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

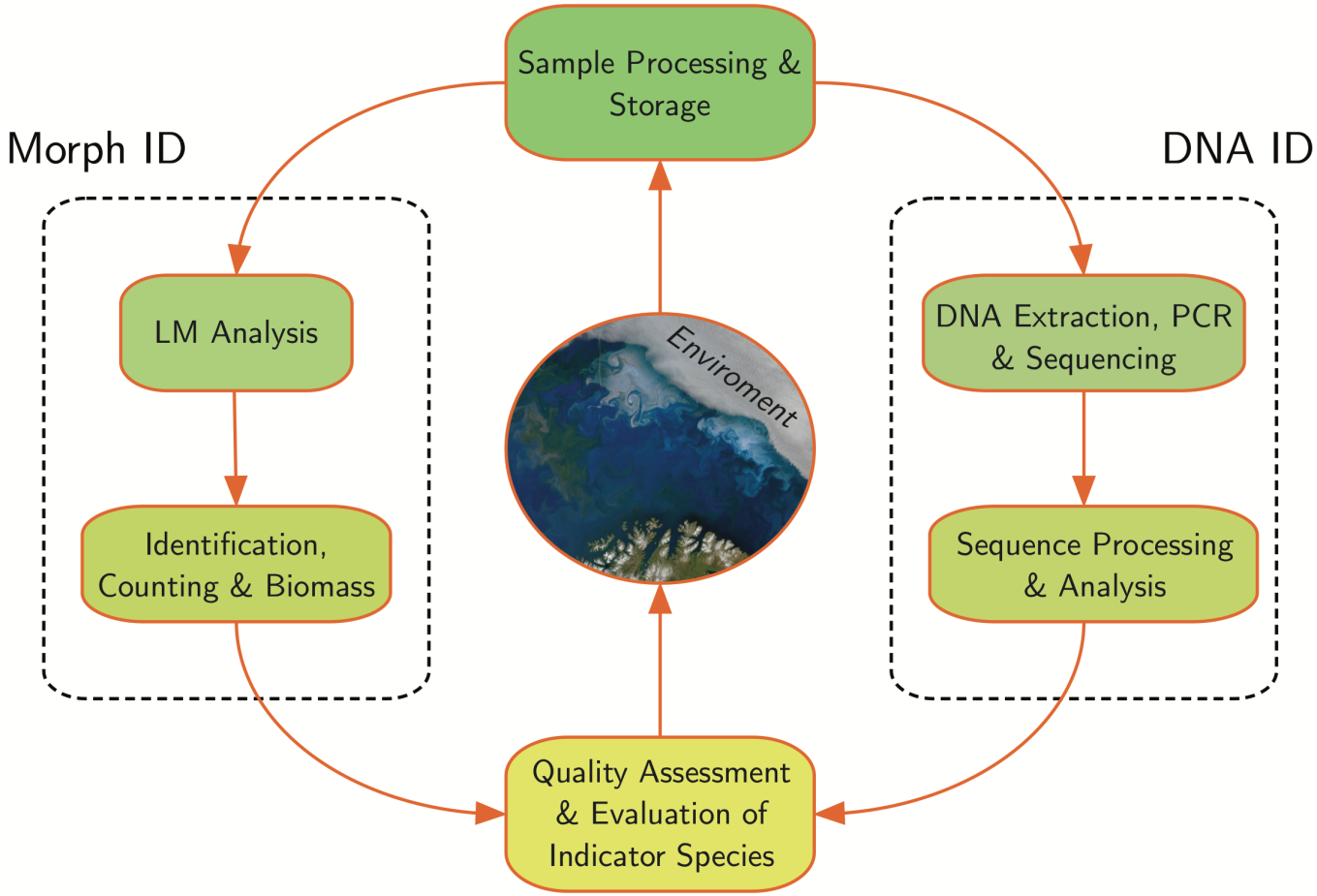


Figure 1: Complementing processing pipeline to study dynamics of freshwater samples.

