Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Microbial diversity characterization of seawater in a pilot study using Oxford Nanopore Technologies long-read sequencing

Michael Liem (m.liem@biology.leidenuniv.nl)
Universiteit Leiden Instituut Biologie Leiden
Tonny Regensburg-Tuïnk
Universiteit Leiden Instituut Biologie Leiden
Christiaan Henkel
nmbu
Hans Jansen
Future Genomics Technology
Herman Spaink
Universiteit Leiden Instituut Biologie Leiden

Research note

Keywords: Metagenomics, Oxford Nanopore Technology, MinION sequencing, oceanic microbiome, k-mer analysis, genome assembly,

Posted Date: November 11th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17068/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published on February 2nd, 2021. See the published version at https://doi.org/10.1186/s13104-021-05457-3.

Abstract

Objective: Currently the majority of non-culturable microbes in sea water are yet to be discovered, Nanopore offers a solution to overcome the challenging tasks to identify the genomes and complex composition of oceanic microbiomes. In this study we evaluate the utility of Oxford Nanopore Technologies (ONT) sequencing to characterize microbial diversity in seawater from multiple locations. We compared the microbial species diversity of retrieved environmental samples from two different locations and time points.

Results: With only three ONT flow cells we were able to identify thousands of organisms, including bacteriophages, from which a large part at species level. It was possible to assemble genomes from environmental samples with Flye. In several cases this resulted in >1 Mbp contigs and in the particular case of a *Thioglobus singularis* species it even produced a near complete genome. k-mer analysis reveals that a large part of the data represents species of which close relatives have not yet been deposited to the database. These results show that our approach is suitable for scalable genomic investigations such as monitoring oceanic biodiversity and provides a new platform for education in biodiversity.

Introduction

Although marine microbes have been studied for multiple decades there is still little knowledge on species diversity in the largest ecological environments of our planet [1-3]. Current database collections are estimated to represent <5% of oceanic microbial communities [4].

Large-scale metagenomics analyses of seawater have been performed already since 2004 showing remarkable species diversity [5]. However, even with availability of abundant sequencing technology resources a complete understanding on the entire diversity remains a challenging task. Recent studies focussing on marine biodiversity show that a variety of sediments harbour different ecosystems that are particularly extreme in deep ocean environments. There have been many exploratory studies of harnessing marine microorganism for the production of bioactive compounds, with versatile medicinal, industrial, or agricultural applications [6].

Microbial diversity characterization has primarily relied on traditional high-throughput short-read sequencing methods, such as Illumina [7-12] or 454 sequencing [5]. Even though Pacific Biosciences single-molecule long-read sequencing has been used to catalogue the diversity of coral-associated microbial communities, these studies require amplification and 16S rRNA homology to position microbes taxonomically [5,7, 9-11, 13-15].

In this pilot study we evaluate the utility of Oxford Nanopore Technologies (ONT) sequencing to characterize microbial diversity in seawater. Our strategy is based on a method to analyze riverine samples [33] and aims to classify microbial diversification directly from environmental samples with minimal computational and financial cost over a relatively short time span. This will facilitate future

scalable investigations such as monitoring oceanic biodiversity and landscape the time and space dynamics these microbes are subject to.

Results

Sample collection, data quality control and verification of microbial content

We collected samples from coastal regions of both the Atlantic Ocean (west part of the English Channel – Roscoff, France, August 2017) and the south part of the North Sea (Wassenaarseslag, the Netherlands, July 2017 and August 2018). From here on, we refer to these as samples 1, 2 and 3, respectively. MinION 48-hour sequencing runs on every sample resulted in three datasets, particularly for sample 1 data statistics appear relatively suboptimal compared to data from laboratory cultures (**Figure 1 A**). We used the top 3 longest reads to assess data quality (**additional file 1**) and used 16S rRNA primers to confirm microbial DNA isolates (**additional file 2**).

Read length and quality distributions of MinION sequencing runs

Figure 1 A Read length and quality distributions of 48-hour run sequencing data for sample 1, 2 and 3 (from left to right). Mean read lengths vary from 1,511 up to 7,983 bp with similar base call qualities (around PHRED 12). Plots are based on NanoPlot plotting [23]. **B** Taxonomic tree on a subset of the data generated from sample 1 data. Every node stands for a taxonomical ID that is supported with at least 831 reads. In red the most abundant species present in all three samples. Dark blue nodes together with the red node highlight the top-5 most abundantly present species in this sample.

Seawater characterization using k-mer classification

Using OneCodex [26] we generated classification trees for the three datasets. These are built from raw sequencing data and indicate the taxonomic relation between the detected microbial classes. This relation is based on taxonomic identifiers (taxids) provided by the NCBI taxonomy database.

