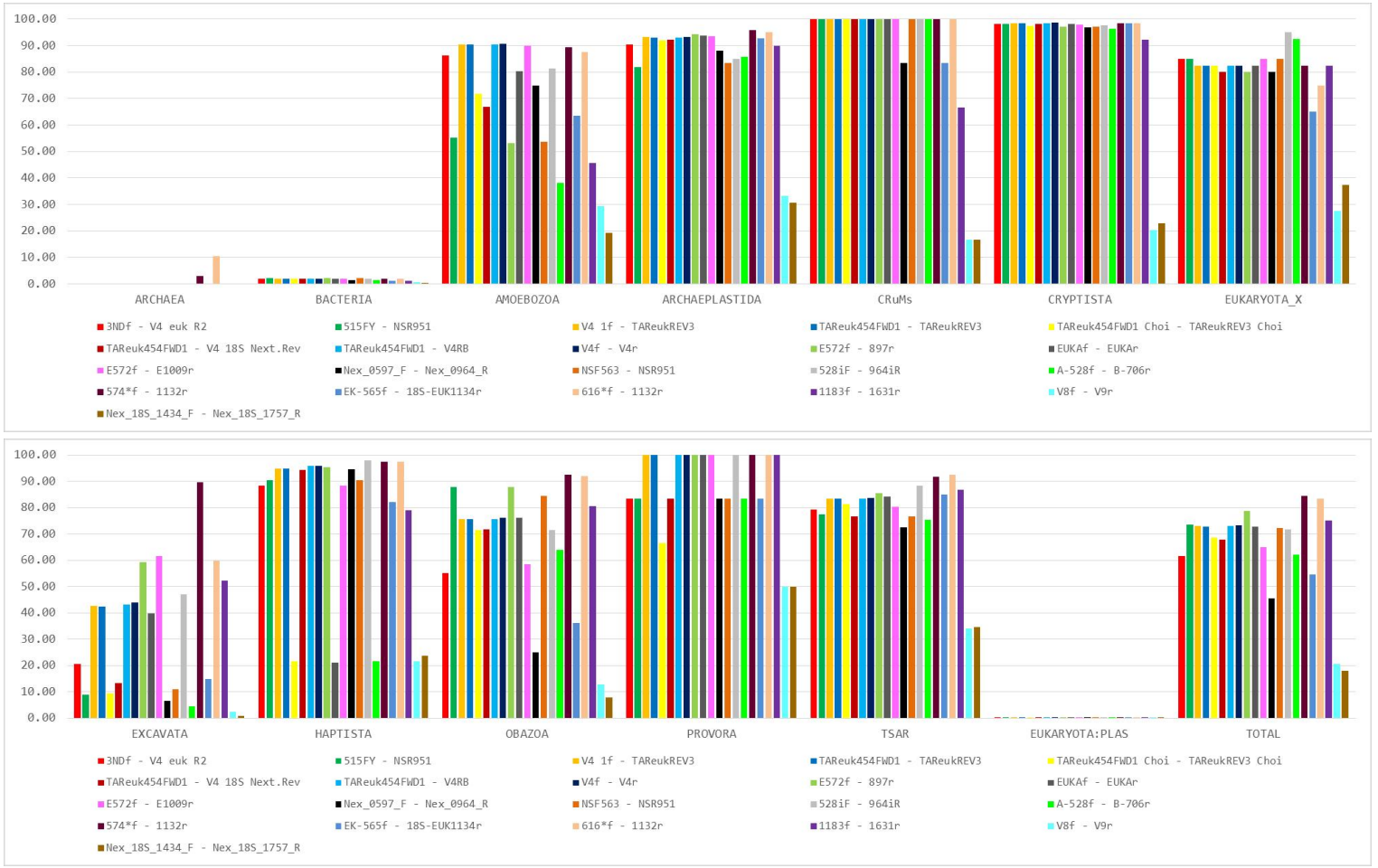


Figure 3. For an easier visualisation of the primer pairs' coverage of different taxonomic groups, this figure was created by joining figures 2A-2N.

amplified by #4, #13 and #17, and for *Schistosoma* amplified by #4, #13, #15, #17 and #24. *Entamoeba histolytica* was not amplified by primer pairs #4, #17, #21 and #26; pairs #22 and #24 produced long amplicons, while #13, #19 and #20 produced very long amplicons (length > 600 bp). *Fasciolopsis buski* was amplified by #3, #4, #13, #15, #17, #20 and #22; however, #3 produced long amplicons, while #20 and #22 produced very long amplicons. *Giardia intestinalis* was amplified only by pairs #14, #20 and #22. *Necator americanus*, *Trichinella spp.* and *Trichuris trichiura* were amplified by all primer pairs but #3, #15, #16, #19, #21 and #26; only #20 produced long amplicons. *Neobalantidium coli* was not amplified by pairs #3, #15 and #19. Both *Strongyloides* species were amplified by pairs #4, #7, #8, #9, #10, #11, #12 and #22.

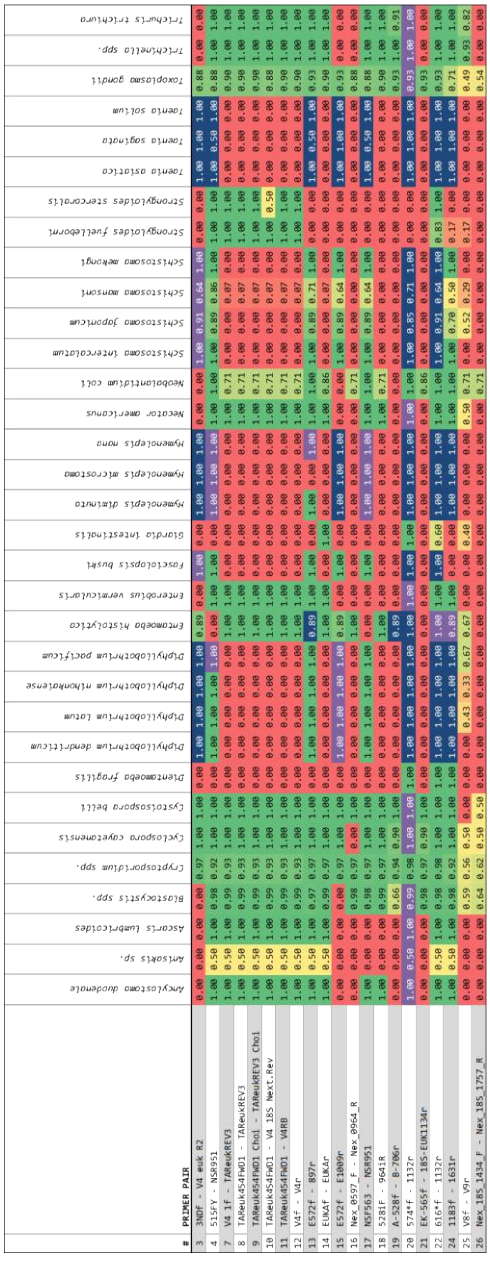


Table 6. This heat map represents all tested primer pairs' coverage on the target eukaryotic human intestinal parasites, along with information about the length of the produced amplicons. The value ($0.00 < x < 1.00$) shown in each cell corresponds to the ratio between the number of predicted amplified sequences and the number of total sequences of a certain parasite included in the PR^2 database. According to their value, cells are coloured with a chromatic scale (0.00 = Red, 0.50 = Yellow, 1.00 = Green) to better visualise the amplification ability of the tested primer pairs. Cells are coloured in purple if the considered primer pair is predicted to amplify the specified parasite's sequence producing amplicons with mean length between 550 bp and 600 bp; cells are coloured in blue if the produced amplicons are longer than 600 bp.
For a better visualisation, this heat map was divided in two parts that are showed in the next two pages.

#	PRIMER PAIR	*Ancylostoma duodenale*	*Anisakis sp.*	*Ascaris lumbricoides*	*Blastocystis spp.*	*Cryptosporidium spp.*	*Cyclospora cayetanensis*	*Cystoisospora belli*	*Dientamoeba fragilis*	*Diphyllobothrium dendriticum*	*Diphyllobothrium latum*	*Diphyllobothrium nihonkaiense*	*Diphyllobothrium pacificum*	*Entamoeba histolytica*	*Enterobius vermicularis*	*Fasciolopsis buski*	*Giardia intestinalis*
3	3NDf - V4 euk R2	0.00	0.00	0.00	0.00	0.97	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.89	0.00	1.00	0.00
4	515FY - NSR951	1.00	0.50	1.00	0.98	0.92	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00
7	V4 1f - TAReukREV3	1.00	0.50	1.00	0.99	0.93	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
8	TAReuk454FWD1 - TAReukREV3	1.00	0.50	1.00	0.99	0.93	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
9	TAReuk454FWD1 Choi - TAReukREV3 Choi	1.00	0.50	1.00	0.99	0.93	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
10	TAReuk454FWD1 - V4 18S Next.Rev	1.00	0.50	1.00	0.99	0.93	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
11	TAReuk454FWD1 - V4RB	1.00	0.50	1.00	0.99	0.93	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
12	V4f - V4r	1.00	0.50	1.00	0.99	0.93	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00
13	E572f - 897r	1.00	0.50	1.00	0.97	0.97	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.89	1.00	1.00	0.00
14	EUKAf - EUKAr	1.00	0.50	1.00	0.99	0.97	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00	1.00
15	E572f - E1009r	0.00	0.00	0.00	0.00	0.97	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.89	0.00	1.00	0.00
16	Nex_0597_F - Nex_0964_R	0.00	0.00	0.00	0.98	0.97	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
17	NSF563 - NSR951	1.00	0.00	0.00	0.98	0.97	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00	0.00
18	528iF - 964iR	1.00	0.00	0.00	0.99	0.97	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
19	A-528f - B-706r	0.00	0.00	0.00	0.66	0.94	0.90	1.00	0.00	0.00	0.00	0.00	0.00	0.89	0.00	0.00	0.00
20	574*f - 1132r	1.00	0.50	1.00	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
21	EK-565f - 18S-EUK1134r	0.00	0.00	0.00	0.98	0.97	0.90	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	616*f - 1132r	1.00	0.50	1.00	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.60
24	1183f - 1631r	1.00	0.50	1.00	0.98	0.92	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	1.00	0.00	0.00
25	V8f - V9r	0.00	0.00	0.00	0.59	0.56	0.50	0.00	0.00	0.00	0.43	0.33	0.67	0.67	0.00	0.00	0.40
26	Nex_18S_1434_F - Nex_18S_1757_R	0.00	0.00	0.00	0.64	0.62	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Hymenolepis diminuta	*Hymenolepis microstoma*	*Hymenolepis nana*	*Necator americanus*	*Neobalantidium coli*	*Schistosoma intercalatum*	*Schistosoma japonicum*	*Schistosoma mansoni*	*Schistosoma mekongi*	*Strongyloides fuelleborni*	*Strongyloides stercoralis*	*Taenia asiatica*	*Taenia saginata*	*Taenia solium*	*Toxoplasma gondii*	*Trichinella spp.*	*Trichuris trichiura*	PRIMER PAIR	#
1.00	1.00	1.00	0.00	0.00	1.00	0.91	0.64	1.00	0.00	0.00	1.00	1.00	1.00	0.88	0.00	0.00	3NDf - V4 euk R2	3
1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.86	1.00	1.00	1.00	1.00	0.50	1.00	0.88	1.00	1.00	515FY - NSR951	4
0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.07	0.00	1.00	1.00	0.00	0.00	0.00	0.90	1.00	1.00	V4 1f - TAReukREV3	7
0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.07	0.00	1.00	1.00	0.00	0.00	0.00	0.90	1.00	1.00	TAReuk454FWD1 - TAReukREV3	8
0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.07	0.00	1.00	1.00	0.00	0.00	0.00	0.90	1.00	1.00	TAReuk454FWD1 Choi - TAReukREV3 Choi	9
0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.07	0.00	1.00	0.50	0.00	0.00	0.00	0.88	1.00	1.00	TAReuk454FWD1 - V4 18S Next.Rev	10
0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.07	0.00	1.00	1.00	0.00	0.00	0.00	0.90	1.00	1.00	TAReuk454FWD1 - V4RB	11
0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.07	0.00	1.00	1.00	0.00	0.00	0.00	0.90	1.00	1.00	V4f - V4r	12
1.00	0.00	1.00	1.00	1.00	1.00	0.89	0.71	1.00	0.00	0.00	1.00	0.50	1.00	0.93	1.00	1.00	E572f - 897r	13
0.00	0.00	0.00	1.00	0.86	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.90	1.00	1.00	EUKAf - EUKAr	14
1.00	1.00	1.00	0.00	0.00	1.00	0.89	0.64	1.00	0.00	0.00	1.00	1.00	1.00	0.93	0.00	0.00	E572f - E1009r	15
0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	Nex_0597_F - Nex_0964_R	16
1.00	1.00	1.00	1.00	1.00	1.00	0.89	0.64	1.00	0.00	0.00	1.00	0.50	1.00	0.88	1.00	1.00	NSF563 - NSR951	17
0.00	0.00	0.00	1.00	0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.90	1.00	1.00	528iF - 964iR	18
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	1.00	0.91	A-528f - B-706r	19
1.00	1.00	1.00	1.00	1.00	1.00	0.85	0.71	1.00	0.00	0.00	1.00	1.00	1.00	0.93	1.00	1.00	574*f - 1132r	20
0.00	0.00	0.00	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	EK-565f - 18S-EUK1134r	21
1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.64	1.00	0.83	1.00	1.00	1.00	1.00	0.93	1.00	1.00	616*f - 1132r	22
1.00	1.00	1.00	1.00	1.00	1.00	0.70	0.50	1.00	0.17	0.00	1.00	1.00	1.00	0.71	1.00	1.00	1183f - 1631r	24
0.00	0.00	0.00	0.50	0.71	0.00	0.52	0.29	0.00	0.17	0.00	0.00	0.00	0.00	0.49	0.93	0.82	V8f - V9r	25
0.00	0.00	0.00	0.00	0.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.54	0.00	0.00	Nex_18S_1434_F - Nex_18S_1757_R	26

Overall, primer pairs #7, #8, #9, #10, #11, #12, #14, #16, #18, #19, #21, #25 and #26 poorly performed and thus were castoff, as they were unable to amplify more than half of the eukaryotic parasites of interest. Primer pairs #3 and #15 were discarded because they were able to amplify about 60% of the target organisms, but produced too many long amplicons. Primer pair #20 amplified 31 out of 33 parasites, but generated long amplicons with 27 of them so it was removed (**Table 7**). By comparing the global performances of the remaining primer pairs (#4, #13, #17, #22 and #24), we observed that #4 (515FY - NSR951), #13 (E572f - 897r) and #22 (616*f - 1132r) were the most efficient in amplifying the target eukaryotic parasites, thus we selected them for the DNA metabarcoding protocol. Specifically, #4 and #13 amplified at least 27 out of 33 parasites, generating few long amplicons; #22 was the only pair predicted to amplify all target organisms, hence it was included even if it produced long amplicons for almost half of them (**Table 8**).

#	PRIMER PAIR	G/T	R/T	P/T	B/T
3	3NDf - V4 euk R2	15.15	39.39	15.15	30.30
4	515FY - NSR951	69.70	9.09	12.12	9.09
7	V4 1f - TAReukREV3	48.48	51.52	0.00	0.00
8	TAReuk454FWD1 - TAReukREV3	48.48	51.52	0.00	0.00
9	TAReuk454FWD1 Choi - TAReukREV3 Choi	48.48	51.52	0.00	0.00
10	TAReuk454FWD1 - V4 18S Next.Rev	48.48	51.52	0.00	0.00
11	TAReuk454FWD1 - V4RB	48.48	51.52	0.00	0.00
12	V4f - V4r	48.48	51.52	0.00	0.00
13	E572f - 897r	69.70	15.15	3.03	12.12
14	EUKAf - EUKAr	45.45	54.55	0.00	0.00
15	E572f - E1009r	30.30	39.39	12.12	18.18
16	Nex_0597_F - Nex_0964_R	18.18	81.82	0.00	0.00
17	NSF563 - NSR951	57.58	24.24	9.09	9.09
18	528iF - 964iR	33.33	66.67	0.00	0.00
19	A-528f - B-706r	21.21	75.76	0.00	3.03
20	574*f - 1132r	12.12	6.06	33.33	48.48
21	EK-565f - 18S-EUK1134r	18.18	81.82	0.00	0.00
22	616*f - 1132r	51.52	0.00	3.03	45.45
24	1183f - 1631r	54.55	12.12	3.03	30.30
25	V8f - V9r	30.30	69.70	0.00	0.00
26	Nex_18S_1434_F - Nex_18S_1757_R	18.18	81.82	0.00	0.00

Table 7. This heat map summarises the results shown in **Table 6**.
"T" equals to the number of target human eukaryotic intestinal parasites considered in this work (i.e., 33): "G", "R", "P" and "B" equal, respectively, to the number of Green, Red, Purple and Blue cells in the considered primer pair's row in **Table 6**. For each primer pair, there are shown the percentage values of the four ratios "G/T" (i.e., the percentage of target parasites which >50% of sequences are predicted to be amplified), "R/T" (i.e., the percentage of target parasites which <50% of sequences are predicted to be amplified), "P/T" (i.e., the percentage of target parasites which are predicted to be amplified, but the produced amplicons' mean length is between 550 bp and 600 bp) and "B/T" (i.e., the percentage of target parasites which are predicted to be amplified, but the produced amplicons are longer than 600 bp). According to their value, cells are coloured with a chromatic scale (for "G/T": 0.00 = Red, 50.00 = Yellow, 100.00 = Green; for "R/T", "P/T" and "B/T": 0.00 = Green, 50.00 = Yellow, 100.00 = Red) to better visualise the amplification ability of the tested primer pairs.

