What is in my Sample? – Challenges and Approaches for Unveiling the Hidden Diversity in Plankton Samples

Marie Hoffmann (MSc Computer Science)

Advisers: **Prof. Knut Reinert, Prof. Michael T. Monaghan**

Department of Mathematics and Computer Science
Free University Berlin

Disputation - **August 23, 2022**

Motivation: Lake Monitoring Project

Conducted at IGB Berlin

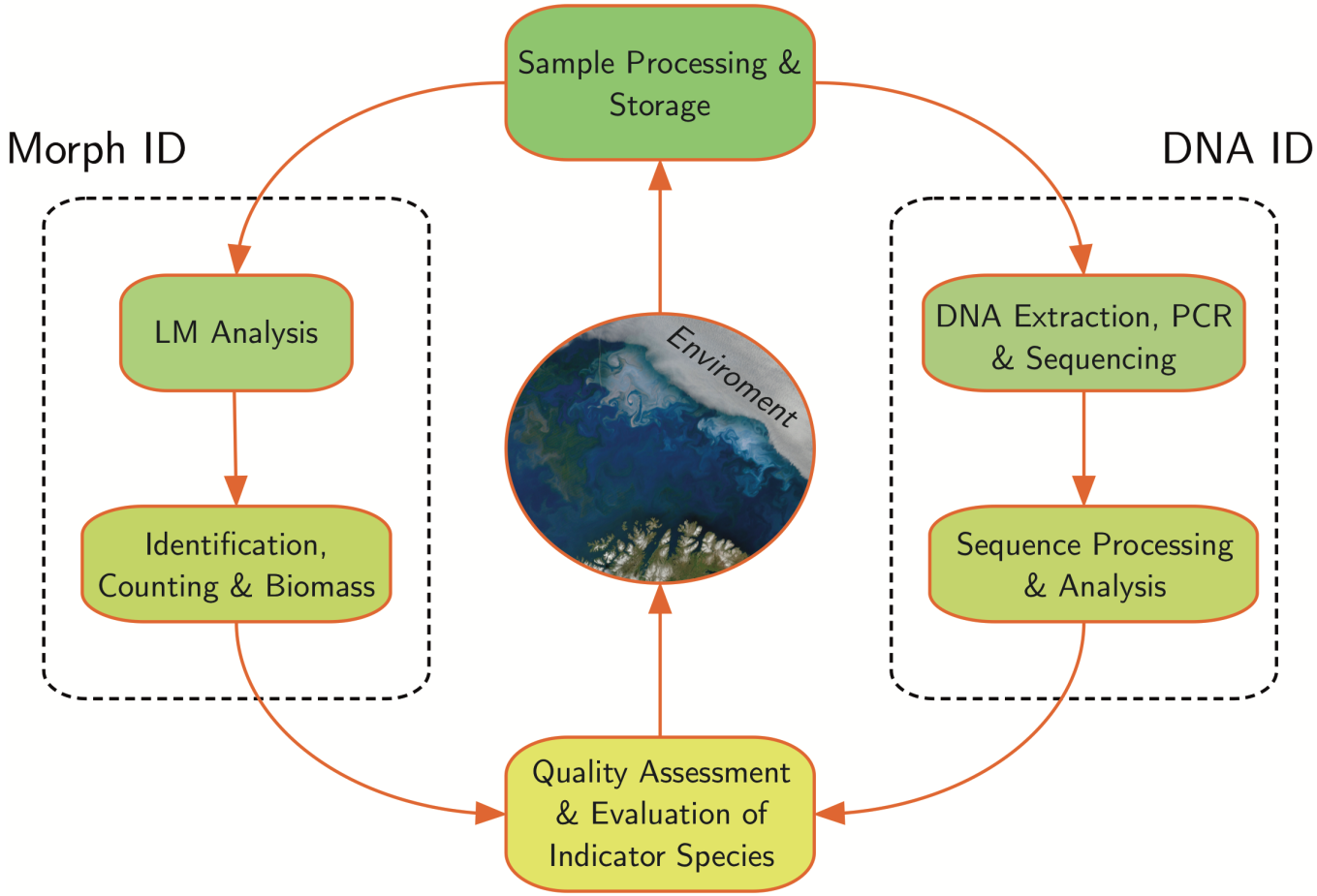


Figure 1: Complementing processing pipeline to study dynamics of freshwater samples.