Despite the fact that a large part of all three datasets could not be classified (47%, 69% and 38% for sample 1, 2 and 3, respectively) (**additional file 7**), all taxonomic trees highlight the complexity of microbial communities present at a single site. None of our three datasets reveal an overall dominant species, the largest differences between samples appear at low abundances. However 4.46% (sample 1), 15.66% (sample 2) and 7.82% (sample 3) of classified reads belong to *Planktomarina temperata* (**Figure 1 B** and **additional file 3**, red node), which is therefore the most abundant species present in the three data sets combined. Please refer to **additional file 3** for more highlights on classification trees of all three samples

Comparison of Onecodex species classification between different locations and time interval and overall identified ranks

Figure 2 A) Venn diagram comparison of identified species by OneCodex, highlighting species that are time and space dependent. **B)** Overall OneCodex classification ranks per dataset, the majority of classified reads have been linked to a species level

The taxonomic levels assigned by OneCodex range from kingdom down to species-specific. Reads that cannot be linked to a particular taxonomic level are labelled 'no rank'. In total 1,750, 3,017 and 2,007 taxids are assigned to the data of sample 1, 2 and 3, respectively. More than half of the ranks that OneCodex was able to classify are assigned to species level **(Figure 2 B**) in all three samples.

Interestingly, at least 484 microbes are identified in all samples (**Figure 2 A**). Some highlights include: 92 different Flavobacteriaceae bacterium and Flavobacteriales bacterium strains; 19 different *Candidatus* Pelagibacter strains; 18 Pelagibacteraceae bacterium and 6 SAR strains. This indicates that these communities are less time and location dependent compared to the 262 and 1,127 species that were found exclusively in France or Dutch areas, respectively. Furthermore, 607 and 129 species are exclusively observed in the Netherlands. As they exist at different times, they provide an initial impression of the time-dependent dynamics of these local communities. Finally, 135 and 77 species could be identified that are present at both locations, however only detectable at particular times. This could be an indication that even over large areas microbes are subject to time regulated dynamics.

Metagenomics assembly on raw sequencing data and blast verification on the top-3 longest contigs

In an attempt to verify OneCodex classification results as well as to assess the current metagenomics assemblers capabilities we subsequently assembled the three datasets separately. We have assembled our complex metagenomics datasets with Flye and retrieved 256, 1,735 and 968 contigs with mean coverage of 14x, 13x and 10x from samples 1, 2 and 3, respectively (**Table 1**). Notably, although it has higher coverage, assembly results from sample 2 did not exceed results from sample 3. On the contrary, sample 3 resulted better average contig length, maximum contig length and N50 values compared to sample 2 (**Table 1**).

Table 1 *Flye assembly statistics*

Assembly stats	France (1)	The Netherlands '17 (2)	The Netherlands '18 (3)
contigs	256	1,735	968
length (bp)	8,678,102	107,863,873	94,117,952
min length (bp)	2,432	536	494
mean length (bp)	33,898	62,169	97,229
max length (bp)	219,363	1,098,797	1,648,106
N50	40,621	75,928	153,524

Impressively, Flye was able to reconstruct a full genome from our third sample: 75% of our 1.6 Mbp contig aligns with 80% identity to *Candidatus* Thioglobus singularis of which its complete genome is a single circular chromosome of 1.7 Mbp (**additional file 4**). Additionally we show that OneCodex was able to identify certain species only using assembly results (**additional file 5**).

Data quality of unclassified reads and additional in silico PCR analysis

Poor read quality and relatively short read lengths could be a potential reason explaining why OneCodex was unable to classify taxids. Therefore, we investigated quality and length of unclassified reads (**additional file 6**). These statistics indicate that, in theory, these reads should provide OneCodex with sufficient information to resolve classifications. That OneCodex was not able to classify these reads, even to the most general taxonomic levels (such as kingdom or phylum) adds to the notion that these reads originate from species that are novel.

Inspection of low complexity regions in unclassified reads using tandem repeat analysis

An additional circumstance that might explain why reads are left unclassified is the presence of low complexity regions such as repeat elements. We have analysed the presence of repeat elements with Tandem Repeat Finder [22] in raw sequencing data and compared these to repeat counts of the unclassified reads. In none of our samples did we observe an increased presence of repetitive elements, on the contrary, the repetitive element count is lowered in every case (**additional file 8**).