As a matter of fact, we required a primer pair that could amplify all target parasites in order to assure detecting every pathogen contaminating the human faecal samples, in particular those unseen by the other exploited pairs. Therefore, all samples were examined with each selected pair; then, <550 bp amplicons and >550 bp amplicons were analysed with different bioinformatics pipelines and the results were compared to check their correspondence.

#	PRIMER PAIR	PR2	EUKARYOTES	PARASITES	>550 bp
3	3NDf - V4 euk R2	61.63	67.54	60.61	75.00
4	515FY - NSR951	73.65	80.73	90.91	23.33
7	V4 1f - TAReukREV3	73.00	80.02	48.48	0.00
8	TAReuk454FWD1 - TAReukREV3	72.87	79.89	48.48	0.00
9	TAReuk454FWD1 Choi - TAReukREV3 Choi	68.65	75.25	48.48	0.00
10	TAReuk454FWD1 - V4 18S Next.Rev	67.95	74.48	48.48	0.00
11	TAReuk454FWD1 - V4RB	72.98	80.00	48.48	0.00
12	V4f - V4r	73.25	80.30	48.48	0.00
13	E572f - 897r	78.84	86.43	84.85	17.86
14	EUKAf - EUKAr	72.81	79.81	45.45	0.00
15	E572f - E1009r	64.98	71.22	60.61	50.00
16	Nex_0597_F - Nex_0964_R	45.58	49.97	18.18	0.00
17	NSF563 - NSR951	72.16	79.10	75.76	24.00
18	528iF - 964iR	71.66	78.56	33.33	0.00
19	A-528f - B-706r	62.10	68.09	24.24	12.50
20	574*f - 1132r	84.45	92.60	93.94	87.10
21	EK-565f - 18S-EUK1134r	54.64	59.92	18.18	0.00
22	616*f - 1132r	83.46	91.50	100.00	48.48
24	1183f - 1631r	75.07	82.35	87.88	37.93
25	V8f - V9r	20.52	22.50	30.30	0.00
26	Nex_18S_1434_F - Nex_18S_1757_R	17.97	19.71	18.18	0.00

Table 8. This heat map summarises all information obtained by testing each selected primer pair by *in silico* PCR: the "PR2" column reports the percentage of sequences of the PR^2 database that are amplified by the tested primer pair; the "EUKARYOTES" column reports the percentage of eukaryotic sequences (excluding "Eukaryota:plas" sequences) that are amplified by the tested primer pair; the "PARASITES" column reports the percentage of target parasites which sequences are amplified by the tested primer pair, independently on the length of the produced amplicons (= "G/T" + "P/T" + "B/T"); the ">550 bp" column reports the percentage of target parasites sequences producing amplicons longer than 550 bp (= ("P/T" + "B/T") / ("G/T" + "P/T" + "B/T")). According to their value, cells are coloured with a chromatic scale (for "PR2", "EUKARYOTES" and "PARASITES": 0.00 = Red, 50.00 = Yellow, 100.00 = Green; for ">550 bp": 0.00 = Green, 50.00 = Yellow, 100.00 = Red) to better visualise the amplification ability of the primer pairs.

5.2. Primer pairs amplification test

The selected and 5'-modified (see chapter 4.2) primer pairs efficiently amplified the laboratory-available positive controls (see chapter 4.4), confirming that they could work well under the standard PCR amplification conditions (see chapter 4.3). In addition, the height of each band in the 1.5% agarose gel reflected the predicted length of the amplicons generated by each primer pair (**Figure 4**).

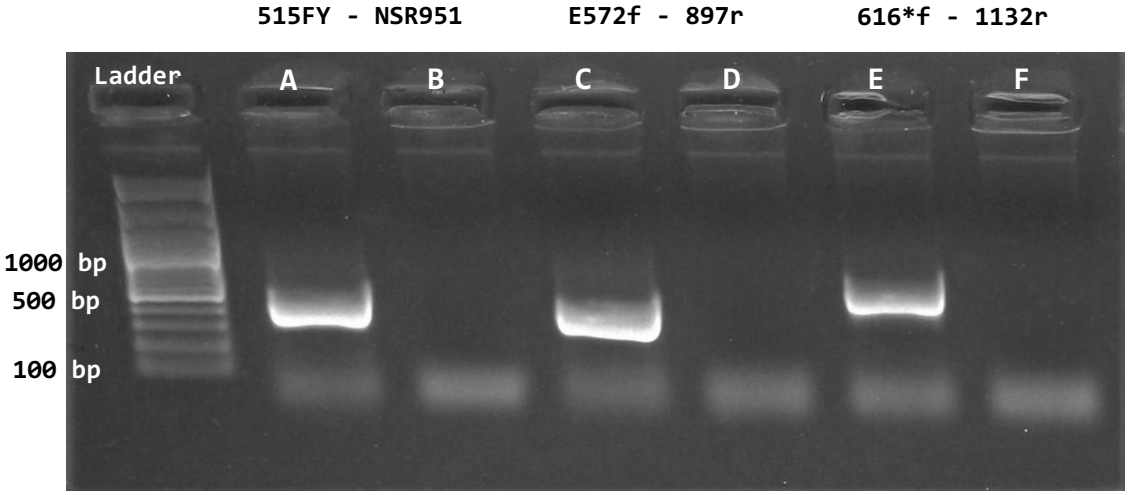


Figure 4. 1.5% agarose gel showing the products of the standard PCR amplification of eukaryotic genomes using the selected primer pairs. Lanes A and B correspond to positive and negative controls amplified with primer pair 515FY - NSR951; lanes C and D correspond to positive and negative controls amplified with primer pair E572f - 897r; lanes E and F correspond to positive and negative controls amplified with primer pairs 616*f - 1132r.

5.3. Optimisation of PCR amplification

Total DNA extracted from the formalin-fixed faecal samples (A-E, **Table 2**) via the standard DNA extraction protocol (see chapter 4.6) was subjected to standard PCR amplification with the selected primer pairs, but no amplicon was produced. Therefore, we tried increasing the amount of DNA in the PCR reaction (11.2 µL of DNA added to 13.8 µL of MM*) to overcome potential sensitivity issues, but amplification failed again.

Since faeces are generally rich in PCR inhibitors, we purified the extracted DNA (see chapter 4.7) and repeated the standard PCR amplification; however, it still gave negative results.

After those failings, we decided to pre-treat the formalin-fixed faecal samples to assure removing any trace of the formalin fixative (see chapter 4.8) that could inhibit PCR amplification; then, we manually extracted total DNA (see chapter 4.9) and amplified it with the standard reaction. Unfortunately, no amplicon was produced again.

However, by increasing the amount of DNA in the PCR reaction (11.2 μL of DNA added to 13.8 μL of MM*) we detected a very poor amplification, and thus we decided to also augment the number of PCR cycles so that we could generate a higher amount of amplicons. Overall, an efficient amplification of the target metabarcodes was achieved with all selected primer pairs by exploiting the optimised PCR amplification protocol (see chapter 4.10).

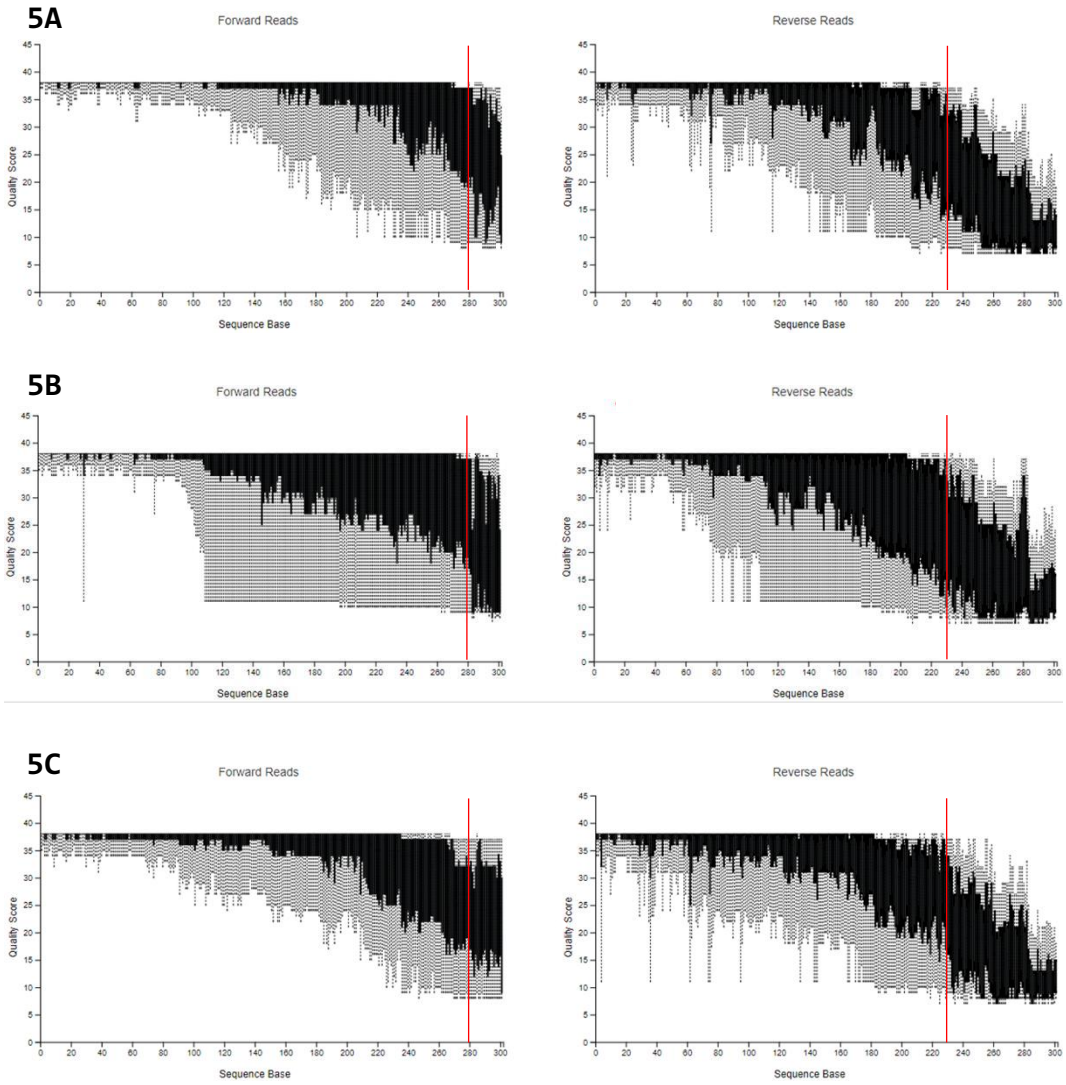
Before preparing the sequencing libraries (see chapter 4.11), we assigned to each PCR product (generated by amplifying the target metabarcodes of a faecal sample with a selected primer pair) a unique ID number. Products of samples A-E amplified by the 515FY - NSR951 primer pair were identified as 1460764-1460768; products of samples A-E amplified by the E572f - 897r primer pair were identified as 1460769-1460773; products of samples A-E amplified by the 616*f - 1132r primer pair were identified as 1460774-1460778 (**Table 2**).

5.4. Sequencing quality control

Analysing the sequencing results, we found an average Phred quality score (Q) above 30 (i.e., 1 error every 1000 sequenced bases) for more than 70% of reads in all the Illumina MiSeq runs (70% Q30 reads is the threshold proposed by Illumina to define a good 600-cycles sequencing run). In particular, 78.82% of reads were above Q30 when sequencing samples 1460764-1460768, 78.36% of reads were above Q30 when sequencing samples 1460769-1460773, and 75.76% of reads were above Q30 when sequencing samples 1460774-1460778. These results, obtained through the FastQC reports, were however expected.

5.5. Bioinformatics analysis of <550 bp amplicons

Raw reads were uploaded within the QIIME 2 environment and subsequently processed with the q2-cutadapt plugin to remove the amplification primers. Moreover, after the demultiplexing and trimming steps, we had to further truncate the low-quality 3'-end of both forward and reverse reads to obtain high-quality sequences that could be efficiently overlapped. Therefore, we evaluated again the mean quality of the demultiplexed and trimmed reads (**Figures 5A-5C**) and decided to truncate all forward reads at position 280 and all reverse reads at position 230. By using these parameters and considering the fact that the q2-dada2 plugin merges forward and reverse reads only if they overlap for at least 12 nucleotides, the longest metabarcodes that could be analysed with this bioinformatics pipeline (i.e., QIIME 2) were 498 bp long.



Figures 5A-5C. Each figure represents the quality profiles of 10000 randomly collected forward and reverse reads, after the step of primer removal. Plots A, B and C correspond respectively to the batch of samples 1460764-1460768, 1460769-146077 and 1460774-1460778. The red lines indicate where the reads were subsequently truncated during processing using the q2-dada2 plugin.

Overall, 42.8% of reads belonging to samples 1460764-1460768, 56.1% of reads belonging to samples 1460769-1460773 and 14.5% of reads belonging to samples 1460774-1460778 passed all DADA2 filters and were merged into ASVs representing the sequenced metabarcodes (≤498 bp) (**Tables 9A-9C**). The last percentage value was very low but not surprising, as primer pair 616*f - 1132r was anticipated to produce amplicons longer than 550 bp, thus the correspondent forward and reverse reads were not overlapping and could not be merged. As a consequence, the lowest amount of ASVs was identified in samples 1460774-1460778. Noteworthy, the longest identified ASV belonged to those samples and was 498 bp long (**Table 9D**), thus as long as the longest metabarcode that could be analysed with this approach.

9A

SAMPLE	SAMPLE ID	MERGED READS	ASVs
A	1460764	72,357	106
B	1460765	74,435	163
C	1460766	37,628	106
D	1460767	58,166	121
E	1460768	44,200	122
		286,786	469

9B

SAMPLE	SAMPLE ID	MERGED READS	ASVs
A	1460769	59,769	122
B	1460770	53,109	105
C	1460771	95,272	149
D	1460772	53,785	115
E	1460773	82,135	156
		344,070	496

9C

SAMPLE	SAMPLE ID	MERGED READS	ASVs
A	1460774	7,467	149
B	1460775	26,336	92
C	1460776	8,248	69
D	1460777	12,419	85
E	1460778	20,795	60
		75,265	379

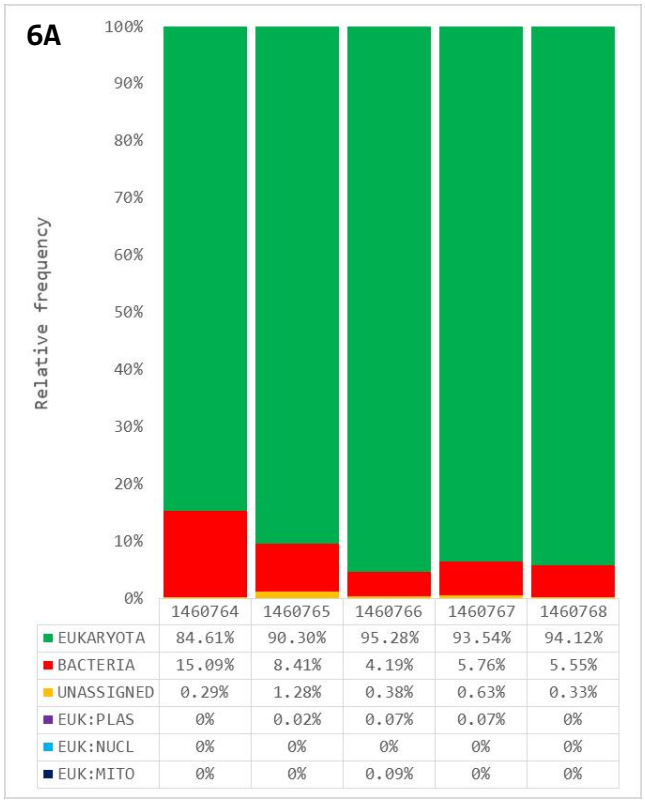
9D

PRIMER PAIR	MIN LENGTH	MAX LENGTH	MEAN LENGTH
515FY - NSR951	280 bp	490 bp	345.23 bp
E572f - 897r	280 bp	457 bp	310.62 bp
616*f - 1132r	303 bp	498 bp	471.55 bp

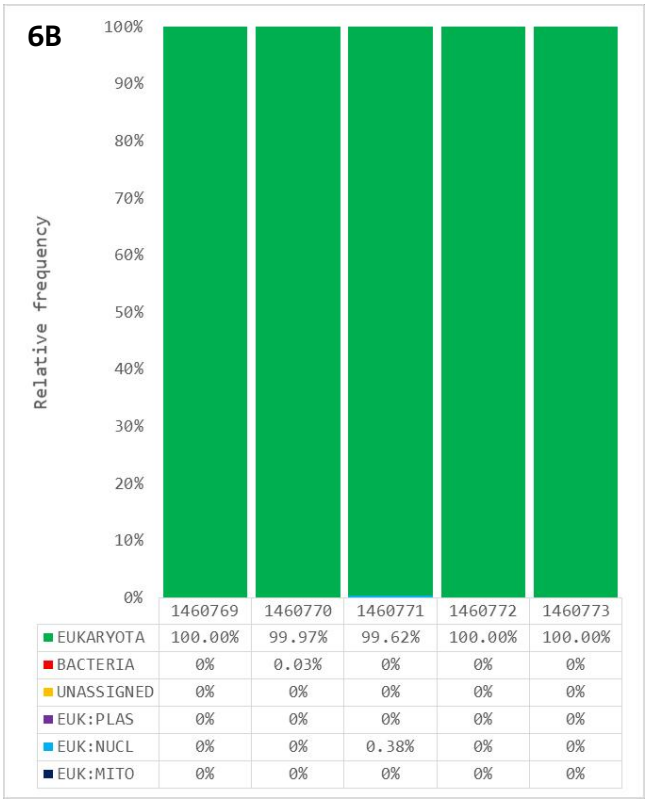
Tables 9A-9C show the results of the processing step using the q2-dada2 plugin. For each sample, it is reported the amount of merged reads that have passed all DADA2 filters and the amount of ASVs that have been identified. Below the "MERGED READS" column, the sum of all samples' merged reads is reported, while below the "ASVs" column, the total number of unique ASVs (not the sum of all samples' ASVs because some are shared between different samples) can be appreciated.
Table 9D shows the minimum, maximum and mean lengths of the ASVs that belong to all the samples amplified with the indicated primer pair.