Motivation: Lake Monitoring Project

Necessity to Intensify Environmental Sampling

- Conditions can change rapidly: see River Oder
  - Human factors (pesticides, waste water, ...)
  - Bloom of toxic Golden algae species
- Inherent limitation to modeling dynamics
  - "*Chaos is not rare in natural ecosystems*" by Rogers et al.[5]
  - Highest prevalence for plankton and insects

Figure 2: About 200 tons of dead fish have been removed so far. © dpa/Patrick Pleul

Motivation: Lake Monitoring Project

Challenge: Species Diversity

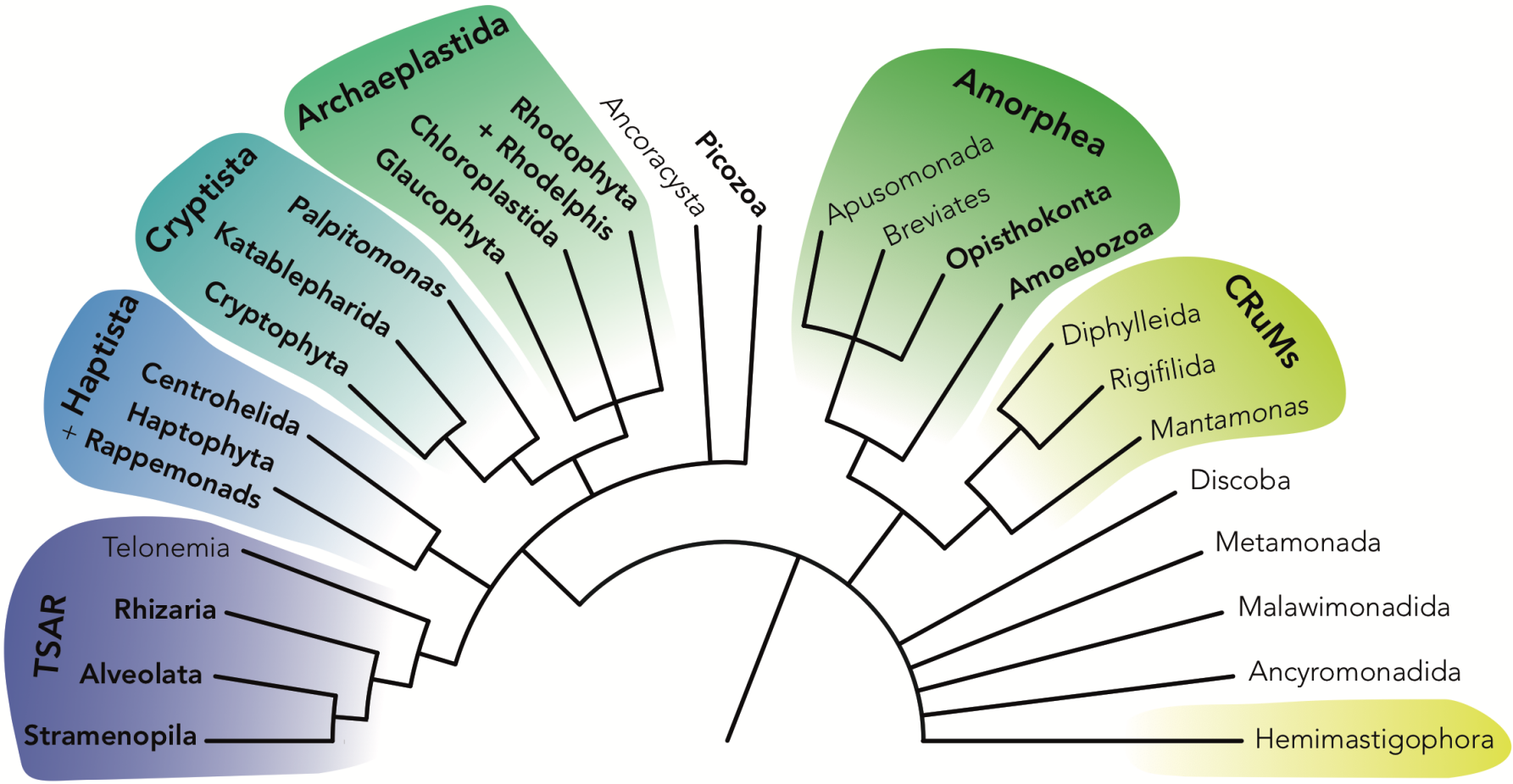


Figure 3: Dendrogram of the eukaryote tree of life proposed by [1].