Materials And Method

Please refer to **additional file 9** for descriptions on 1) sample collection and DNA isolation, 2) OneCodex k-mer based characterization 3) repetitive content analysis and 4) data visualisation

DNA library preparation, sequencing, data quality control and statistics

DNeasy powerwater kit (Qiagen) was used to isolate DNA, according to manufacturer's protocol with three additional enzymes (**additional file 9**). We used R9.4 flow cells for sequencing all three seawater samples. Libraries were prepared using rapid kits (SQK-RAD004) according to the manufacturer protocols available at that time (Oxford Nanopore Technologies, Oxford, UK). Data acquisition and base-calling were performed by MinKNOW (v19.06.8).

Using in silico *PCR analysis to verify microbial genomes*

To highlight the presence of microbial genomes FastPCR [24] was used to perform *in silico* PCR analysis using primer pair sequences for identification of bacteria and archaea. FastPCR allows users to upload a set of primer sequences and reports, among others, positions and length of hits found on the input data. We used the currently 'best available' rRNA primer pair, primer 1 and 2 are 17 and 21 bp long, respectively, with a total amplicon size of 464 bp (primer 1: 5'-CCTACGGGNGGCNGCAG-3', primer 2: 5'-GACTACNNGGGTATCTAATCC-3').

Assembly of long read metagenomics samples using the Flye assembler

Flye [27] is currently one of the few *de novo* assembly pipelines that allows genomic reconstruction of complex metagenomics samples with coverage as low as 2x. We have downloaded the assembly

software from the GitHub repository (v2.6), used the metagenome default settings and provided the raw sequencing data.

Discussion

In this study, we have investigated the use of Nanopore sequencing for seawater metagenomics. Our main aims were to investigate the effectiveness of DNA isolation from samples directly obtained from the environment, optimize laboratory protocols for maximum sequencing results and evaluation of current metagenomics identification and assembly software. We used multiple isolation procedures, several different storage methods and subjected the data to a set of different analysis software. With only three ONT flow cells we were able to identify thousands of organisms, including bacteriophages, from which a large part at species level. It was possible to assemble genomes from environmental samples with Flye. In several cases this resulted in >1 Mbp contigs and in the particular case of a *Thioglobus singularis* species it even produced a near complete genome.

While OneCodex was able to identify the diversity of a substantial amount of our samples, it could not resolve any classification for a large part of our data. The large k-mer size is most probably a crucial factor for unclassified data, due to the relatively low quality (approximately 10% error) of long-read data 10 bp would be a more suitable k-mer size. We confirmed that the data quality of these reads (both read length and quality distributions) are within acceptable bounds and observed no particular repetitive element enrichment compared to the reads that contributed to classifications

Despite the fact that these experiments are pilot studies, we have observed promising results for both laboratory protocols and species identifications analysis. As described above, sample collection, DNA isolation and species identification is still hindered by both technical and biological difficulties. However, this study provides a good impression that the elegance of the method originates from simplicity. We have performed equivalent experiments in student field practical assignments with similar marine samples, and students showed that even under more restricted conditions (12-hour sequencing runs) large biodiversity could still be detected.

Please refer to **additional file 10** for additional discussion

Limitations

This study focusses on the applicability of long read sequencing data and downstream analysis tools, further studies should take into consideration that; higher coverage data sets would contribute to a deeper understanding of oceanic microbial diversity. Additionally, strategically chosen locations and seasonal or fixed time points would provide a more relevant overview of the microbial diversity landscape and its dynamics. We have not performed comparative analysis for different sequencing platforms.

Abbreviations

ONT: Oxford Nanopore Technology

Taxids: taxonomic identifiers

Declarations

ACKNOWLEDGEMENT

We would like to express our gratitude to OneCodex for answering questions on the available genome selection and the help with the CLI, and IBL, Ing Mark Arendsthorst department Microbiology for sharing their equipment.