Each ASV was assigned a taxonomy by comparing its sequence to those of the PR^2 database. Primer pairs 515FY - NSR951 and 616*f - 1132r amplified some bacterial sequences (**Figures 6A and 6C**), while only one off-target sequence was amplified by primer pair E572f - 897r (**Figure 6B**).
Focusing on the Bacteria found in the analysed samples (**Supplementary Figures S1A-S1C**), we observed that the results were in contrast with our predictions, as each primer pair amplified several bacterial sequences that were missed during the *in silico* PCR tests. Moreover, 5 bacterial species found in the samples by primer pairs 515FY - NSR951 and 616*f - 1132r were not detected by the third pair, even if it amplified those species *in silico*. Anyway, these results were of difficult interpretation as almost all detected Bacteria corresponded to environmental microorganisms that should have not been found in faecal matter.
On what concerns the detected eukaryotes, the results better resembled our predictions, as the number of eukaryotic species amplified by each primer pair both *in vivo* and *in silico* was 10-fold higher than the number of species amplified *in vivo* but not *in silico*. However, the overall quality of the results was not as good as we foresaw. Indeed, we found organisms that usually do not populate human faeces, as they were mainly environmental fungi, few algae and plants, and very few arthropods and animal-parasitizing fungi.

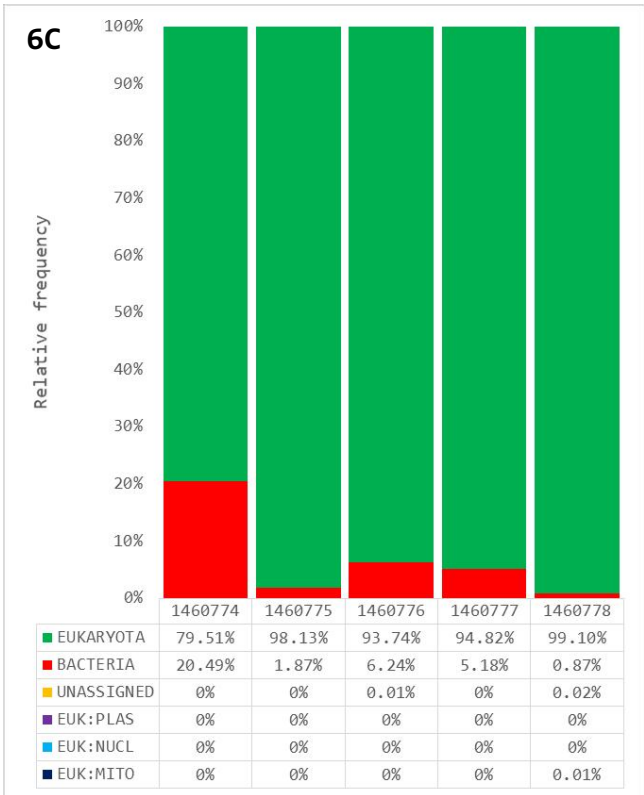
In addition to this, we could not detect any human intestinal parasite that UKNEQAS Parasitology declared to contaminate the analysed faecal samples (**Table 2**).



	1460764	1460765	1460766	1460767	1460768
EUKARYOTA	84.61%	90.30%	95.28%	93.54%	94.12%
BACTERIA	15.09%	8.41%	4.19%	5.76%	5.55%
UNASSIGNED	0.29%	1.28%	0.38%	0.63%	0.33%
EUK:PLAS	0%	0.02%	0.07%	0.07%	0%
EUK:NUCL	0%	0%	0%	0%	0%
EUK:MITO	0%	0%	0.09%	0%	0%



	1460769	1460770	1460771	1460772	1460773
EUKARYOTA	100.00%	99.97%	99.62%	100.00%	100.00%
BACTERIA	0%	0.03%	0%	0%	0%
UNASSIGNED	0%	0%	0%	0%	0%
EUK:PLAS	0%	0%	0%	0%	0%
EUK:NUCL	0%	0%	0.38%	0%	0%
EUK:MITO	0%	0%	0%	0%	0%



	1460774	1460775	1460776	1460777	1460778
EUKARYOTA	79.51%	98.13%	93.74%	94.82%	99.10%
BACTERIA	20.49%	1.87%	6.24%	5.18%	0.87%
UNASSIGNED	0%	0%	0.01%	0%	0.02%
EUK:PLAS	0%	0%	0%	0%	0%
EUK:NUCL	0%	0%	0%	0%	0%
EUK:MITO	0%	0%	0%	0%	0.01%

Figures 6A-6C. Barplots show the distribution of taxa in the samples, whose values are summarised in the table below.

5.6. Bioinformatics analysis of amplicons of any length

Pre-processing of the raw reads using fastp resulted in high quality forward and reverse reads that represented the 5' and 3' -ends of the sequenced metabarcodes, respectively (**Tables 10A-10C**). These unmerged reads were then paired-end aligned to the sequences of the PR^2 database using Bowtie 2, leading to the identification of archaeal, bacterial and eukaryotic organisms in all samples.

10A

SAMPLE	SAMPLE ID	% FILTERED READS	MEAN LENGTH F/R
A	1460764	94.48	246 bp / 222 bp
B	1460765	91.59	244 bp / 220 bp
C	1460766	93.13	260 bp / 232 bp
D	1460767	89.78	256 bp / 229 bp
E	1460768	93.25	259 bp / 232 bp

10B

SAMPLE	SAMPLE ID	% FILTERED READS	MEAN LENGTH F/R
A	1460769	95.78	262 bp / 232 bp
B	1460770	93.96	262 bp / 232 bp
C	1460771	93.90	262 bp / 232 bp
D	1460772	87.04	262 bp / 232 bp
E	1460773	78.55	262 bp / 232 bp

10C

SAMPLE	SAMPLE ID	% FILTERED READS	MEAN LENGTH F/R
A	1460774	89.87	262 bp / 232 bp
B	1460775	88.20	262 bp / 232 bp
C	1460776	88.94	262 bp / 232 bp
D	1460777	89.27	262 bp / 232 bp
E	1460778	89.18	262 bp / 232 bp

Tables 10A-10C. Each table shows the results of reads pre-processing using fastp. For each sample, it is reported the percentage of raw reads that have passed all filters and the mean lengths of the resulting forward and reverse pre-processed reads.

Focusing on the found prokaryotes, we observed that more than 90% of the species to which the reads were paired-end aligned were not amplified *in silico* by the different primer pairs. 102 bacterial species were identified in samples 1460764-1460768, but only 7 of them were amplified *in silico* by primer pair 515FY - NSR951; 68 bacterial species were identified in samples 1460769-1460773, but only 6 of them were amplified *in silico* by primer pair E572f - 897r; 109 archaeal and bacterial species were identified in samples 1460774-1460778, but only 9 of them were amplified *in silico* by primer pair 616*f - 1132r. However, differently from the results obtained from the analyses with QIIME 2, many identified genera corresponded to bacteria that could have been detected in a human faecal sample, being commensals (e.g., *Collinsella*, *Enterococcus*, *Faecalibacterium*, *Prevotella*, etc.) or pathogens (e.g., *Campylobacter*, *Klebsiella*, *Pseudomonas*, *Shigella*, etc.) of the human gastrointestinal tract. The others were environmental species.
Our predictions were again more reliable regarding the detected eukaryotes; indeed, the number of eukaryotic species amplified by each primer pair both *in vivo* and *in silico* was 2 to 4 -fold higher than the number of species amplified only *in vivo*. Unfortunately, almost 99% of the detected species were environmental fungi, algae, plants, metameric land worms, insects and free-living micro-eukaryotes, thus organisms that should not be found in a human faecal sample. And again, no human intestinal parasite was found.

The same unmerged reads, pre-processed with fastp, were paired-end aligned to the user-produced database composed of 18S rRNA gene sequences of the human eukaryotic intestinal parasites of interest. Only two target organisms were identified: *Cyclospora cayetanensis* in all samples but 1460775, and *Neobalantidium coli* in samples 1460765-1460768, 1460771, 1460773-1460775 and 1460778. However, none of the two was expected to be found in the faecal samples, as they did not correspond to the intestinal parasites identified by UKNEQAS Parasitology via microscopic analyses (**Table 2**).

5.7. Comparing the results of the bioinformatics analyses

334 taxonomies were identified in the faecal samples using QIIME 2: 12.6% of them were shared across all samples amplified with the three selected primer pairs, 24% were unique to samples 1460764-1460768, 17.3% were unique to samples 1460769-1460773 and 17.3% were unique to samples 1460774-1460778 (**Figure 7**).
3504 taxonomies were identified in the faecal samples using Bowtie 2 + featureCounts: 13.7% of them were shared across all samples amplified with the three selected primer pairs, 21.2% were unique to samples 1460764-1460768, 5.4%

were unique to samples 1460769-1460773 and 34% were unique to samples 1460774-1460778 (**Figure 8**).

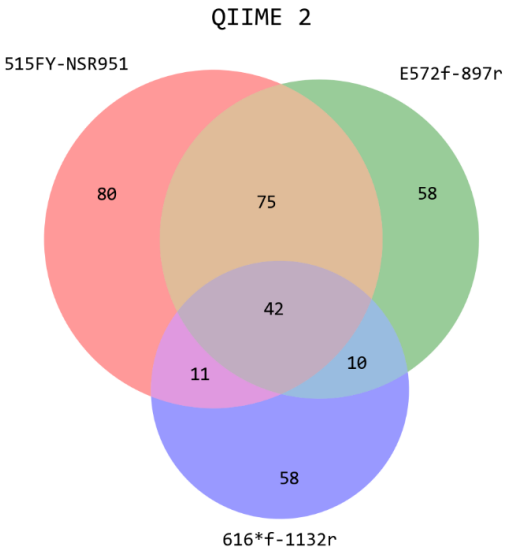


Figure 7. The area-weighted Venn diagram represents the number of exclusive or shared taxa, assigned with QIIME 2, that were detected in the samples amplified by the three primer pairs.

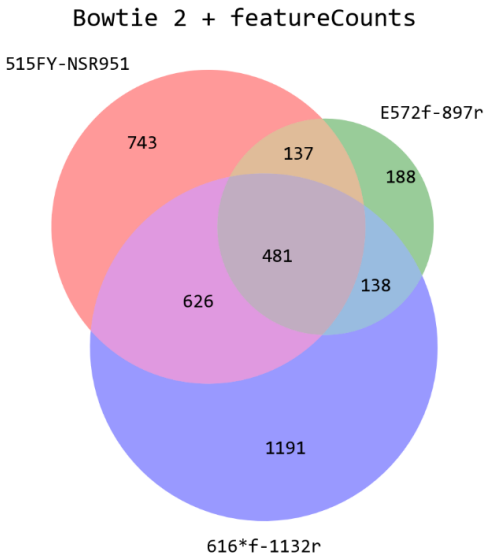
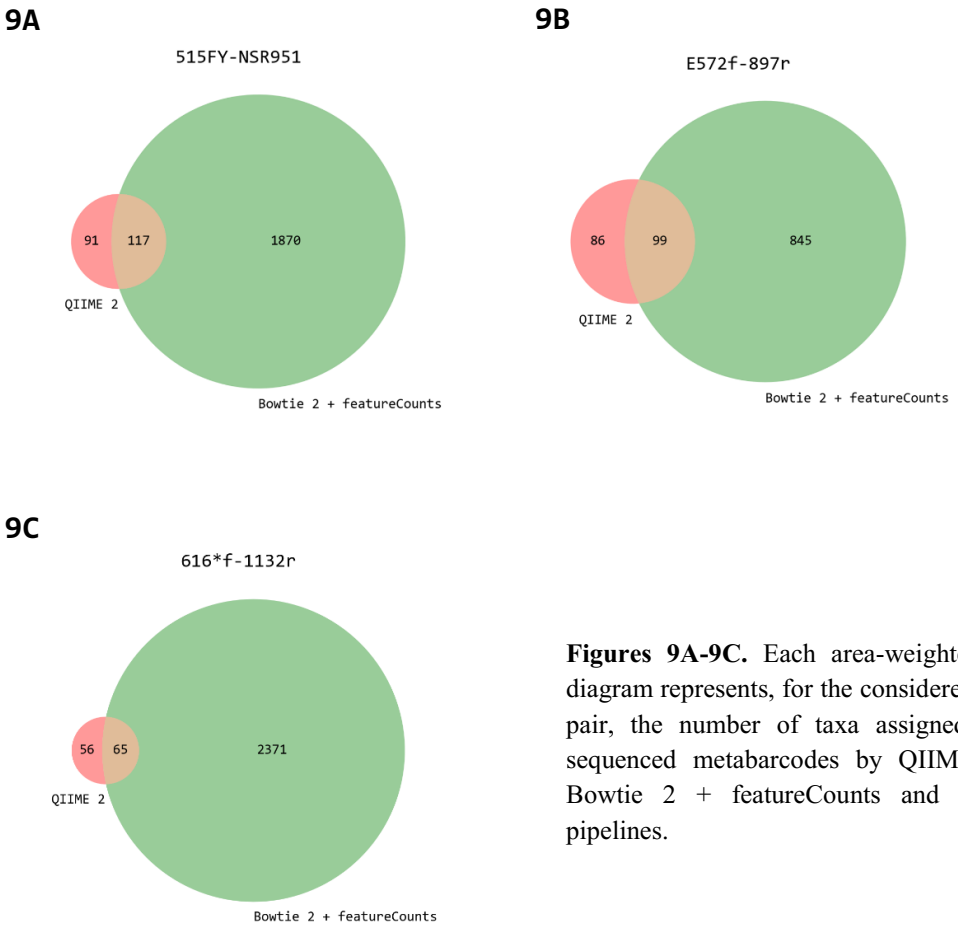


Figure 8. The area-weighted Venn diagram represents the number of exclusive or shared taxa, assigned with Bowtie 2 + featureCounts, that were detected in the samples amplified by the three primer pairs.

Overall, the highest number of taxonomies identified with QIIME 2 was in samples 1460764-1460768 (208), followed by samples 1460769-1460773 (185) and then by samples 1460774-1460778 (121). On the other hand, Bowtie 2 + featureCounts allowed detecting 2436 species in samples 1460774-1460778, 1987 species in samples 1460764-1460768 and 944 species in samples 1460769-1460773 (**Figures 9A-9C**). By comparing the organisms detected with the two bioinformatics pipelines, we observed that the number of taxonomies uniquely identified with QIIME 2 was strongly biased by the fact that most of them were not nine-level taxonomies, and thus they were not defined as identical to any of the nine-level taxonomies identified with Bowtie 2 + featureCounts. With further analyses, we verified that almost each non-nine-level taxonomy of QIIME 2 was "contained" in at least one nine-level taxonomy of Bowtie 2 + featureCounts, thus we approximated them as shared results between the two bioinformatics pipelines. Only 14, 21 and 4 genera or species were uniquely identified with QIIME 2 in samples 1460764-1460768, 1460769-1460773 and 1460774-1460778, respectively; therefore, Bowtie 2 + featureCounts allowed detecting at least 82.5% of the organisms found with QIIME 2.