Motivation: Lake Monitoring Project

Study Design

Pooled samples underwent both

- Morph ID
- DNA ID with three markers

Marker ID	Name	Ref.
EUK15	TAReuk454FWD1 TAReukREV3	[6]
EUK14	F-566a R-1200	[3]
DIV4	DIV4for DIV4rev3	[7]

Table 1: Primers for metabarcoding on freshwater plankton samples.

Figure 4: Integrated water sampler. ©HYDRO-BIOS Apparatebau GmbH

Motivation: Lake Monitoring Project

Study Results: Number of Taxa

	Taxa/OTU	Genus Level ID	Species Level ID
Morph ID	235	206 (88 %)	146 (62 %)
EUK15 ID	325	277 (69 %)	128 (40 %)
EUK14 ID	506	268 (53 %)	166 (33 %)
DIV4 ID	543	344 (63 %)	231 (43 %)

Table 2: Number of taxa (Morph ID) and OTUs (molecular ID) found and identified to genus or species level.

Key Findings Morph ID vs DNA ID

Morph ID

- number of
  species $\in \{\text{Morph ID}\} \backslash \{\text{DNA ID}\} = 141$
- Unbiased towards abundant organisms
- Allows biomass estimation

DNA ID

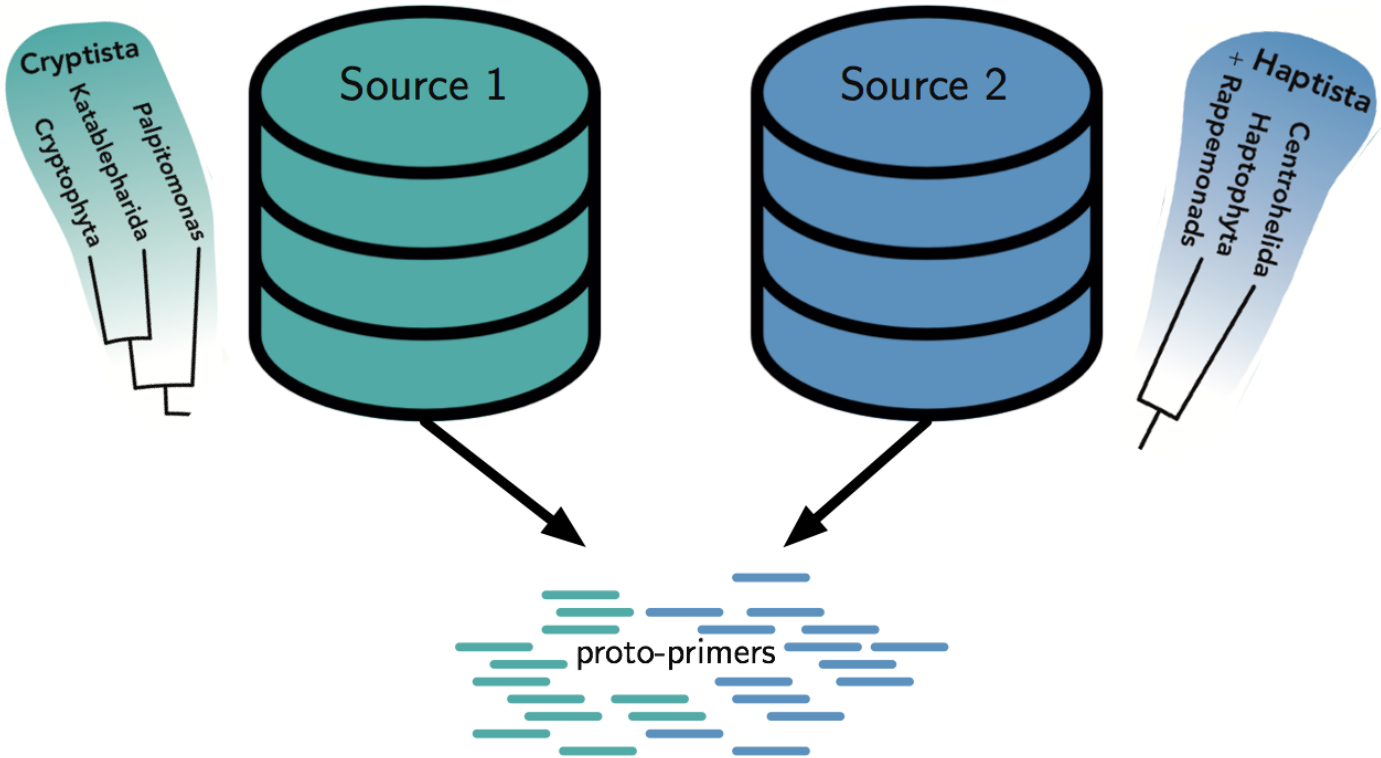
- better presence/absence test of species
- sensitive to spurious, invasive species, fungi
- Species remain undetected if
  - primers insufficiently match
  - similarity threshold low for OTU formation
  - **Phylum of rotifers undetected**

