

University of South Bohemia

Faculty of Science

**Bioinformatic determination of the contamination in
environmental DNA from water**

Bachelor thesis

Max Höller

Supervisor: RNDr. Petr Blabolil, Ph.D.

České Budějovice 2022

Bibliographical Detail

Höller M., 2022: Bioinformatic determination of the contamination in environmental DNA from water. Bc. Thesis, in English. – 53 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

Annotation

The aim of this study is to determine the bycatch in the samples collected and processed at 10 different sampling sites (Klicava, Rimov, Zlutice, Laz, Šumava, Pravíkov, Kladiny, Kalich, Tuska, Písek) in the summer and autumn/winter seasons. The results were analysed by comparing the number of sequences assigned to fish, non-fish vertebrates, humans, and bacteria. Results are discussed in the context of published information in biological, ecological, and bioinformatics articles.

Declaration

I declare that I am the author of this qualification thesis and that in writing it I have used the sources and literature displayed in the list of used sources only.

In České Budějovice on 06.12.2022

Abstract

When using environmental DNA (eDNA) for biomonitoring of specific taxa, bycatch often makes up a significant portion of the analysed data. If only the targeted sequences are used, much sequencing depth, time, and money is lost. The goal of this study is to determine the amount of bycatch and possible contamination, and to monitor their origin. The study was conducted in ten different waters in the Czech Republic, sampled in summer and autumn/winter between 2018 and 2022. It also examined the amount and structure of bycatch obtained from eDNA metabarcoding of water bodies using Tapirs bioinformatics pipeline and custom Python codes. Different fish taxa, other vertebrate taxa, bacterial taxa, and humans were distinguished. Based on the analysed data, graphs were created to show the distribution of the different taxa in each sample of a sampling campaign. The obtained results have important implications for applying this method to noninvasive biomonitoring initiatives.

Acknowledgements

I would like to express my deepest gratitude to my family, Chiara, R. dos Santos and all my friends who always stood by me and gave me advice if I had setbacks. Further I want to thank my supervisor RNDr. Petr Blabolil, Ph.D. for the help on how to properly write a scientific study. The work was supported by the projects MSM200961901 "The true picture of eDNA", and by CAS within the program of the Strategy AV 21 "Land conservation and restoration".

Contents

1. Introduction	1
2. Work Aims	2
3. Material and Methods	3
3.1. Study sites	3
3.2. eDNA sampling & extraction	4
3.3. eDNA metabarcoding	4
3.4. Bioinformatics analysis	4
3.5. Further Data analysis	5
4. Results	6
4.1. Klíčava	6
4.2. Rimov	11
4.3. Zlutice	16
4.5. Sumava	22
4.6. Pravikov	24
4.7. Kalich	27
4.8. Kladiny	29
4.9. Tuska	32
4.10. Pisek	33
5. Discussion	34
5.1. Comparison between different water bodies and seasons	34
5.2. Highest and lowest diversity in defined taxa	35
5.3. Possible contamination in control samples	36
5.4 Bioinformatics	36
6. Conclusion	37
References	38
List of tables	42
List of figures	42
List of Appendices	44

1. Introduction

In the last two decades, a continuous decline in global biodiversity has been observed (Butchart, et al., 2010). Consequently, increasing efforts are made by organisations like the World Wildlife Fund (WWF), Fauna & Flora International (FFI), or United Nation Environment Program (UNEP) in taxa, populations and communities preservation, since the stability of our ecosystem increases with the number of present taxa (Jing & Taiyi, 2015). Precise information about biota is essential for proper management (Angermeier & Karr, 1994). Freshwater ecosystems are amongst the most vulnerable ones on Earth. According to recent estimates, taxa populations in freshwater ecosystems are very sensitive to anthropogenic pressures (Vilmi, et al., 2017). Traditional methods for fish sampling are often expensive, laborious, invasive, and sometimes even destructive and selective, which can lead to imprecise data (Kubečka, et al., 2009).

A modern way to obtain data about biological elements is based on the analysis of environmental DNA (eDNA). The principle is the detection of an organismal footprint. The samples of environmental DNA can be collected from air, soil, water etc. and it is fully non-invasive (Taberlet et al., 2012). This method was first used for microorganisms. Later on, vertebrates were included as well (Díaz-Ferguson & Moyer, 2014). Nowadays eDNA is used in many different fields, such as forensics, quality of food products or faecal pollution (Steward, 2019). The central part of the used method is Polymerase Chain Reaction (PCR) amplification with primers targeting specific DNA fragments. There are two types of possible analysis: Either the specification to a single taxa (barcoding) or to a whole taxonomic group (metabarcoding). The eDNA metabarcoding has a tremendous potential in biomonitoring as it is capable of describing communities and biodiversity from just a very small part of an environment such as a drop of water (Ficetola, et al., 2014). This is one of the reasons why there is a shift from looking at individual taxa to looking at whole communities (Cristescu & Hebert, 2018).

Short and standardised genetic markers, called DNA metabarcodes, are used to determine the specific taxonomy of an isolated specimen. They are specialised for specific organisms, so markers with different properties must be considered when analysing a wide range of taxa. Because these markers are generally short, they provide hypervariability between taxa (Riaz et al., 2011). The reason why so many different biological groups can be detected using these sequences is that the sequences are present in mitochondria, which were originally bacteria and have been incorporated into the cells of most eukaryotic organisms. Despite this hypervariability, not all taxa can be correctly separated. This is because the performance of PCR varies by taxa (Taberlet, Bonin, Zinger, & Coissac, 2018). In addition, because PCR and sequencing are artificial processes, errors can always occur. One example of improving the performance of sequencing is to add the artificial control sequence PhiX to monitor quality (Illumina, Inc., CA, USA).

In this work a bioinformatic pipeline was used on sequencing data collected from 2018 to 2022 from different water bodies in the Czech Republic. Further processing was done with custom python and R codes.

2. Work Aims

The aim of this Bachelor Thesis was to differentiate the proportion of the target fish sequences from potential laboratory contamination (human DNA) and the bycatch ('other vertebrate' and bacterial DNA). The ratio and diversity of the groups detected in different campaigns are important for stakeholders to efficiently utilise the method for biodiversity monitoring.

3. Material and Methods

3.1. Study sites

The water samples were collected at ten different water bodies in the Czech Republic in summer, autumn, and/or winter periods in the years 2018 to 2022. Most of the samples came from the drinking water reservoirs Klicava, Rimov and Zlutice where precise data about the biota are crucial for high water quality management. To verify the consistency of eDNA metabarcoding detections, fully harvestable fishponds were included in the study systems. Finally, a few more water bodies with different trophic statuses and biota were included to test the eDNA workflow of the main types of Czech water bodies. (Table 1)

Table 1 Basic characteristics and sampling effort in studied water bodies.								
Water body	Name	District	Surface area (ha)	maximum depth (m)	Trophic status	Sampling year	Sampling season	Number of samples*
Reservoir	Kličava	Rakovník	62	31	oligotrophic	2018-2020	summer, winter, autumn	29/30 // 30/30 // 0/30
Reservoir	Řimov	České Budějovice	206	41	meso-eutrophic	2018-2020	summer, winter, autumn	38/35 // 38/35 // 0/38
Reservoir	Žlutice	Karlovy Vary	167	22	eutrophic	2018-2020	summer, winter, autumn	28/29 // 29/29 // 0/29
Reservoir	Láz	Příbram	17	13	oligotrophic	2020	autumn	4
Lake, brooks	Laka lake, Jezerní brook	Klatovy	2,5	3,8	oligotrophic	2020-2021/22	autumn	19 // 27
Pond	Horní Pravíkov	Pelhřimov	2,6	2,0	eutrophic	2019, 2020	summer, autumn	55/56 // 0/60
Pond	Kalich	Pelhřimov	9,5	4,1	eutrophic	2019	summer, autumn	55/55
Pond	Dolní Kladiny	Pelhřimov	29	3,3	eutrophic	2019, 2020	summer, autumn	55/64 // 0/60
Pond	Tuska	Pelhřimov	10	3,9	eutrophic	2020	autumn	60
Pond	Lusk, Sirotčí	Písek	0.7	1,8	eutrophic	2020	summer, autumn	19
*The numbers are listed as summer/autumn(winter) 1 st year // summer/autumn(winter) 2 nd year // ...								

3.2. eDNA sampling & extraction

In each waterbody, the number of samples was dependent on the primary research project (Blabolil, et al., 2021) (Blabolil, et al., 2022), reflecting the size and depth of the waterbody. Each sample was represented by 2L of collected water, collected by pooling five 400mL subsamples. Field blanks (0.5L purified water per sampling campaign) were taken in the field and handled as regular samples.

In a decontaminated laboratory, that is not used for handling fish specimens or DNA, all samples were filtered within 24 hours of collection. One litre of each sample (2x 0.5L) was filtered through two sterile mixed cellulose acetate and cellulose nitrate filters (47mm diameter, 0.45µm porosity) using Nalgene filtration units in combination with a vacuum pump. In 2018, the filters were folded using sterile tweezers, placed in 1.5mL Eppendorf tubes filled with 96% ethanol, and stored in the dark at 4°C. In following years, filters were frozen at -20°C. Field blanks also served as filtration blanks that were run during filtration to monitor potential contamination.

The filters in ethanol were removed onto individual sterile Petri dishes to dry for approximately 30 min. Two sterile filters washed in 99% ethanol were processed alongside the sample filters to act as evaporation blanks and detect possible contamination for each of the sampling campaigns. All dried filters were then stored at -20°C for one week until eDNA extraction was performed.

The eDNA was extracted following the Mu-DNA water protocol, which consists of the five steps: lysis, inhibitor removal, silica binding, wash and elution (Sellers, Di Muri, Gómez, & Hänfling, 2018). To minimise the risk of contamination between water bodies for each sampling campaign, the eDNA extraction was performed separately. For each sampling campaign the DNA was extracted from all filters, field and filtration blanks, evaporation blanks and extraction blanks (i.e. extraction buffers only). An extraction blank was included for each sampling campaign. Until PCR amplification was performed, the extracted DNA was stored at -20°C for one week.

2.3. eDNA metabarcoding

The primer pair 12S-V5-F (5'-ACTGGGATTAGATACCCC-3') and 12S-V5-R (5'-TAGAACAGGCTCCTCTAG-3'), which amplify a 73-110 bp fragment of the mitochondrial 12S ribosomal RNA (rRNA) gene in vertebrates (Riaz, Shehzad, Viari, Pompanon, & Taberlet, 2011), and a ~106 bp fragment in european fishes (Hänfling, et al., 2016), was used to produce the PCR amplicons from which the sequencing library was generated. In order to detect possible false positives and false negatives, PCR negative controls (molecular grade water) and PCR positive controls were included within each sub-library. The latter used tissue DNA (0.05ng/µL) from *Maylandia Zebra* (Boulenger, 1899), a cichlid from Lake Malawi, which is not naturally present in waters of the Czech Republic. Two step PCR was conducted to prepare genetic libraries for sequencing. The libraries were run at 13pM with 10% PhiX Control v3 on an Illumina MiSeq® using a MiSeq Reagent Kit v3 (600-cycle) (Illumina, Inc., CA, USA).

3.4. Bioinformatics analysis

The raw sequence reads resulting from Illumina MiSeq were demultiplexed by a custom python script. Further analysis was done by using a bioinformatics pipeline called Tapirs (Tapirs, 2020/2022). While making use of the Snakemake workflow manager, Tapirs is separated into various modules. Each module performs a particular step of the total workflow. The three main steps of Tapirs workflow are “Quality Control”, ”Taxonomic Assignment”, and “Reports and Reproducibility”.

In order to get appropriate results from the sampled data, a custom blast database was created and added to the Tapirs configuration file instead of the standard blast database. The method of analysis was set to only “blast” and the taxonomic ranks were taxa as lowest and order as highest rank. The Fastp trimming and merging was done with a set maximum amplicon length (FASTP_max_len1,

FASTP_max_len2) of 106, the forward and reverse primer lengths (FASTP_trim_font1, FASTP_trim_font2) were set to 20. The quality value (FASTP_qual_phred) was set to 30 and the maximum percentage of unqualified percentage (FASTP_unqualified_percent_limit) to 40. Further the minimum 3' poly g and poly x (x = any base) bases to be trimmed (FASTP_goly_g_min, FASTP_goly_x_min) were set to 10. The minimum overlap required for merging (FASTP_min_overlap) was 20, the maximum percentage of difference (FASTP_diff_percent_limit) as well as the maximum base differences (FASTP_diff_limit) required for merging got set to 5, the maximum allowed length which is required to keep the read after merging (FASTP_length_limit) was 106. The blast variable BLAST_max_target_seqs was set to 50, BLAST_min_perc_ident to 75, the minimal E-Value (BLAST_min_evalue) was set to 1e-20 and the number of BLAST_threads to 2.

3.5. Further Data analysis

After performing the analysis with Tapirs, the results were used for further processing. Since this study includes not just vertebrates, but also bacteria, which the used Tapirs database did not target, the unassigned sequences were extracted with a python code and aligned against the 16S_ribosomal_RNA database of National Center for Biotechnology Information (National Library of Medicine, 2022). The resulting files were then reworked to get uniform data and to be merged with the Tapirs results. Further the total number of reads got recalculated. A 0.1% threshold was applied to the data to further minimise the number of false positives. The outcoming data was used to differentiate between fish, other vertebrates and bacteria. This was done using a custom python program which uses the analysed data to create four graphs each showing the numbers of different taxa of fish, human, other vertebrates and bacteria for each sample of a certain probe. It is important to mention that 'Homo sapiens' was treated separately from 'other vertebrates'.

4. Results

The results show the numbers of reads of detected taxa in different samples within a sampling campaign. They provide information about the number of fish, other vertebrates, humans and bacteria. In the following graphs one can observe the regular samples followed by controls dry blank (D), extraction blank (E), field blank (F), PCR negative (N) and PCR positive control (P). Each figure displays a sampling campaign and contains four graphs showing the number of reads assigned to either fishes, humans, other vertebrates or bacteria.

4.1. Klíčava

The Regular samples of Klicava-2018-summer include 15 different taxa of fish, 13 different taxa of other vertebrates, one kind of bacteria, and humans. Regular sample '07' shows the highest amount of total reads (305,154), with the vast majority being fish (~220,000). The sample that contains the smallest proportion of fish compared to other vertebrates and humans is '01', in which over 100,000 reads of '*Branta bernicla*' were assigned, while just ~25,000 reads got assigned to fish. Just very few bacteria DNA reads (51) were detected in sample '06'. The fish taxon, which is represented the most in the regular samples is '*Perca fluviatilis*'. The positive control sample shows only '*Maylandia zebra*' (fish), while negative control, extraction blank, and field blank only contain human reads. The total mean for human is nearly 20,000 reads per sample, while other vertebrates only reach ~10,000 reads. (Figure 1)