Figures 9A-9C. Each area-weighted Venn diagram represents, for the considered primer pair, the number of taxa assigned to the sequenced metabarcodes by QIIME 2, by Bowtie 2 + featureCounts and by both pipelines.

This work suggests that the two bioinformatics pipelines may provide slightly different results for the same sets of samples. In particular, even if QIIME 2 was designed to support DNA metabarcoding studies (e.g., on prokaryotes), it fails in analysing long amplicons (> 550 bp); on the other hand, the custom-made Bowtie 2 + featureCounts pipeline, which was assembled in this work specifically to overcome the above limitations of QIIME 2, seems to be more consistent and includes almost all QIIME 2 results.

6. DISCUSSION

Nowadays, diagnosis of human intestinal infections caused by parasitic protozoa and helminths still relies on the poorly sensitive and specific microscopic examination of faecal samples. Clinicians can employ molecular methods (e.g., the PCR-based test) to improve the analytical sensitivity and specificity of the diagnostic procedure, but since modern techniques are not standardised yet and have some major drawbacks, the traditional approach is still considered as the unreplaceable gold standard.
DNA metabarcoding is an efficient approach that several laboratories exploit to investigate the taxonomic composition of the human gut microbiota, primarily its bacteriome collection, by analysing faecal samples. With this in mind, we wanted to evaluate whether it could be possible to develop a metabarcoding protocol that would allow the detection of the most common intestinal pathogenic protozoa and helminths in human stool. Indeed, we think that it could be a user-friendly, cost-effective and standardised method that could support the traditional diagnostic technique in a more efficient way than the available molecular procedures.

Notwithstanding the advantages of this approach, we had to solve different problems to complete the development of the protocol: (I) several universal primer pairs targeting the eukaryotic SSU rRNA gene are available, but not all of them efficiently amplify all or most of the clinically relevant intestinal eukaryotes; (II) DNA extraction, amplification and analysis are efficiently performed starting from high-quality samples, but contaminated faeces are frequently fixed in formalin to inactivate the pathogens, and this agent impairs the aforementioned protocol steps; (III) DNA metabarcoding is prevalently performed to characterise bacteria and fungi, exploiting validated bioinformatics pipelines that provide reliable results, but they cannot be always applied to the analysis of other metabarcodes (e.g., the long amplicons produced by amplifying the V4 region of the 18S rRNA gene of some eukaryotes). The following paragraphs describe how we overcame these issues.

In silico PCR was exploited to select 3 universal primer pairs amplifying the V4 hypervariable region of the 18S rRNA gene of the target human intestinal parasites. We required the highest possible coverage of the pathogens of interest, but also the production of amplicons shorter than 550 bp, as we exploited the Illumina MiSeq 300PE sequencing strategy and longer amplicons would be more difficult to analyse. The best compromise was achieved with primer pairs 515FY - NSR951, E572f - 897r and 616*f - 1132r. Their amplification efficiency was checked using an optimised PCR protocol on positive controls composed of purified eukaryotic genomes.

Total DNA extracted from the formalin-fixed faecal samples could not be efficiently amplified, even by increasing the amount of template or by purifying the DNA. This was due to the formalin fixative, which causes the formation of crosslinks between nucleic acids and proteins that strongly inhibit PCR amplification by keeping DNA in a closed conformation (Vitošević et al., 2018). Indeed, amplicons were proficiently generated only after pre-treating the faecal samples to remove formalin, confirming its inhibitory activity.
Preparation of the sequencing library followed the standard protocol, and sequencing errors were very limited.

Bioinformatics analyses using QIIME 2 provided negative results. We expected to identify few taxonomies in samples 1460774-1460778, as primer pair 616*f - 1132r was predicted to generate many >550 bp amplicons that could not be analysed with this bioinformatics pipeline. However, the detected organisms were very different from what we could foresee. As a matter of fact, almost all of them corresponded to environmental bacteria and eukaryotes that should not be found in human stool, and the exploited primer pairs amplified sequences of species that were missed during the preliminary *in silico* PCR tests. Moreover, none of the intestinal parasites declared to contaminate the faecal samples was detected.
Similar results were obtained by analysing the sequenced metabarcodes with fastp + Bowtie 2 + featureCounts. This time, we foresaw detecting the most species in samples 1460774-1460778, since primer pair 616*f-1132r has the highest coverage of the PR^2 database among the selected primer pairs and we exploited a bioinformatics pipeline that allows analysing metabarcodes of any length. Nevertheless, 90% of the found prokaryotes were not amplified *in silico*, and none of the identified eukaryotes was a commensal or parasite of the human gastrointestinal tract.

Only two human eukaryotic intestinal parasites (i.e., *Cyclospora cayetanensis* and *Neobalantidium coli*) were detected in the faecal samples by aligning the pre-processed reads to our custom-made database of parasites, but they were certainly false positives. As a matter of fact, they were not identified when aligning the same reads to the complete PR^2 database, thus they were the outcomes of the only poor-quality alignments that Bowtie 2 could perform with the parasites' database.

Since we exploited validated primer pairs, optimised the PCR amplification reaction, sequenced the target metabarcodes with a high-accuracy Illumina platform and analysed the produced reads with robust bioinformatics pipelines, the most plausible explanation to these negative results is the poor quality of the analysed faecal samples. Indeed, they were stored for several weeks and fixed with formalin, which can severely damage the DNA by breaking the phosphodiester bonds of its backbone (Vitošević et al., 2018) and inducing base alterations (i.e., N-hydroxymethyl mono-adducts on adenine, cytosine and guanine (Kawanishi et al., 2014)) in a time-dependent manner. Therefore, the samples may have contained fragmented DNA molecules rich of altered bases that biased the metabarcoding analyses, as they were aligned to sequences of prokaryotic and eukaryotic species that should not be identified in human faeces and thus did not allow the detection of the declared intestinal parasites.

Anyway, even if the obtained results were mostly artefacts, our *in silico* predictions about the primer pairs' performances seemed to be valid. Of the total amount of organisms detected by each primer pair, 63.9% were predicted for the 515FY - NSR951 pair, 70.8% were predicted for the E572f - 897r pair and 77.4% were predicted for the 616*f - 1132r pair.

7. CONCLUSION AND FUTURE PERSPECTIVES

In the present thesis, we designed and evaluated a metabarcoding approach for the detection of eukaryotic intestinal parasites in human stool. Unfortunately, since the contaminated faecal samples we could analyse were fixed with formalin and stored for a long period, the contained DNA molecules were probably fragmented and rich in base alterations, thus the results were mostly artefacts and we could not identify the contaminating intestinal parasites.
Nevertheless, the developed protocol should be robust: we selected the universal primer pairs by performing several *in silico* analyses and validated them; we optimised the PCR amplification protocol to efficiently produce the target amplicons; we followed standard protocols for the sequencing library preparation and sequenced the DNA molecules on the Illumina MiSeq platform; we exploited validated bioinformatics tools to analyse the sequencing results. Therefore, it would be interesting to repeat the described analyses on fresh, contaminated faecal samples to better evaluate the feasibility and potential success of this method. In addition, the protocol could be tested on mock communities of spiked samples (i.e., faecal matter artificially added with the target parasites' genomes (https://www.nature.com/articles/s41598-022-14176-z#Sec12)), while bioinformatics pipelines performance could be even assessed by using simulated data.

When performing a metabarcoding experiment, we recommend exploiting the fastp + Bowtie 2 + featureCounts pipeline for the bioinformatics analyses, as it is more versatile and identifies almost all species detected with QIIME 2, which, on the other hand, is limited to the analysis of forward and reverse reads that have to overlap at their 3'-ends.
Moreover, we highly discourage aligning the pre-processed reads to custom-made databases composed of few target sequences, as the identified organisms may be the outcome of low-quality alignments, thus corresponding to false positives.

An alternative version of this DNA metabarcoding protocol could be developed by using multiple primer pairs that are specific for different parasites' taxa. In particular, it would require modifying only the first PCR amplification step, where the extracted DNA would be amplified via several PCR reactions (one for each primer pair); then, if generated, the resulting amplicons would be pooled together to create a unique sequencing library that would be sequenced using an NGS platform and finally analysed by exploiting the described bioinformatics pipelines. With this approach, the first PCR amplification would become an initial control step revealing the possible presence or absence of the searched parasites.

As a matter of fact, if no amplicon was generated with a primer pair, it would probably mean the absence of the parasites that should be amplified by the same primer pair; on the other hand, if PCR products were produced, they would be further analysed to reveal which pathogenic species were contaminating the faecal sample. This control step could be helpful during clinical investigations, as negative PCR results would allow rapidly excluding some parasitic infections and help the clinician decide how to initiate the anti-parasitic treatment.

Regardless of how the first PCR amplification step would be performed, if clinicians needed to store the faecal samples, we would recommend using conservatives different from formalin (e.g., guanidine isothiocyanate, a chaotropic agent that inhibits DNases and RNases) to avoid DNA alterations and thus allow unbiased molecular analyses.

The work presented in this thesis, even if consists in a preliminary evaluation of DNA metabarcoding application to detect eukaryotic intestinal pathogens in formalin-fixed human faecal samples, could be exploited as support material to the future development of a clinically relevant diagnostic technique that solves the here-cited limitations and provides sensitive and specific diagnoses sustaining or even replacing those of the traditional microscopic examination of faeces.

8. ACKNOWLEDGEMENTS

I am grateful to the BMR Genomics S.r.l. team and in particular to Professor Giorgio Valle, PhD, for the opportunity to carry out my internship in this company. Special thanks go to Loris Bertoldi, PhD, and Eleonora Sattin, PhD, for their constant support and for teaching and explaining me the bioinformatics techniques and biological insights at the basis of this thesis. Finally yet importantly, I would like to thank Dr. Valeria Michela Besutti (AOP - UOC Microbiologia e Virologia) for kindly providing the samples analysed in this work.

9. SUPPLEMENTARY INFORMATION

9.1. BIOINFORMATICS PIPELINES

9.1.1. Bioinformatics analysis on QIIME 2 (v. 2023.5)

Working in the directory where the raw reads (FASTQ format files, named as [sample]_S[n]_L001_R[1,2]_001.fastq.gz) were downloaded (reads_directory), convert the raw reads in FastQC reports:

```
fastqc *fastq.gz -o fastqc_reports_output_directory
```

Working in reads_directory, activate QIIME 2 and create the required MANIFEST V2 file:

```
conda activate qiime2-2023.5
```

```
printf "sample-id\tforward-absolute-filepath\treverse-absolute-filepath\n" > manifest_v2.tsv
```

```
for r1 in *R1*; do sample=$( echo $r1 | cut -f 1 -d "_" ); r2=$( echo $r1 | sed 's/R1/R2/' ); printf "$sample\t$PWD/$r1\t$PWD/$r2\n"; done >> manifest_v2.tsv
```

Working in reads_directory, import and demultiplex the raw reads:

```
qiime tools import --type 'SampleData[PairedEndSequencesWithQuality]' --input-path reads_directory_absolute_filepath/manifest_v2.tsv --output-path raw_reads.qza --input-format PairedEndFastqManifestPhred33V2
```

```
qiime demux summarize --i-data raw_reads.qza --o-visualisation raw_reads.qzv
```

```
qiime tools export --input-path raw_reads.qzv --output-path demultiplexed_reads_output_directory
```

Working in reads_directory, trim the raw reads:

```
qiime cutadapt trim-paired --i-demultiplexed-sequences raw_reads.qza --p-front-f forward_adapter_and_primer_sequence --p-front-r reverse_adapter_and_primer_sequence --o-trimmed-sequences raw_reads.trimmed.qza
```

```
qiime demux summarize --i-data raw_reads.trimmed.qza --o-visualisation raw_reads.trimmed.qzv
```

```
qiime tools export --input-path raw_reads.trimmed.qzv --output-path trimmed_reads_output_directory
```

Working in reads_directory, perform trimmed reads filtering, dereplication, denoising and merging, and chimera detection:

```
qiime dada2 denoise-paired --i-demultiplexed-seqs raw_reads.trimmed.qza --p-trunc-len-f read1_truncation_position --p-trunc-len-r read2_truncation_position --p-chimera-method consensus --o-table feature_ASV_table --o-representative-sequences features_ASVs --o-denoising-stats dada2_results.stats
```

```
qiime metadata tabulate --m-input-file dada2_results.stats.qza --o-visualisation dada2_results.stats.qzv
```

```
qiime tools export --input-path dada2_results.stats.qzv --output-path dada2_results_output_directory
```

Working in reads_directory, create the interactive reports of the ASV table and of the ASVs:

```
qiime feature-table summarize --i-table feature_ASV_table.qza --o-visualisation feature_ASV_table.qzv
```

```
qiime tools export --input-path feature_ASV_table.qzv --output-path feature_ASV_table_output_directory
```

```
qiime feature-table tabulate-seqs --i-data features_ASVs.qza --o-visualisation features_ASVs.qzv
```

```
qiime tools export --input-path features_ASVs.qzv --output-path features_ASVs_output_directory
```

Working in reads_directory, save the ASV table as a TSV format file and the ASVs as a FASTA format file:

```
qiime tools export --input-path feature_ASV_table.qza --output-path feature_ASV_table_tsv_output_directory
```

```
biom convert --input-fp feature_ASV_table_tsv_output_directory/feature-table.biom --output-fp feature_ASV_table_tsv_output_directory/feature_ASV_table.tsv --to-tsv
```

```
qiime tools export --input-path features_ASVs.qza --output-path features_ASVs_FASTA_output_directory
```

Working in the directory where the sequences and taxonomy files of the PR^2 database were downloaded (pr2_directory), convert those files in QZA format files:

```
qiime tools import --type 'FeatureData[Sequence]' --input-path pr2_version_5.0.0_SSU_mothur.fasta --output-path pr2_version_5.0.0_SSU_mothur_seqs.qza
```

```
qiime tools import --type 'FeatureData[Taxonomy]' --input-format
HeaderlessTSVTaxonomyFormat --input-path
pr2_version_5.0.0_SSU_mothur.tax --output-path
pr2_version_5.0.0_SSU_mothur_tax.qza
```

Working in a subdirectory (pr2_subdirectory), extract the target metabarcode sequence from each sequence of the PR^2 database:

```
qiime feature-classifier extract-reads --i-sequences
pr2_directory/pr2_version_5.0.0_SSU_mothur_seqs.qza --p-f-primer
forward_primer_sequence --p-r-primer reverse_primer_sequence --p-
min-length 0 --o-reads primer_pair_name_extracted_pr2_reads.qza
```

Working in pr2_subdirectory, dereplicate the extracted metabarcode sequences and the taxonomy of the PR^2 database:

```
qiime rescript dereplicate --i-sequences
primer_pair_name_extracted_pr2_reads.qza --i-taxa
pr2_directory/pr2_version_5.0.0_SSU_mothur_tax.qza --p-rank-
handles 'disable' --p-mode uniq --o-dereplicated-sequences
primer_pair_name_extr_derep_pr2_reads.qza --o-dereplicated-taxa
primer_pair_name_derep_pr2_tax.qza
```

Working in pr2_subdirectory, train the taxonomic classifier on the dereplicated, extracted metabarcode sequences of the PR^2 database:

```
qiime rescript evaluate-fit-classifier --i-sequences
primer_pair_name_extr_derep_pr2_reads.qza --i-taxonomy
primer_pair_name_derep_pr2_tax.qza --output-dir
classifier_output_directory
```

Working in classifier_output_directory, create the interactive report of the classifier:

```
qiime tools export --input-path evaluation.qzv --output-path
classifier_evaluation_output_directory
```

Working in pr2_subdirectory, create the interactive reports describing the extracted and dereplicated metabarcode sequences:

```
qiime rescript evaluate-seqs --i-sequences
primer_pair_name_extracted_pr2_reads.qza --o-visualisation
primer_pair_name_extracted_pr2_reads.qzv
```

```
qiime tools export --input-path
primer_pair_name_extracted_pr2_reads.qzv --output-path
extracted_reads_output_directory
```

```
qiime rescript evaluate-seqs --i-sequences
primer_pair_name_extr_derep_pr2_reads.qza --o-visualisation
primer_pair_name_extr_derep_pr2_reads.qzv
```

```
qiime tools export --input-path primer_pair_name_extr_derep_pr2_reads.qzv --output-path extracted_dereplicated_reads_output_directory
```