Figure 5: Rotifer. CC BY-SA 3.0 de from Wikipedia.

Motivation: Lake Monitoring Project

Desiderium for Ongoing Protocol Enhancement

1. Robust *de novo* primer discovery and *in silico* evaluation
   - Automated proto-primer search in uncurated databases
   - Robust to sequencing errors
   - Preference for high-frequent proto-primers



Robust Primer Discovery with PriSeT [PriSeT]

How to make it computational feasible?

Idea: Avoid MSA, but search for frequent k-mers (=proto-primers)

1. Build FM-index [2] on $T = R_1 \circ R_2 \circ \cdots \circ R_m$
   - FM-index provides

$$\begin{aligned} \text{locate}(T, \text{kmer}) &:= A[l : r] \\ \text{frequency}(T, k) &:= [r_i - l_i + 1]_{i \in [1:|T|]} \end{aligned}$$

   with l_i, r_i denoting the ranges for k-mer $T[i : i + k - 1]$
2. Lookup k-mers with minimal frequency threshold

R_i reference sequence

k target length for proto-primer

kmer = proto-primer

A suffix array derived from text T

Robust Primer Discovery

How to make it computational feasible?

Primer fitness test: Annealing, CG, T_m, Runs, $(A|T)$-tail, etc.

1. Two-bit encoding scheme A $\mapsto$ 00, C $\mapsto$ 01, G $\mapsto$ 10, T $\mapsto$ 11

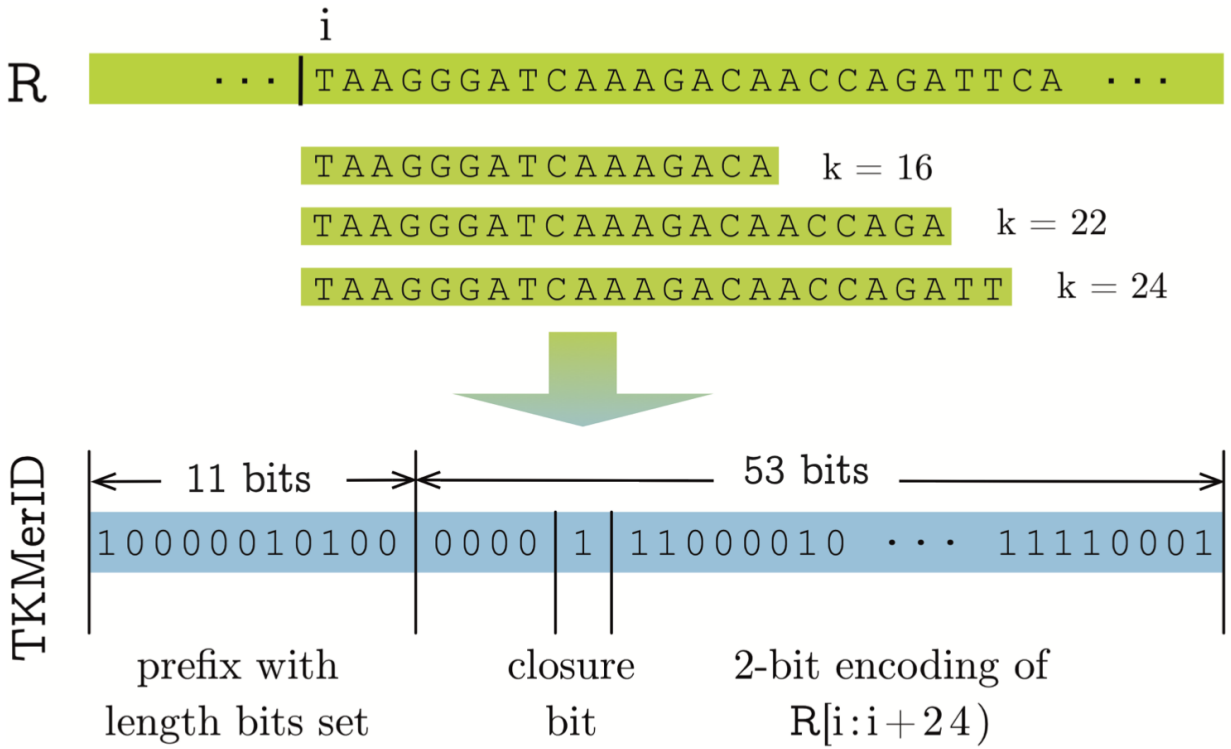


Figure 6: K-mer compression scheme.

Robust Primer Discovery

How to make it computational feasible?

Primer fitness test: Annealing, CG, T_m, Runs, $(A|T)$-tail, etc.

1. Two-bit encoding scheme A $\mapsto$ 00, C $\mapsto$ 01, G $\mapsto$ 10, T $\mapsto$ 11
2. Bit-parallelism

sequence with closure bit	0	1	A		C		G		T	
code	0	1	0	0	0	1	1	0	1	1
p = code & $(01)_4$	0	0	0	0	0	1	0	0	0	1
p = p << 1	1	0	0	0	1	0	0	0	1	0
q = code & $(10)_4$	0	0	0	0	0	0	1	0	1	0
x = p XOR q	0	0	0	0	1	0	1	0	0	0
popcount(x)					2					

Figure 4: Bit-parallelized counting of CG.