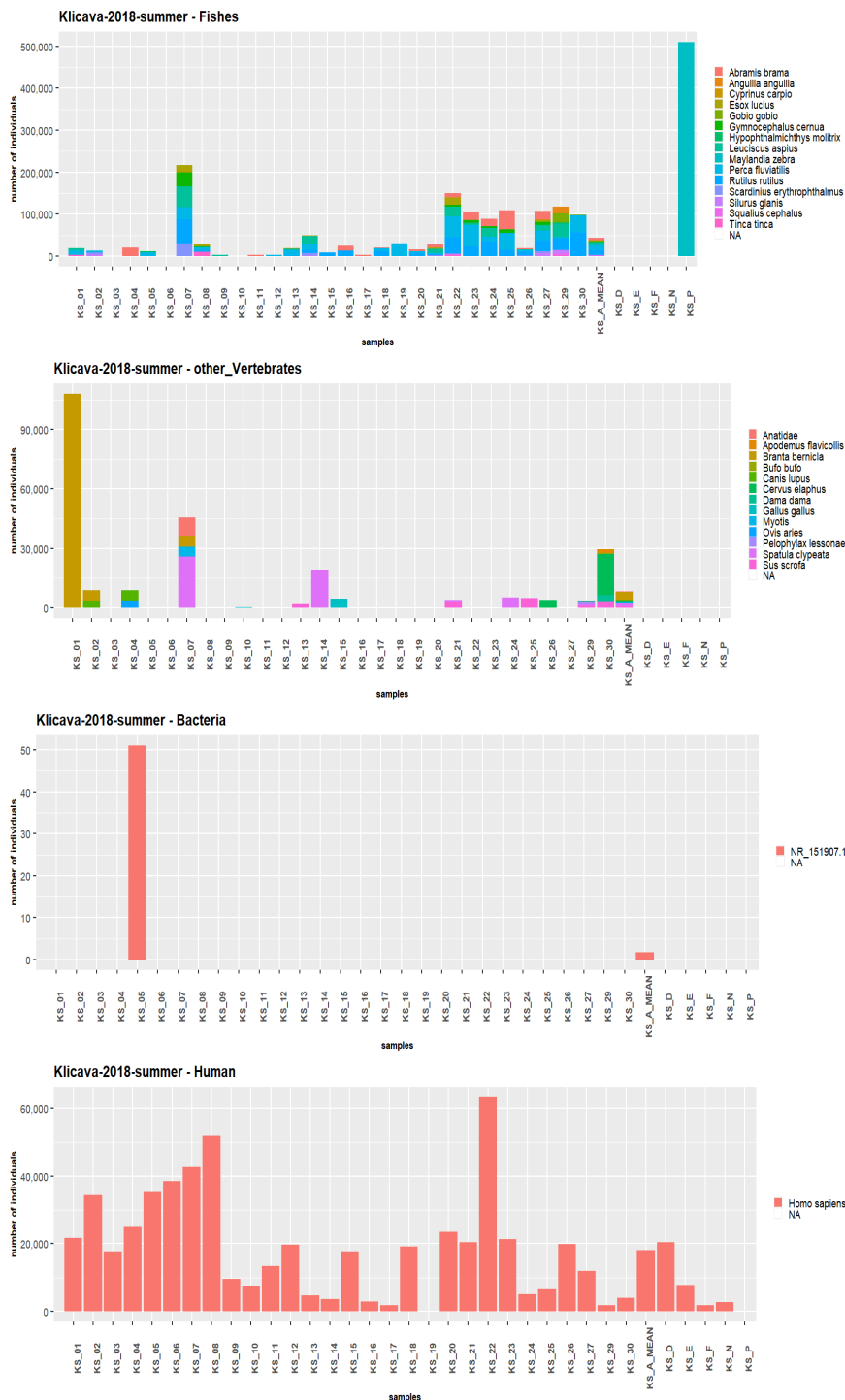
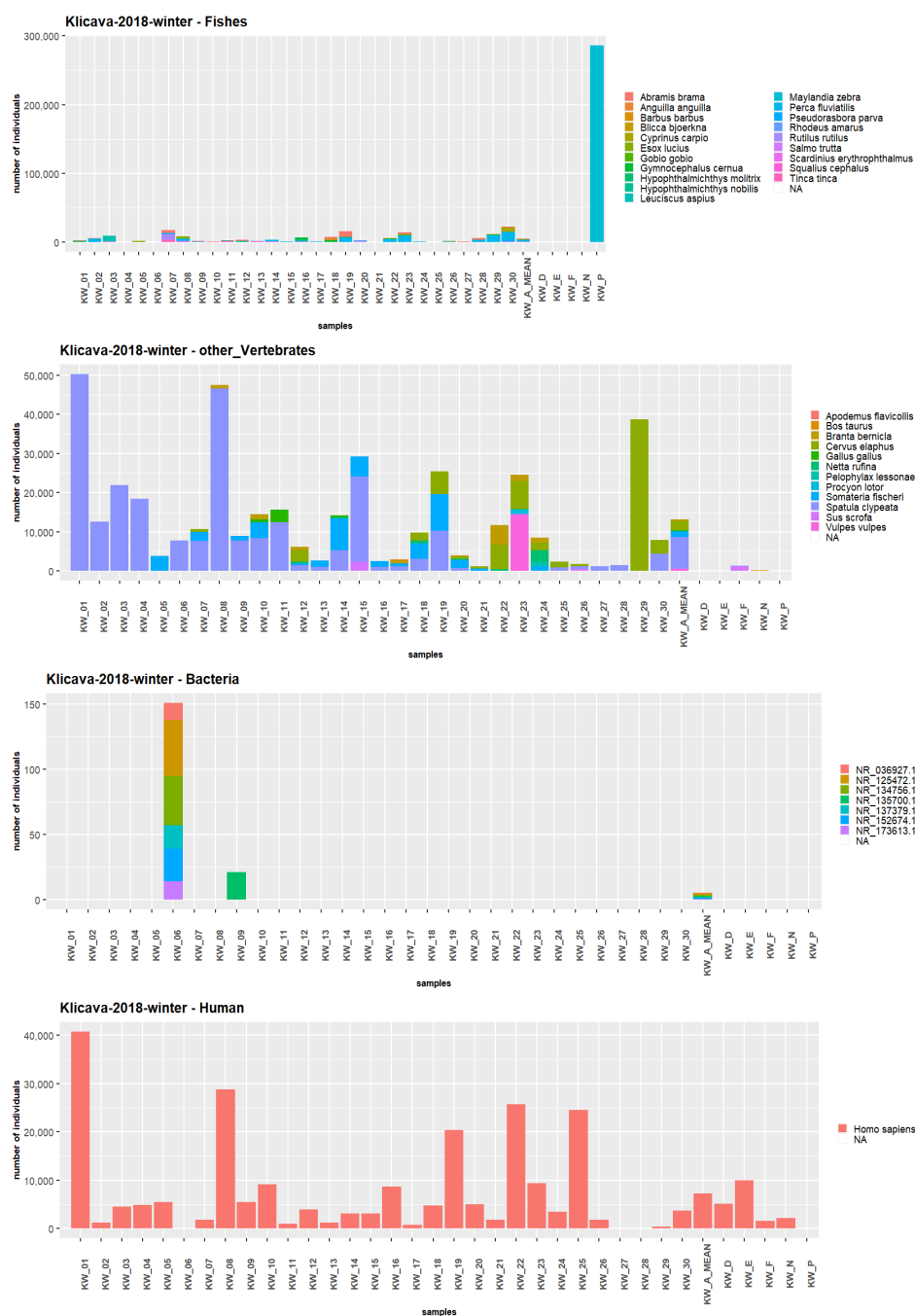


Figure 1 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2018-summer sampling event.



The regular samples of Klicava-2018-winter show a low number of fish throughout the whole sampling campaign, even though there is a high variety of taxa (20). In particular, as in a large number of samples, more ‘other vertebrates’ taxa than fish were detected. The biggest difference can be observed in sample ‘01’, in which 50,287 other vertebrates, 40,733 human and not even 3,000 fish reads were detected. Both, the highest amount of different bacterial taxa and highest total number of bacteria is present in sample ‘06’. The vast majority of reads assigned to ‘*Spatula clypeata*’. A few samples show a relatively high amount of human reads (01, 08, 22, 25). (Figure 2)

Figure 2 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2018-winter sampling event.



Figure 3 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2019-summer sampling event.

Across the regular samples of Klicava-2019-summer' one can observe a high number of reads assigned to fish, especially to '*Rutilus rutilus*', of which over 600,000 reads were detected. The number of other vertebrates on the other hand is small throughout most regular samples with an absolute high of ~37,000 reads in sample '5ZL'. The bacteria chart shows a small bacterial contamination in a few samples mostly by '*Haloferrula helveola*' (NR_041673.1). Sample '17' shows a majority of human DNA. '41' consists of human DNA to 50% as well. (Figure 3)



Figure 4 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2019-autumn sampling event.

Most of the samples in the Klicava-2019-autumn sampling campaign show low numbers of fish reads. The most fish are found in samples '51', '5KL', and '5ZT', in which between ~68,000 and ~120,000 reads were assigned. The only two samples that contain a high amount of 'other vertebrates' are '1P' and '5KL' with over 80,000 assigned reads each. In general only few reads were assigned in this sampling campaign's regular samples. There is a mean of ~600 reads assigned to human DNA in regular samples. (Figure 4)

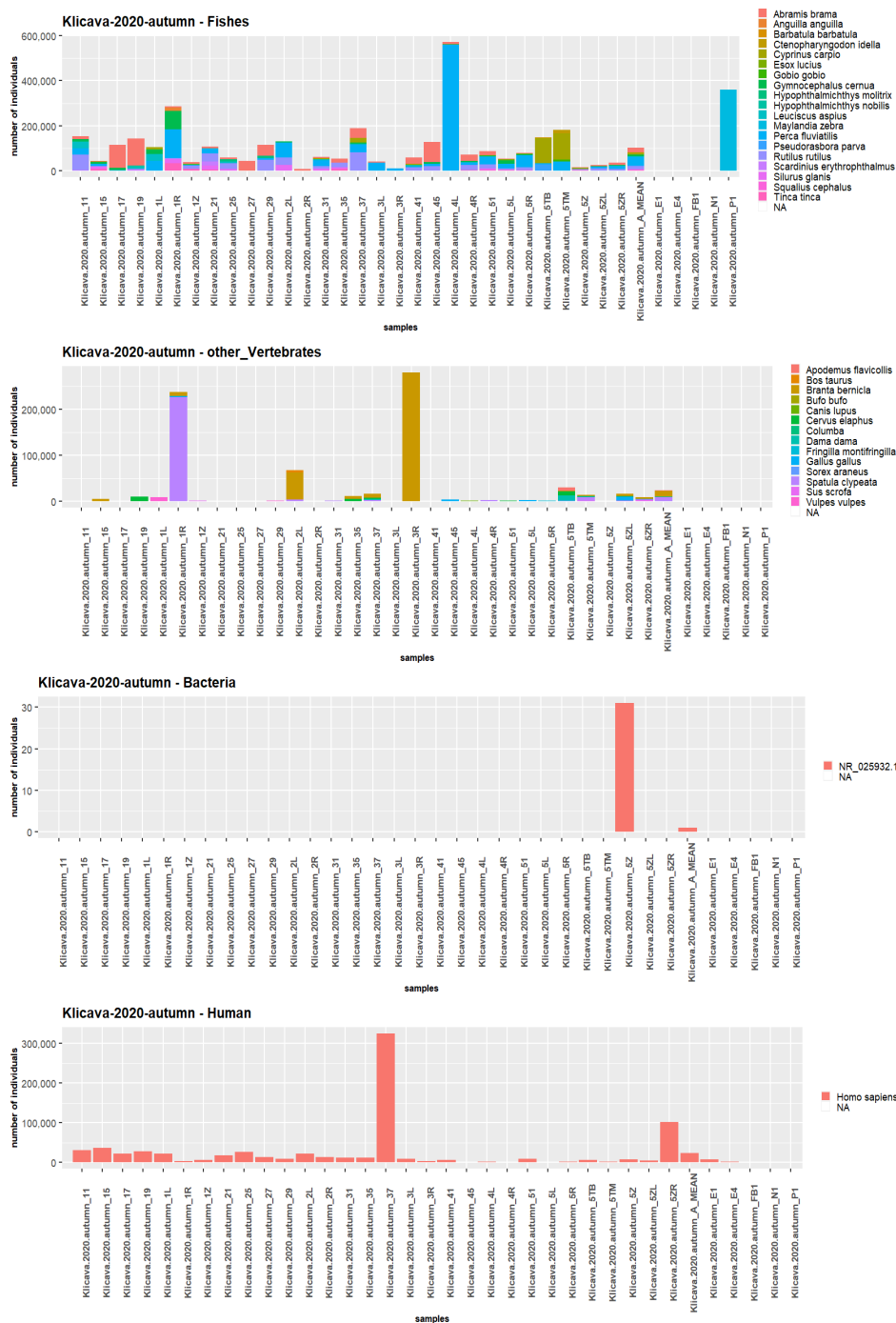
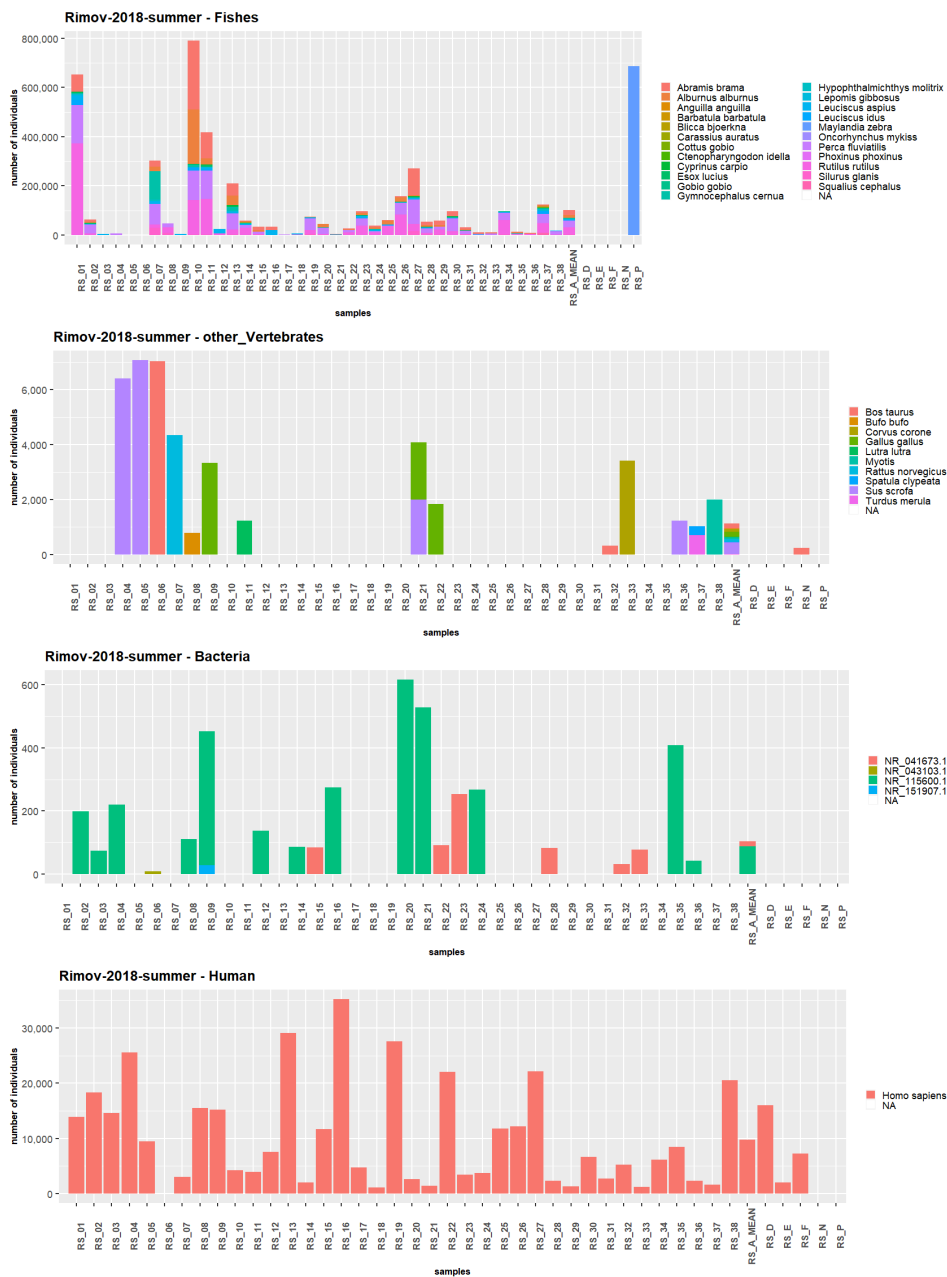


Figure 5 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2020-autumn sampling event.

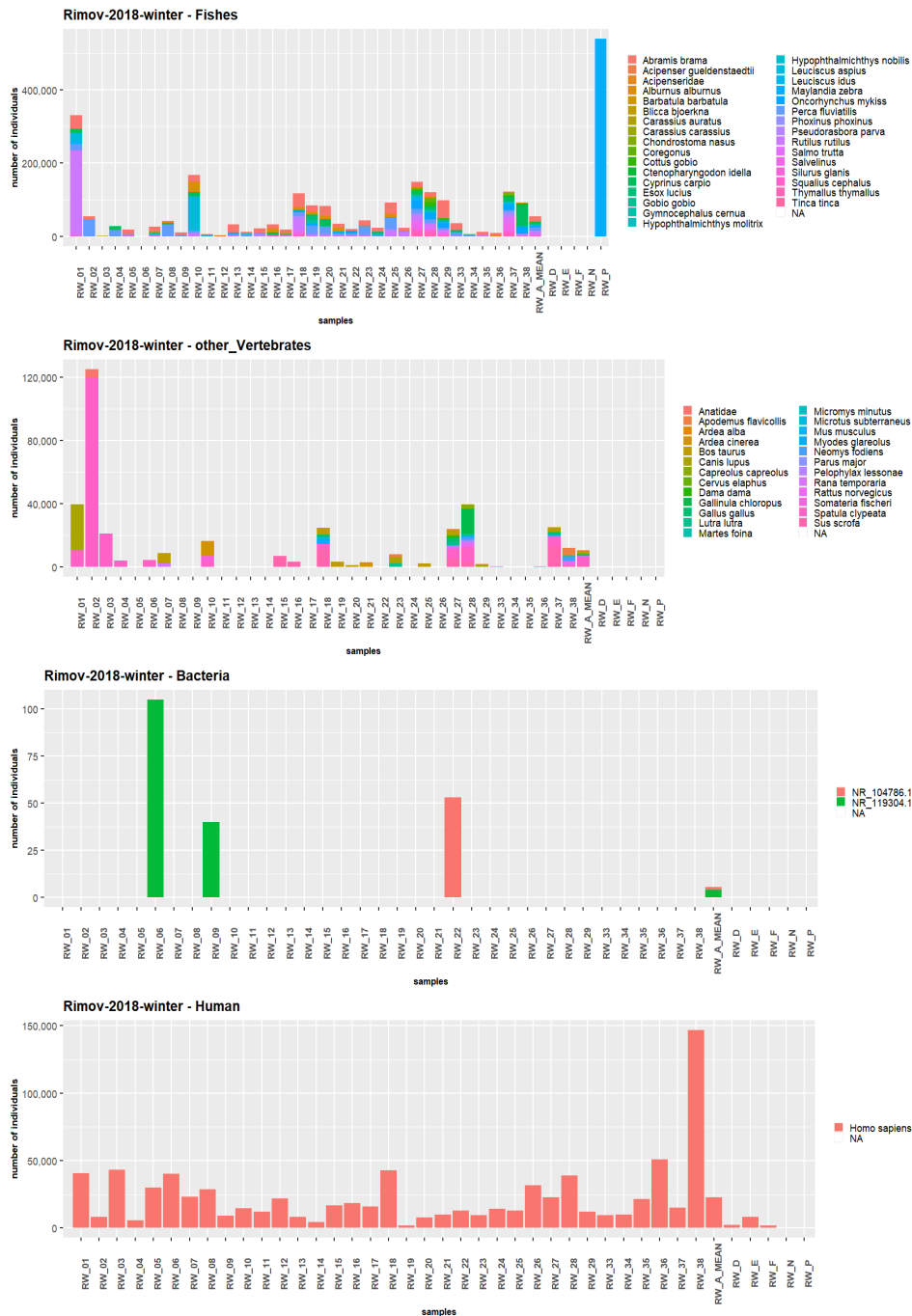
Klicava-2020-autumn shows higher total numbers of reads assigned. 19 different taxa of fish were detected. The fish taxa with the highest mean over all regular samples is ‘*Perca fluviatilis*’, which has its highest representation in sample ‘4L’, where 558,351 reads were assigned. Even though three samples stand out with ~280,000 (1R), ~60,000 (2L), and ~290,000 (3R), most of the regular samples show no, or just low amounts of reads assigned to ‘other_Vertebrates’. In many samples, more than human than ‘other_vertrebrate’ reads were assigned. Additionally, a high number of human DNA was detected in sample ‘37’ (324,697 reads), and 100,000 reads were assigned to humans in sample ‘5ZR’. (Figure 5)