Working in reads_directory, assign the taxonomy to the ASVs:

```
qiime feature-classifier classify-sklearn --i-classifier classifier_output_directory_absolute_filepath/classifier.qza --i-reads features_ASVs.qza --o-classification taxonomy_ASVs.qza
```

```
qiime metadata tabulate --m-input-file taxonomy_ASVs.qza --o-visualisation taxonomy_ASVs.qzv
```

```
qiime tools export --input-path taxonomy_ASVs.qzv --output-path ASVs_taxonomy_output_directory
```

Working in reads_directory, create the interactive barplots of the ASVs taxonomy:

```
qiime taxa barplot --i-table feature_ASV_table.qza --i-taxonomy taxonomy_ASVs.qza --o-visualisation taxonomy_ASVs_barplots.qzv
```

```
qiime tools export --input-path taxonomy_ASVs_barplots.qzv --output-path barplots_output_directory
```

9.1.2. Bioinformatics analysis on fastp (v. 0.23.2) + Bowtie 2 (v. 2.5.1) + featureCounts (v. 2.0.1) (with full PR^2 database (v. 5.0.0))

Working in reads_directory, pre-process the raw reads with fastp:

```
for r1 in *R1*; do r2=$( echo $r1 | sed 's/R1/R2/' ); out1=$( echo $r1 | sed 's/L001_R1_001/R1_prep/' ); out2=$( echo $r1 | sed 's/L001_R1_001/R2_prep/' ); sample=$( echo $r1 | cut -f 1 -d "_" ); fastp -i $r1 -I $r2 -o $out1 -O $out2 -h $sample.html -q phred_quality -u unqualified_limit -e average_phred_score --detect_adapter_for_pe -f read1_trim_front_position -t read1_trim_tail_position -F read2_trim_front_position -T read2_trim_tail_position; done
```

```
cp *prep* fastp_results_output_directory
```

Working in a subdirectory (reads_subdirectory), create the PR^2 database index required by Bowtie 2:

```
bowtie2-build pr2_directory_absolute_filepath/pr2_version_5.0.0_SSU_mothur.fasta pr2_index
```

Working in fastp_results_output_directory, align the pre-processed reads to the PR^2 database index and create sorted BAM format files:

```
for r1 in *R1*; do r2=$( echo $r1 | sed 's/R1/R2/' ); sample=$( echo $r1 | cut -f 1 -d "_" ); bowtie2 -x reads_subdirectory/pr2_index -1 $r1 -2 $r2 alignment_option -I min_fragment_length -X max_fragment_length | samtools view -b | samtools sort > reads_subdirectory/${sample}_sorted.bam; done
```

Working in reads_subdirectory, create the index of each BAM format file:

```
for file in *sorted*; do samtools index $file; done
```

Working with Python (ipython3) in reads_subdirectory, create the annotation file required by featureCounts:

```
import sys
```

```
dictionary = {}
```

```
with open("pr2_directory_absolute_filepath/pr2_version_5.0.0_SSU_mothur.tax") as taxonomy:
    for line in taxonomy:
        line = line.strip()
        splitted_line = line.split("\t")
        seqID = splitted_line[0]
        tax = splitted_line[1]
        dictionary[seqID] = tax
```

```
annot = open("annotation_file.tsv", "w")
```

```
header_line = "GeneID\tChr\tStart\tEnd\tStrand\n"
```

```
annot.write(header_line)
```

```
with
open("pr2_directory_absolute_filepath/pr2_version_5.0.0_SSU_mothur
.fasta") as sequences:
      for entry in sequences:
            entry = entry.strip()
            if entry.startswith(">"):
                  seqID = entry.replace(">", "")
            else:
                  seq = entry
                  annot.write(f"{dictionary[seqID]}\t{seqID}\t{1}\
                  t{len(seq)}\t{'+'}\n")
```

```
annot.close()
```

Working in `reads_subdirectory`, count the aligned paired-end reads:

```
featureCounts -a annotation_file.tsv -o feature_counts.tsv *.bam -
B -C -d min_fragment_length -D max_fragment_length -F SAF -M -p
```

9.1.3. Bioinformatics analysis on fastp (v. 0.23.2) + Bowtie 2 (v. 2.5.1) + featureCounts (v. 2.0.1) (with user-produced database)

The used-produced database (`parasites_18S_rRNA_pr2_database.fasta`) was downloaded in `reads_subdirectory_2`, and then we modified it to produce two files: a FASTA format file with the 18S rRNA gene sequences and a TSV format file with the correspondent taxonomies.

```
cut -f 1 -d "|" parasites_18S_rRNA_pr2_database.fasta >
parasites_18S_rRNA_pr2_seq_database.fasta
```

```
cut -f 1-8 -d "|" -s parasites_18S_rRNA_pr2_database.fasta | sed
's/>//' | sed 's/|/\t/' | sed 's/|/;/g' >
parasites_18S_rRNA_pr2_taxonomy_database.tsv
```

Working in `reads_subdirectory_2`, we created the user-produced database index required by Bowtie 2:

```
bowtie2-build parasites_18S_rRNA_pr2_seq_database.fasta
small_database_index
```

Working in `fastp_results_output_directory`, we aligned the pre-processed reads to the user-produced database index and created sorted BAM format files:

```
for r1 in *R1*; do r2=$( echo $r1 | sed 's/R1/R2/' ); sample=$( echo $r1 | cut -f 1 -d "_" ); bowtie2 -x reads_subdirectory_2/small_database_index -1 $r1 -2 $r2 alignment_option -I min_fragment_length -X max_fragment_length | samtools view -b | samtools sort > reads_subdirectory_2/${sample}_sorted.bam; done
```

Working in `reads_subdirectory_2`, we created the index of each BAM format file:

```
for file in *sorted*; do samtools index $file; done
```

Working with Python (`ipython3`) in `reads_subdirectory_2`, we created the annotation file required by featureCounts:

```
import sys
```

```
small_dictionary = {}
```

```
with open("parasites_18S_rRNA_pr2_taxonomy_database.tsv") as taxon:
    for row in taxon:
        row = row.strip()
        splitted_row = row.split("\t")
        seqsID = splitted_row[0]
        taxa = splitted_row[1]
        small_dictionary[seqsID] = taxa
```

```
annotation = open("new_annotation_file.tsv", "w")
```

```
header = "GeneID\tChr\tStart\tEnd\tStrand\n"
```

```
annotation.write(header)
```

```
with open("parasites_18S_rRNA_pr2_seq_database.fasta") as sequenc:
    for lines in sequenc:
        lines = lines.strip()
        if lines.startswith(">"):
            seqsID = lines.replace(">", "")
        else:
            seqs = lines
            annotation.write(f"{small_dictionary[seqsID]}\t{seqsID}\t{1}\t{len(seqs)}\t{'+'}\n")
```

```
annotation.close()
```

Working in `reads_subdirectory_2`, we counted the aligned paired-end reads:

```
featureCounts -a new_annotation_file.tsv -o new_feature_counts.tsv *.bam -B -C -d min_fragment_length -D max_fragment_length -F SAF -M -p
```

9.1.4. Comparison of the results obtained from the analyses with the 3 selected primer pairs

Working in `home_directory`, which is above the three `reads_directory` (one for each analysis with a selected primer pair), we uploaded the "`venn_qiime2.py`" file and created the Venn diagram describing how many taxonomies were assigned by QIIME 2 for each analysis with a selected primer pair:

```
python venn_qiime2.py reads_directory(of-primer-pair-515FY-NSR951)/ASVs_taxonomy_output_directory/metadata.tsv reads_directory(of-primer-pair-E572f-897r)/ASVs_taxonomy_output_directory/metadata.tsv reads_directory(of-primer-pair-616*f-1132r)/ASVs_taxonomy_output_directory/metadata.tsv
```

The "`venn_qiime2.py`" file contained the following script:

```
#!/usr/bin/env python
```

```
import sys
```

```
from matplotlib import font_manager
consolas_url=[font for font in font_manager.findSystemFonts(fontpaths=None, fontext='ttf') if "Consolas-Regular" in font][0]
consolas_title=font_manager.FontProperties(fname=consolas_url, size=22)
consolas_label=font_manager.FontProperties(fname=consolas_url, size=16)
```

```
input_file_1=sys.argv[1]
```

```
list_515FY_NSR951=[]
```

```
metadata_1=open(input_file_1, "r")
for line in metadata_1:
      if line.startswith("Feature") or line.startswith("#"):
            continue
      field=line.split("\t")
      taxonomy=field[1]
      list_515FY_NSR951.append(taxonomy)
metadata_1.close()
```

```
input_file_2=sys.argv[2]
```

```
list_E572f_897r=[]
```

```
metadata_2=open(input_file_2, "r")
for lines in metadata_2:
      if lines.startswith("Feature") or lines.startswith("#"):
            continue
      fields=lines.split("\t")
```

```
      taxon=fields[1]
      list_E572f_897r.append(taxon)
metadata_2.close()
```

```
input_file_3=sys.argv[3]
```

```
list_616f_1132r=[]
```

```
metadata_3=open(input_file_3, "r")
for linea in metadata_3:
      if linea.startswith("Feature") or linea.startswith("#"):
            continue
      campo=linea.split("\t")
      tax=campo[1]
      list_616f_1132r.append(tax)
metadata_3.close()
```

```
import matplotlib.pyplot as plt
from matplotlib_venn import venn3
fig=plt.figure(figsize=(8,8), dpi=150)
set1=set(list_515FY_NSR951)
set2=set(list_E572f_897r)
set3=set(list_616f_1132r)
diagram=venn3([set1, set2, set3], ('515FY-NSR951', 'E572f-897r',
'616*f-1132r'))
for text in diagram.set_labels:
      text.set_fontproperties(consolas_label)
for text in diagram.subset_labels:
      text.set_fontproperties(consolas_label)
plt.title("QIIME 2", fontproperties=consolas_title)
fig.savefig("venn_qiime2_results.png")
```

Working in `home_directory`, we uploaded the "`venn_featurecounts.py`" file and created the Venn diagram describing how many taxonomies were assigned by Bowtie 2 + featureCounts for each analysis with a selected primer pair:

```
python venn_featurecounts.py reads_directory(of-primer-pair-515FY-
NSR951)/reads_subdirectory/feature_counts.tsv  reads_directory(of-
primer-pair-E572f-897r)/reads_subdirectory/feature_counts.tsv
reads_directory(of-primer-pair-616*f-
1132r)/reads_subdirectory/feature_counts.tsv
```

The "`venn_featurecounts.py`" file contained the following script:

```
#!/usr/bin/env python
```

```
import sys
```

```
from matplotlib import font_manager
consolas_url=[font             for             font             in
font_manager.findSystemFonts(fontpaths=None,   fontext='ttf')   if
"Consolas-Regular" in font][0]
```

```
consolas_title=font_manager.FontProperties(fname=consolas_url,
size=22)
consolas_label=font_manager.FontProperties(fname=consolas_url,
size=16)
```

```
input_file_1=sys.argv[1]
```

```
list_515FY_NSR951=[]
```

```
feature_counts_1=open(input_file_1, "r")
for line in feature_counts_1:
    if line.startswith("#") or line.startswith("Geneid"):
        continue
    field=line.split("\t")
    count=[int(n) for n in field[6:]]
    tot=sum(count)
    if tot>0:
        taxonomy=field[0]
        list_515FY_NSR951.append(taxonomy)
feature_counts_1.close()
```

```
input_file_2=sys.argv[2]
```

```
list_E572f_897r=[]
```

```
feature_counts_2=open(input_file_2, "r")
for lines in feature_counts_2:
    if lines.startswith("#") or lines.startswith("Geneid"):
        continue
    fields=lines.split("\t")
    counts=[int(m) for m in fields[6:]]
    total=sum(counts)
    if total>0:
        taxon=fields[0]
        list_E572f_897r.append(taxon)
feature_counts_2.close()
```

```
input_file_3=sys.argv[3]
```

```
list_616f_1132r=[]
```

```
feature_counts_3=open(input_file_3, "r")
for linea in feature_counts_3:
    if linea.startswith("#") or linea.startswith("Geneid"):
        continue
    campo=linea.split("\t")
    conta=[int(p) for p in campo[6:]]
    totale=sum(conta)
    if totale>0:
        tax=campo[0]
        list_616f_1132r.append(tax)
feature_counts_3.close()
```

```
import matplotlib.pyplot as plt
```

```
from matplotlib_venn import venn3
fig=plt.figure(figsize=(8,8), dpi=150)
set1=set(list_515FY_NSR951)
set2=set(list_E572f_897r)
set3=set(list_616f_1132r)
diagram=venn3([set1, set2, set3], ('515FY-NSR951', 'E572f-897r',
'616*f-1132r'))
```

```
for text in diagram.set_labels:
    text.set_fontproperties(consolas_label)
for text in diagram.subset_labels:
    text.set_fontproperties(consolas_label)
plt.title("Bowtie 2 + featureCounts",
fontproperties=consolas_title)
fig.savefig("venn_featurecounts_results.png")
```

9.1.5. Comparison of the results obtained from the analyses at 9.1.1 and 9.1.2

Working in `reads_directory`, we collapsed the ASV table to the species level, created the interactive report of the collapsed ASV table and saved the collapsed ASV table as a TSV format file:

```
qiime taxa collapse --i-table feature_ASV_table.qza --i-taxonomy
taxonomy_ASVs.qza --p-level 9 --o-collapsed-table
collapsed_feature_table.qza
```

```
qiime feature-table summarize --i-table collapsed_feature_table.qza
--o-visualisation collapsed_feature_table.qzv
```

```
qiime tools export --input-path collapsed_feature_table.qzv --
output-path collapsed_feature_table_output_directory
```

```
qiime tools export --input-path collapsed_feature_table.qza --
output-path collapsed_feature_table_tsv_output_directory
```

```
biom convert --input-fp
collapsed_feature_table_tsv_output_directory/feature-table.biom --
output-fp
collapsed_feature_table_tsv_output_directory/collapsed_feature_tab
le.tsv --to-tsv
```

Working in `reads_directory`, we uploaded the "`final_comparison.py`" file and compared the taxonomies assigned by QIIME 2 and by Bowtie 2 + featureCounts:

```
python final_comparison.py reads_subdirectory/feature_counts.tsv
collapsed_feature_table_tsv_output_directory/collapsed_feature_tab
le.tsv > final_comparison_primer_pair_name.tsv
```

The "final_comparison.py" file contained the following script:

```
#!/usr/bin/env python
```

```
import sys
```

```
input_file_1=sys.argv[1]
```

```
dictionary_feature_counts={}
```

```
feature_counts=open(input_file_1, "r")
for l,line in enumerate(feature_counts):
      line=line.strip("\n")
      if line.startswith("#") or line.startswith("Geneid"):
            continue
      field=line.split("\t")
      count=[int(n) for n in field[6:]]
      tot=sum(count)
      if tot>0:
            dictionary_feature_counts[field[0]]=tot
feature_counts.close()
```

```
input_file_2=sys.argv[2]
```

```
collapsed_feature_table=open(input_file_2, "r")
for lines in collapsed_feature_table:
      lines=lines.strip("\n")
      if lines.startswith("#"):
            continue
      fields=lines.split("\t")
      counts=[float(m) for m in fields[1:]]
      total=sum(counts)
      if fields[0] in dictionary_feature_counts:
            print(fields[0], dictionary_feature_counts[fields[0]],
            total)
```

Working in reads_directory, we uploaded the "venn_finalcomparison_primer_pair_name.py" file and created the Venn diagram describing how many taxonomies were assigned by QIIME 2, how many taxonomies were assigned by Bowtie 2 + featureCounts, and how many taxonomies were assigned by both:

```
python venn_finalcomparison_primer_pair_name.py
ASVs_taxonomy_output_directory/metadata.tsv
reads_subdirectory/feature_counts.tsv
```

The "venn_finalcomparison_primer_pair_name.py" file contained the following script:

```
#!/usr/bin/env python
```

```
import sys
```

```
from matplotlib import font_manager
consolas_url=[font             for             font             in
font_manager.findSystemFonts(fontpaths=None,   fontext='ttf')   if
"Consolas-Regular" in font][0]
consolas_title=font_manager.FontProperties(fname=consolas_url,
size=22)
consolas_label=font_manager.FontProperties(fname=consolas_url,
size=16)
```

```
input_file_1=sys.argv[1]
```

```
list_primer_pair_name=[]
```

```
metadata=open(input_file_1, "r")
for line in metadata:
      if line.startswith("Feature") or line.startswith("#"):
            continue
      field=line.split("\t")
      taxonomy=field[1]
      list_primer_pair_name.append(taxonomy)
metadata.close()
```