Robust Primer Discovery

Evaluation on Plankton Dataset: Library

	Clade	Name	Taxa	Covered	Accs	Lib Size
Phytoplankton	33849	Bacillariophyta	2,060	1,724	3,474	4.98 MB
	304574	Charophyceae	153	138	350	0.42 MB
	3041	Chlorophyta	10,490	9,466	15,377	31.79 MB
	2825	Chrysophyceae	428	339	507	0.89 MB
	3027	Cryptophyta	396	344	653	1.97 MB
	2864	Dinophyceae	6,147	4,630	7,151	6.24 MB
	33682	Euglenozoa	1,912	1,710	3,254	19.29 MB
	5747	Eustigmatophyceae	250	215	344	1.4 MB
Zooplankton	554915	Amoebozoa	3,211	2,817	3,898	4.63 MB
	33651	Bicosoecida	101	79	119	0.15 MB
	28009	Choanoflagellata	131	88	186	0.28 MB
	136419	Cercozoa	1,221	953	1,562	2.34 MB
	5878	Ciliophora	4,101	2,977	4,868	6.94 MB
	6657	Crustacea	45,058	25,643	50,163	38.85 MB
	6231	Nematoda	13,954	12,086	20,975	82.44 MB
	27999	Perkinsidae	81	75	114	0.13 MB
	10190	Rotifera	1,429	1,254	1,727	1.57 MB
Fungi	451864	Dikarya	141,097	129,254	209,449	518.44 MB
	112252	Fungi i. s.	7,169	5,948	9,149	10.21 MB

Figure 5: Reference sequences sampled from plankton taxa as found in GenBank's nt dataset.

Robust Primer Discovery

Evaluation on Plankton Dataset: comparison of proto-primer

Clade	Primer	Frequency	Coverage ↑	Variation	Primer	Frequency	Coverage	Variation ↑
33849	SSU	768	0.43 (734/1724)	631	EUK14	755	0.42	651
	b099967d0f5ac180	750	0.39 (673/1724)	27	33c14baf2ac76276	675	0.38	591
304574	eb59a790ece1766	35	0.23 (32/138)	13	d8d47dc9b873d02b	31	0.22	25
	EUK14	30	0.21 (29/138)	25	EUK14	30	0.21	25
⋮								
10190	57bc43fe1080644d	224	0.18 (224/1254)	1	EUKA	69	0.05	66
	EUKA	69	0.05 (66/1254)	66	bbcb9dc15a5fc34c	216	0.17	40

Figure 6: Excerpt from table of computed proto-primers (16-digits) compared to published primer pairs (SSU, EUK14, EUKA). Coverage: number of sequences covered. Variation: number of unique amplicon.