Motivation: Lake Monitoring Project

Challenge: Species Diversity I

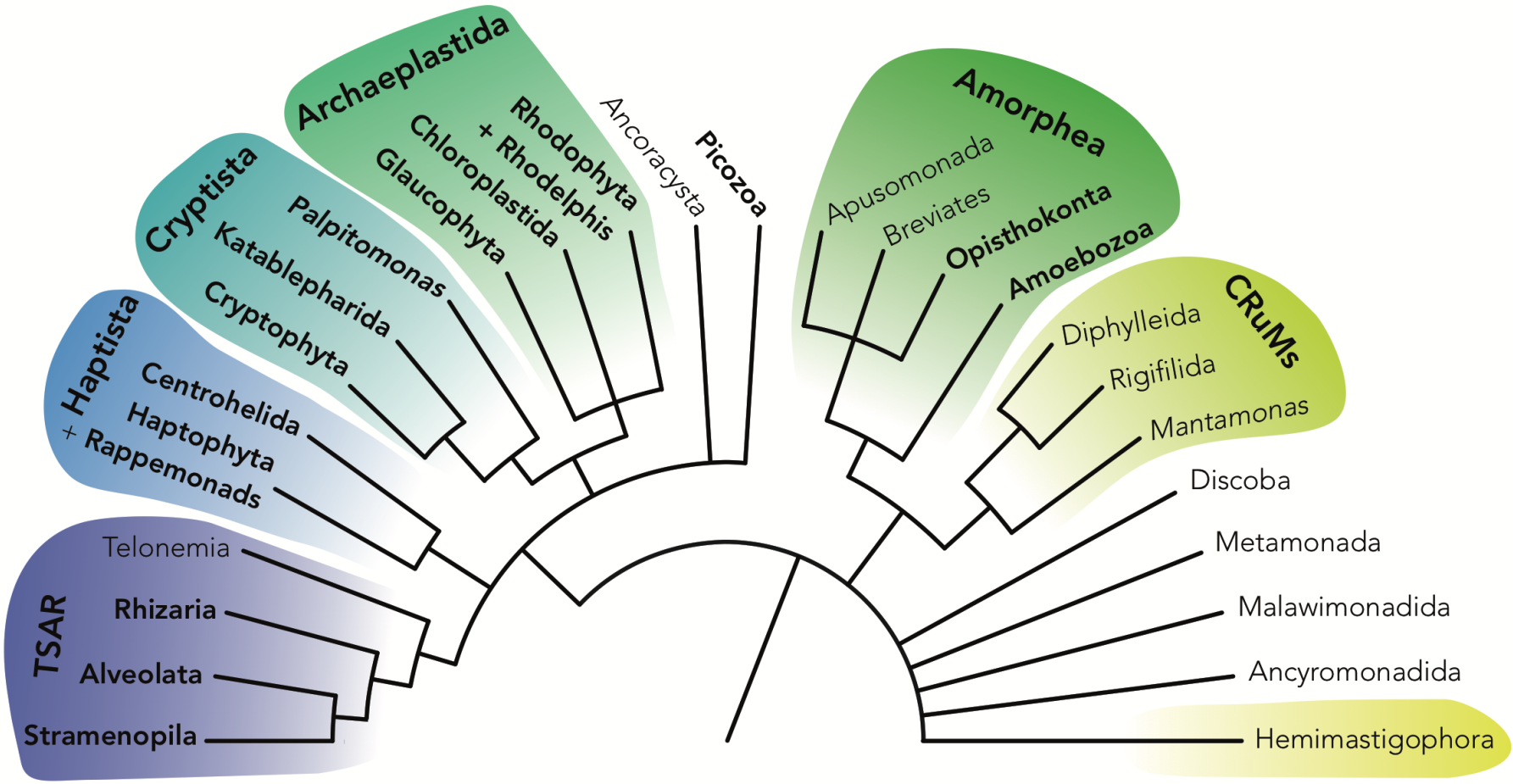


Figure 2: Dendrogram of the eukaryote tree of life proposed by [3].