DATA AVAILABILITY

https://www.ncbi.nlm.nih.gov/bioproject/PRJNA611514

References

1. Zobell, C. E., Marine Microbiology, Chronica Botanica Co, Waltham, Mass., USA, 1946, p. 240.
2. Velankar, N. K., Bacteria isolated from seawater and marine mud off Mandapam (Gulf of Mannar and Palk Bay). Indian J. Fish., 1957, 4, 208–227.
3. Wood, E. J. F., Some aspects of marine microbiology. J. Mar. Biol. Assoc. India, 1959, 1, 26–32.
4. Salazar, Sunagawa. Marine microbial diversity. Curr Biol. 2017 Jun 5;27(11):R489-R494; doi: 10.1016/j.cub.2017.01.01
5. Rusch DB et al., The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007 Mar;5(3):e77; doi: 10.1371/journal.pbio.0050077
6. Dipesh Dhakal et al., Marine Rare Actinobacteria: Isolation, Characterization, and Strategies for Harnessing Bioactive Compounds (review). Front Microbiol. 2017; doi: 10.3389/fmicb.2017.01106
7. Ghai R et al., Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Sci Rep. 2013;3:2471; doi: 10.1038/srep02471
8. Planes S et al., The Tara Pacific expedition-A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean PLoS Biol. 2019 Sep 23;17(9):e3000483; doi: 10.1371/journal.pbio.3000483
9. Hamdan HZ et al., Characterization of the microbial community diversity and composition of the coast of Lebanon: Potential for petroleum oil biodegradation. Mar Pollut Bull. 2019 Dec;149:110508; doi: 10.1016/j.marpolbul.2019.110508
10. Tobias-Hünefeldt SP et al., Depth and location influence prokaryotic and eukaryotic microbial community structure in New Zealand fjords. Sci Total Environ. 2019 Nov 25;693:133507; doi: 10.1016/j.scitotenv.2019.07.313

11. Gong B et al., High-throughput sequencing and analysis of microbial communities in the mangrove swamps along the coast of Beibu Gulf in Guangxi, China. Sci Rep. 2019 Jun 28;9(1):9377; doi: 10.1038/s41598-019-45804-w
12. Venter JC et al., Environmental Genome Shotgun Sequencing of the Sargasso Sea. 2004 Apr 2;304(5667):66-74; doi: 10.1126/science.1093857
13. Martín-Cuadrado AB et al., Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. PLoS One. 2007 Sep 19;2(9):e914; doi: 10.1371/journal.pone.0000914
14. Pootakham W et al., High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. Sci Rep. 2017 Jun 5;7(1):2774; doi: 10.1038/s41598-017-03139-4
15. Willis C et al., Influence of 16S rRNA variable region on perceived diversity of marine microbial communities of the Northern North Atlantic. FEMS Microbiol Lett. 2019 Jul 1;366(13). pii: fnz152; doi: 10.1093/femsle/fnz152
16. Rohit Ghai et al., Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Sci Rep. 2013; doi: 10.1038/srep02471
17. Stephen J. Giovannoni et al., Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science 19 Aug 2005: Vol. 309, Issue 5738, pp. 1242-1245; doi: 10.1126/science.1114057
18. Rich VI et al., Time-series analyses of Monterey Bay coastal microbial picoplankton using a 'genome proxy' microarray. Environ Microbiol. 2011 Jan;13(1):116-134. doi: 10.1111/j.1462-2920.2010.02314.x
19. McCarren J et al., Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. Environ Microbiol. 2007 Apr;9(4):846-58; doi: 10.1111/j.1462-2920.2006.01203.x
20. de la Torre JR et al., Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc Natl Acad Sci U S A. 2003 Oct 28;100(22):12830-5. Epub 2003 Oct 17; doi: 10.1073/pnas.2133554100
21. Gómez-Pereira PR et al., Genomic content of uncultured Bacteroidetes from contrasting oceanic provinces in the North Atlantic Ocean. Environ Microbiol. 2012 Jan;14(1):52-66; doi: 10.1111/j.1462-2920.2011.02555.x
22. Benson,"Tandem repeats finder: a program to analyze DNA sequences" Nucleic Acids Research (1999) Vol. 27, No. 2, pp. 573-580.
23. Nanoplot Github webpage. https://github.com/wdecoster/NanoPlot, Accessed 30 January 2020.
24. Kalendar R, Khassenov B, Ramankulov Y, Samuilova O, Ivanov KI 2017. FastPCR: an in silico tool for fast primer and probe design and advanced sequence analysis. Genomics, 109: 312-319. DOI: 10.1016/j.ygeno.2017.05.005
25. Anna Klindworth et al., Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 2013 Jan; 41(1): e1; Published online 2012 Aug 28. doi: 10.1093/nar/gks808

26. One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification Samuel S Minot, Niklas Krumm, Nicholas B. GreenfieldPublished 2015 doi:10.1101/027607
27. Kolmogorov M. et al. (2018) Assembly of Long Error-Prone Reads Using Repeat Graphs. Bioinformatics, 35(13), 2019, 2303–2305; doi: 10.1093/bioinformatics/bty956
28. Marshall KT et al., Genome Sequence of "Candidatus Thioglobus singularis" Strain PS1, a Mixotroph from the SUP05 Clade of Marine Gammaproteobacteria. Genome Announc. 2015 Oct 22;3(5). pii: e01155-15. doi: 10.1128/genomeA.01155-15
29. Zhao Y et al., Abundant SAR11 viruses in the ocean. 2013 Feb 21;494(7437):357-60. doi: 10.1038/nature11921
30. Moreau H et al., Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J Virol. 2010 Dec;84(24):12555-63. doi: 10.1128/JVI.01123-10
31. Bartelme RP1 et al., Complete Genome Sequence of the Fish Pathogen Flavobacterium columnare Strain C#2 Genome Announc. 2016 Jun 23;4(3). pii: e00624-16. doi: 10.1128/genomeA.00624-16
32. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016
33. Kate Reddington et al., Metagenomic analysis of planktonic riverine microbial consortia using nanopore sequencing reveals insight into river microbe taxonomy and function, *GigaScience*, Volume 9, Issue 6, June 2020, giaa053,https://doi.org/10.1093/gigascience/giaa053