Results: at least one new primer pair identified by PriSeT

- Ranking by coverage: for 11 / 19 clades
- Ranking by amplicon variation: for 7 / 19 clades

Robust Primer Discovery

Evaluation on Plankton Dataset: Runtime of proto-primer generation

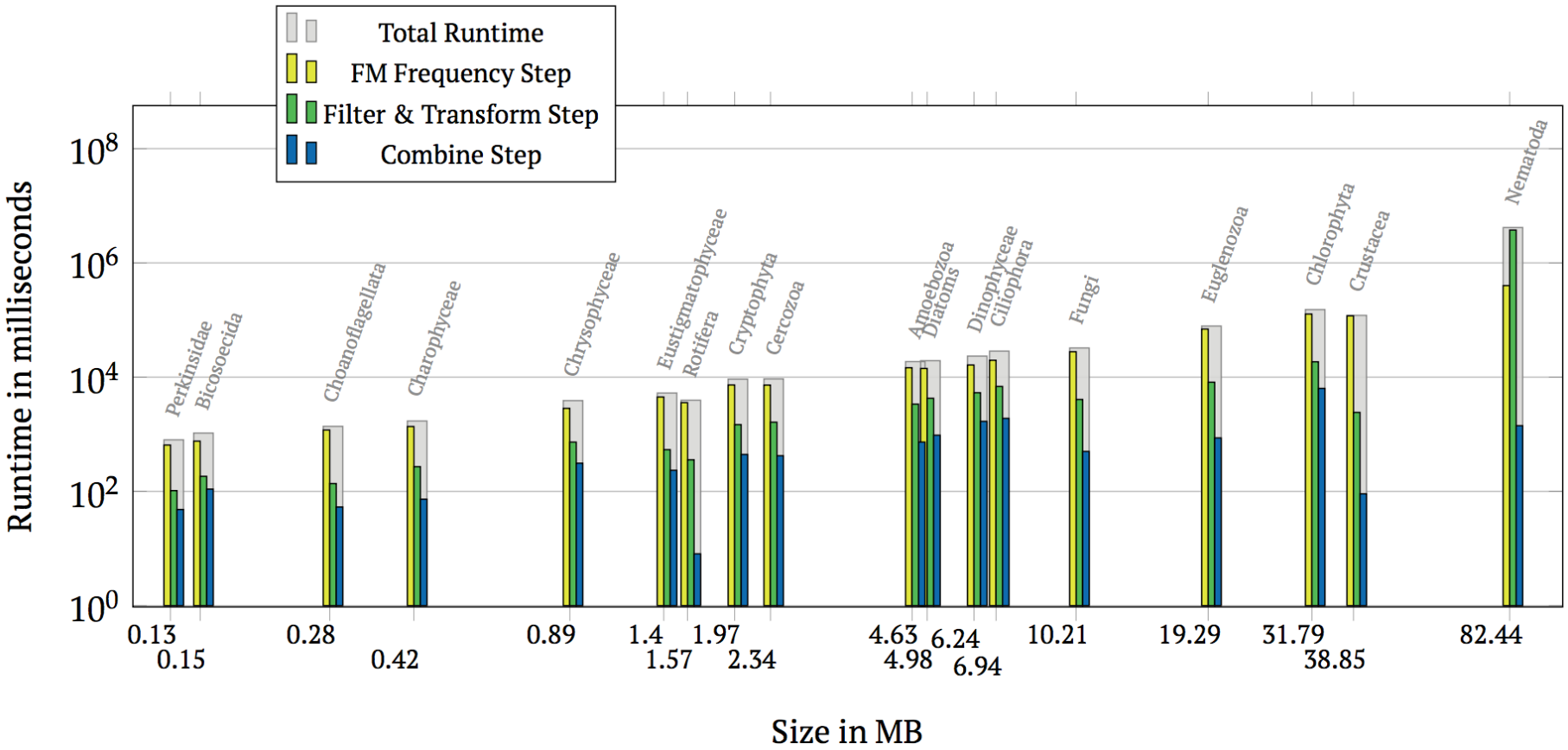


Figure 7: Runtimes for plankton clades vary from about 1 sec (Perkinsidae) to 15 min (Nematoda).

