

Supplementary Material 1

Profiling the eukaryotic diversity over a year in a lake ecosystem through short- and long-read metabarcoding

Chauvet M^{1*}, Terzian P¹, Moné A¹, Debroas D¹, Bronner G¹, Lepère C¹

¹CNRS, Laboratoire Microorganismes : Génome Et Environnement, Université Clermont Auvergne,
Clermont-Ferrand 63000, France

* Corresponding author

Table S1. Number of reads remaining after each major step of the short-read bioinformatic pipeline.

Sample	Correspondence	Reads.in	Merged	Quality	Reads.out	Final_reads
18d10	A-01-04-19-G	88486	34415	26255	16185	15824
18d32	A-01-07-19-G	57139	40911	30994	23452	23190
18d31	A-01-07-19-P	77250	59205	44001	43804	42179
18d36	A-02-09-19-G	36293	25105	20213	17478	17233
18d42	A-02-12-19-G	93805	73271	55763	52409	51719
18d20	A-03-06-19-G	78279	54649	42296	40199	39705
18d19	A-03-06-19-P	86300	64209	48035	47756	46912
18d40	A-04-11-19-G	63813	50524	39061	29819	29367
18d39	A-04-11-19-P	30689	19727	14761	14728	14176
18d13	A-06-05-19-P	56437	44712	33672	33386	32445
18d22	A-06-06-19-G	87190	57006	40165	36083	35471
18d21	A-06-06-19-P	50554	37287	27156	26731	26116
18d44	A-07-01-20-G	85057	68585	48323	43784	43265
18d43	A-07-01-20-P	86425	67476	48384	48155	46299
18d38	A-08-10-19-G	38901	21418	16481	12375	12011
18d37	A-08-10-19-P	26894	17070	12726	12677	12234
18d24	A-11-06-19-G	79138	52538	39121	36838	36339
18d23	A-11-06-19-P	83485	59205	44907	44754	43992
18d6	A-14-01-19-G	74407	48301	34311	28025	27296
18d5	A-14-01-19-P	67034	54401	42242	41977	40548
18d26	A-14-06-19-G	87269	9635	7428	2178	1945
18d25	A-14-06-19-P	86235	60905	46197	45807	44921
18d28	A-17-06-19-G	87644	66084	53640	42626	42028
18d27	A-17-06-19-P	66228	49661	38976	38603	36731
18d4	A-17-12-18-G	83477	64153	53124	26457	25914
18d3	A-17-12-18-P	72542	57064	45369	43920	42301
18d12	A-18-04-19-G	72741	32703	25080	17517	17167
18d11	A-18-04-19-P	64012	45144	34886	34408	33304
18d2	A-19-11-18-G	82503	61413	47925	16180	15593
18d8	A-20-02-19-G	57880	41258	29656	27478	26509
18d30	A-20-06-19-G	68238	36693	29803	25401	25197
18d29	A-20-06-19-P	70911	54658	41383	41199	39552
18d16	A-27-05-19-G	66455	47529	33595	28751	28130
18d15	A-27-05-19-P	76029	57145	41204	40857	39858
18d18	A-29-05-19-G	93505	63843	44326	39439	38761
18d17	A-29-05-19-P	76925	53945	37069	36670	35697
18d34	A-29-07-19-G	29835	5299	4223	1945	1800
18d33	A-29-07-19-P	30482	16871	11577	11545	10842

This table reports the number of reads remaining after each major step of the short-read bioinformatic pipeline.

Column descriptions

Sample: Identifier of demultiplexed sequencing files.

Correspondence: Sample identifier used throughout the study.

Reads.in: Total number of raw reads before any processing.

Merged: Number of reads after paired-end merging.

Quality: Number of reads retained after quality filtering using the PANAM pipeline (sequences with ambiguous 'N' bases, those below 200 bp and chimeras were removed).

Reads.out: Number of reads per sample after taxonomic cleaning (removal of metazoans in particular).

Final_reads: Final number of reads after applying an additional abundance-based filtering step (OTUs representing less than 0.005% of the total reads were removed). This represents the effective number of reads used in downstream ecological and diversity analyses.

Table S2. Number of reads remaining after each major step of the long-read bioinformatic pipeline.