Figures

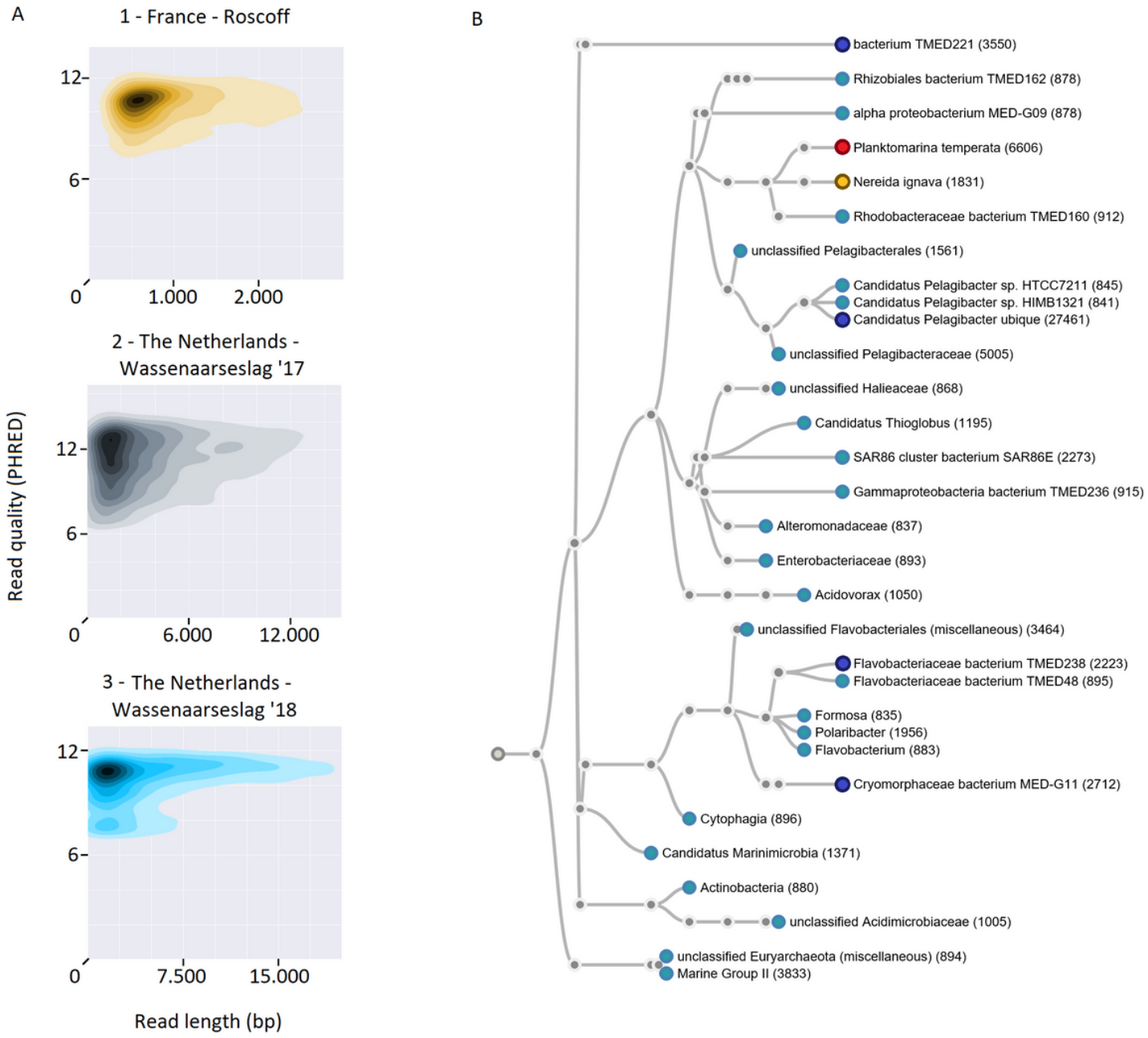

Figure 1

A Read length and quality distributions of 48-hour run sequencing data for sample 1, 2 and 3 (from left to right). Mean read lengths vary from 1,511 up to 7,983 bp with similar base call qualities (around PHRED 12). Plots are based on NanoPlot plotting [23]. B Taxonomic tree on a subset of the data generated from sample 1 data. Every node stands for a taxonomical ID that is supported with at least 831 reads. In red the most abundant species present in all three samples. Dark blue nodes together with the red node highlight the top-5 most abundantly present species in this sample.