Motivation: Lake Monitoring Project

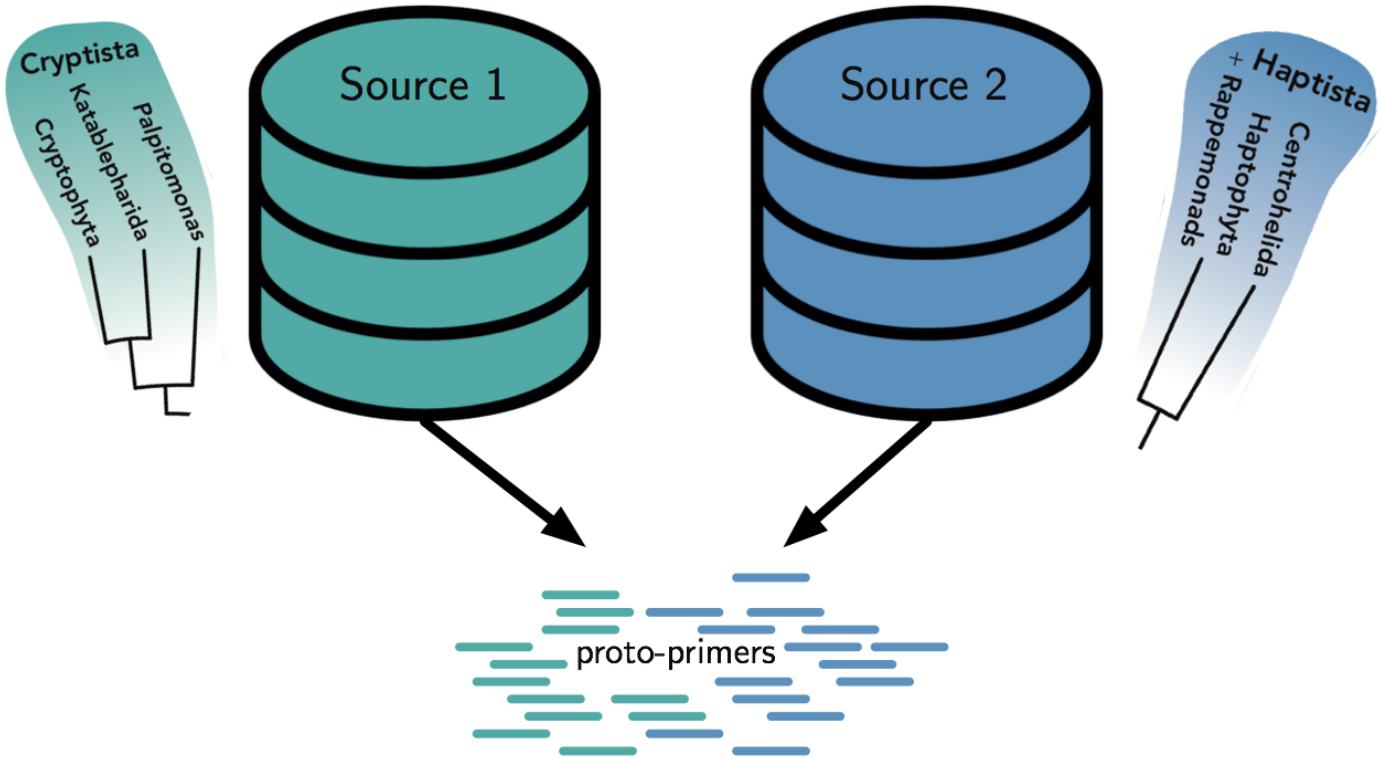
Challenge: Capturing Diversity via NGS

1. Primer Design: taxonomic *width* versus *depth*
2. Optimization: Missing ground truth
3. Resolution: Databases are sparse, erroneous, uncurated
4. Consolidation: No unique, generally accepted taxonomic tree!
5. PCR: Same difficulties as single organisms DNA analysis: repeats, intra-species variation, ...

Motivation: Lake Monitoring Project

Desiderium for Ongoing Protocol Enhancement

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



I Robust Primer Discovery

How to make it computational feasible?

Avoid multiple sequence alignment computation

1. Build FM-index [4] 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

I Robust Primer Discovery

How to make it computational feasible?