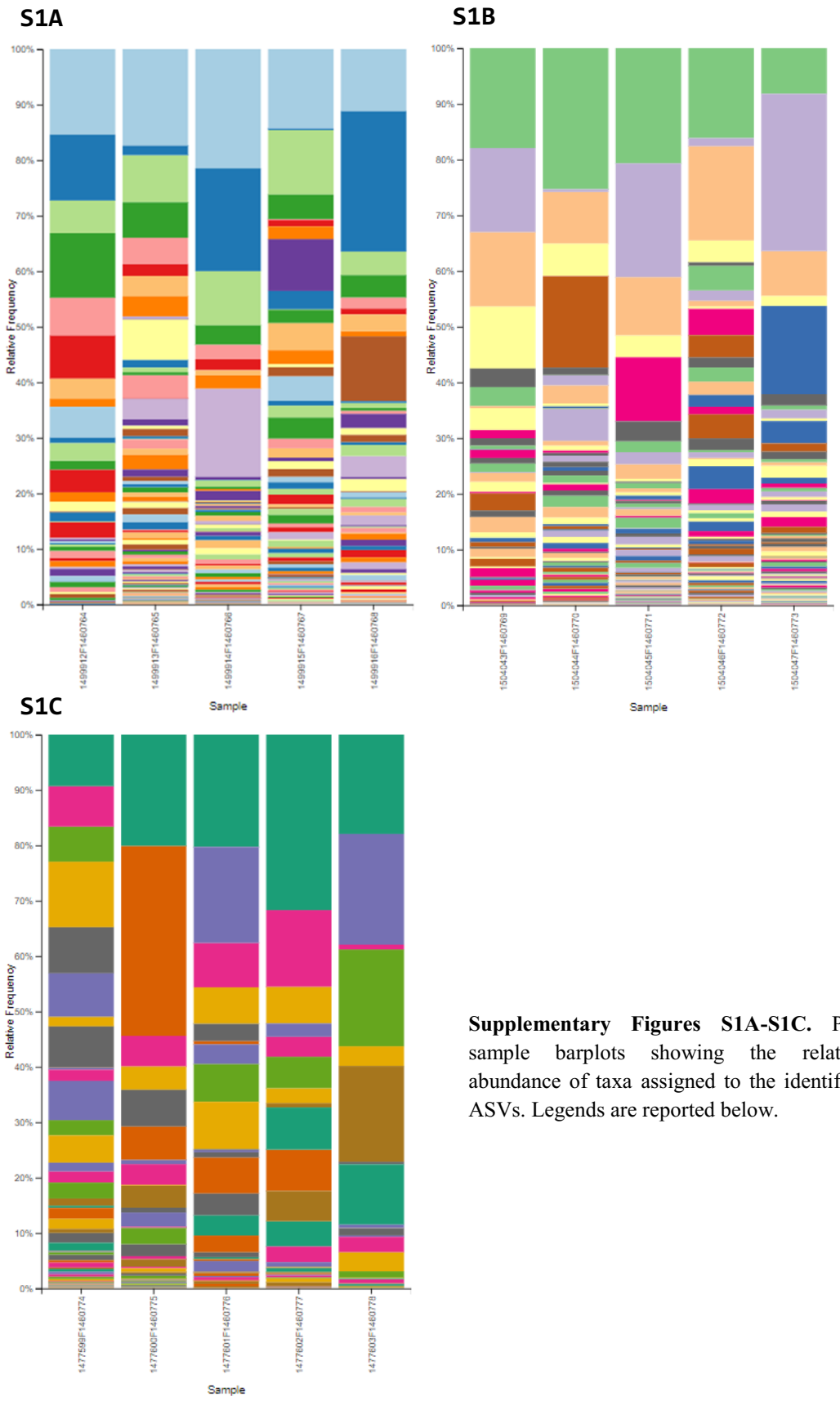
```
input_file_2=sys.argv[2]
```

```
lista_primer_pair_name=[]
```

```
feature_counts=open(input_file_2, "r")
for linea in feature_counts:
      if linea.startswith("#") or linea.startswith("Geneid"):
            continue
      campo=linea.split("\t")
      conta=[int(n) for n in campo[6:]]
      tot=sum(conta)
      if tot>0:
            taxon=campo[0]
            lista_primer_pair_name.append(taxon)
feature_counts.close()
```

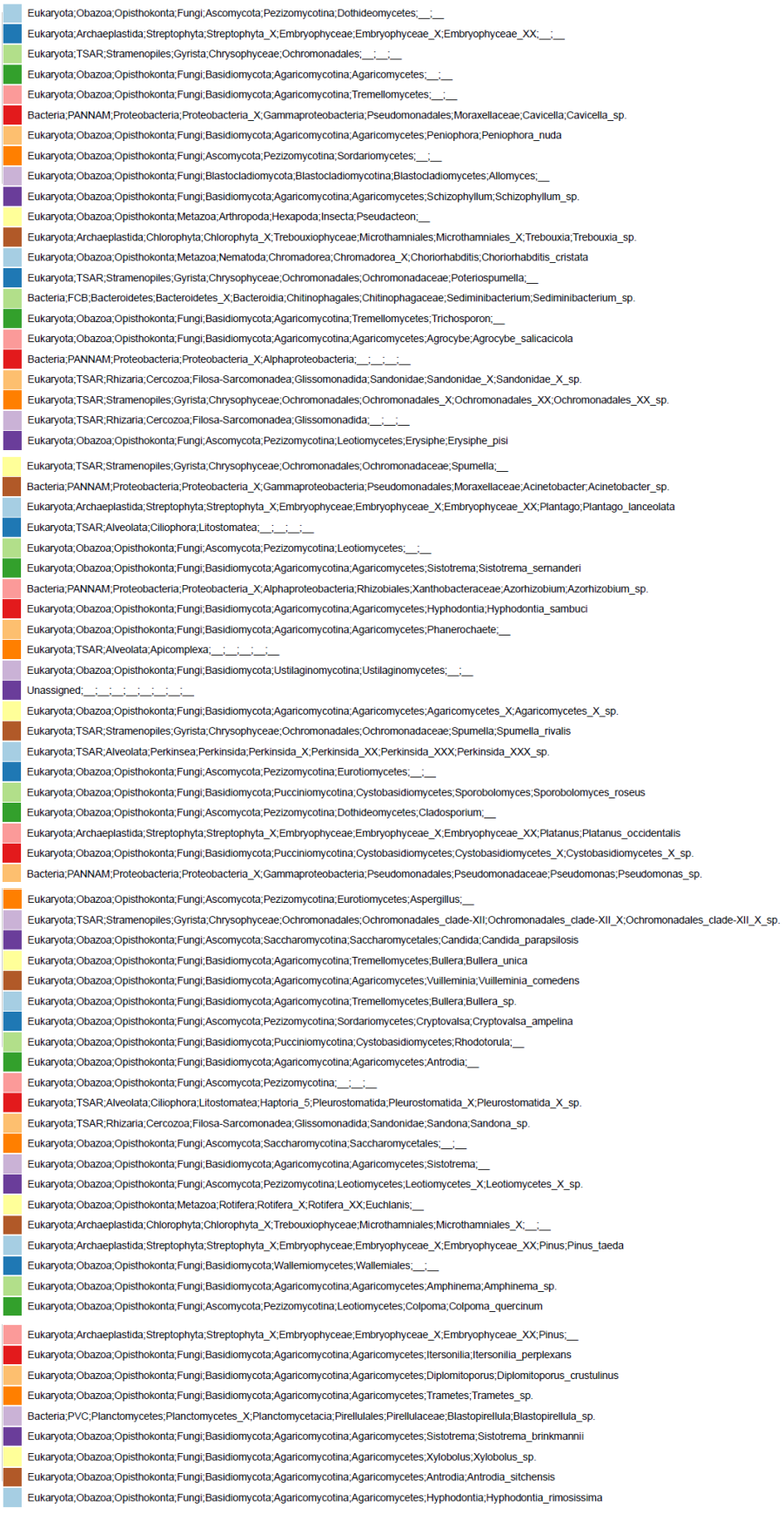
```
import matplotlib.pyplot as plt
from matplotlib_venn import venn2
fig=plt.figure(figsize=(8,8), dpi=150)
set1=set(list_primer_pair_name)
set2=set(lista_primer_pair_name)
diagram=venn2([set1,   set2],   ('QIIME   2',   'Bowtie   2   +
featureCounts'))
for text in diagram.set_labels:
      text.set_fontproperties(consolas_label)
for text in diagram.subset_labels:
      text.set_fontproperties(consolas_label)
plt.title("primer_pair_name", fontproperties=consolas_title)
fig.savefig("venn_result_primer_pair_name.png")
```

9.2. SUPPLEMENTARY FIGURES



Supplementary Figures S1A-S1C. Per-sample barplots showing the relative abundance of taxa assigned to the identified ASVs. Legends are reported below.

Legend Figure S1A




Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Ophiostoma;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Auricularia;__
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Crustacea;Maxillopoda;Cyclopidae;Cyclopidae_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Hymenostomatia;Ophryoglenida;Ichthyophthirius;Ichthyophthirius_multifiliis
Eukaryota;Obazoa;Opisthokonta;Metazoa;Annelida;Annelida_X;Annelida_XX;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Cryptococcus;Cryptococcus_carnescens
Bacteria;FCB;Bacteroidetes;Bacteroidetes_X;Bacteroidia;Chitinophagales;Chitinophagaceae;__;__
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Ulvophyceae;Ulotrichales;Ulotrichales_X;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Oligoporus;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Leotiomycetes;Blumeria;Blumeria_graminis
Eukaryota;Obazoa;Opisthokonta;Pluriformea;Pluriformea_X;Pluriformea_XX;Pluriformea_XXX;Syssomonas;Syssomonas_multiformis
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Saccharomycotina;Saccharomycetales;Wickerhamomyces;Wickerhamomyces_canadensis
Eukaryota;Amoebozoa;Tubulinea;Tubulinea_X;Elardia;Euamoebida;Vermamoebidae;__;__
Eukaryota;TSAR;Alveolata;Chrompodellids;Colpodellidea;Colpodellida;Colpodellaceae;Colpodellidae2;Colpodellidae2_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Leotiomycetes;Helotiales;Helotiales_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Trechispora;Trechispora_alnicola
Eukaryota;TSAR;Stramenopiles;Gyrista;Peronosporomycetes;Peronosporomycetes_X;Peronosporomycetes_XX;Plasmopara;Plasmopara_viticola
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Cystobasidiomycetes;__;__
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Chlorophyceae;Oedogoniales;Oedogoniales_X;Oedogonium;__
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;Paraphysomonadales;Paraphysomonadaceae;Paraphysomonas;Paraphysomonas_uniformis
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Sarocladium;Sarocladium_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Lentinellus;__
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Sorghum;Sorghum_bicolor
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Melanconis;__
Eukaryota;TSAR;Stramenopiles;Bigyra;Opalozoa;Opalozoa_X;MAST-12;MAST-12_X;MAST-12_X_sp.
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhizobiales;Xanthobacteraceae;Xanthobacteraceae_X;Xanthobacteraceae_X_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Cystobasidiomycetes;Rhodotorula;Rhodotorula_mucilaginosa
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Prunus;Prunus_persica
Eukaryota;TSAR;__;__;__;__;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Ophiognomonia;Ophiognomonia_clavigignenti-juglandacearum
Eukaryota;__;__;__;__;__;__;__;__;__
Eukaryota;TSAR;Stramenopiles;Gyrista;Mediophyceae;Thalassiosirales;Stephanodiscaceae;Stephanodiscus;__
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Scuticociliatia_2;Scuticociliatia_2_X;Protocyclidium;Protocyclidium_citrullus
Eukaryota;Obazoa;Opisthokonta;Fungi;Fungi_X;Fungi_XX;Fungi_XXX;Fungi_XXXX;Fungi_XXXX_sp.
Eukaryota;TSAR;Alveolata;__;__;__;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Rozellomycota;Rozellomycota_X;Rozellomycota_XX;Rozellomycota_XXX;Rozellomycota_XXX_sp.
Eukaryota;Cryptista;Cryptophyta;Cryptophyta_X;Cryptophyceae;Cryptomonadales;Cryptomonadales_X;Cryptomonas;Cryptomonas_curvata
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Coprinellus;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Botryobasidium;Botryobasidium_subcoronatum
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Microdochium;Microdochium_nivale
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Athelia;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Filobasidium;Filobasidium_uniguttulatum
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Agaricostilbomycetes;Bensingtonia;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Coprinopsis;Coprinopsis_atramentaria
Eukaryota;TSAR;Alveolata;Ciliophora;Colpodea;Colpodea_X;Colpodida;Colpoda;Colpoda_steinii
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Haptoria_6;Lacrymariidae;Lacrymaria;Lacrymaria_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Heterotrichea;Heterotrichea_X;Stentoridae;Stentor;Stentor_coeruleus
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Agaricus;Agaricus_campestris
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;Ochromonadales;Ochromonadales_clade-XI;Ochromonadales_clade-XI_X;Ochromonadales_clade-XI_X_sp.
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Zygnemophyceae;Zygnemophyceae_X;Zygnemophyceae_XX;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Cystobasidiomycetes;Rhodotorula;Rhodotorula_bacarum
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Mycena;__
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas;Sphingomonas_sp.
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Trebouxiophyceae;Microthamniales;Microthamniales_X;Trebouxia;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Cryptococcus;__
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Hymenostomatia;Tetrahymenida;Tetrahymenida_X;Tetrahymenida_X_sp.
Eukaryota;Obazoa;Opisthokonta;Metazoa;Porifera;Porifera_X;Demospongiae;Ephydatia;Ephydatia_fluviatilis
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Sistotrema;Sistotrema_oblongisporum
Eukaryota;Obazoa;Opisthokonta;Fungi;__;__;__;__;__;__
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Haptoria_5;Pleurostomatida;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Gymnopus;Gymnopus_luxurians
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Dioszegia;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Xylodon;Xylodon_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Hypotrichia;Oxytrichidae;Stylonychia;Stylonychia_lemnae
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Taphrinomycotina;Taphrinomycetes;Taphrinomycetes_X;Taphrinomycetes_X_sp.
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;Paraphysomonadales;Lepidochromonadaceae;Lepidochromonas;Lepidochromonas_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Oligohymenophorea_X;Oligohymenophorea_XX;Oligohymenophorea_XXX;Oligohymenophorea_XXX_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Peritrichia_2;Sessilida;Vorticella;Vorticella_campanula
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Oxyporus;__
Bacteria;CCD;Patescibacteria;Patescibacteria_X;Parcubacteria;Candidatus_Kaiserbacteria;Candidatus_Kaiserbacteria_X;Candidatus_Kaiserbacteria_XX;Candidatus_Kaiserbacteria_XX_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Phlebia;Phlebia_sp.
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;Ochromonadales;Ochromonadales_clade-XIII;Ochromonadales_clade-XIII_X;Ochromonadales_clade-XIII_X_sp.
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Gammaproteobacteria;__;__;__;__
Eukaryota;TSAR;Stramenopiles;Bigyra;Bicoecea;Pseudodendromonadales;Pseudodendromonadales_X;Pseudodendromonadales_XX;Pseudodendromonadales_XX_sp.
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Acidimicrobiia;Microtrichales;Microtrichaceae;Sva0996_marine_group;Sva0996_marine_group_sp.