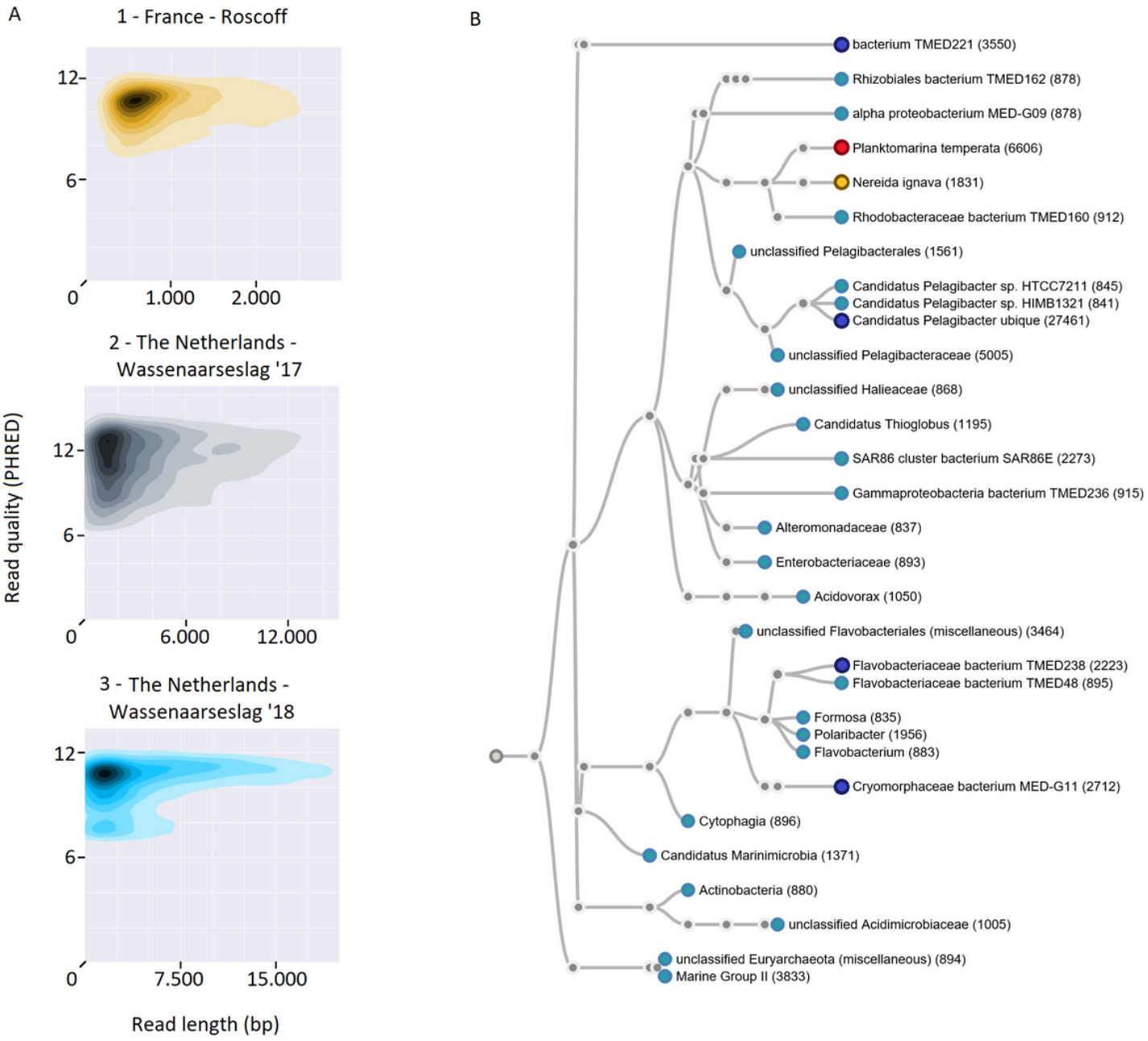

Figure 1

A Read length and quality distributions of 48-hour run sequencing data for sample 1, 2 and 3 (from left to right). Mean read lengths vary from 1,511 up to 7,983 bp with similar base call qualities (around PHRED 12). Plots are based on NanoPlot plotting [23]. B Taxonomic tree on a subset of the data generated from sample 1 data. Every node stands for a taxonomical ID that is supported with at least 831 reads. In red the most abundant species present in all three samples. Dark blue nodes together with the red node highlight the top-5 most abundantly present species in this sample.