Primer fitness test

1. Two-bit encoding scheme

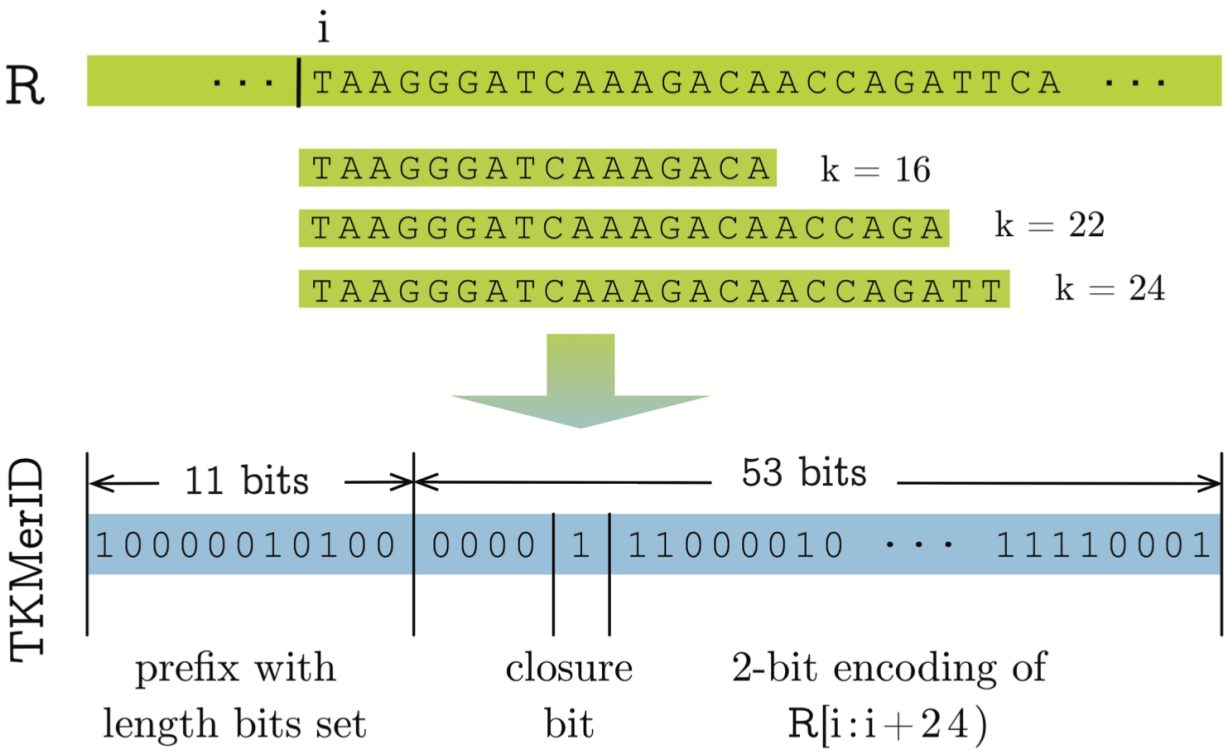


Figure 3: K-mer compression scheme.

Primer fitness test

1. Two-bit encoding scheme
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.

Presented techniques implemented in PriSeT [1].

I 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.