Eukaryota;TSAR;Alveolata;Ciliophora;Heterotrichea;Heterotrichea_X;Stentoridae;Stentor;Stentor_muelleri
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Bjerkandera;__
Bacteria;CCD;Patescibacteria;Patescibacteria_X;Gracilibacteria;Absconditabacteriales_(SR1);Absconditabacteriales_(SR1)_X;Absconditabacteriales_(SR1)_XX;Absconditabacteriales_(SR1)_XX_sp.
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhizobiales;Xanthobacteraceae;Xanthobacter;Xanthobacter_sp.
Bacteria;PVC;Planctomycetes;Planctomycetes_X;Planctomycetacia;Pirellulales;Pirellulaceae;Pirellula;Pirellula_sp.
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;Frankiales;__;__;__
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Chlorophyceae;Sphaeropleales;Sphaeropleales_X;__;__
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Halodurantibacterium;Halodurantibacterium_sp.
Eukaryota;Obazoa;Opisthokonta;Metazoa;Rotifera;Rotifera_X;Rotifera_XX;__;__
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Gammaproteobacteria;Alteromonadales;Alteromonadaceae;Rheinheimera;Rheinheimera_sp.
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Trebouxiophyceae;Watanabea-Clade;Watanabea-Clade_X;Apatococcus;__
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium;Methylobacterium_sp.
Eukaryota;TSAR;Alveolata;Alveolata_X;Alveolata_XX;Alveolata_XXX;Alveolata_XXXX;Alveolata_XXXXX;Alveolata_XXXXX_sp.
Eukaryota;TSAR;Rhizaria;Cercozoa;Filosa-Thecofilosea;Filosa-Thecofilosea_X;Chlamydophryidae;Lecythium;Lecythium_hyalinum
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Tulasnella;Tulasnella_pruinosa
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Novosphingobium;Novosphingobium_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Hypotrichia;Holostichidae;Holosticha;Holosticha_diademata
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Euplotia;Euplotidae;Euplotes;Euplotes_daidaleos
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Ustilaginomycotina;Exobasidiomycetes;Malassezia;Malassezia_restricta
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Ustilaginomycotina;Exobasidiomycetes;Entyloma;Entyloma_ficariae
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Bosea;Bosea_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Minimedusa;Minimedusa_pubescens
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Chelicerata;Arachnida;Cheletomimus;Cheletomimus_wellsi
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;Micrococcales;Microbacteriaceae;__;__
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Oligohymenophorea_X;Urocentridae;Urocentrum;Urocentrum_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Hyphodontia;Hyphodontia_crustosa
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;Corynebacteriales;Dietziaceae;Dietzia;Dietzia_sp.
Eukaryota;TSAR;Stramenopiles;Gyrista;Bacillariophyceae;Cymbellales;Gomphonemataceae;Encyonema;__
Eukaryota;Obazoa;Opisthokonta;__;__;__;__;__;__;__
Eukaryota;TSAR;Alveolata;Ciliophora;Heterotrichea;Heterotrichea_X;Stentoridae;Stentor;Stentor_roeselii
Eukaryota:plas;TSAR:plas;Stramenopiles:plas;Gyrista:plas;__;__;__;__;__
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;Micrococcales;Micrococcaceae;Paenarthrobacter;Paenarthrobacter_sp.
Eukaryota:mito;Archaeplastida:mito;Streptophyta:mito;Streptophyta_X:mito;Embryophyceae:mito;Embryophyceae_X:mito;Embryophyceae_XX:mito;Dioon:mito;Dioon_edule:mito
Bacteria;FCB;Bacteroidetes;Bacteroidetes_X;Bacteroidia;Flavobacteriales;Weeksellaceae;Cloacibacterium;Cloacibacterium_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Dothideomycetes;Lophiostoma;Lophiostoma_fuckelii
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Microvirga;Microvirga_sp.
Eukaryota;Cryptista;Cryptophyta;Cryptophyta_X;Cryptophyceae;Cryptomonadales;Cryptomonadales_X;__;__
Eukaryota:plas;Archaeplastida:plas;Streptophyta:plas;Streptophyta_X:plas;Embryophyceae:plas;Embryophyceae_X:plas;Embryophyceae_XX:plas;Epifagus:plas;Epifagus_virginiana:plas
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;__;__;__
Eukaryota;Obazoa;Opisthokonta;Metazoa;Rotifera;Rotifera_X;Rotifera_XX;Notholca;Notholca_acuminata
Eukaryota;Excavata;Discoba;Euglenozoa;Kinetoplastea;Neobodonida;Neobodonidae;Neobodo;Neobodo_sp.
Eukaryota;Amoebozoa;__;__;__;__;__;__;__
Eukaryota;TSAR;Stramenopiles;Gyrista;Peronosporomycetes;Peronosporomycetes_X;Saprolegniales;Aphanomyces;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Monoblepharomycota;Sanchytriales;Sanchytriaceae;__;__
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Haptoria_5;Pleurostomatida;Hemiophrys;Hemiophrys_procera
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;Micrococcales;Microbacteriaceae;Agrococcus;Agrococcus_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Peritrichia_2;Sessilida;Vorticella;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Saccharomycotina;Stephanoascales;Trichomonascus;Trichomonascus_ciferrii
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;__;__;__;__
Eukaryota:plas;Cryptista:plas;Cryptophyta:plas;Cryptophyta_X:plas;Cryptophyceae:plas;Cryptomonadales:plas;Cryptomonadales_X:plas;Cryptomonas:plas;Cryptomonas_pyrenoidifera:plas
Bacteria;FCB;Bacteroidetes;Bacteroidetes_X;Bacteroidia;Chitinophagales;Chitinophagaceae;Lacibacter;Lacibacter_sp.
Bacteria;FCB;Bacteroidetes;Bacteroidetes_X;Bacteroidia;Sphingobacteriales;Sphingobacteriaceae;__;__
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Peniculia;Frontoniidae_2;Frontoniidae_2_X;Frontoniidae_2_X_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Ustilaginomycotina;Exobasidiomycetes;Tilletiopsis;__
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Crustacea;Maxillopoda;Eucyclops;__
Eukaryota;TSAR;Rhizaria;Cercozoa;Endomyxa;Vampyrellida;Leptophryidae;Leptophrys;Leptophrys_vorax
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Hexapoda;Insecta;__;__
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;Micrococcales;Dermabacteraceae;Brachybacterium;Brachybacterium_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Phyllopharyngea;Suctoria;PHYLL_2;PHYLL_2_X;PHYLL_2_X_sp.
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;Ochromonadales;Ochromonadaceae;Poteriospumella;Poteriospumella_lacustris

Legend Figure S1B


Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Sistotrema;Sistotrema_oblongisporum
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Hydnochaete;Hydnochaete_duportii
Eukaryota;Obazoa;Opisthokonta;Fungi;Fungi_X;Fungi_XX;Fungi_XXX;Fungi_XXXX;Fungi_XXXX_sp.
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Populus;Populus_trichocarpa
Eukaryota;Obazoa;Opisthokonta;__;__;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Mycena;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Sistotrema;Sistotrema_brinkmannii
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Hymenostomatia;Ophryoglenida;Ichthyophthirius;Ichthyophthirius_multifiliis
Eukaryota;TSAR;Stramenopiles;Gyrista;Peronosporomycetes;Peronosporomycetes_X;Peronosporomycetes_XX;Plasmopara;Plasmopara_viticola
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Crustacea;Maxillopoda;Eucyclops;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Ophiostoma;__
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Euplotia;Euplotidae;Euplotes;Euplotes_daidaleos
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Haptoria_5;Pleurostomatida;__;__
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Trebouxiophyceae;Microthamniales;Microthamniales_X;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Ustilaginomycotina;Exobasidiomycetes;Malassezia;Malassezia_restricta
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Sarocladium;Sarocladium_sp.
Eukaryota:nucl;Cryptista:nucl;Cryptophyta:nucl;Cryptophyta_X:nucl;Cryptophyceae:nucl;Cryptomonadales:nucl;Cryptomonadales_X:nucl;Cryptomonas:nucl;Cryptomonas_curvata:nucl
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Crustacea;Maxillopoda;Cyclopidae;Cyclopidae_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Peritrichia_2;Sessilida;Vorticella;Vorticella_campanula
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Hannaella;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Vuilleminia;Vuilleminia_comedens
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Bullera;Bullera_unica
Eukaryota;TSAR;Alveolata;Ciliophora;Phyllopharyngea;Suctoria;PHYLL_2;PHYLL_2_X;PHYLL_2_X_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Filobasidium;__
Eukaryota;Obazoa;Opisthokonta;Metazoa;Porifera;Porifera_X;Demospongiae;Ephydatia;Ephydatia_fluviatilis
Eukaryota;Amoebozoa;Discosea;Discosea_X;Centramoebia;Himatismenida;Parvamoebidae;Parvamoeba;Parvamoeba_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Hypotrichia;Holostichidae;Holosticha;Holosticha_diademata
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Ustilaginomycotina;Exobasidiomycetes;Malassezia;Malassezia_globosa
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Hymenostomatia;Tetrahymenida;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Bullera;Bullera_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Ustilaginomycotina;Exobasidiomycetes;Tilletiopsis;__
Eukaryota;TSAR;Alveolata;Alveolata_X;Alveolata_XX;Alveolata_XXX;Alveolata_XXXX;Alveolata_XXXXX;Alveolata_XXXXX_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Hypotrichia;Hypotrichia_X;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Cystobasidiomycetes;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Saccharomycotina;Saccharomycetales;Wickerhamomyces;Wickerhamomyces_canadensis
Eukaryota;Obazoa;Opisthokonta;Metazoa;Annelida;Annelida_X;Annelida_XX;__;__
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Hypotrichia;Oxytrichidae;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Ceratobasidium;Ceratobasidium_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Basidiomycota_X;Basidiomycota_XX;Basidiomycota_XXX;Basidiomycota_XXX_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Pucciniomycetes;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Antrodia;Antrodia_sitchensis
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;Ochromonadales;Ochromonadales_clade-XI;Ochromonadales_clade-XI_X;Ochromonadales_clade-XI_X_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Litostomatea_X;Litostomatea_XX;__;__
Eukaryota;Obazoa;Opisthokonta;Choanoflagellata;Choanoflagellatea;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Xylobolus;Xylobolus_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Cystobasidiomycetes;Rhodotorula;Rhodotorula_bacarum
Eukaryota;TSAR;Alveolata;Ciliophora;Colpodea;Colpodea_X;Bursariomorphida;Bursariomorphida_X;Bursariomorphida_X_sp.
Eukaryota;TSAR;Rhizaria;Cercozoa;Filosa-Sarcomonadea;Glissomonadida;__;__;__
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Trebouxiophyceae;Watanabea-Clade;Watanabea-Clade_X;Apatococcus;__
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Hexapoda;Insecta;Liposcelis;Liposcelis_brunnea
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Coprinopsis;Coprinopsis_atramentaria
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Eurotiomycetes;Penicillium;Penicillium_brevicompactum
Eukaryota;Amoebozoa;Tubulinea;Tubulinea_X;Elardia;Euamoebida;Vermamoebidae;Hartmannella;Hartmannella_vermiformis
Eukaryota;TSAR;Alveolata;Ciliophora;Phyllopharyngea;Suctoria;Ephelotidae;Ephelota;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Auricularia;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Ophiognomonia;Ophiognomonia_clavigignenti-juglandacearum
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Microdochium;Microdochium_nivale
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Hyphodontia;Hyphodontia_rimosissima
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Cunninghamia;Cunninghamia_lanceolata_var._konishii
Eukaryota;Obazoa;Opisthokonta;Pluriformea;Pluriformea_X;Pluriformea_XX;Pluriformea_XXX;Syssomonas;Syssomonas_multiformis
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Dioszegia;Dioszegia_crocea
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Cryptovalsa;Cryptovalsa_ampelina
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Wallemiomycetes;Wallemiales;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Athelia;Athelia_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Cystobasidiomycetes;Rhodotorula;Rhodotorula_mucilaginosa
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Amphinema;Amphinema_sp.
Eukaryota;Obazoa;Opisthokonta;Metazoa;Arthropoda;Chelicerata;Arachnida;Teutonia;__
Eukaryota;TSAR;Stramenopiles;Gyrista;Bacillariophyceae;Naviculales;Naviculaceae;Navicula;__
Eukaryota;Obazoa;Opisthokonta;Ichthyosporea;Ichthyophonida;Pseudoperkinsidae;Pseudoperkinsidae_X;Pseudoperkinsidae_XX;Pseudoperkinsidae_XX_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Haptoria_5;Pleurostomatida;Loxophyllum;__
Eukaryota;TSAR;Rhizaria;Cercozoa;Filosa-Thecofilosea;Filosa-Thecofilosea_X;Chlamydophryidae;Lecythium;Lecythium_hyalinum
Eukaryota;TSAR;Stramenopiles;Bigyra;Bicoecea;Pseudodendromonadales;Pseudodendromonadales_X;Pseudodendromonadales_XX;Pseudodendromonadales_XX_sp.
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Chlorophyceae;Sphaeropleales;Sphaeropleales_X;__;__
Eukaryota;Obazoa;Opisthokonta;Choanoflagellata;Choanoflagellatea;Craspedida;Monosigidae_Group_A;Codosiga;Codosiga_hollandica
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Scuticociliatia_1;Philasterida;Philasterida_X;Philasterida_X_sp.


Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Ulvophyceae;Cladophorales;Cladophorales_X;Cladophora;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Dothideomycetes;Alternaria;__
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Hymenostomatia;Glaucomidae;Glaucomidae_X;Glaucomidae_X_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Dothideomycetes;Lophiostoma;Lophiostoma_fuckelii
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Eurotiomycetes;Phaeomoniella;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Eurotiomycetes;Penicillium;__
Eukaryota;TSAR;Alveolata;Ciliophora;Phyllopharyngea;Suctoria;Evaginogenida;Evaginogenida_X;Evaginogenida_X_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Hypotrichia;__;__;__
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Haptoria_5;Pleurostomatida;Hemiophrys;Hemiophrys_macrostoma
Eukaryota;TSAR;Alveolata;Ciliophora;Phyllopharyngea;Cyrtophoria_2;Lynchellidae_1;Chlamydonella;Chlamydonella_irregularis
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Dothideomycetes;Torula;Torula_herbarum
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Poria;Poria_cocos
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Agaricostilbomycetes;Bensingtonia;__
Eukaryota;Cryptista;Cryptophyta;Cryptophyta_X;Cryptophyceae;Cryptomonadales;Cryptomonadales_X;Cryptomonas;__
Eukaryota;TSAR;Stramenopiles;Gyrista;Peronosporomycetes;Peronosporomycetes_X;Saprolegniales;__;__
Eukaryota;Obazoa;Opisthokonta;Metazoa;Rotifera;Rotifera_X;Rotifera_XX;Euchlanis;__
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Brassica;Brassica_oleracea
Eukaryota;Excavata;Discoba;Euglenozoa;Kinetoplastea;Kinetoplastea_X;Kinetoplastea_XX;Rhynchobodo;Rhynchobodo_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Litostomatea;Haptoria_6;Lacrymariidae;Lacrymaria;Lacrymaria_sp.
Eukaryota;TSAR;Alveolata;Ciliophora;Spirotrichea;Hypotrichia;Oxytrichidae;Stylonychia;Stylonychia_lemnae
Eukaryota;Obazoa;Opisthokonta;Metazoa;Rotifera;Rotifera_X;Rotifera_XX;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Leotiomycetes;Colpoma;Colpoma_quercinum
Eukaryota;TSAR;Alveolata;Chrompodellids;Colpodellidea;Colpodellida;Colpodellaceae;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Pezizomycotina_X;Microthyrium;Microthyrium_microscopicum
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Pinus;Pinus_taeda
Eukaryota;Excavata;Discoba;Discoba_X;Heterolobosea;Tetramitia_III;Vahlkampfiidae_III;Naegleria;Naegleria_sp.
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Microvirga;Microvirga_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Chytridiomycota;Rhizophydiales;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Oligoporus;__
Eukaryota;Obazoa;Opisthokonta;Metazoa;Cnidaria;Cnidaria_X;Hydrozoa;Craspedacusta;Craspedacusta_sowerbyi
Eukaryota;TSAR;Rhizaria;Cercozoa;Filosa-Sarcomonadea;Glissomonadida;Allapsidae;Viridiraptor;Viridiraptor_invadens
Eukaryota;TSAR;Alveolata;Ciliophora;Oligohymenophorea;Peritrichia_2;Sessilida;Epistylis;Epistylis_wuhanensis
Eukaryota;Obazoa;Opisthokonta;Choanoflagellata;Choanoflagellatea;Choanoflagellatea_X;Choanoflagellata_XX_Clade-2;Sphaeroeca;Sphaeroeca_leprechaunica
Eukaryota;TSAR;Alveolata;Ciliophora;Phyllopharyngea;Cyrtophoria_4;Chilodonellidae;Pseudochilodonopsis;Pseudochilodonopsis_sp.
Eukaryota;Excavata;Discoba;Euglenozoa;Kinetoplastea;Eubodonida;Bodonidae;Bodo;Bodo_uncinatus
Eukaryota;TSAR;Stramenopiles;Gyrista;Bacillariophyceae;Naviculales;Diadesmidaceae;Luticola;Luticola_permuticopsis
Eukaryota;Obazoa;Opisthokonta;Metazoa;__;__;__;__;__