Robust Primer Discovery

Theoretical Runtime and Space Consumption

	FM-Index	FM Frequency
Runtime	$\mathcal{O}(N)$	$O(N(\kappa_{\max} + occ))$
Space	$\mathcal{O}(\log \lvert\Sigma\rvert N) + o(\log \lvert\Sigma\rvert^2 N)$	$\mathcal{O}(N)$
	Filter & Transform	**Combine**
Runtime	$\mathcal{O}(\kappa_{\max} N)$	$\mathcal{O}(n\kappa_{\max}(\frac{N}{n} - \tau_{min})(\tau_{max} - \tau_{min}))$
Space	$\mathcal{O}(N)$	$\mathcal{O}(n(\frac{N}{n} - \tau_{min})(\tau_{max} - \tau_{min}))$

Table 3: Runtime classes and space occupation module-wise with N as the total library size, $|\Sigma|$ is four, because of the underlying alphabet being $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, n the number of references per library, $\kappa_{\max}$ the largest k-mer length, ω the window width, and $\tau_{\min/\max}$ the amplicon length limits, and occ the expected number of k-mer occurrences.

Robust Primer Discovery

Evaluation on SARS-CoV-2

PriSeT: 286 proto-primers in total, 114 with unique transcripts

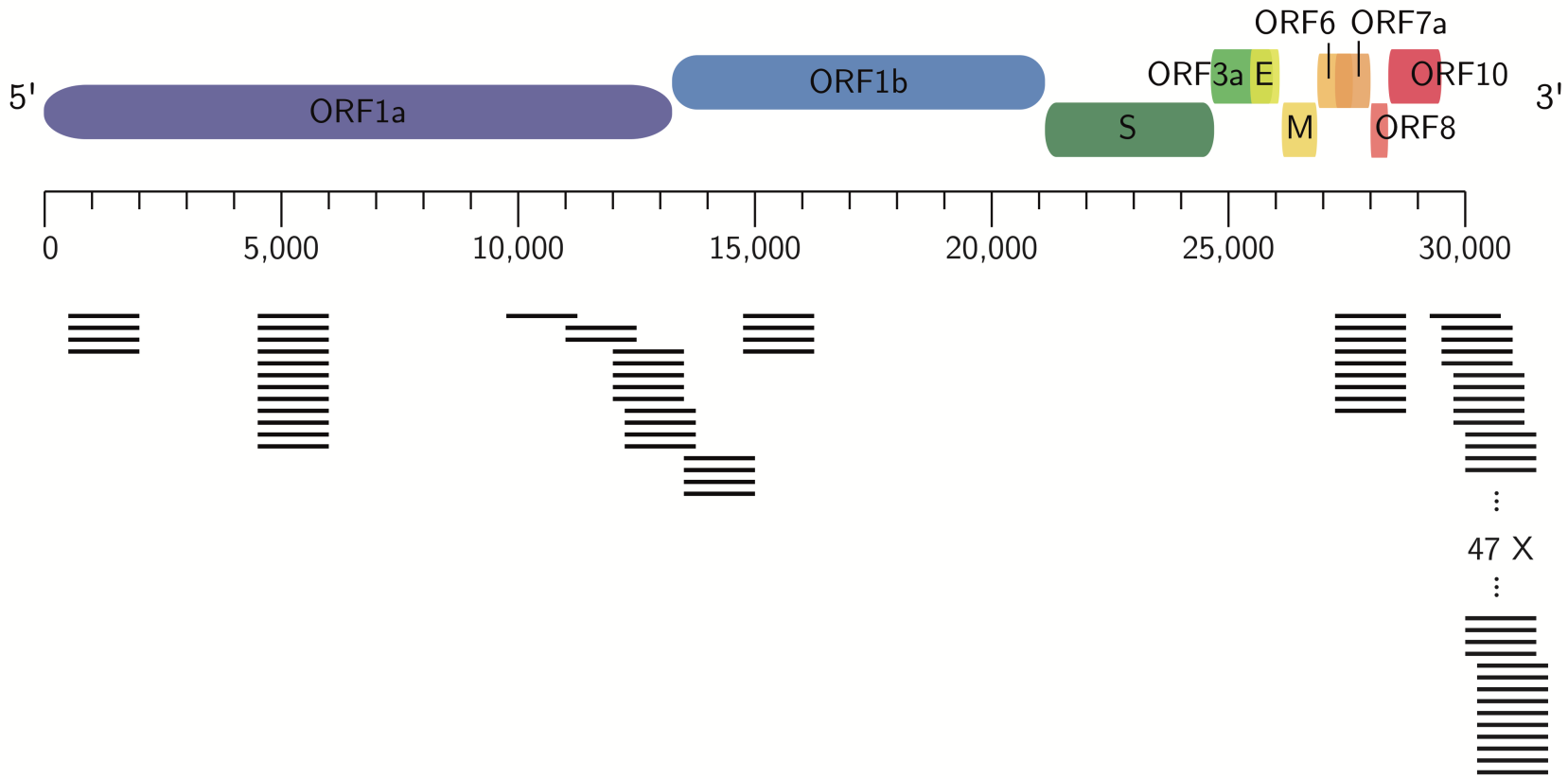


Figure 8: Top: Genome of SARS-CoV-2 based on GenBank MN908947.3 and NCBI's ORFfinder. Bottom: Transcript positions for proto-primers.

List of References

[1] Fabien Burki et al. "The New Tree of Eukaryotes". In: *Trends in Ecology & Evolution* 35 (1 2020), pp. 43–55. DOI: 10.1016/j.tree.2019.08.008.

[2] Paolo Ferragina and Giovanni Manzini. "Indexing Compressed Text". In: *J. ACM* 52.4 (July 2005), pp. 552–581. DOI: 10.1145/1082036.1082039.

[3] Kenan Hadziavdic et al. "Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers". In: *PLOS ONE* 9.2 (Feb. 2014), pp. 1–10. DOI: 10.1371/journal.pone.0087624.

List of References