4.2. Rimov



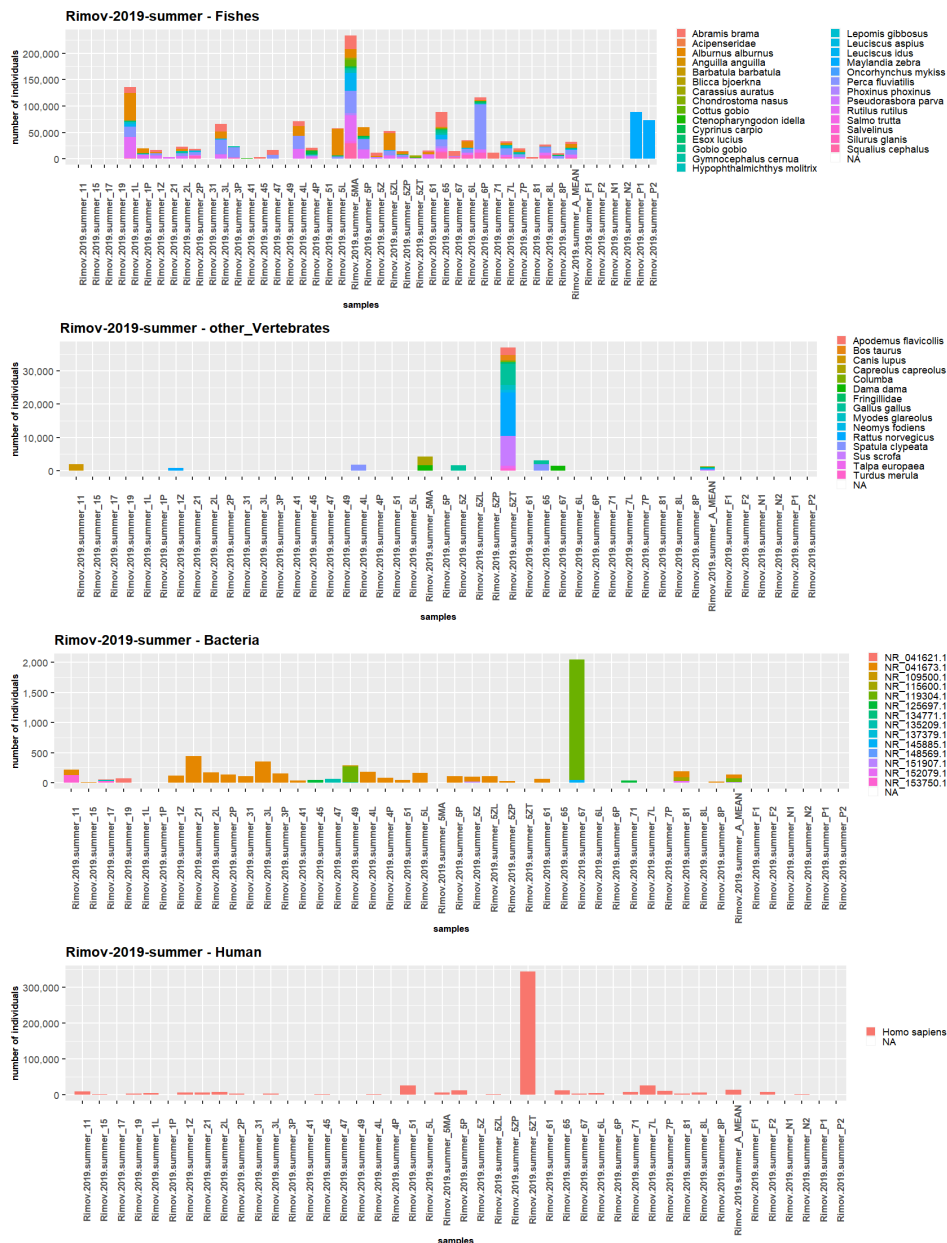
In Rimov-2018-summer there is a high variability in total numbers of reads assigned. It is important to point out the vast amount of fish (~3.3million) compared to the low number of human (~380,000) and other vertebrates (~45,000) in the case of total reads assigned, as well as in the number of different taxa (fish: 23; other vertebrates: 10). The mean of the total number of fish in regular samples is ~100,000, even though one of the samples contains 794,055 assigned reads. In sample '20' and '21', the ratio of bacteria to fish, human and other vertebrates ($616 : 44375 / 2660 / 0$) is elevated, compared to other samples. (Figure 6)

Figure 6 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2018-summer sampling event.



Rimov-2018-winter shows a very high variability of taxa (fish: 32, other vertebrates: 26). Sample '01' has the highest number of fish reads with over 230,000 reads of '*Rutilus rutilus*'. In Sample '02', the great majority of the reads got assigned to 'other_Vertebrates', as '*Spatula clypeata*' claims nearly 120,000 reads. A sample which shows a very distinct assignment is '06', because only 40,201 human reads, 4,585 other vertebrate reads as well as 105 bacterial reads were assigned. In '38' one can also observe a high amount of human activity, as 146,698 out of the total 250,846 reads are assigned to '*Homo sapiens*'. (Figure 7)

Figure 7 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2018-winter sampling event.



A high variability of assigned reads is given in the Rimov-2019-summer sampling campaign. Samples ‘11’, ‘15’, ‘17’, and ‘19’ all have less than (‘15’, ‘17’, ‘19’) or just over (‘11’) 10,000 assigned reads in total, while others have more than 100,000 reads. In general, a rather small number of samples have reads assigned to ‘other_Vertebrates’ and if they do there are less than 5,000 assigned. In sample ‘67’ one can observe an unusually high number of reads assigned to bacteria (2001). The positive control detected only fish, the negative control and the field blanks only human DNA. (Figure 8)

Figure 8 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2019-summer sampling event.

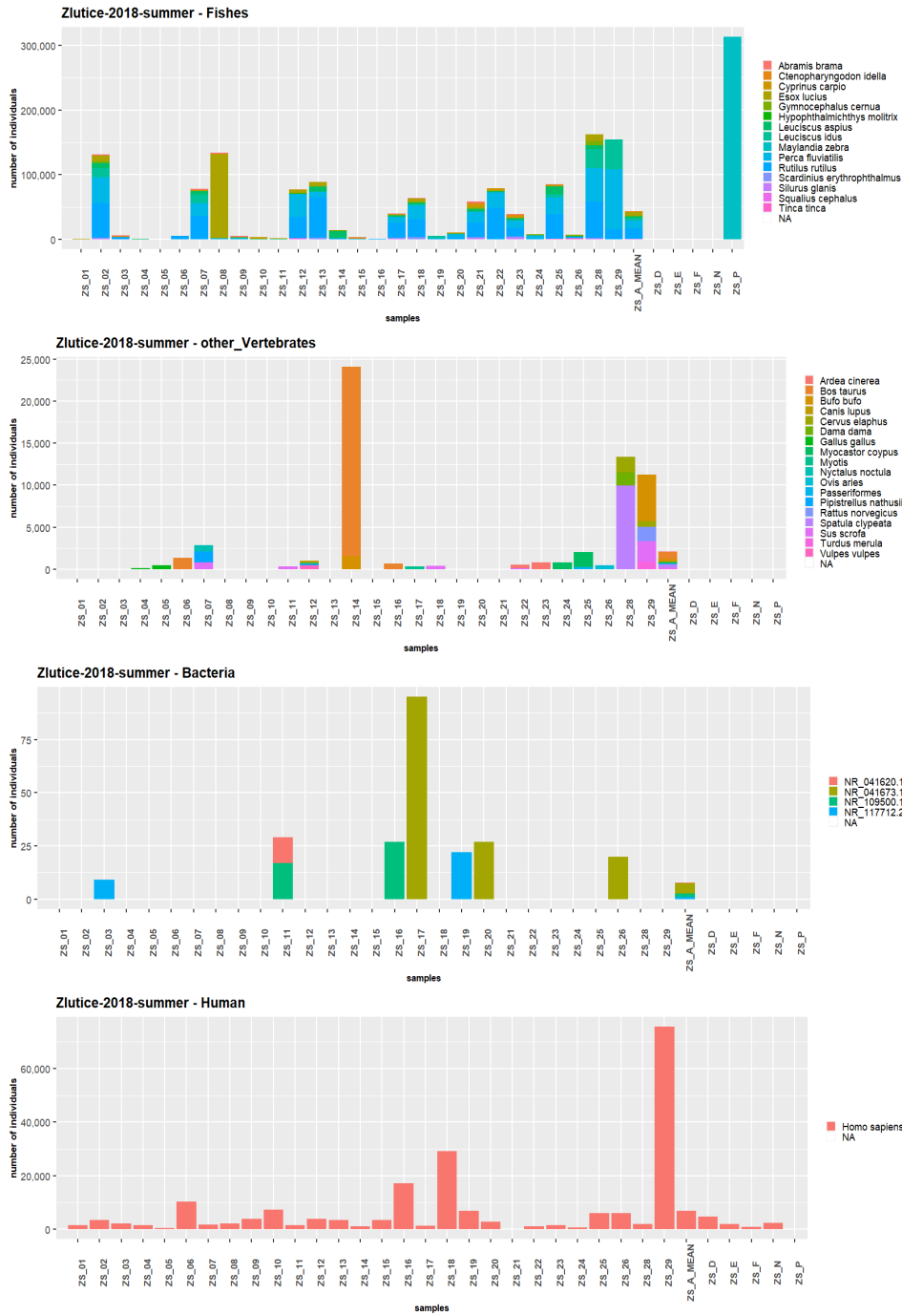


The samples of Rimov-2019-autumn sampling campaign contain a high variability in taxa (fish: 33, other_vertrebrates: 25, bacteria: 4) and numbers of detected reads. Sample ‘1L’ contains a majority of ‘other_Vertebrates’ as ‘*Spatula clypeata*’ claims 50,051 assigned reads. The mean shows that in all regular samples ‘*Spatula clypeata*’ is the most represented taxa of ‘other_Vertebrates’ (~5,000 assigned reads). But the mean on the fish graphs shows that the taxon which is most represented in all samples is the fish ‘*Abramis brama*’ (>20,000 assigned reads).

(Figure 9)

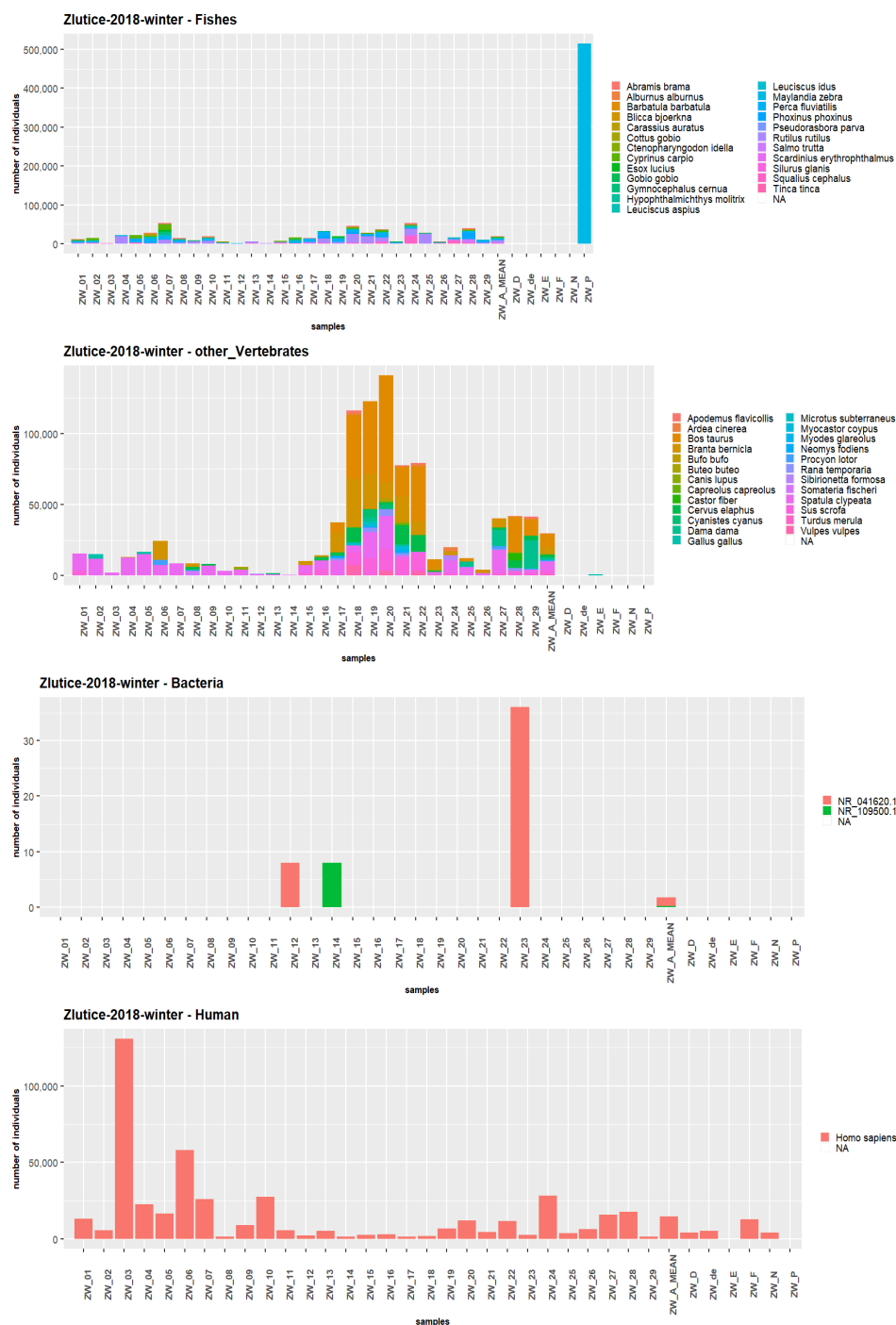
Figure 9 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2019-autumn sampling event.

4.3. Zlutice



About half of the samples of the Zlutice-2018-summer sampling event are dominated by the presence of the fish '*Perca fluviatilis*' and '*Rutilus rutilus*'. In sample '08', nearly all the assigned reads belong to '*Esox lucius*' (129,927 out of 135,527). The only sample in which 'other_Vertebrates' got significant numbers of reads assigned is sample '14' as '*Bos taurus*' claims more than half of the total reads (22,488 out of 38,795). In sample '29', human DNA (75,556), next to fish (~154,000) makes up a big portion of assigned reads, while other vertebrates only hold ~11,000 reads. (Figure 11)

Figure 11 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2018-summer sampling event.



The most noticeable result in the regular samples of Zlutice-2018-winter is the high distortion of the ‘Fishes’ graph, due to large amounts of reads in the positive control sample. It is also observable that in samples ‘17’ to ‘23’, more than half of the assigned reads are assigned to ‘other_Vertebrates’, specifically ‘*Bos taurus*’, ‘*Branta bernicla*’ and ‘*Spatula clypeata*’. The peak of sequences assigned to ‘*Homo sapiens*’ can be detected in sample ‘03’, where it is nearly the only taxa represented (130,687 reads out of 133,650). Bacteria, present only in three samples, have no significant impact on the total numbers. Finally, one could notice the control sample ‘de’ which was to test for diesel and was included in this sampling campaign. (Figure 12)

Figure 12 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2018-winter sampling event.



The sampling campaign of Zlutice-2019-summer yields a high variability of assigned reads in regular samples, reaching from 6,637 ('31') up to 299,426 ('5L'). Out of the outstanding 299,426 reads, 217,961 were assigned to '*Rutilus rutilus*', which also results in the highest mean in 'Fishes'. In comparison to the previous year, a higher number of different taxa was detected. In all the regular samples, 'other_Vertebrates', behind Bacteria, are the least represented group with a mean just over 1,000 reads. In sample '19' one can observe a majority of human DNA (22,913 out of 35996 reads). (Figure 13)

Figure 13 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2019-summer sampling event.