Legend Figure S1C


Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Caulobacter;Caulobacter_sp.
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Lysobacter;Lysobacter_sp.
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Hydrogenophaga;Hydrogenophaga_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Fungi_X;Fungi_XX;Fungi_XXX;Fungi_XXXX;Fungi_XXXX_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Eurotiomycetes;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Leotiomycetes;__;__
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Trebouxiophyceae;Microthamniales;Microthamniales_X;Trebouxia;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Schizophyllum;__
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Pinus;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Sordariomycetes;Escovopsis;Escovopsis_sp.
Bacteria;FCB;Bacteroidetes;Bacteroidetes_X;Bacteroidia;Cytophagales;Spirosomaceae;Arcicella;Arcicella_sp.
Bacteria;PANNAM;Proteobacteria;Proteobacteria_X;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Bosea;Bosea_sp.
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Schizophyllum;Schizophyllum_commune
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Dothideomycetes;Alternaria;Alternaria_alternata
Bacteria;Terrabacteria;Actinobacteria;Actinobacteria_X;Actinobacteria_XX;Propionibacteriales;Nocardioidaceae;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Hydnum;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Eurotiomycetes;Aspergillus;__
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Trebouxiophyceae;Microthamniales;Microthamniales_X;__;__
Unassigned;__;__;__;__;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Agaricomycetes_X;Agaricomycetes_X_sp.
Eukaryota:mito;Obazoa:mito;Opisthokonta:mito;Fungi:mito;Ascomycota:mito;Pezizomycotina:mito;Sordariomycetes:mito;Ophiocordyceps:mito;Ophiocordyceps_prolifica:mito
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Botryobasidium;Botryobasidium_subcoronatum
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Filobasidium;__
Eukaryota;TSAR;__;__;__;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Hyphodontia;Hyphodontia_crustosa
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Mycena;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Dothideomycetes;Pyrenophora;Pyrenophora_teres
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Eurotiomycetes;Phaeomoniella;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Pezizomycotina;Eurotiomycetes;Penicillium;Penicillium_janthinellum
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Saccharomycotina;Saccharomycetales;Candida;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Agrocybe;__
Eukaryota;Obazoa;Opisthokonta;__;__;__;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Saccharomycotina;Saccharomycetales;Candida;Candida_sp.
Eukaryota;Archaeplastida;Streptophyta;Streptophyta_X;Embryophyceae;Embryophyceae_X;Embryophyceae_XX;Platanus;Platanus_occidentalis
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;Ochromonadales;Ochromonadaceae;__;__
Eukaryota;TSAR;Stramenopiles;Gyrista;Chrysophyceae;__;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Sistotrema;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Sistotrema;Sistotrema_biggsiae
Eukaryota;Obazoa;Opisthokonta;Fungi;Ascomycota;Saccharomycotina;__;__;__
Eukaryota;Cryptista;Cryptophyta;Cryptophyta_X;Cryptophyceae;Cryptomonadales;Cryptomonadales_X;Cryptomonas;Cryptomonas_parapyrenoidifera
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Agaricomycetes;Sistotrema;Sistotrema_resinicystidium
Eukaryota;Archaeplastida;Chlorophyta;Chlorophyta_X;Trebouxiophyceae;Microthamniales;Microthamniales_X;Trebouxia;Trebouxia_impressa
Eukaryota;Obazoa;Opisthokonta;Ichthyosporea;Ichthyophonida;Pseudoperkinsidae;Pseudoperkinsidae_X;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Agaricomycotina;Tremellomycetes;Tremellomycetes_X;Tremellomycetes_X_sp.
Eukaryota;TSAR;Stramenopiles;Gyrista;Peronosporomycetes;Peronosporomycetes_X;__;__;__
Eukaryota;Obazoa;Opisthokonta;Fungi;Basidiomycota;Pucciniomycotina;Cystobasidiomycetes;__;__

10. BIBLIOGRAPHY

AMCLI-CoSP, 2022. Percorso Diagnostico delle Parassitosi Intestinali (versione gennaio 2022). https://www.amcli.it/wp-content/uploads/2022/03/Percorso-diagnostico-delle-Parassitosi-Intestinali.pdf

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Adam, R.D., 2021. Giardia duodenalis: Biology and Pathogenesis. Clin Microbiol Rev 34, e0002419. https://doi.org/10.1128/CMR.00024-19

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., Huse, S.M., 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS One 4, e6372. https://doi.org/10.1371/journal.pone.0006372

Balzano, S., Abs, E., Leterme, S., 2015. Protist diversity along a salinity gradient in a coastal lagoon. Aquatic Microbial Ecology 74, 263–277. https://doi.org/10.3354/ame01740

Bass, D., Silberman, J., Brown, M., Tice, A., Jousset, A., Geisen, S., Hartikainen, H., 2016. Coprophilic amoebae and flagellates, including Guttulinopsis, Rosculus and Helkesimastix, characterise a divergent and diverse rhizarian radiation and contribute to a large diversity of faecal-associated protists. Environmental Microbiology 18. https://doi.org/10.1111/1462-2920.13235

Bates, S., Berg-Lyons, D., LAUBER, C., Walters, W., Knight, R., Fierer, N., 2012. A preliminary survey of lichen associated eukaryotes using Pyrosequencing. The Lichenologist 44. https://doi.org/10.1017/S0024282911000648

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Cheetham, R.K., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M.J., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M.D., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Catenazzi, M.C.E., Chang, S., Cooley, R.N., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fajardo, K.V.F., Furey, W.S., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Jones, T.A.H., Kang, G.-D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I.,

Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ng, B.L., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, Andrew C., Pike, Alger C., Pinkard, D.C., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Rodriguez, A.C., Roe, P.M., Rogers, John, Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Sohna, J.E.S., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., vandeVondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, Jane, Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J., 2008. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. Nature 456, 53–59. https://doi.org/10.1038/nature07517

Berrilli, F., Di Cave, D., Cavallero, S., D'Amelio, S., 2012. Interactions between parasites and microbial communities in the human gut. Front Cell Infect Microbiol 2, 141. https://doi.org/10.3389/fcimb.2012.00141

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A., Gregory Caporaso, J., 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6, 90. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible,

interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bradley, I.M., Pinto, A.J., Guest, J.S., 2016. Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. Appl Environ Microbiol 82, 5878–5891. https://doi.org/10.1128/AEM.01630-16

Bråte, J., Logares, R., Berney, C., Ree, D.K., Klaveness, D., Jakobsen, K.S., Shalchian-Tabrizi, K., 2010. Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. ISME J 4, 1144–1153. https://doi.org/10.1038/ismej.2010.39

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

Carnegie, R.B., Meyer, G.R., Blackbourn, J., Cochennec-Laureau, N., Berthe, F.C.J., Bower, S.M., 2003. Molecular detection of the oyster parasite Mikrocytos mackini, and a preliminary phylogenetic analysis. Dis Aquat Organ 54, 219–227. https://doi.org/10.3354/dao054219

Chabé, M., Lokmer, A., Ségurel, L., 2017. Gut Protozoa: Friends or Foes of the Human Gut Microbiota? Trends Parasitol 33, 925–934. https://doi.org/10.1016/j.pt.2017.08.005

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Choi, J., Park, J.S., 2020. Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform. Sci Rep 10, 6519. https://doi.org/10.1038/s41598-020-63561-z

Clemente, J.C., Ursell, L.K., Parfrey, L.W., Knight, R., 2012. The impact of the gut microbiota on human health: an integrative view. Cell 148, 1258–1270. https://doi.org/10.1016/j.cell.2012.01.035

Comeau, A.M., Li, W.K.W., Tremblay, J.-É., Carmack, E.C., Lovejoy, C., 2011. Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. PLoS One 6, e27492. https://doi.org/10.1371/journal.pone.0027492

Del Campo, J., Pons, M.J., Herranz, M., Wakeman, K.C., Del Valle, J., Vermeij, M.J.A., Leander, B.S., Keeling, P.J., 2019. Validation of a universal set of primers to study animal-associated microeukaryotic communities. Environ Microbiol 21, 3855–3861. https://doi.org/10.1111/1462-2920.14733

Diniz, W.J.S., Canduri, F., 2017. REVIEW-ARTICLE Bioinformatics: an overview and its applications. Genet Mol Res 16. https://doi.org/10.4238/gmr16019645

Dubik, M., Pilecki, B., Moeller, J.B., 2022. Commensal Intestinal Protozoa-Underestimated Members of the Gut Microbial Community. Biology (Basel) 11, 1742. https://doi.org/10.3390/biology11121742

Enberg, S., Majaneva, M., Autio, R., Blomster, J., Rintala, J.-M., 2018. Phases of microalgal succession in sea ice and the water column in the Baltic Sea from autumn to spring. Marine Ecology Progress Series 599, 19–34.

Fadeev, E., Salter, I., Schourup-Kristensen, V., Nöthig, E.-M., Metfies, K., Engel, A., Piontek, J., Boetius, A., Bienhold, C., 2018. Microbial Communities in the East and West Fram Strait During Sea Ice Melting Season. Frontiers in Marine Science 5, 429. https://doi.org/10.3389/fmars.2018.00429

Guérin, A., Striepen, B., 2020. The Biology of the Intestinal Intracellular Parasite Cryptosporidium. Cell Host & Microbe 28, 509–515. https://doi.org/10.1016/j.chom.2020.09.007

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana, R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R., Stoeck, T., Vaulot, D., Zimmermann, P., Christen, R., 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research 41, D597–D604. https://doi.org/10.1093/nar/gks1160

Hadziavdic, K., Lekang, K., Lanzen, A., Jonassen, I., Thompson, E.M., Troedsson, C., 2014. Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. PLoS One 9, e87624. https://doi.org/10.1371/journal.pone.0087624

Hamad, I., Raoult, D., Bittar, F., 2016. Repertory of eukaryotes (eukaryome) in the human gastrointestinal tract: taxonomy and detection methods. Parasite Immunol 38, 12–36. https://doi.org/10.1111/pim.12284

Haque, R., 2007. Human intestinal parasites. J Health Popul Nutr 25, 387–391.

Hu, T., Chitnis, N., Monos, D., Dinh, A., 2021. Next-generation sequencing technologies: An overview. Human Immunology, Next Generation Sequencing and its Application to Medical Laboratory Immunology 82, 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Hugerth, L.W., Muller, E.E.L., Hu, Y.O.O., Lebrun, L.A.M., Roume, H., Lundin, D., Wilmes, P., Andersson, A.F., 2014. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. PLoS One 9, e95567. https://doi.org/10.1371/journal.pone.0095567

Iebba, V., Totino, V., Gagliardi, A., Santangelo, F., Cacciotti, F., Trancassini, M., Mancini, C., Cicerone, C., Corazziari, E., Pantanella, F., Schippa, S., 2016. Eubiosis and dysbiosis: the two sides of the microbiota. New Microbiol 39, 1–12.

Jong, E., 2002. Intestinal parasites. Prim Care 29, 857–877. https://doi.org/10.1016/s0095-4543(02)00047-7

Kawanishi, M., Matsuda, T., Yagi, T., 2014. Genotoxicity of formaldehyde: molecular basis of DNA damage and mutation. Frontiers in Environmental Science 2.

Kim, E., Sprung, B., Duhamel, S., Filardi, C., Kyoon Shin, M., 2016. Oligotrophic lagoons of the South Pacific Ocean are home to a surprising number of novel eukaryotic microorganisms. Environ Microbiol 18, 4549–4563. https://doi.org/10.1111/1462-2920.13523

Kounosu, A., Murase, K., Yoshida, A., Maruyama, H., Kikuchi, T., 2019. Improved 18S and 28S rDNA primer sets for NGS-based parasite detection. Sci Rep 9, 15789. https://doi.org/10.1038/s41598-019-52422-z

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., Galand, P.E., 2019. Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. ISME J 13, 388–401. https://doi.org/10.1038/s41396-018-0281-z

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Lee, M.F., Lindo, J.F., Auer, H., Walochnik, J., 2019. Successful extraction and PCR amplification of Giardia DNA from formalin-fixed stool samples. Experimental Parasitology 198, 26–30. https://doi.org/10.1016/j.exppara.2019.01.010

Leung, J.M., Graham, A.L., Knowles, S.C.L., 2018. Parasite-Microbiota Interactions With the Vertebrate Gut: Synthesis Through an Ecological Lens. Front Microbiol 9, 843. https://doi.org/10.3389/fmicb.2018.00843

Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lokmer, A., Cian, A., Froment, A., Gantois, N., Viscogliosi, E., Chabé, M., Ségurel, L., 2019. Use of shotgun metagenomics for the identification of protozoa in the gut microbiota of healthy individuals from worldwide populations with various industrialization levels. PLoS One 14, e0211139. https://doi.org/10.1371/journal.pone.0211139

Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., Knight, R., 2012. Diversity, stability and resilience of the human gut microbiota. Nature 489, 220–230. https://doi.org/10.1038/nature11550

Lukeš, J., Stensvold, C.R., Jirků-Pomajbíková, K., Wegener Parfrey, L., 2015. Are Human Intestinal Eukaryotes Beneficial or Commensals? PLoS Pathog 11, e1005039. https://doi.org/10.1371/journal.ppat.1005039

Mangot, J.-F., Domaizon, I., Taib, N., Marouni, N., Duffaud, E., Bronner, G., Debroas, D., 2013. Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. Environmental Microbiology 15, 1745–1758. https://doi.org/10.1111/1462-2920.12065

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200

Mathison, B.A., Sapp, S.G.H., 2021. An annotated checklist of the eukaryotic parasites of humans, exclusive of fungi and algae. Zookeys 1069, 1–313. https://doi.org/10.3897/zookeys.1069.67403

Moreno, Y., Moreno-Mesonero, L., Amorós, I., Pérez, R., Morillo, J.A., Alonso, J.L., 2018. Multiple identification of most important waterborne protozoa in surface water used for irrigation purposes by 18S rRNA amplicon-based metagenomics. Int J Hyg Environ Health 221, 102–111. https://doi.org/10.1016/j.ijheh.2017.10.008

Nathan, N.N., Philpott, D.J., Girardin, S.E., 2021. The intestinal microbiota: from health to disease, and back. Microbes Infect 23, 104849. https://doi.org/10.1016/j.micinf.2021.104849

Parfrey, L.W., Walters, W.A., Knight, R., 2011. Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. Front Microbiol 2, 153. https://doi.org/10.3389/fmicb.2011.00153

Parfrey, L.W., Walters, W.A., Lauber, C.L., Clemente, J.C., Berg-Lyons, D., Teiling, C., Kodira, C., Mohiuddin, M., Brunelle, J., Driscoll, M., Fierer, N., Gilbert, J.A., Knight, R., 2014. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. Front Microbiol 5, 298. https://doi.org/10.3389/fmicb.2014.00298

Partida-Rodríguez, O., Serrano-Vázquez, A., Nieves-Ramírez, M.E., Moran, P., Rojas, L., Portillo, T., González, E., Hernández, E., Finlay, B.B., Ximenez, C., 2017. Human Intestinal Microbiota: Interaction Between Parasites and the Host Immune Response. Arch Med Res 48, 690–700. https://doi.org/10.1016/j.arcmed.2017.11.015

Piewngam, P., De Mets, F., Otto, M., 2020. Intestinal microbiota: The hidden gems in the gut? Asian Pac J Allergy Immunol 38, 215–224. https://doi.org/10.12932/AP-020720-0897

Piredda, R., Tomasino, M.P., D'Erchia, A.M., Manzari, C., Pesole, G., Montresor, M., Kooistra, W.H.C.F., Sarno, D., Zingone, A., 2017. Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. FEMS Microbiol Ecol 93, fiw200. https://doi.org/10.1093/femsec/fiw200

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41, D590–D596. https://doi.org/10.1093/nar/gks1219

Robeson, M.S., O'Rourke, D.R., Kaehler, B.D., Ziemski, M., Dillon, M.R., Foster, J.T., Bokulich, N.A., 2021. RESCRIPt: Reproducible sequence taxonomy reference database management. PLoS Comput Biol 17, e1009581. https://doi.org/10.1371/journal.pcbi.1009581

Robinson, C.J., Bohannan, B.J.M., Young, V.B., 2010. From structure to function: the ecology of host-associated microbial communities. Microbiol Mol Biol Rev 74, 453–476. https://doi.org/10.1128/MMBR.00014-10

Simon, M., Jardillier, L., Deschamps, P., Moreira, D., Restoux, G., Bertolino, P., López-García, P., 2015. Complex communities of small protists and unexpected occurrence of typical marine lineages in shallow freshwater systems. Environ Microbiol 17, 3610–3627. https://doi.org/10.1111/1462-2920.12591

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., Herndl, G.J., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc Natl Acad Sci U S A 103, 12115–12120. https://doi.org/10.1073/pnas.0605127103

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breiner, H.-W., Richards, T.A., 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19 Suppl 1, 21–31. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Taberlet, P., Bonin, A., Zinger, L., Coissac, É., 2018. Environmental DNA: For Biodiversity Research and Monitoring, Environmental DNA: For Biodiversity Research and Monitoring. https://doi.org/10.1093/oso/9780198767220.001.0001

van Lieshout, L., Verweij, J.J., 2010. Newer diagnostic approaches to intestinal protozoa. Curr Opin Infect Dis 23, 488–493. https://doi.org/10.1097/QCO.0b013e32833de0eb

Vaulot, D., Geisen, S., Mahé, F., Bass, D., 2022. pr2-primers: An 18S rRNA primer database for protists. Mol Ecol Resour 22, 168–179. https://doi.org/10.1111/1755-0998.13465

Vitošević, K., Todorović, M., Varljen, T., Slović, Ž., Matić, S., Todorović, D., 2018. Effect of formalin fixation on pcr amplification of DNA isolated from healthy autopsy tissues. Acta Histochem 120, 780–788. https://doi.org/10.1016/j.acthis.2018.09.005

Xu, Z., Li, Yifan, Lu, Y., Li, Yingdong, Yuan, Z., Dai, M., Liu, H., 2020. Impacts of the Zhe-Min Coastal Current on the biogeographic pattern of microbial eukaryotic communities. Progress in Oceanography 183, 102309. https://doi.org/10.1016/j.pocean.2020.102309

Zhan, A., Hulak, M., Sylvester, F., Huang, X., Adebayo, A., Abbott, C., Adamowicz, S., Heath, D., Cristescu, M., MacIsaac, H., 2013. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. Methods in Ecology and Evolution 4, 558–565. https://doi.org/10.1111/2041-210X.12037