[4] Marie Hoffmann, Michael T. Monaghan, and Knut Reinert. "PriSeT: Efficient de Novo Primer Discovery". In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '21. Gainesville, Florida: Association for Computing Machinery, 2021. ISBN: 9781450384506. DOI: 10.1145/3459930.3469546. URL: https://doi.org/10.1145/3459930.3469546.

[5] Tanya L Rogers, Bethany J Johnson, and Stephan B Munch. "Chaos is not rare in natural ecosystems". In: *Nature Ecology & Evolution* 6.8 (2022), pp. 1105–1111. ISSN: 2397-334X. DOI: 10.1038/s41559-022-01787-y. URL: https://doi.org/10.1038/s41559-022-01787-y.

[6] Thorsten Stoeck et al. "Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water". In: *Molecular Ecology* 19.s1 (2010), pp. 21–31. DOI: 10.1111/j.1365-294X.2009.04480.x.

List of References

[7] Joana Amorim Visco et al. “Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data”. In: *Environmental Science & Technology* 49.13 (2015), pp. 7597–7605. DOI: 10.1021/es506158m.

Thanks to

- Commission: Knut Reinert, Michael T. Monaghan, Katharina Jahn, Sandro Andreotti
- Bioinformatics Group: SeqAn Team
- BeGenDiv: Tatiana Semenova-Nelson, Camilla Mazzoni, Felix Heeger
- IGB: Rita Adrian, Justyna Wollinska, Ursula Newen
- any many more

Appendix Motivation: Lake Monitoring Project

Study Results: Unidentified Taxa for DNA ID

$$|\text{species} \in \{\text{Morph ID}\} \backslash \{\text{DNA ID}\}| = 141$$

Marker ID	Species		*in silico* PCR	
	Morphologically identified	Reference available/missing	Success	Failure
EUK15	144	70/74	68 (97 %)	2 (3 %)
EUK14	144	70/74	67 (96 %)	3 (4 %)
DIV4	144	70/74	48 (68 %)	22 (32 %)

Table 4: Results of *in silico* PCR for each primer set tested.

Appendix Evaluation on SARS-CoV-2

qPCR Settings

Parameter	Settings SARS-CoV-2
k	[18 : 24]
τ [nt]	[60 : 150]
Tm [°C]	[55 : 63]
GC [%]	[50 : 60]
4-Runs of C or G	yes
Self-Annealing	on
ΔTm [K]	5
Cross-Annealing	on

Table 5: Settings for qPCR.