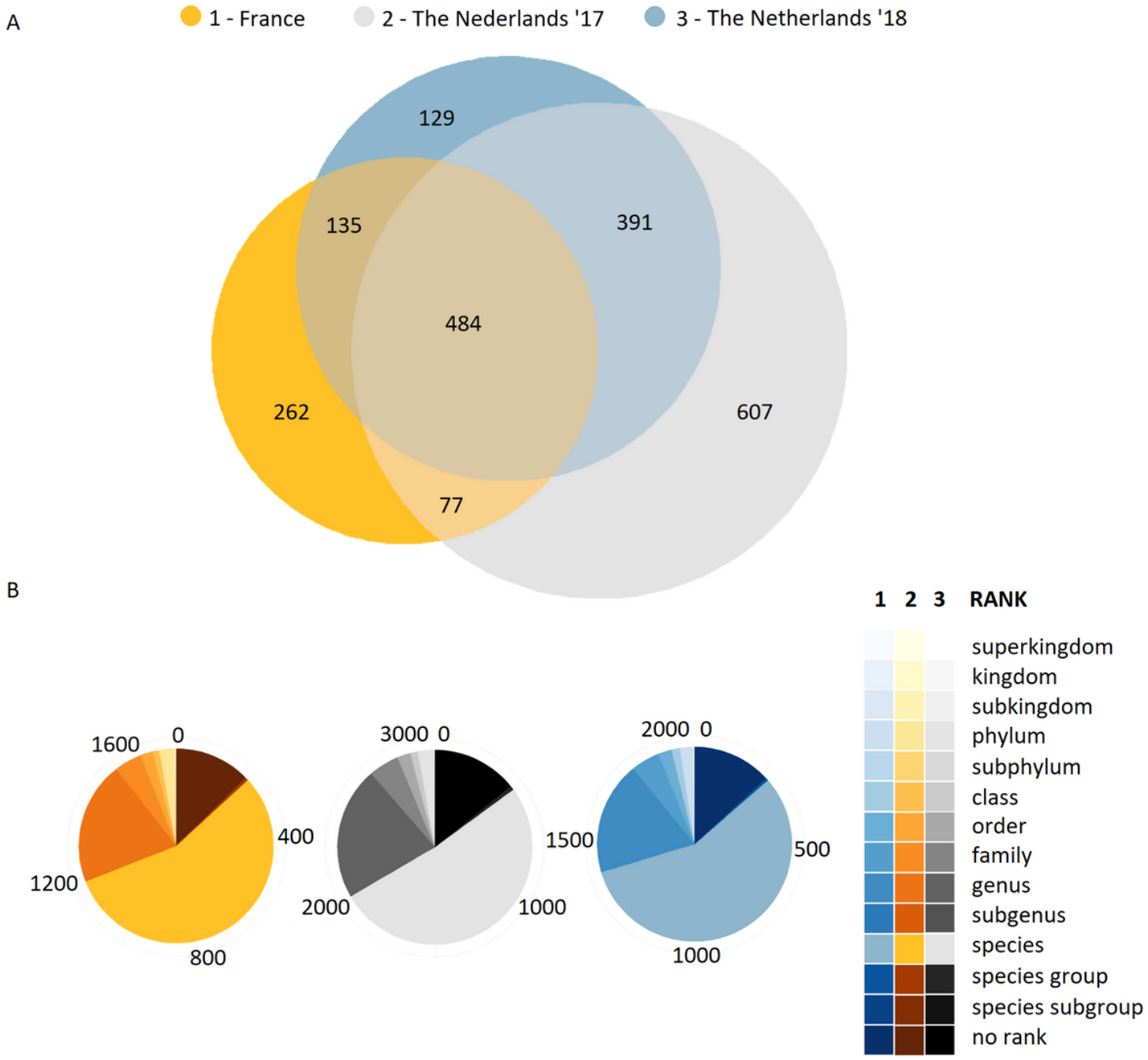

Figure 2

A) Venn diagram comparison of identified species by OneCodex, highlighting species that are time and space dependent. B) Overall OneCodex classification ranks per dataset, the majority of classified reads have been linked to a species level