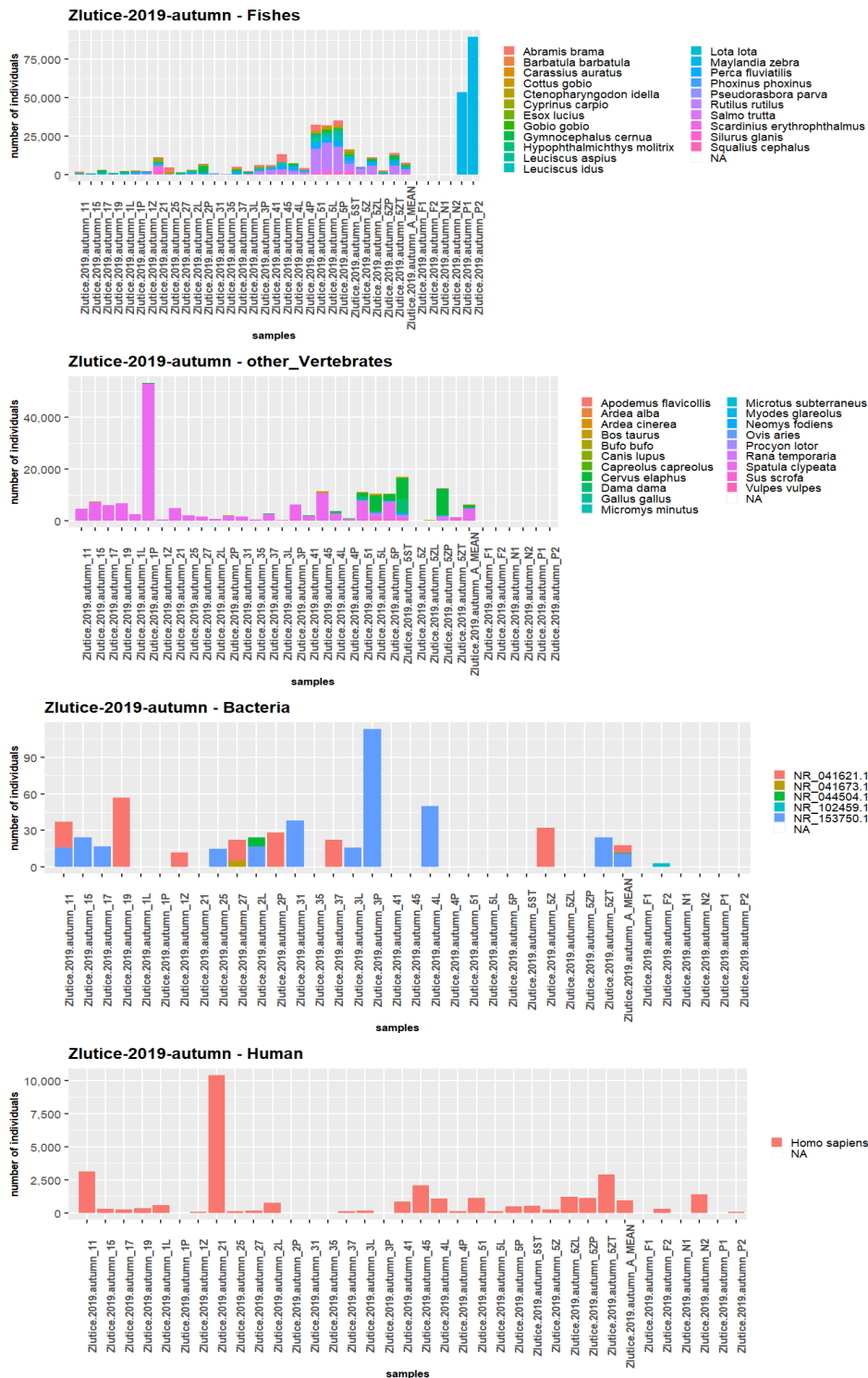


Figure 14 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2019-autumn sampling event.

Zlutice-2019-autumn shows a generally low number of reads assigned per sample. Many of them contain less than 10,000 assigned reads. Sample ‘1P’ is particularly interesting due to an extreme high of ‘*Spatula clypeata*’ (‘other_Vertebrates’) with over 53,154 out of 55,796 assigned reads. In general, ‘*Spatula clypeata*’ is exceptionally represented in Zlutice-2019-autumn. A rather uncommon portion of assigned reads in sample ‘3P’ is representing bacterium ‘*Cronbergia siamensis*’ (NR_153750.1), as it nearly contains 1% of the total assigned reads. (Figure 14)



Figure 15 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2020-autumn sampling event.

In Zlutice-2020-autumn the mean reaches nearly 80,000 assigned reads of which the vast majority is fish. Sample '5R' is particularly important since it has over 400,000 fish, over 10,000 'other_Vertebrates' and more than 3,000 human reads assigned. Out of the reads assigned to fish, 332,339 are assigned to '*Rutilus rutilus*'. In general, the majority of reads got assigned to fish rather than human, other vertebrates or bacteria. The only sample that contains a salient number of reads assigned to humans is '5Z' with nearly 60,000 assigned reads. (Figure 15)



Figure 16 Numbers of different taxa of fish, other vertebrates, bacteria and human in Laz-2020 sampling event.

4.4. Laz

In the Laz-2020 sampling campaign just four samples were taken. Just two of the samples ('03', '04') contain DNA of 'other_Vertebrates'. Most of the present DNA got assigned to '*Leuciscus aspius*' (209,850), a fish taxa. Just in sample '01', humans have a similar high number of reads as fish, as over 30,000 of the 61,415 reads were assigned to humans. (Figure 16)

4.5. Sumava



Figure 17 Numbers of different taxa of fish, other vertebrates, bacteria and human in Sumava-2020 sampling event.

The sampling campaign of Sumava-2020 does not yield a high number of different fish taxa. The amount of different taxa in ‘other_Vertebrates’ is significantly higher. The fish taxon that is most represented is ‘*Salmo trutta*’, which has a mean of over 75,000 in the regular samples. The vertebrate which has the highest representation is ‘*Castor fiber*’ with a mean of ~9,000 assigned reads. Noticeable is the peak of 82,400 reads assigned to humans in sample ‘17’. Bacteria can only be detected in three samples (‘10’, ‘16’, ‘18’). (Figure 17)

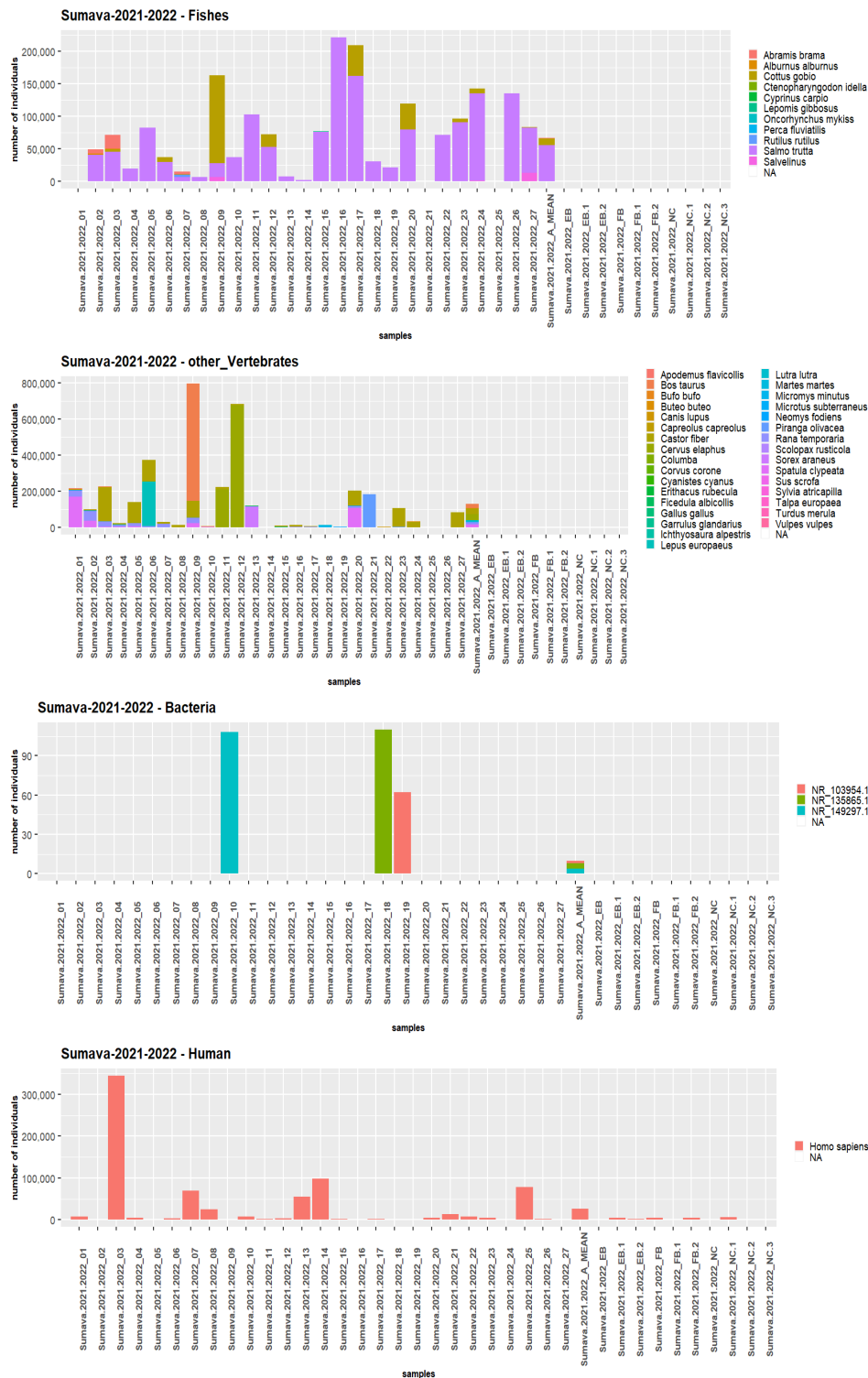
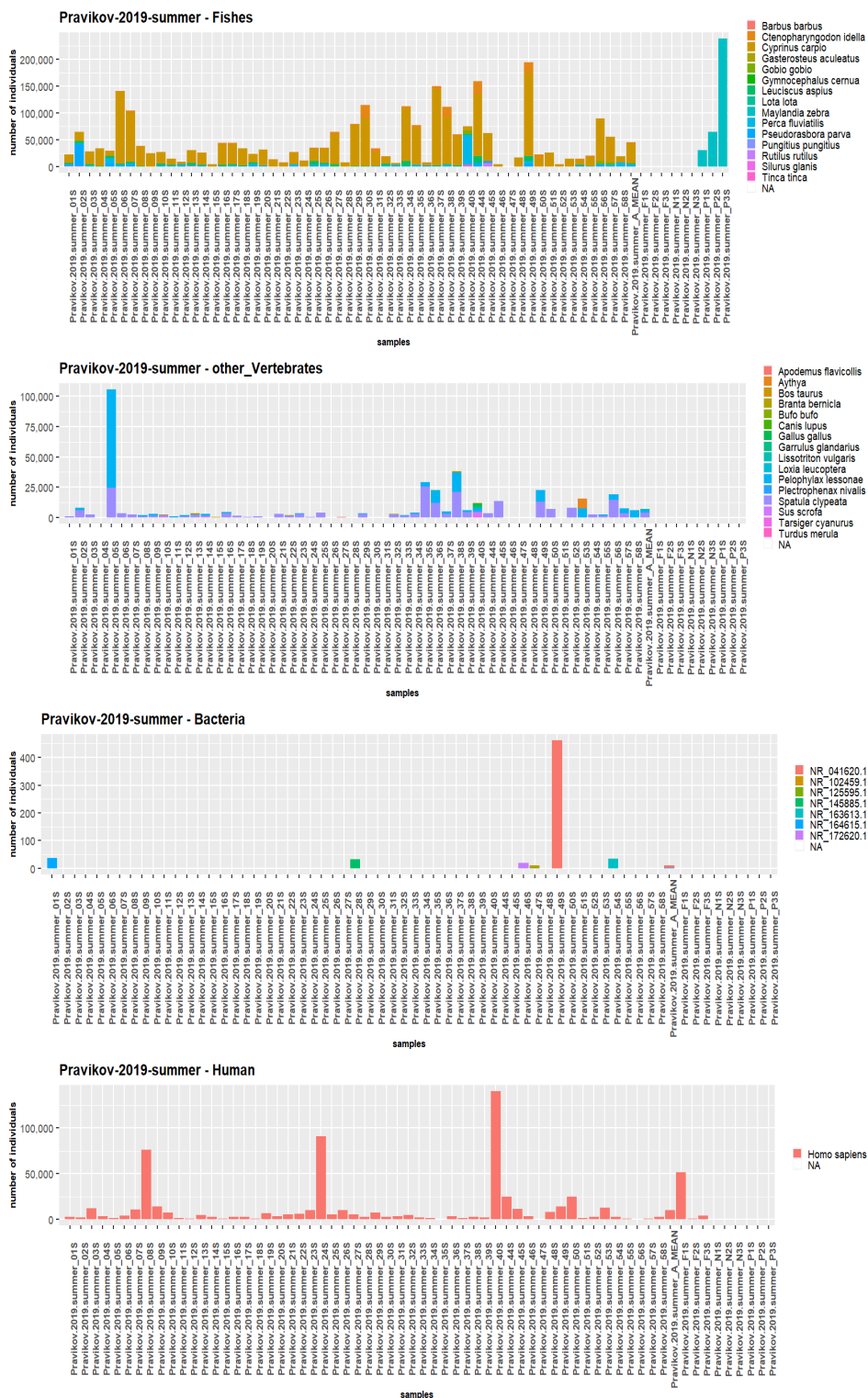


Figure 18 Numbers of different taxa of fish, other vertebrates, bacteria and human in Sumava-2021-2022 sampling event.

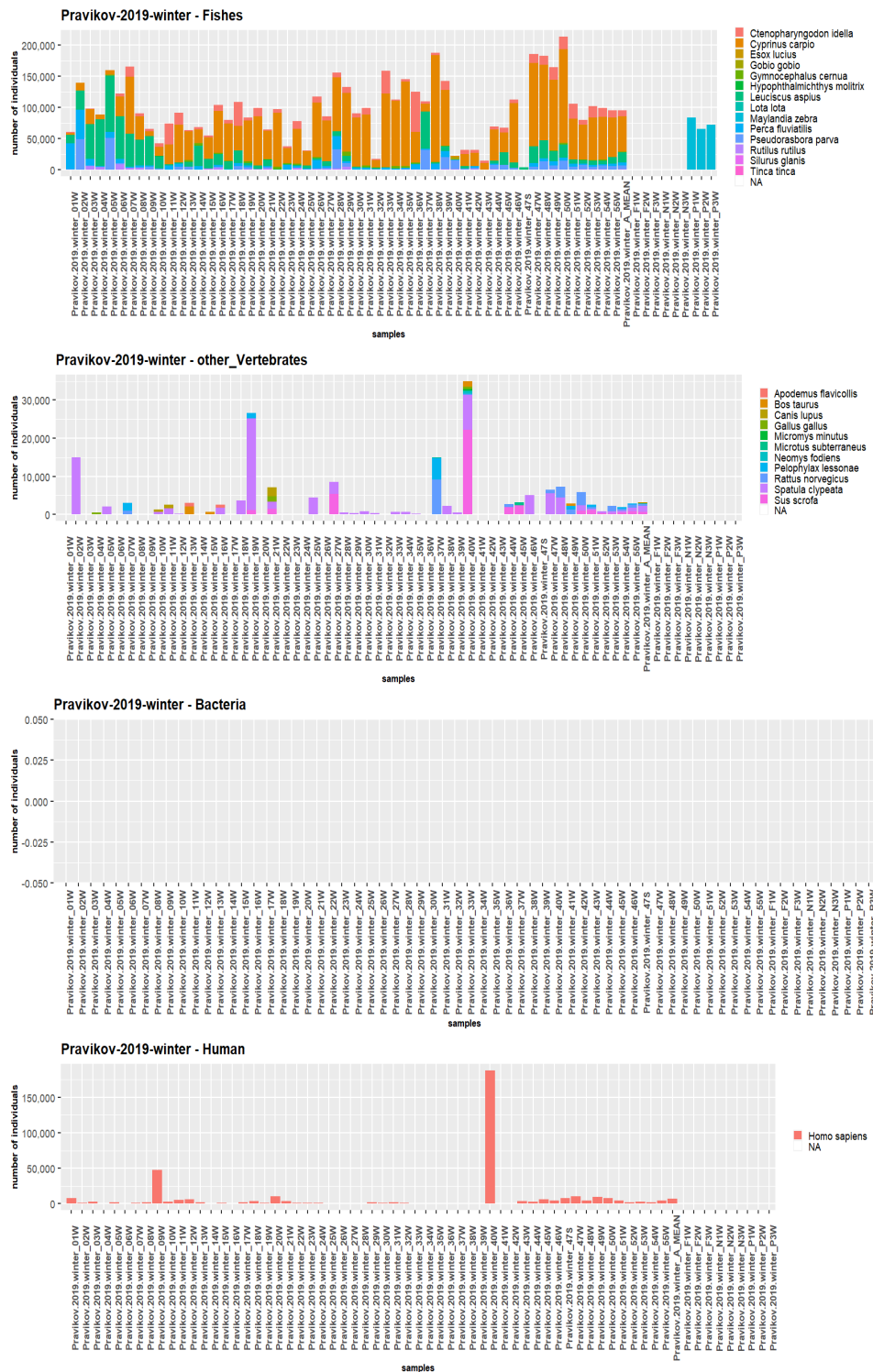
In Sumava-2021-2022, the number of fish taxa (11) is significantly lower than the one of ‘other_Vertebrates’ (32). High amounts of ‘*Salmo trutta*’ can be observed. Sample ‘09’ yields 134,699 reads of ‘*cottus gobio*’ (fish) and 648,579 reads of ‘*Bos taurus*’ (‘other_Vertebrates’). In sample ‘03’ one can observe a high amount of human DNA (344,907 assigned reads). Just in three samples, bacteria were found (‘10’, ‘18’, ‘19’) to take a small share of the reads. (Figure 18)

4.6. Pravikov



The Pravikov-2019-summer sampling campaign has high numbers of reads assigned to the fish taxon ‘*Cyprinus carpio*’. Additionally, ‘other_Vertebrates’ have low numbers of reads assigned. However sample ‘05’ can be named as an exception as 80,839 reads were assigned to ‘*Pelophylax lessonae*’ and 24,371 were assigned to ‘*Spatula clypeata*’. In sample ‘49S’, 461 reads were assigned to the bacterium ‘*Persicirhabdus*’ (NR_041620.1). The samples ‘08S’, ‘24S’ and ‘40S’ show high numbers of reads assigned to ‘*Homo sapiens*’ (76,005 / 90540 / 140,075). (Figure 19)

Figure 19 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pravikov-2019-summer sampling event.



In the samples of Pravikov-2019-winter the level of detected fish DNA is well balanced. The majority of observed fish taxa is assigned to ‘*Cyprinus carpio*’. A small number of reads was assigned to ‘other_Vertebrates’. Bacteria were not assigned in this campaign. In sample ‘40W’, the number of assigned reads to humans (188,007) is far higher than the one assigned to fish (~30,000). (Figure 20)

Figure 20 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pravikov-2019-winter sampling event.