I 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

I Robust Primer Discovery

Evaluation on Plankton Dataset: Runtime of proto-primer generation

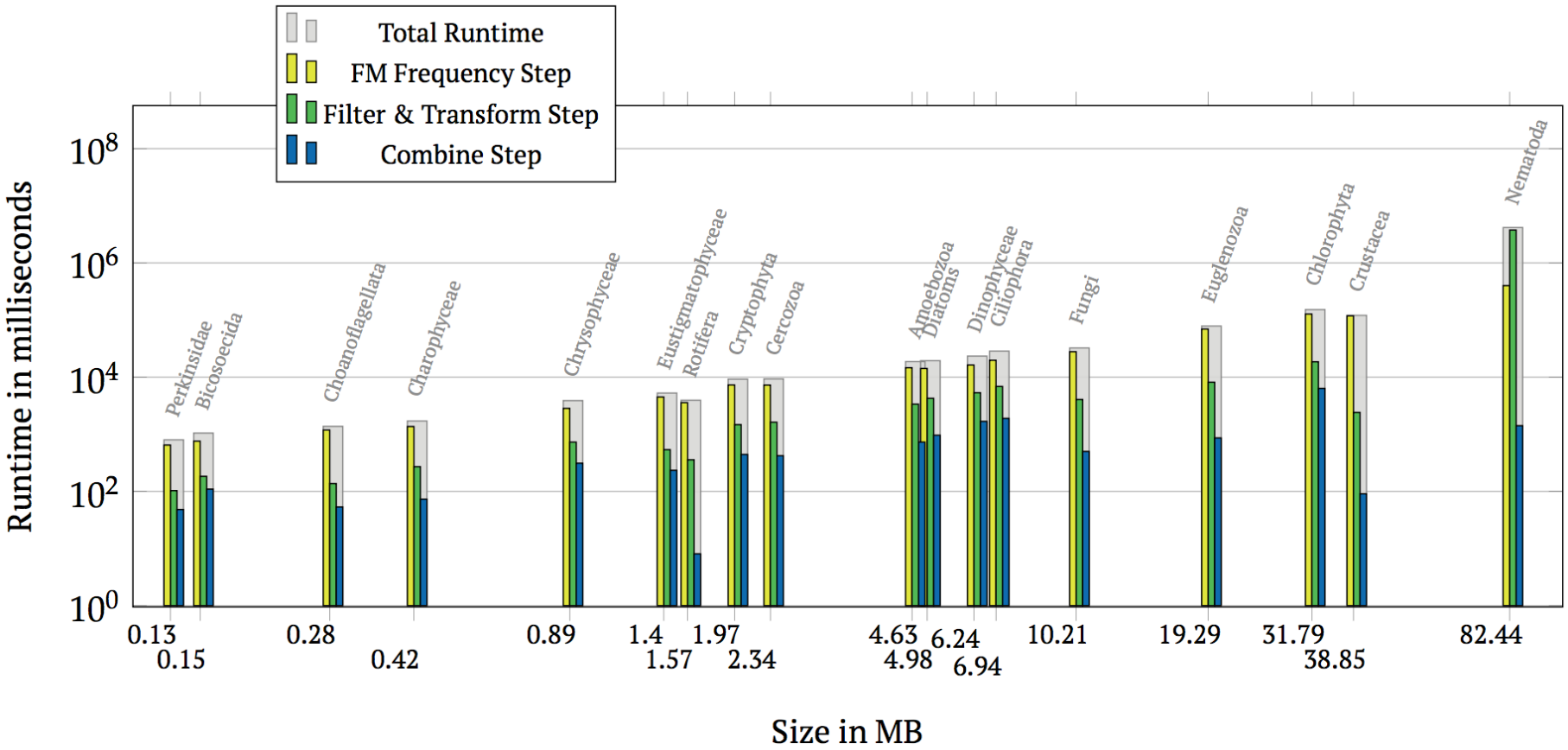


Figure 7: Runtimes for plankton clades vary from about 1 sec (Perkinsidae) to 15 min (Nematoda).

Motivation: Lake Monitoring Project

Desiderium for Ongoing Protocol Enhancement II

II Structured digitalization of protocols, past studies

- Allow for meta-analyses
- Shorten ramp-up period of new hires
- Improve reproducability of analyses