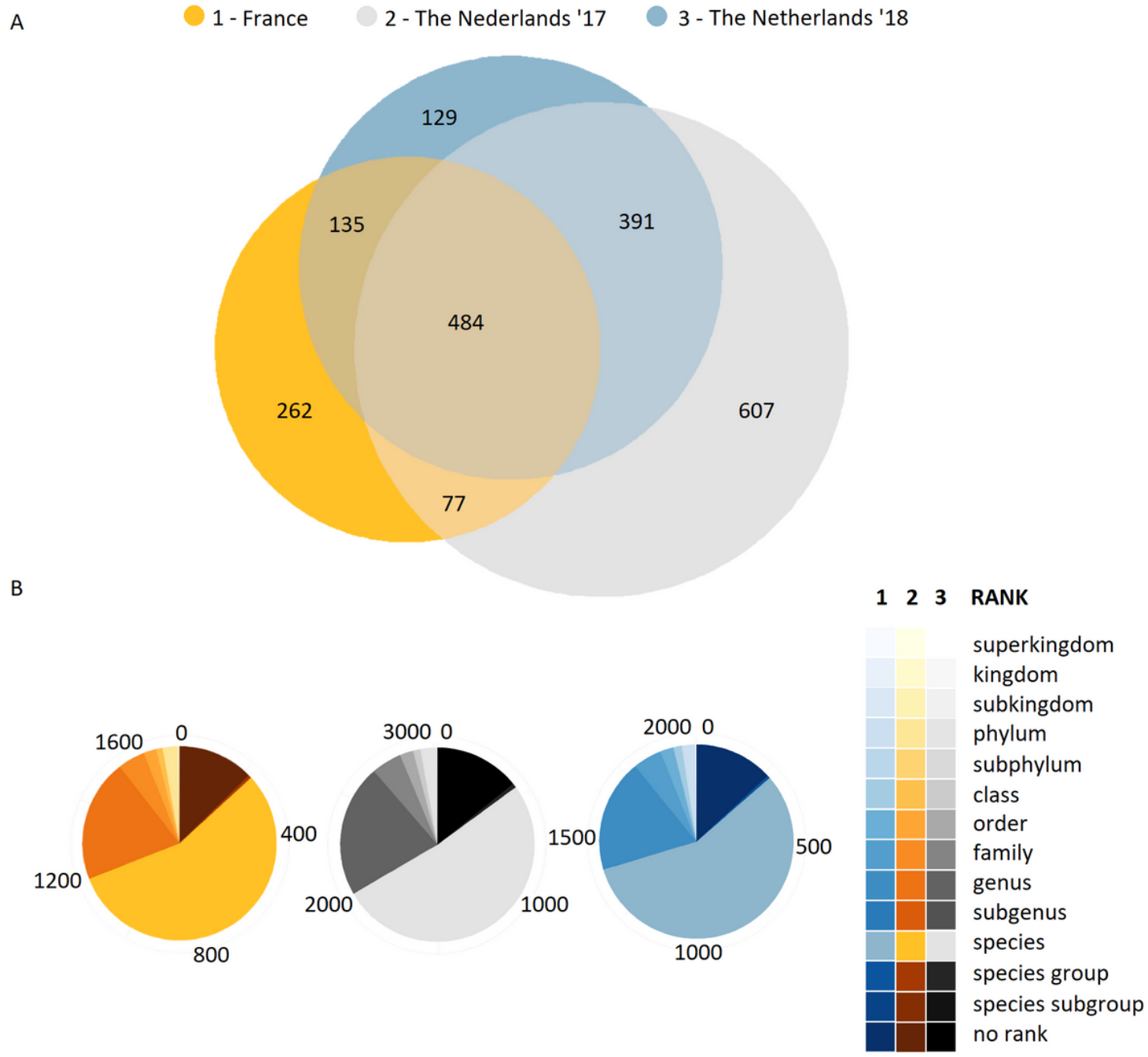

Figure 2

A) Venn diagram comparison of identified species by OneCodex, highlighting species that are time and space dependent. B) Overall OneCodex classification ranks per dataset, the majority of classified reads have been linked to a species level

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1AFTable1.docx
- additionalfile1AFTable1.docx
- additionalfile2AFTable2.docx
- additionalfile2AFTable2.docx
- additionalfile4AFTable3.docx
- additionalfile4AFTable3.docx
- additionalfile5AFFigure3.docx
- additionalfile5AFFigure3.docx
- additionalfile3AFFigure12.docx
- additionalfile3AFFigure12.docx
- additionalfile6AFFigure4.docx
- additionalfile6AFFigure4.docx
- additionalfile7AFTable4.docx
- additionalfile7AFTable4.docx
- additionalfile8AFFigure5.docx
- additionalfile8AFFigure5.docx
- additionalfile9AFFigure6.docx
- additionalfile9AFFigure6.docx