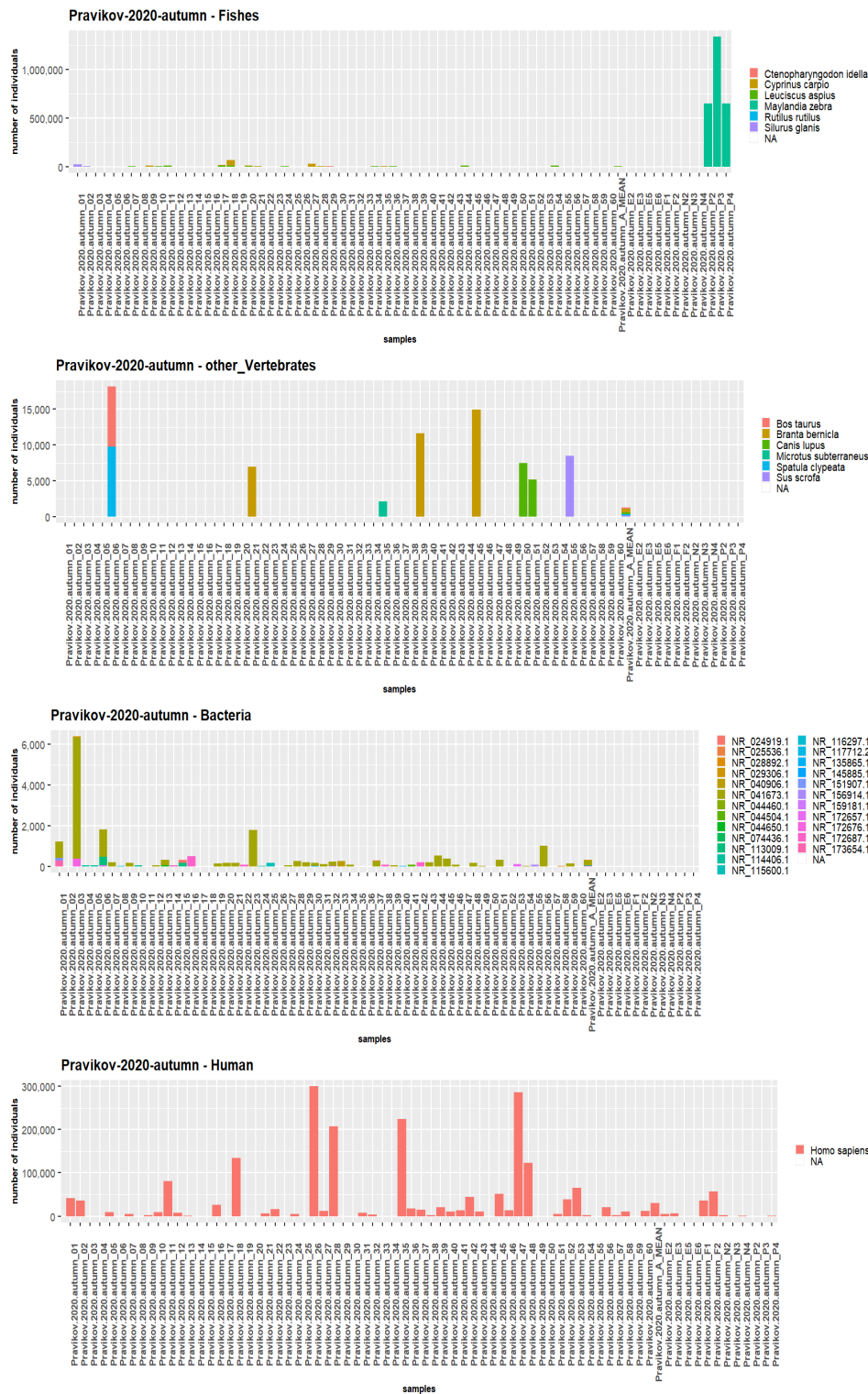


Figure 21 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pravikov-2020-autumn sampling event.

In Pravikov-2020-autumn it is important to focus on the high amount of human DNA that got detected as well as the high distortion in the ‘Fishes’ graph due to the high numbers of reads in positive control samples (up to 1,331,402 reads assigned to ‘*Maylandia zebra*’). There are six fish as well as six ‘other_Vertebrates’ taxa and 24 different bacteria. Even though there is a big distortion in ‘Fishes’, it can be observed that human is the taxon, which is assigned to most reads in the regular samples. Samples ‘26’ and ‘47’, which nearly reach 299,174 and 286,398 assigned reads are particularly important. Sample three holds over 6,000 bacteria reads, most of them being assigned to ‘*Haloferula helveola*’ (‘NR_041673.1’), which is the most represented Bacterium in this sampling campaign. (Figure 21)

4.7. Kalich

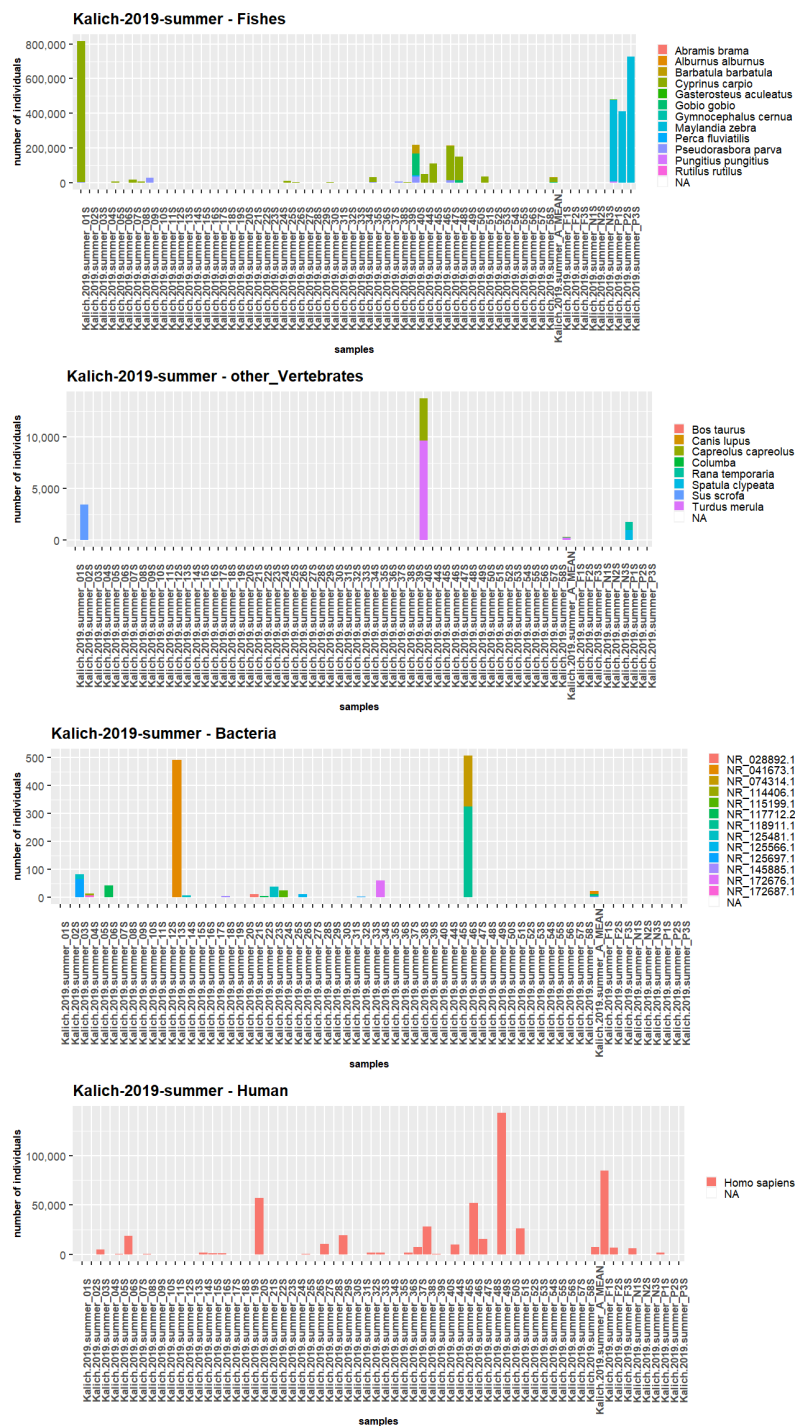
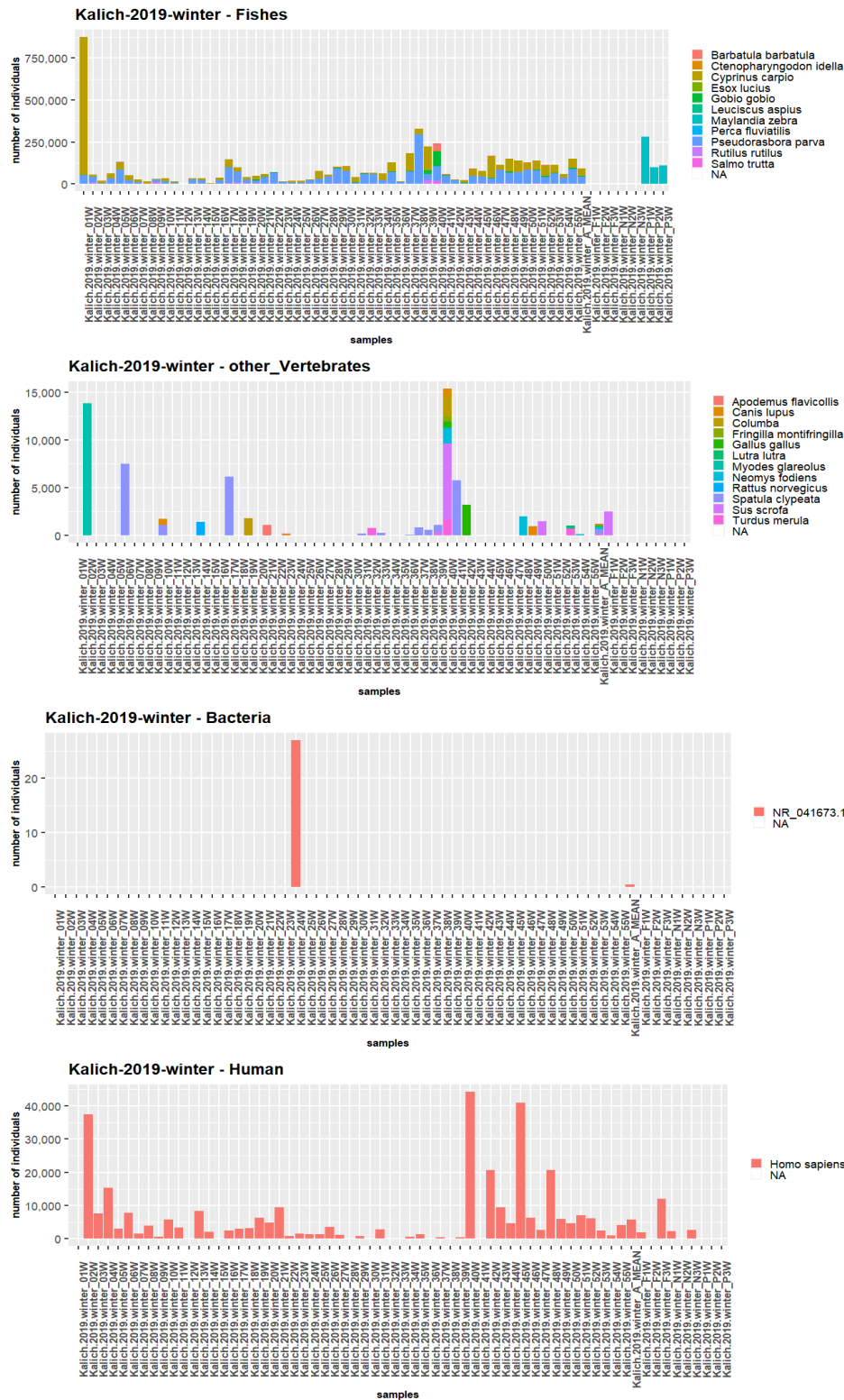


Figure 22 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kalich-2019-summer sampling event.

Most of the regular samples of Kalich-2019-summer sampling campaign have low numbers of assigned reads. The exception with the highest number of reads assigned is sample '01S' where 811,124 reads were assigned to 'Cyprinus carpio' ('Fishes'). Samples '40S', '45S', '47S' and '48S' show over 100,000 assigned reads. The sample which contains most 'other_Vertebrates' is '40S' (~18,000 reads). In samples '13S' and '46S', 493 and 507, reads were assigned to bacteria. In sample '49S', 142,831 reads of human were detected. (Figure 22)



Kalich-2019-winter contains high amounts of fish with a mean of ~100,000 reads assigned. One can observe more humans than other vertebrates in most of the samples. Bacteria are represented in one sample ('24W'; 27 reads). An outstanding sample is '01W' as 875,805 reads are assigned to fish with most of them being assigned to the taxon '*Cyprinus carpio*'. Still, the taxon assigned the most in total is '*Perca fluviatilis*'. (Figure 23)

Figure 23 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kalich-2019-winter sampling event.

4.8. Kladiny



Figure 24 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kladiny-2019-summer sampling event.

In Kladiny-2019-summer a high variety of fish taxa (20) and a low variety of ‘other_Vertebrates’ (5) and bacteria were detected. The variety also represents the total number of reads assigned. While most of the samples show a majority of fish, sample ‘56S’ contains 109,686 reads which were assigned to humans. The sample with the most overall reads is ‘09S’ as 328,155 reads got assigned. Bacteria are found in nearly all the samples, while ‘other_Vertebrates’ are sparsely distributed. In the category ‘Bacteria’, many samples show more than 500 assigned reads, with a maximum of 2,977 in sample ‘02S’. (Figure 24)



Figure 25 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kladiny-2019-winter sampling event.

Kladiny-2019-winter shows high numbers of reads assigned to fish, as the mean reaches over 75,000. Even though ‘other_Vertebrates’ has assigned reads in just a few samples, one of them (‘5ot’) stands out as ~100,000 reads were detected. Out of these ~100,000, 60,800 got assigned to ‘*Bos taurus*’. In ‘Human’, two samples are noteworthy as samples ‘10W’ and ‘5ot’ show 146,107 and 81,748 reads assigned to ‘*Homo sapiens*’. No bacteria were found in this sampling campaign. (Figure 25)



The graphs of Kladiny-2020-autumn show that a high amount of fish as well as human DNA was detected. The ‘other_Vertebrates’ graph contains just three samples with over 10,000 reads assigned, while most of the samples hit the 10,000 reads mark in ‘Fishes’ and ‘Human’. In ‘Bacteria’ sample ‘34’ stands out, as 2,123 reads were assigned to the bacterium ‘*Oceanifera marina*’ (‘NR_173618.1’). The fish taxon, which is represented the most in Kladiny-2020-autumn, is ‘*Cyprinus carpio*’ with a mean of ~30,000 reads assigned in the regular samples. Most of the control samples contain a majority of reads assigned to ‘*Homo sapiens*’. (Figure 26)

Figure 26 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kladiny-2020-autumn sampling event.

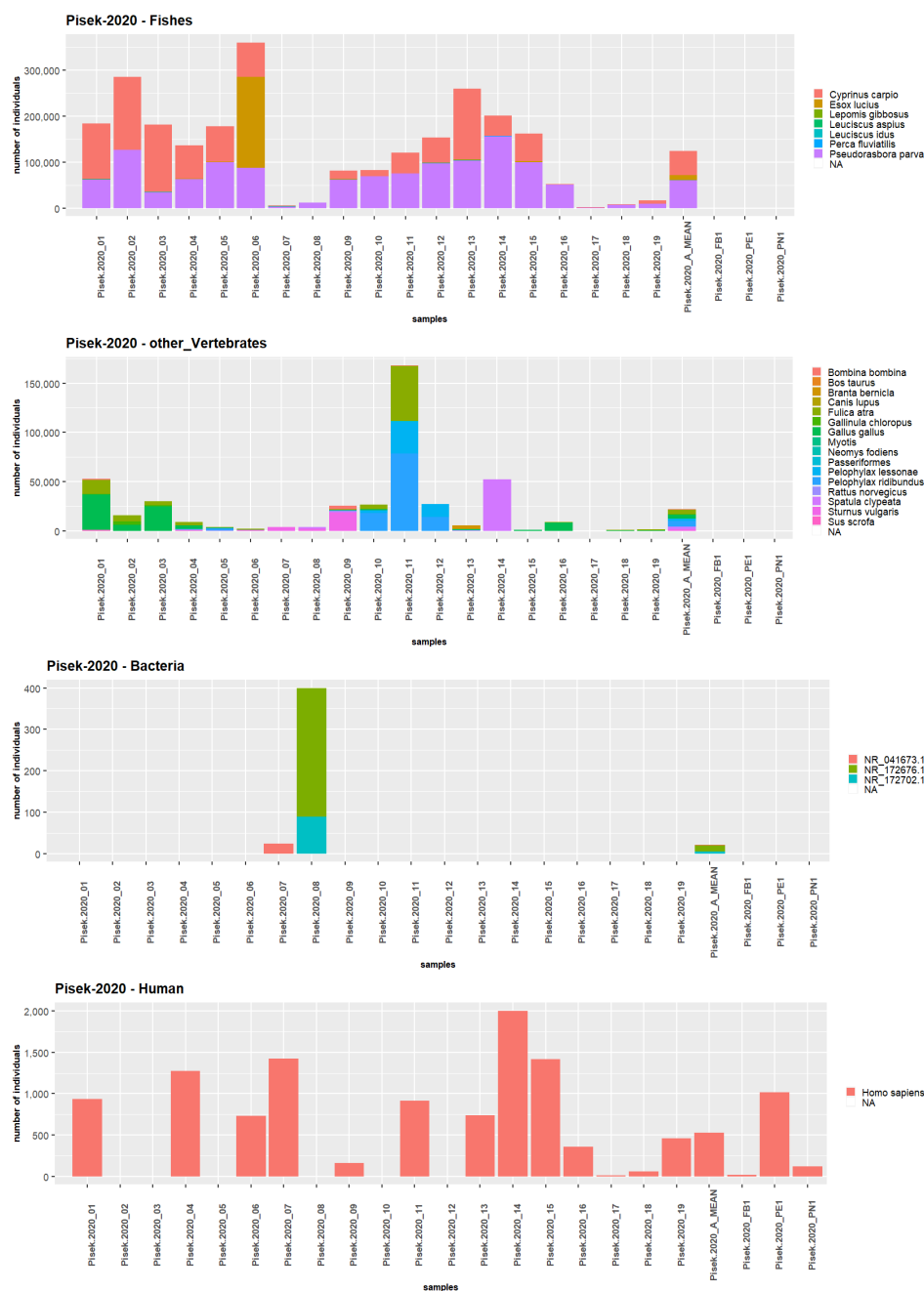
4.9. Tuska



In Tuska-2020-autumn sampling campaign the most reads in total are assigned to different fish taxa. Nevertheless, ‘other_Vertebrates’ have a higher variety of taxa than ‘Fishes’. Only in sample ‘40’, ‘other_Vertebrates’ surpass the 25,000 mark as over 100,000 reads are assigned. Four samples have more than 1,000 reads assigned to bacteria (samples: ‘01’, ‘12’, ‘29’). In 18 samples more reads got assigned to human than to fish or ‘other_Vertebrates’ (‘02’, ‘04’, ‘14’, ‘15’, ‘17’, ‘23’, ‘24’, ‘36’, ‘37’, ‘45’, ‘47’, ‘48’, ‘49’, ‘52’, ‘54’, ‘56’, ‘57’, ‘59’). (Figure 27)

Figure 27 Numbers of different taxa of fish, other vertebrates, bacteria and human in Tuska-2020-autumn sampling event.