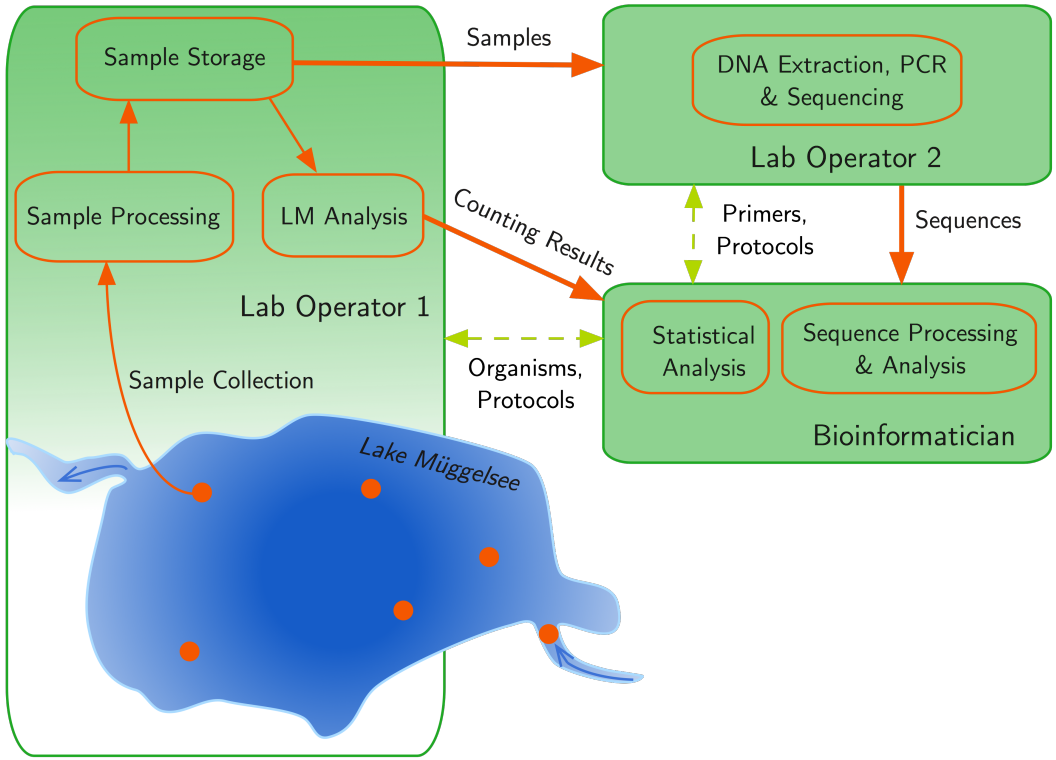


Figure 8: Data transfer between lab operators and bioinformaticians.

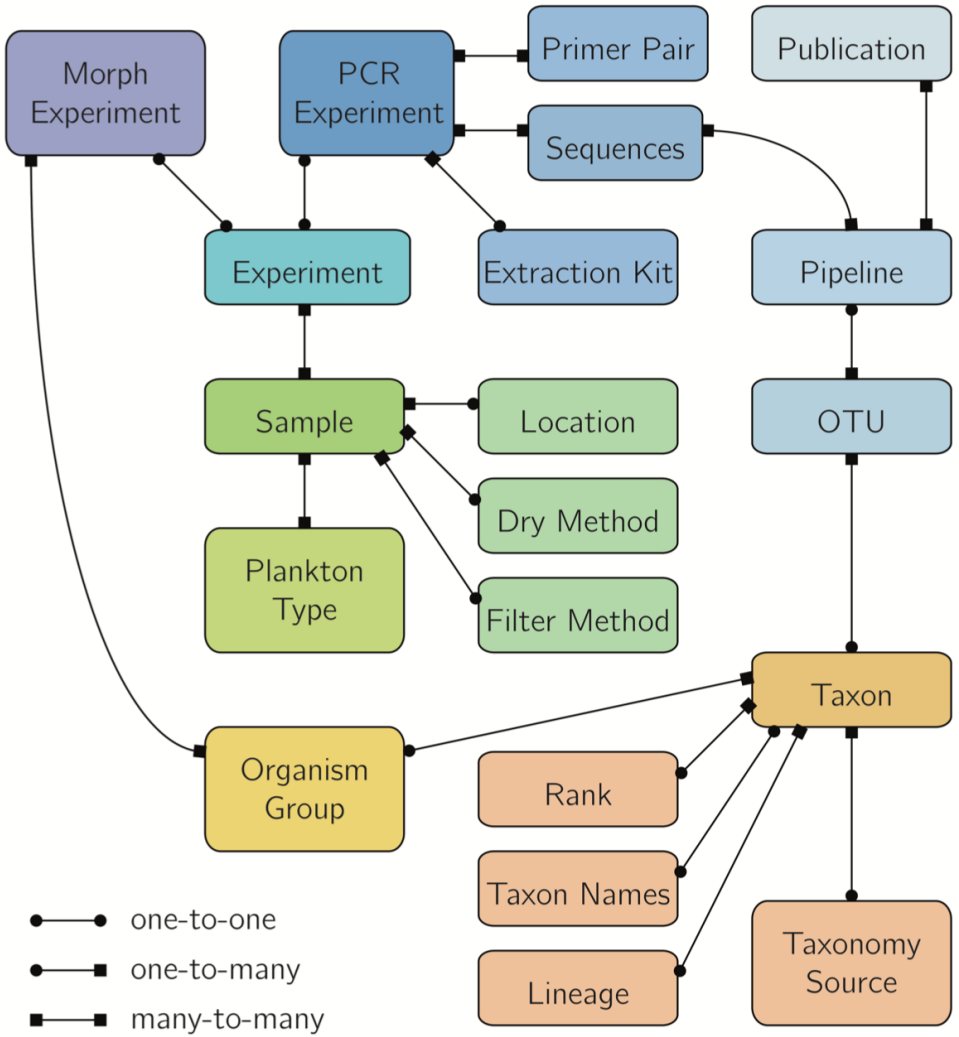


Figure 9: Database schema [2]

List of References

[1] URL: https://github.com/mariehoffmann/priSeT.

[2] URL: https://github.com/mariehoffmann/plankton_database.

[3] 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.

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

[5] 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.

Thanks to

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