4.10. Pisek



Most of the samples in the Pisek-2020 sampling campaign have high numbers of detected reads. Most of them were assigned to the fish taxa '*Pseudorasbora parva*' and '*Cyprinus carpio*', each having a mean of ~60,000 reads assigned in the regular samples. The only sample in which 'other_Vertebrates' contains more reads than 'Fishes' is sample '11', as ~170,000 reads were assigned to 'other_Vertebrates'. Just two samples show the presence of bacteria ('07', '08'). Overall a small number of reads were assigned to humans. (Figure 28)

Figure 28 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pisek-2020 sampling event.

5. Discussion

The eDNA metabarcoding method has proven to provide high-resolution data on the species diversity of different taxonomic groups in different aquatic ecosystems. Based on the primers used, a DNA fragment specific to a narrow (e.g., single species or genus) or broad (e.g., subphylum) taxonomic group can be detected and the target species or multiple taxonomic groups can be bioinformatically identified. In addition, in some cases, non-target taxa may be detected, reducing the sequencing depth and affecting the interpretability of the results. Evaluating the different taxonomic contributions within individual samples is the first step in understanding the interpretability of eDNA metabarcoding data. Therefore this bachelor thesis has an important implication for applying this method to aid management and policy initiatives.

5.1. Comparison between different water bodies and seasons

The samples taken from Klicava all have similar numbers of taxa represented. Most show a majority of fish. An exception is the sampling campaign of winter in the year 2018 as fish only have ~12.5% of the total mean while in other sampling campaign the fish mean holds over ~72% (summer-2018), ~75% (summer-2019), ~68% (autumn-2019) and ~66% (autumn-2020). Rimov contains a high number of different fish taxa each year and season, while ‘other_Vertebrates’ are less represented in summer. All sampling campaigns in Rimov show a high percentage of fish DNA (~90%, 2018-summer; ~58%, 2018-winter ; ~81%, 2019-summer; ~62%, 2019-autumn; ~88%, 2020-autumn).

On the second sampling site, Zlutice, the majority of sampling campaigns show that most reads were assigned to fish. The only campaign where this does not apply is, again, the winter of 2018 as fish only are ~32% of total reads. The higher amount of other vertebrate DNA can be caused by the fact that in autumn and early winter, before freezing, tending waterbirds typically rest on the water (Laursen et al., 2005). Another cause could be that fish are poikilotherms which means they cannot regulate their body temperature. This means their activity is temperature dependent. In colder conditions, most species are less active and thus less eDNA gets into the water (Persson, 1986). The number of different taxa can be roughly compared to the one from Rimov.

The sampling campaign of 2020 which was done in Laz on the other hand has a low amount of different taxa which probably is caused by the low number of only four samples, as well as the low organic productivity of the water since it is oligotrophic (Jeppesen et al., 2000). There was a much higher amount of fish than human or other vertebrates detected as over 86% are fish. Further, in Sumava 2020 a rather small number of different fish taxa and a high number of ‘other_Vertebrates’ taxa are shown. The cause for that could be the extreme conditions in the spring area as only some species are adapted to the conditions featuring cold water and low amount of food (Ventura et al., 2017). The trend for the most reads assigned is the opposite, as an average of ~80,000 reads were assigned to fish and only an average of ~24,000 to humans and ‘other_Vertebrates’. In 2021-2022 on the other hand, humans (~15%) and other vertebrates (~58%) together make up a bigger portion than fish (~27%).

In the sampling campaigns in Pravikov-2019-summer / winter, even when 14, 15 different taxa were assigned to fish, an average of 80% / 50% of fish reads got assigned to ‘*Cyprinus carpio*’ as it is the most common fish species in Czech ponds (Pechar, 2000). In general there is a higher population density in fish ponds than in natural water which can be observed as well (Blabolil et al., 2022). Both campaigns in 2019 show a majority in fish (~74% and ~87%). The year after though, there are just very few taxa with low numbers of reads detected in ‘Fishes’ as well as ‘other_Vertebrates’. It is notable that in sample ‘03’ a vast majority of assigned reads were from bacteria (5936 out of 6382). This is caused by the nearby town which releases its wastewater not cleaned properly. It is common for eDNA samples to be contaminated by wastewater (Staley et al., 2018). Another cause for the high amounts of bacteria

detected could be that fertilisation by livestock raises the trophic status, provides an increased microbial loop in the food web and introduces a higher amount of bacteria (Boyd, 2018). A lot of samples contain a majority of human DNA. Both these indices could be a sign for heavy contamination in the samples of Pravikov-2020-autumn sampling campaign.

Interestingly, in Kalich in summer 2019, in the first sample 811,124 reads got assigned to '*Cyprinus carpio*' which is severely represented in the mean. Most of the samples though, have very little detected reads. Thus it is interesting that three samples have over 50,000 and up to 140,000 reads assigned to humans. Most probably this indicates laboratory contamination as well. It is not uncommon that contamination occurs in the laboratory when the filters have to be handled manually (Sepulveda et al., 2020). In all three sampling campaigns in Kladiny an average vast majority of reads got assigned to different fish, while in summer the majority were '*Cyprinus carpio*', in winter, most reads got assigned to '*Gymnocephalus aculeatus*'. While in winter of 2019 there were no bacteria detected at all, in summer, nearly every sample contains bacteria. In Tuska in autumn of 2020 an average of over 50,000 fish (most of it '*Cyprinus carpio*') got assigned which makes over 60% of the total average. Interestingly, sample '40' has an extremely high amount of reads assigned to 'other_Vertebrates', while nearly no reads got assigned to fish. Many of the samples show the presence of bacteria. The highest bacterial contaminations are present in sample '01', '12' and '29'. In Pisek-2020 as in most other sampling campaigns fish makes with ~125,000 assigned reads (~83%) the main part of the total average. There is just very little human or bacterial contamination detected in the analysed samples.

5.2. Highest and lowest diversity in defined taxa

The highest diversity of fish taxa could be observed in Rimov, as in 2019-autumn 33 different taxa were detected. Whereas the opposite counts for the sampling campaigns in 2020 in Laz and Sumava as only six different fish taxa were detected. In Pisek as well a low number of taxa was detected as just seven could be observed. This can be led back to the, compared to other sampling sites, smaller water bodies (Mehner et al., 2005). Interestingly even though Pravikov normally showed a higher number of taxa, in autumn of 2020, as well just six got detected. This could be, as stated before, a sign that something went wrong in the process of sampling or the sample processing.

The highest diversity of other vertebrates on the other hand was detected in Sumava in 2021/22 as reads were assigned to 32 different taxa. This outstanding number of different taxa could be due to the fact that the water bodies (natural lake and brooks) are located in the area of the Sumava National park where nature can evolve more freely. (Siikamäki et al., 2015) The lowest diversity of other vertebrates was detected, as was the case for fish, in autumn of 2020 in Pravikov (six taxa) and 2020 in Laz (three taxa). It would be likely that the causes of the low diversities of other vertebrates are the same as for fish.

The highest diversity of bacteria was found in Pravikov 2020 autumn, as 24 different taxa were detected. This high contamination with bacteria as well indicates laboratory contamination. The lowest diversity of bacteria can be found in Pravikov and Kladiny in 2019 in winter. In both sampling campaigns, no bacteria were detected. The cause for that could be the season, as bacteria are much less active in colder conditions. (Russell, 2000)

The highest amount of reads assigned to humans can be found in the sampling campaign of 2020 autumn in Pravikov. Many samples show human reads, in four of them even over 200,000 reads got assigned to humans. This, together with the low number of fish and other vertebrate taxa and a high amount of bacterial taxa, raises a very strong suspicion towards contamination in the field or laboratory.

Compared to the total number of reads, the sampling campaign Tuska-2020-autumn also yielded a lot of reads assigned to humans. The lowest amount of human DNA detected was in the sampling campaign of Pisek in 2020 as no sample shows more than 2,000 reads assigned to humans

even if there was a fairly high number of reads assigned in total. It is probable that this is because there are no human settlements nearby.

5.3. Possible contamination in control samples

Most of the PCR positive control samples do not show signs of contamination as the used DNA template was highly concentrated and specifically added to prove the PCR run. Nevertheless, even though eDNA based techniques are highly sensitive, it is known that these methods are not flawless at detecting species (Ficetola, et al., 2014). It could have happened that ‘Malaysia zebra’ was detected in other samples as well due to tag jumping or other sources.

In ten of the 28 sampling campaigns, contamination was detected. In five of the ten contaminated sampling campaigns, only humans were detected (Zlutice-2020-autumn, Zlutice-2019-autumn, Pravikov-2020-autumn, Klicava-2019-autumn, Kladiny-2019-winter). In contrast the other samples showing contamination included different fish or other vertebrates. Most of the time though the contamination is just very minor which could be because of sequencing errors. In some of the PCR positive control samples, no reads were assigned at all. This result depicts the overall standard as most of the positive controls have no contamination and if, it is just very little. (Gehri et al., 2021; Holman et al., 2022)

The PCR negative on the other hand, in most samples (24) at least small amounts of contamination can be observed. The only sampling campaigns containing no contamination are ‘Rimov-2018-winter’, ‘Pravikov-2019-winter’, ‘Klicava-2020-autumn’ and ‘Laz-2020’. 17 of the 24 sampling campaigns having contaminated PCR negative control samples, are contaminated just by human DNA. The sampling campaigns Zlutice-2020-autumn, Tuska-2020-autumn, Sumava-2021-2022, Pravikov-2019-summer, Klicava-2018-winter, Kladiny-2020-autumn only show more than two different taxa assigned. In Rimov-2018-summer only ‘*Bos taurus*’ (‘other_Vertebrates’) was detected. It is not uncommon to have contamination in negative control samples when doing a metabarcoding study since ~25% of metabarcoding studies show contamination. (Sepulveda, et al., 2020) Nevertheless, a big problem of contaminated negative control samples containing other taxa than ‘*Maylandia zebra*’ is that they can further contaminate the regular samples as well.

In 26 out of the 28 sampling campaigns, the field and filtration blanks contain human DNA. Nine of them as well got reads assigned to other vertebrates or fish. In Klicava-2019-autumn no human but ‘*Bos taurus*’ (other vertebrate) was detected. The only Sampling campaign which is not included in the aforementioned is Sumava-2020 as no field blank was included.

14 sampling campaigns (Klicava-2018-summer, Klicava-2018-winter, Klicava-2020-autumn, Pisek-2020, Pravikov-2020-autumn, Rimov-2018-summer, Rimov-2018-winter, Rimov-2020-autumn, Sumava-2020, Sumava-2021-2022, Tuska-2020-autumn, Zlutice-2018-summer, Zlutice-2018-winter, Zlutice-2020-autumn) contained extraction blanks. In all of them, human reads were detected. five of them (Klicava-2020-autumn, Pisek-2020, Tuska-2020-autumn, Zlutice-2018-winter, Zlutice-2020-autumn) even contained DNA of fish or other vertebrates.

5.4 Bioinformatics

As earlier mentioned Tapirs has quite a few parameters one could alter. Most probably this would have an impact on the results given. It would be interesting to see what kind of changes in the results occur, if the Tapirs settings were different. In this study, I used standard default environment settings except for the ‘FASTP_length_limit’ which was changed from 110 to 106 (Tapirs, 2020/2022; <https://github.com/EvoHull/Tapirs/blob/master/config/config.yaml>). If the minimal E-Value (BLAST_min_evalue) would have been set to a higher E-Value than 1e-20, most probably this would

lead to an increased number of false detection since everything that got assigned with a higher E-Value than $1e-20$ was not further processed. Another parameter that would increase false detection is 'FASTP_len_required' which for this study was set to 90. This parameter is for discarding all reads which are shorter than 90 bases. If one would lower the parameter, the amount of false positives would rise since the probability of a correct hit rises with the length of the read. For further studies one could think of including bacteria in the reference database as it would remove the laborious part of extracting and once more analysing unassigned reads, as well as minimising the risk of errors in the aforementioned process.

6. Conclusion

One can expect fish eDNA will be the most common substance in water. The results demonstrate, natural water is a cocktail of eDNA from water and terrestrial sources. As usual, finishing one research topic will open several other issues. The results bring very important information for biologists aiming in eDNA metabarcoding utilisation. The method is highly sensitive to contamination. The sampling should be precisely planned in terms of number of samples and timing. In the subsequent study I would like to focus in more detail on samples with high contamination and try to find some predictable pattern in terms of sampling and processing. It will be very useful to try different primers such as MiFish (Miya et al., 2015) and Teleo (Valentini et al., 2016), use PCR blocking primers designed to block the amplification of some potential bycatch such as human (Rojahn et al., 2021) and bioinformatic pipelines (dos Santos, 2021). In an ideal case the reference database should be local and at least collection of local samples, sequencing and creation of local databases should be started (Elbrecht et al., 2016). In terms of data processing, I plan to have a look at diversity indices and multivariate statistical analysis using ordination methods to reveal more ecological patterns and prove the usefulness of eDNA metabarcoding for biological monitoring (Šmilauer & Lepš, 2014).

References

- Angermeier, P. L., & Karr, J. R. (1994). Biological integrity versus biological diversity as policy directives: Protecting biotic resources. In *Ecosystem management* (pp. 264–275). Springer.
- Blabolil, P., Griffiths, N. P., Hänfling, B., Jůza, T., Draštík, V., Knežević-Jarić, J., dos Santos, R., Mrkvička, T., & Peterka, J. (2022). The true picture of environmental DNA, a case study in harvested fishponds. *Ecological Indicators*, 142, 109241. <https://doi.org/10.1016/j.ecolind.2022.109241>
- Blabolil, P., Harper, L. R., Řičanová, Š., Sellers, G., Di Muri, C., Jůza, T., Vašek, M., Sajdlová, Z., Rychtecký, P., Znachor, P., Hejzlar, J., Peterka, J., & Hänfling, B. (2021). Environmental DNA metabarcoding uncovers environmental correlates of fish communities in spatially heterogeneous freshwater habitats. *Ecological Indicators*, 126, 107698. <https://doi.org/10.1016/j.ecolind.2021.107698>
- Boulenger, G. A. (1899). A Revision of the African and Syrian Fishes of the Family Cichlida. Part II.1. *Proceedings of the Zoological Society of London*, 67(1), 98–144. <https://doi.org/10.1111/j.1469-7998.1899.tb06852.x>
- Boyd, C. E. (2018). Aquaculture pond fertilization. *CABI Reviews*, 2018, 1–12. <https://doi.org/10.1079/PAVSNNR201813002>
- Butchart, S. H. M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J. P. W., Almond, R. E. A., Baillie, J. E. M., Bomhard, B., Brown, C., Bruno, J., Carpenter, K. E., Carr, G. M., Chanson, J., Chenery, A. M., Csirke, J., Davidson, N. C., Dentener, F., Foster, M., Galli, A., Galloway, J. N., Genovesi, P., Gregory, R. D., Hockings, M., Kapos, V., Lamarque, J. F., Leverington, F., Loh, J., McGeoch, M. A., McRae, L., Minasyan, A., Hernández Morcillo, M., Oldfield, T. E. E., Pauly, D., Quader, S., Revenga, C., Sauer, J. R., Skolnik, B., Spear, D., Stanwell-Smith, D., Stuart, S. N., Symes, A., Tierney, M., Tyrrell, T. D., Vié, J. C., Watson, R. (2010). Global Biodiversity: Indicators of Recent Declines. *Science*, 328(5982), 1164–1168. <https://doi.org/10.1126/science.1187512>
- Cristescu, M. E., & Hebert, P. D. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics*, 49(1), 209–230.
- Díaz-Ferguson, E. E., & Moyer, G. R. (2014). History, applications, methodological issues and perspectives for the use environmental DNA (eDNA) in marine and freshwater environments. *Revista de Biología Tropical*, 62(4), 1273–1284.
- dos Santos, R. (2021). *Comparison of bioinformatics pipelines for eDNA metabarcoding data analysis of fish populations in Czech reservoirs* [Jihočeská univerzita v Českých Budějovicích, Přírodovědecká fakulta]. <https://theses.cz/id/9003oy/>

- Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J.-N., Coissac, E., Boyer, F., & Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, 4, e1966. <https://doi.org/10.7717/peerj.1966>
- Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguët-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15(3), 543–556. <https://doi.org/10.1111/1755-0998.12338>
- Gehri, R. R., Larson, W. A., Gruenthal, K., Sard, N. M., & Shi, Y. (2021). EDNA metabarcoding outperforms traditional fisheries sampling and reveals fine-scale heterogeneity in a temperate freshwater lake. *Environmental DNA*, 3(5), 912–929. <https://doi.org/10.1002/edn3.197>
- Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., & Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, 25(13), 3101–3119. <https://doi.org/10.1111/mec.13660>
- Holman, L. E., Chng, Y., & Rius, M. (2022). How does eDNA decay affect metabarcoding experiments? *Environmental DNA*, 4(1), 108–116. <https://doi.org/10.1002/edn3.201>
- Jeppesen, E., Peder Jensen, J., Søndergaard, M., Lauridsen, T., & Landkildehus, F. (2000). Trophic structure, species richness and biodiversity in Danish lakes: Changes along a phosphorus gradient. *Freshwater Biology*, 45(2), 201–218. <https://doi.org/10.1046/j.1365-2427.2000.00675.x>
- Jing, S., & Taiyi. (2015). *International Organizations that Help Conserve Biodiversity and Ecosystem*. Personal Project in Global Perspectives Cempaka International School Damansara Presented by Shui Jing & Taiyi. <http://globalperspectives-sjty.weebly.com/1/post/2015/07/international-organizations-that-help-protect-biodiversity-and-ecosystem.html>
- Kubečka, J., Hohnová, E., Matěna, J., Peterka, J., Amarasinghe, U. S., Bonar, S. A., Hateley, J., Hickley, P., Suuronen, P., Tereschenko, V., Welcomme, R., & Winfield, I. J. (2009). The true picture of a lake or reservoir fish stock: A review of needs and progress. *Fisheries Research*, 96(1), 1–5. <https://doi.org/10.1016/j.fishres.2008.09.021>
- Laursen, K., Kahlert, J., & Frikke, J. (2005). Factors affecting escape distances of staging waterbirds. *Wildlife Biology*, 11(1), 13–19. [https://doi.org/10.2981/0909-6396\(2005\)11\[13:FAEDOS\]2.0.CO;2](https://doi.org/10.2981/0909-6396(2005)11[13:FAEDOS]2.0.CO;2)
- Mehner, T., Diekmann, M., Brämick, U., & Lemcke, R. (2005). Composition of fish communities in German lakes as related to lake morphology, trophic state, shore structure and human-use intensity. *Freshwater Biology*, 50(1), 70–85. <https://doi.org/10.1111/j.1365-2427.2004.01294.x>

- MiSeq System* | Focused power for targeted gene and small genome sequencing. (2022). Retrieved 23 November 2022, from <https://www.illumina.com/systems/sequencing-platforms/miseq.html>
- Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. *Royal Society Open Science*, 2(7), 150088. <https://doi.org/10.1098/rsos.150088>
- NCBI. (2022). Retrieved 23 November 2022, from <https://www.ncbi.nlm.nih.gov/home/download/>
- Persson, L. (1986). Temperature-induced shift in foraging ability in two fish species, roach (*Rutilus rutilus*) and perch (*Perca fluviatilis*): Implications for coexistence between poikilotherms. *The Journal of Animal Ecology*, 829–839.
- Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, 39(21), e145. <https://doi.org/10.1093/nar/gkr732>
- Rojahn, J., Gleeson, D. M., Furlan, E., Haeusler, T., & Bylemans, J. (2021). Improving the detection of rare native fish species in environmental DNA metabarcoding surveys. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 31(4), 990–997. <https://doi.org/10.1002/aqc.3514>
- Russell, N. J. (2000). Toward a molecular understanding of cold activity of enzymes from psychrophiles. *Extremophiles*, 4(2), 83–90. <https://doi.org/10.1007/s007920050141>
- Sellers, G. S., Di Muri, C., Gómez, A., & Hänfling, B. (2018). Mu-DNA: A modular universal DNA extraction method adaptable for a wide range of sample types. *Metabarcoding and Metagenomics*, 2, e24556.
- Sepulveda, A. J., Hutchins, P. R., Forstchen, M., McKeefry, M. N., & Swigris, A. M. (2020). The Elephant in the Lab (and Field): Contamination in Aquatic Environmental DNA Studies. *Frontiers in Ecology and Evolution*, 8. <https://www.frontiersin.org/articles/10.3389/fevo.2020.609973>
- Siikamäki, P., Kangas, K., Paasivaara, A., & Schroderus, S. (2015). Biodiversity attracts visitors to national parks. *Biodiversity and Conservation*, 24(10), 2521–2534. <https://doi.org/10.1007/s10531-015-0941-5>
- Šmilauer, P., & Lepš, J. (2014). *Multivariate Analysis of Ecological Data using CANOCO 5*. Cambridge University Press.
- Staley, Z. R., Chuong, J. D., Hill, S. J., Grabuski, J., Shokralla, S., Hajibabaei, M., & Edge, T. A. (2018). Fecal source tracking and eDNA profiling in an urban creek following an extreme rain event. *Scientific Reports*, 8(1), Article 1. <https://doi.org/10.1038/s41598-018-32680-z>

- Stewart, K. A. (2019). Understanding the effects of biotic and abiotic factors on sources of aquatic environmental DNA. *Biodiversity and Conservation*, 28(5), 983–1001.
- Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford University Press.
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050. <https://doi.org/10.1111/j.1365-294X.2012.05470.x>
- Tapirs. (2022). [Python]. EvoHull. <https://github.com/EvoHull/Tapirs> (Original work published 2020)
- Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.-M., Peroux, T., Crivelli, A. J., Olivier, A., Acqueberge, M., Le Brun, M., Møller, P. R., Willerslev, E., Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4), 929–942. <https://doi.org/10.1111/mec.13428>
- Ventura, M., Tiberti, R., Buchaca, T., Buñay, D., Sabás, I., & Miró, A. (2017). Why should we preserve fishless high mountain lakes? In *High mountain conservation in a changing world* (pp. 181–205). Springer, Cham.
- Vilmi, A., Karjalainen, S. M., & Heino, J. (2017). Ecological uniqueness of stream and lake diatom communities shows different macroecological patterns. *Diversity and Distributions*, 23(9), 1042–1053. <https://doi.org/10.1111/ddi.12594>
- Illumina, Inc, CA, USA. Retrieved 25 November 2022, from <https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html>

List of tables

Table 1 Basic characteristics and sampling effort in studied water bodies.

3

List of figures

Figure 1 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2018-summer sampling event. **Error! Bookmark not defined.**

Figure 2 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2018-winter sampling event. **Error! Bookmark not defined.**

Figure 3 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2019-summer sampling event. **Error! Bookmark not defined.**

Figure 4 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2019-autumn sampling event. **Error! Bookmark not defined.**

Figure 5 Numbers of different taxa of fish, other vertebrates, bacteria and human in Klicava-2020-autumn sampling event. **Error! Bookmark not defined.**

Figure 6 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2018-summer sampling event. **Error! Bookmark not defined.**

Figure 7 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2018-winter sampling event. **Error! Bookmark not defined.**

Figure 8 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2019-summer sampling event. **Error! Bookmark not defined.**

Figure 9 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2019-autumn sampling event. **Error! Bookmark not defined.**

Figure 10 Numbers of different taxa of fish, other vertebrates, bacteria and human in Rimov-2020-autumn sampling event. **Error! Bookmark not defined.**

Figure 11 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2018-summer sampling event. **Error! Bookmark not defined.**

Figure 12 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2018-winter sampling event. **Error! Bookmark not defined.**

Figure 13 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2019-summer sampling event. **Error! Bookmark not defined.**

Figure 14 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2019-autumn sampling event. **Error! Bookmark not defined.**

Figure 15 Numbers of different taxa of fish, other vertebrates, bacteria and human in Zlutice-2020-autumn sampling event. **Error! Bookmark not defined.**

Figure 16 Numbers of different taxa of fish, other vertebrates, bacteria and human in Laz-2020 sampling event. **Error! Bookmark not defined.**

Figure 17 Numbers of different taxa of fish, other vertebrates, bacteria and human in Sumava-2020 sampling event. **Error! Bookmark not defined.**

Figure 18 Numbers of different taxa of fish, other vertebrates, bacteria and human in Sumava-2021-2022 sampling event. **Error! Bookmark not defined.**

Figure 19 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pravikov-2019-summer sampling event. **Error! Bookmark not defined.**

Figure 20 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pravikov-2019-winter sampling event. **Error! Bookmark not defined.**

Figure 21 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pravikov-2020-autumn sampling event. **Error! Bookmark not defined.**

Figure 22 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kalich-2019-summer sampling event. **Error! Bookmark not defined.**

Figure 23 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kalich-2019-winter sampling event. **Error! Bookmark not defined.**

Figure 24 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kladiny-2019-summer sampling event. **Error! Bookmark not defined.**

Figure 25 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kladiny-2019-winter sampling event. **Error! Bookmark not defined.**

Figure 26 Numbers of different taxa of fish, other vertebrates, bacteria and human in Kladiny-2020-autumn sampling event. **Error! Bookmark not defined.**

Figure 27 Numbers of different taxa of fish, other vertebrates, bacteria and human in Tuska-2020-autumn sampling event. **Error! Bookmark not defined.**

Figure 28 Numbers of different taxa of fish, other vertebrates, bacteria and human in Pisek-2020 sampling event. **Error! Bookmark not defined.**

List of Appendices

Appendix 1 Code to extract sequences not assigned by tapirs workflow:

```
#!/usr/bin/env python3
import os
import csv
import argparse
import datetime
from tqdm import tqdm
from Bio import SeqIO
from pathlib import Path

"""
Extract Unassigned:

Extract Unassigned sequences from Tapirs
pipeline using the LCA and FASTA files.
"""

DATETIME = datetime.datetime.now().strftime("%d%m%Y_%H%M%S")

PATH_RESULT = "{DATETIME}_Extract_Unassigned".format(DATETIME=DATETIME)

def getArguments():
    """Get arguments from terminal

    This function gets arguments from terminal via argparse

    Returns
    -----
    arguments: Namespace
        Namespace object with all arguments
    """

    parser = argparse.ArgumentParser(
        description='Extract Unassigned sequences from the Tapirs pipeline result')
    parser.add_argument('folder_fasta', type=str,
                        help='A folder with FASTA files')
    parser.add_argument('folder_lca', type=str,
                        help='A folder with LCA files')

    return parser.parse_args()

def extractUnassigned(fasta, lca):
    record_dict = SeqIO.to_dict(SeqIO.parse(fasta, "fasta"))

    with open(lca, newline="") as lca_file:
        reader = csv.DictReader(lca_file, delimiter='\t')
        for row in reader:
            del record_dict[row['query']]

    name = os.path.basename(fasta).split(".")[0]
    with open("{PATH_RESULT}/{name}.fasta".format(PATH_RESULT=PATH_RESULT, name=name), "w") as handle:
        SeqIO.write(record_dict.values(), handle, "fasta")

if __name__ == "__main__":
    args = getArguments()

    Path(PATH_RESULT).mkdir(parents=True, exist_ok=True)

    fasta_files = sorted(Path(args.folder_fasta).rglob("*.nc.fasta"))
```

```
lca_files = sorted(Path(args.folder_lca).rglob("*.lca.tsv"))
for fasta, lca in tqdm(zip(fasta_files, lca_files), desc="Extracting Unassigned"):
    extractUnassigned(fasta, lca)
```

Appendix 2 Code to create compatible file with bacteria:

```
#reworking extracted bacteria files
for (file in list.files('F:/Uni/Thesis/final_files/'))
{
  file_list <- c()
  for (filename in list.files(paste('F:/Uni/Thesis/outs/', strsplit(file, '_')[[1]][1], sep = '')))
  {
    file_list <- append(file_list, strsplit(filename, '_')[[1]][1])
  }
  #additional items for creating dataframe
  file_list <- c('Accession_Number', file_list)
  file_list <- append(file_list, 'taxonomy')
  m <- data.frame(matrix(nrow = 80, ncol = (length(file_list))))
  colnames(m) <- file_list
  counter <- 1
  #looping through list of files
  for (bact_file in list.files(paste('F:/Uni/Thesis/outs/', strsplit(file, '_')[[1]][1], sep = '')))
  {
    sample_name <- strsplit(bact_file, '_')[[1]][1]
    file_size <- file.info(paste('F:/Uni/Thesis/outs/', strsplit(file, '_')[[1]][1], '/', bact_file, sep = ''))$size
    if (!file_size == 0)
    {
      #read file
      read_bact_file <- read.table(paste('F:/Uni/Thesis/outs/', strsplit(file, '_')[[1]][1], '/', bact_file, sep = ''))
      #adding items to the dataframe
      if (!is.numeric(m[counter, sample_name]))
      {
        for (a in 1:nrow(read_bact_file))
        {
          the_row <- which(m$Accession_Number == read_bact_file$V2[a])
          if (!(read_bact_file$V2[a] %in% m$Accession_Number))
          {
            m$Accession_Number[counter] <- read_bact_file$V2[a]
            bacteria_size <- tail(strsplit(tail(strsplit(read_bact_file$V1[a], ':')[[1]], 1), '=')[[1]], 1)
            m[counter, sample_name] = as.integer(bacteria_size)
            m$taxonomy[counter] <- 'bacteria'
            counter <- counter + 1
          }
        }
      }
      else
      {
        if (!is.numeric(m[the_row, sample_name]))
        {
          bacteria_size <- tail(strsplit(tail(strsplit(read_bact_file$V1[a], ':')[[1]], 1), '=')[[1]], 1)
          m[the_row, sample_name] = as.integer(bacteria_size)
        }
      }
    }
  }
}
m <- m[rowSums(is.na(m)) != ncol(m),]
m[is.na(m)] <- 0
#saving data to tsv file
write.table(m, file = paste('F:/Uni/Thesis/new_outs/', paste(strsplit(file, '_')[[1]][1], '_bact_denoise.tsv', sep = '')), col.names = F,
```

```
row.names = F, sep = '\t')
}
```

Appendix 3 Code to calculate the number of total reads:

```
library(rChoiceDialogs)
library(data.table)

tables.path = list.files('E:/Uni/Thesis/final_merged_files/', pattern = "*.tsv", full.names = TRUE ,recursive = TRUE)
for (file.name in tables.path) {

# Parse base name and dir name
base.name = basename(file.name)
dir.name = dirname(file.name)

# Read file
dat = read.csv(file.name, sep = '\t', header=T, row.names = 1, check.names=F)
dat = dat[, !colnames(dat) %in% "taxonomy",]

# Transpose data
dat = as.data.frame(t(dat))

#calculate total
dat$total = rowSums(dat[, !colnames(dat) %in% c("assigned", "total")])
dat = as.data.frame(t(dat))

#write to file
file.result.name <- paste(dir.name, '/totalled_', base.name, sep = '')
fwrite(x=dat, file=file.result.name, sep = "\t", row.names = T)
}
```

Appendix 4 Code to filter threshold:

```
library(rChoiceDialogs)
library(data.table)

# Threshold of 0.1%
# It needs to be divided by 100 in R
threshold = 0.001

# Get table file names
tables.path = list.files('E:/Uni/Thesis/final_merged_files/total_added/', pattern = "*.tsv", full.names = TRUE ,recursive = TRUE)

for (file.name in tables.path) {

# Parse base name and dir name
base.name = basename(file.name)
dir.name = dirname(file.name)

# Read file
dat = read.csv(file.name, sep = '\t', header=T, row.names = 1, check.names=F)
# Remove last column "taxonomy" if you did not remove it#
#dat = dat[, !colnames(dat) %in% "taxonomy",]

# Transpose data
dat = as.data.frame(t(dat))

# Create 'datt': a proportion reads data frame#
datt = dat / dat$total
datt[is.na(datt)] <- 0
}
```

```

# Apply threshold to proportions of datt to dat#
dat[datt < threshold] = 0

# Calculate assigned
dat$assigned = rowSums(dat[, !colnames(dat) %in% c("unassigned", "assigned", "total")])

# Calculate unassigned
dat$unassigned = dat$total - dat$assigned

# Transpose back
dat = as.data.frame(t(dat))

# Remove species that sum zero
dat = dat[rowSums(dat) > 0,]

# Export dataframe to CSV
file.result.name <- paste(dir.name, 'Filtered_', base.name, sep = "")
fwrite(x=dat, file=file.result.name, sep = "\t", row.names = T)
}

```

Appendix 5 Code to distribute species and create graphs:

```

library(taxize)
#taxize::use_entrez()
#usethis::edit_r_environ()
library(ggplot2)
library(cowplot)
library(scales)

#creating a list of files to be processed
tables_path = list.files('F:/Uni/Thesis/final_merged_files/total_added/finished/', pattern = "*.tsv", full.names = TRUE, recursive = TRUE)
#tables_path = 'F:/Uni/Thesis/final_merged_files/total_added/finished/Final_Zlutice-2020-autumn.tsv'
for (file_name in tables_path)
{
  #Parse base name and dir name
  base_name <- basename(file_name)
  dir_name <- dirname(file_name)
  #read file
  my_file <- read.csv(file_name, sep = "\t", row.names = 1)
  #transpose data
  my_file <- as.data.frame(t(my_file))
  row_names <- rownames(my_file)
  file_names_list <- c()
  species_names_list <- c()
  species_idenfier_list <- c()
  species_count_list <- c()
  count <- 0
  temp_var <- 'try_me'
  temp_name <- 'try_you'

  #loop through rows(samples)
  for (row in 1:nrow(my_file))
  {
    #variables to check presence of group in particular sample
    fish_test <- 0
    homo_test <- 0
    vert_test <- 0
    bact_test <- 0
    species_counter <- 0
    #checking if column belongs to regular or control sample
    if (!grepl("[0-9]", strsplit(strsplit(row_names[row], '_')[[1]][2], split = "')[[1]][1]))
    {

```

```

#check that calculation of the mean only happens once
while (count < 1)
{
  count = count + 1
  mean_df <- data.frame(a = species_identifyer_list, b = species_names_list, c = species_count_list)
  mean_fish <- c()
  mean_vert <- c()
  mean_homo <- c()
  mean_bact <- c()
  already_done_list <- c()

  #calculating the mean of each taxon over all regular samples
  for (mean_row in 1:nrow(mean_df))
  {
    taxonomy <- mean_df$a[mean_row]
    spec_name <- mean_df$b[mean_row]
    #calculate mean for fish
    if ((taxonomy == 'fish') & (!(spec_name %in% already_done_list)) & (!is.na(spec_name)))
    {
      some_list_fish <- c()
      already_done_list <- append(already_done_list, spec_name)
      for (other_row in 1:nrow(mean_df))
      {
        if ((mean_df$b[other_row] == spec_name) & (!is.na(mean_df$b[other_row])))
        {
          some_list_fish <- append(some_list_fish, mean_df$c[other_row])
        }
      }
      some_list_fish <- append(some_list_fish, rep(0, row-length(some_list_fish)))
      mean_fish <- append(mean_fish, mean(some_list_fish))
      species_count_list <- append(species_count_list, mean(some_list_fish))
      species_names_list <- append(species_names_list, spec_name)
      species_identifyer_list <- append(species_identifyer_list, 'fish')
      file_names_list <- append(file_names_list, paste(strsplit(rownames(my_file)[row], '_')[1], '_A_MEAN', sep = ''))
    }
    #calculate mean for human
    else if ((taxonomy == 'human') & (!(spec_name %in% already_done_list)) & (!is.na(spec_name)))
    {
      some_list_homo <- c()
      already_done_list <- append(already_done_list, spec_name)
      for (other_row in 1:nrow(mean_df))
      {
        if ((mean_df$b[other_row] == spec_name) & (!is.na(mean_df$b[other_row])))
        {
          some_list_homo <- append(some_list_homo, mean_df$c[other_row])
        }
      }
      some_list_homo <- append(some_list_homo, rep(0, row-length(some_list_homo)))
      mean_homo <- append(mean_homo, mean(some_list_homo))
      species_count_list <- append(species_count_list, mean(some_list_homo))
      species_names_list <- append(species_names_list, spec_name)
      species_identifyer_list <- append(species_identifyer_list, 'human')
      file_names_list <- append(file_names_list, paste(strsplit(rownames(my_file)[row], '_')[1], '_A_MEAN', sep = ''))
    }
    #calculate mean for other vertebrates
    else if ((taxonomy == 'other vertebrates') & (!(spec_name %in% already_done_list)) & (!is.na(spec_name)))
    {
      some_list_vert <- c()
      already_done_list <- append(already_done_list, spec_name)
      for (other_row in 1:nrow(mean_df))
      {
        if ((mean_df$b[other_row] == spec_name) & (!is.na(mean_df$b[other_row])))
        {
          some_list_vert <- append(some_list_vert, mean_df$c[other_row])
        }
      }
    }
  }
}

```

```

some_list_vert <- append(some_list_vert, rep(0, row-length(some_list_vert)))
mean_vert <- append(mean_vert, mean(some_list_vert))
species_count_list <- append(species_count_list, mean(some_list_vert))
species_names_list <- append(species_names_list, spec_name)
species_identifyer_list <- append(species_identifyer_list, 'other vertebrates')
file_names_list <- append(file_names_list, paste(strsplit(rownames(my_file)[row], '_')[1][1], '_A_MEAN', sep = ''))
}
#calculate mean for bacteria
else if ((taxonomy == 'bacteria') & !(spec_name %in% already_done_list)) & (!is.na(spec_name)))
{
  some_list_bact <- c()
  already_done_list <- append(already_done_list, spec_name)
  for (other_row in 1:nrow(mean_df))
  {
    if ((mean_df$B[other_row] == spec_name) & (!is.na(mean_df$B[other_row])))
    {
      some_list_bact <- append(some_list_bact, mean_df$c[other_row])
    }
  }
  some_list_bact <- append(some_list_bact, rep(0, row-length(some_list_bact)))
  mean_bact <- append(mean_bact, mean(some_list_bact))
  species_count_list <- append(species_count_list, mean(some_list_bact))
  species_names_list <- append(species_names_list, spec_name)
  species_identifyer_list <- append(species_identifyer_list, 'bacteria')
  file_names_list <- append(file_names_list, paste(strsplit(rownames(my_file)[row], '_')[1][1], '_A_MEAN', sep = ''))
}
}
}
#control samples
for (col in 1:(ncol(my_file)-3))
{
  if (!(my_file[row, col] == 0))
  {
    species_name <- colnames(my_file)[col]
    if (!grepl('[0-9]', strsplit(species_name, '_')[1][2]))
    {
      if (grepl('_', species_name))
      {
        species_name_uid <- paste(strsplit(species_name, '_')[1][1], strsplit(species_name, '_')[1][2], sep = ' ')
      }
      else
      {
        species_name_uid <- species_name
      }
    }
    try <- 0
    #handling recurring 'http 400 bad request' error
    while (try < 1)
    {
      tryCatch(
        expr =
        {
          #getting taxonomic information from NCBI
          uid <- get_uid(species_name_uid)
          out <- data.frame(classification(uid)[1])
          out_col <- colnames(out)[1]
        },
        error = function(e)
        {
          Sys.sleep(1)
          uid <- get_uid(species_name_uid)
          out <- data.frame(classification(uid)[1])
          out_col <- colnames(out)[1]
        },
        finally = {}
      )
    }
    if (temp_name == species_name | !temp_var == out_col)

```

```

    {
      try = try + 1
    }
  }
  temp_var <- out_col
  temp_name <- species_name
  #adding values to specific lists for data frame creation
  if ('Actinopteri' %in% out[, out_col] | 'Hyperoartia' %in% out[, out_col])
  {
    #list with species names
    species_names_list <- append(species_names_list, species_name_uid)
    #list with number of reads assigned to species
    species_count_list <- append(species_count_list, my_file[row, col])
    #identifier to further differentiate between fish, human, other vertebrates and bacteria
    species_identifyer_list <- append(species_identifyer_list, 'fish')
    species_counter = species_counter + 1
    #update variable to check if fish present in sample
    fish_test = fish_test + 1
  }
  else if ('Vertebrata' %in% out[, out_col])
  {
    if ('Homo sapiens' %in% out[, out_col])
    {
      species_names_list <- append(species_names_list, species_name_uid)
      species_count_list <- append(species_count_list, my_file[row, col])
      species_identifyer_list <- append(species_identifyer_list, 'human')
      species_counter = species_counter + 1
      homo_test = homo_test + 1
    }
    else if ('Aves' %in% out[, out_col] | 'Reptilia' %in% out[, out_col] | 'Amphibia' %in% out[, out_col] | 'Mammalia' %in% out[,
out_col])
    {
      species_names_list <- append(species_names_list, species_name_uid)
      species_count_list <- append(species_count_list, my_file[row, col])
      species_identifyer_list <- append(species_identifyer_list, 'other vertebrates')
      species_counter = species_counter + 1
      vert_test = vert_test + 1
    }
  }
}
else
{
  species_names_list <- append(species_names_list, species_name)
  species_count_list <- append(species_count_list, my_file[row, col])
  species_identifyer_list <- append(species_identifyer_list, 'bacteria')
  species_counter = species_counter + 1
  bact_test = bact_test + 1
}
}
}
#regular samples
else
{
  #loop through columns(species) of given row(sample)
  for (col in 1:(ncol(my_file)-3))
  {
    if (!(my_file[row, col] == 0))
    {
      species_name <- colnames(my_file)[col]
      if (!grepl('[0-9]', strsplit(species_name, '_')[[1]][2]))
      {
        if (grepl('_', species_name))
        {
          species_name_uid <- paste(strsplit(species_name, '_')[[1]][1], strsplit(species_name, '_')[[1]][2], sep = '')
        }
      }
    }
  }
}

```

```

else
{
  species_name_uid <- species_name
}
#handling recurring 'http 400 bad request' error
try <- 0
while (try < 1)
{
  tryCatch(
    expr =
    {
      #get taxonomic information from NCBI
      uid <- get_uid(species_name_uid)
      out <- data.frame(classification(uid)[1])
      out_col <- colnames(out)[1]
    },
    error = function(e)
    {
      Sys.sleep(1)
      uid <- get_uid(species_name_uid)
      out <- data.frame(classification(uid)[1])
      out_col <- colnames(out)[1]
    },
    finally = {}
  )
  if (temp_name == species_name | !temp_var == out_col)
  {
    try = try + 1
  }
}
temp_var <- out_col
temp_name <- species_name
#adding values to specific lists for data frame creation
if ('Actinopteri' %in% out[, out_col] | 'Hyperoartia' %in% out[, out_col])
{
  species_names_list <- append(species_names_list, species_name_uid)
  species_count_list <- append(species_count_list, my_file[row, col])
  species_identifyer_list <- append(species_identifyer_list, 'fish')
  species_counter = species_counter + 1
  fish_test = fish_test + 1
}
else if ('Vertebrata' %in% out[, out_col])
{
  if ('Homo sapiens' %in% out[, out_col])
  {
    species_names_list <- append(species_names_list, species_name_uid)
    species_count_list <- append(species_count_list, my_file[row, col])
    species_identifyer_list <- append(species_identifyer_list, 'human')
    species_counter = species_counter + 1
    homo_test = homo_test + 1
  }
  else if ('Aves' %in% out[, out_col] | 'Reptilia' %in% out[, out_col] | 'Amphibia' %in% out[, out_col] | 'Mammalia' %in% out[,
out_col])
  {
    species_names_list <- append(species_names_list, species_name_uid)
    species_count_list <- append(species_count_list, my_file[row, col])
    species_identifyer_list <- append(species_identifyer_list, 'other vertebrates')
    species_counter = species_counter + 1
    vert_test = vert_test + 1
  }
}
}
else
{
  species_names_list <- append(species_names_list, species_name)
  species_count_list <- append(species_count_list, my_file[row, col])

```

```

    species_identifyer_list <- append(species_identifyer_list, 'bacteria')
    species_counter = species_counter + 1
    bact_test = bact_test + 1
  }
}
}
}
#adding NAs and 0 to lists if sample empty for given group
if (fish_test == 0)
{
  species_names_list <- append(species_names_list, NA)
  species_count_list <- append(species_count_list, 0)
  species_identifyer_list <- append(species_identifyer_list, 'fish')
  species_counter = species_counter + 1
}
if (homo_test == 0)
{
  species_names_list <- append(species_names_list, NA)
  species_count_list <- append(species_count_list, 0)
  species_identifyer_list <- append(species_identifyer_list, 'human')
  species_counter = species_counter + 1
}
if (vert_test == 0)
{
  species_names_list <- append(species_names_list, NA)
  species_count_list <- append(species_count_list, 0)
  species_identifyer_list <- append(species_identifyer_list, 'other_vertebrates')
  species_counter = species_counter + 1
}
if (bact_test == 0)
{
  species_names_list <- append(species_names_list, NA)
  species_count_list <- append(species_count_list, 0)
  species_identifyer_list <- append(species_identifyer_list, 'bacteria')
  species_counter = species_counter + 1
}
file_names_list <- append(file_names_list, rep(rownames(my_file)[row], species_counter))
}
#get a suiting name for the plot
plot_name <- strsplit(strsplit(base_name, '[_]')[[1]][2], '[.]')[[1]][1]
#creating a dataframe for plotting
plot_data <- data.frame(a = file_names_list, b = species_identifyer_list, c = species_names_list, d = species_count_list)
#create multiple data frames for different groups
fish_plot_data <- data.frame()
homo_plot_data <- data.frame()
vert_plot_data <- data.frame()
bact_plot_data <- data.frame()
#loop through df to detect if it should be assigned to fish, human, other vertebrate, bacteria df
for (r in 1:nrow(plot_data))
{
  if (plot_data$b[r] == 'fish')
  {fish_plot_data <- rbind(fish_plot_data, plot_data[r,])}
  else if (plot_data$b[r] == 'human')
  {homo_plot_data <- rbind(homo_plot_data, plot_data[r,])}
  else if (plot_data$b[r] == 'other_vertebrates')
  {vert_plot_data <- rbind(vert_plot_data, plot_data[r,])}
  else if (plot_data$b[r] == 'bacteria')
  {bact_plot_data <- rbind(bact_plot_data, plot_data[r,])}
}
#Drawing the graphs
#fish graph
fish_col <- hue_pal()(length(unique(fish_plot_data$c)))
fish_plot <- ggplot(fish_plot_data, aes(x = a, y = d, fill = c)) +
  geom_bar(position = 'stack', stat = 'identity') +
  theme(axis.text.x = element_text(size = 8), axis.title = element_text(size = 10, face = 'bold')) +
  theme(axis.text.x = element_text(angle = 90, face = 'bold')) +

```

```

theme(legend.title = element_blank()) +
xlab('samples') +
ylab('number of individuals') +
theme(legend.position = 'right', legend.direction = 'vertical') +
theme(legend.key.size = unit(0.3,'cm'), legend.key.height = unit(0.3,'cm'), legend.key.width = unit(0.3,'cm')) +
ggtitle(paste(plot_name, ' - Fishes', sep = "")) +
theme(axis.title = element_text(size = 8), plot.title = element_text(size = 12, face = 'bold')) +
expand_limits( x = c(0,NA), y = c(0,NA)) +
scale_y_continuous(labels = scales::comma) +
#color for NA values == white to include empty samples
scale_fill_manual(values = fish_col, aesthetics = 'fill', breaks = waiver(), na.value = 'grey100')
#bacteria graph
vert_col <- hue_pal()(length(unique(vert_plot_data$c)))
vert_plot <- ggplot(vert_plot_data, aes(x = a, y = d, fill = c)) +
geom_bar(position = 'stack', stat = 'identity') +
theme( axis.text.x = element_text(size = 8), axis.title = element_text(size = 10, face = 'bold')) +
theme(axis.text.x = element_text(angle = 90, face = 'bold')) +
theme(legend.title = element_blank()) +
xlab('samples') +
ylab('number of individuals') +
theme(legend.position = 'right', legend.direction = 'vertical') +
theme(legend.key.size = unit(0.3,'cm'), legend.key.height = unit(0.3,'cm'), legend.key.width = unit(0.3,'cm')) +
ggtitle(paste(plot_name, ' - other_Vertebrates', sep = "")) +
theme(axis.title = element_text(size = 8), plot.title = element_text(size = 12, face = 'bold')) +
expand_limits( x = c(0,NA), y = c(0,NA)) +
scale_y_continuous(labels = scales::comma) +
scale_fill_manual(values = vert_col, aesthetics = 'fill', breaks = waiver(), na.value = 'grey100')
#bacteria graph
bact_col <- hue_pal()(length(unique(bact_plot_data$c)))
bact_plot <- ggplot(bact_plot_data, aes(x = a, y = d, fill = c)) +
geom_bar(position = 'stack', stat = 'identity') +
theme( axis.text.x = element_text(size = 8), axis.title = element_text(size = 10, face = 'bold')) +
theme(axis.text.x = element_text(angle = 90, face = 'bold')) +
theme(legend.title = element_blank()) +
xlab('samples') +
ylab('number of individuals') +
theme(legend.position = 'right', legend.direction = 'vertical', ) +
theme(legend.key.size = unit(0.3,'cm'), legend.key.height = unit(0.3,'cm'), legend.key.width = unit(0.3,'cm')) +
ggtitle(paste(plot_name, ' - Bacteria', sep = "")) +
theme(axis.title = element_text(size = 8), plot.title = element_text(size = 12, face = 'bold')) +
expand_limits( x = c(0,NA), y = c(0,NA)) +
scale_y_continuous(labels = scales::comma) +
scale_fill_manual(values = bact_col, aesthetics = 'fill', breaks = waiver(), na.value = 'grey100')
#human graph
homo_col <- hue_pal()(length(unique(homo_plot_data$c)))
homo_plot <- ggplot(homo_plot_data, aes(x = a, y = d, fill = c)) +
geom_bar(position = 'stack', stat = 'identity') +
theme( axis.text.x = element_text(size = 8), axis.title = element_text(size = 10, face = 'bold')) +
theme(axis.text.x = element_text(angle = 90, face = 'bold')) +
theme(legend.title = element_blank()) +
xlab('samples') +
ylab('number of individuals') +
theme(legend.position = 'right', legend.direction = 'vertical', ) +
theme(legend.key.size = unit(0.3,'cm'), legend.key.height = unit(0.3,'cm'), legend.key.width = unit(0.3,'cm')) +
ggtitle(paste(plot_name, ' - Human', sep = "")) +
theme(axis.title = element_text(size = 8), plot.title = element_text(size = 12, face = 'bold')) +
expand_limits( x = c(0,NA), y = c(0,NA)) +
scale_y_continuous(labels = scales::comma) +
scale_fill_manual(values = homo_col, aesthetics = 'fill', breaks = waiver(), na.value = 'grey100')
#saving graph as png
png(file = paste('F:/Uni/Thesis/plots/onecol/', plot_name, '.png', sep = ""), width = 1700, height = 2300, res = 150)
abc <- plot_grid(fish_plot, vert_plot, bact_plot, homo_plot, ncol = 1, nrow = 4)
print(abc)
dev.off()
}

```