Demux.bc10	Correspondence	Reads.in	Reads.out	99% consensus	Barrnap_SSU	NoChimera	Consensus_filtered	Final_reads
demux.bc1007_bc1035	A-01-04-19-G	13106	9725	931	906	631	115	453
demux.bc1024_bc1045	A-01-07-19-G	9849	6514	469	461	347	66	167
demux.bc1022_bc1045	A-01-07-19-P	5932	4251	474	467	362	60	319
demux.bc1012_bc1054	A-02-09-19-G	10238	6821	603	587	363	41	254
demux.bc1007_bc1056	A-02-12-19-G	3917	2730	271	258	197	32	157
demux.bc1012_bc1044	A-03-06-19-G	6685	4399	628	610	387	74	298
demux.bc1008_bc1044	A-03-06-19-P	5100	2738	412	381	324	67	304
demux.bc1024_bc1054	A-04-11-19-G	6134	3994	742	738	491	78	320
demux.bc1022_bc1054	A-04-11-19-P	4582	2777	723	715	534	86	479
demux.bc1015_bc1035	A-06-05-19-P	4791	3179	358	353	235	26	156
demux.bc1020_bc1044	A-06-06-19-G	4968	3083	676	656	434	59	342
demux.bc1015_bc1044	A-06-06-19-P	3289	1649	585	580	414	78	388
demux.bc1012_bc1056	A-07-01-20-G	5454	2628	772	764	571	67	529
demux.bc1008_bc1056	A-07-01-20-P	1688	951	267	251	200	35	165
demux.bc1020_bc1054	A-08-10-19-G	4338	2643	413	409	319	76	213
demux.bc1015_bc1054	A-08-10-19-P	5028	2919	462	461	405	84	366
demux.bc1024_bc1044	A-11-06-19-G	5779	3820	476	452	322	52	274
demux.bc1022_bc1044	A-11-06-19-P	1124	593	226	222	187	40	166
demux.bc1020_bc1033	A-14-01-19-G	8228	5033	968	942	516	80	400
demux.bc1015_bc1033	A-14-01-19-P	4389	2242	505	491	360	69	333
demux.bc1007_bc1045	A-14-06-19-G	7435	5254	418	412	291	45	214
demux.bc1005_bc1045	A-14-06-19-P	1823	915	231	223	182	42	159
demux.bc1012_bc1045	A-17-06-19-G	16321	11275	566	555	377	61	203
demux.bc1008_bc1045	A-17-06-19-P	2684	1566	412	403	324	62	264
demux.bc1012_bc1033	A-17-12-18-G	11663	6992	1022	1001	723	122	227
demux.bc1008_bc1033	A-17-12-18-P	12576	7107	1038	897	649	111	559
demux.bc1012_bc1035	A-18-04-19-G	6079	3458	497	488	384	61	145
demux.bc1008_bc1035	A-18-04-19-P	4872	2177	359	319	277	61	252
demux.bc1007_bc1033	A-19-11-18-G	7011	3828	750	737	608	106	388
demux.bc1024_bc1033	A-20-02-19-G	6589	4270	618	604	404	48	302
demux.bc1020_bc1045	A-20-06-19-G	13068	9086	530	525	383	63	203
demux.bc1015_bc1045	A-20-06-19-P	2444	1435	319	315	259	53	229
demux.bc1024_bc1035	A-27-05-19-G	4650	2714	495	480	369	55	318
demux.bc1022_bc1035	A-27-05-19-P	3955	2328	479	472	400	70	353
demux.bc1007_bc1044	A-29-05-19-G	8747	5766	1125	1031	678	119	525
demux.bc1005_bc1044	A-29-05-19-P	6098	3831	732	713	570	100	534
demux.bc1007_bc1054	A-29-07-19-G	10286	6332	358	341	299	47	60
demux.bc1005_bc1054	A-29-07-19-P	4328	2149	633	631	553	106	506

This table reports the number of sequences remaining after each major step of the long-read bioinformatic pipeline. Importantly, two fundamentally different units are presented: (i) raw or filtered read counts (abundance-weighted), and (ii) counts of unique sequences obtained after dereplication and consensus-building steps.

Column descriptions

Demux.bc10: Identifier of demultiplexed sequencing files.

Correspondence: Sample identifier used throughout the study.

Reads.in: Total number of raw reads obtained after demultiplexing (abundance-weighted; true read counts).

Reads.out: Number of reads retained after initial quality filtering (abundance-weighted; true read counts).

99% consensus: Number of consensus sequences generated by clustering reads at 99% similarity. At this step, identical or highly similar reads are collapsed into single representative sequences. (not abundance-weighted; number of unique consensus sequences).

Barrnap_SSU: Number of sequences retained after SSU rRNA gene extraction using Barrnap. (not abundance-weighted; unique sequences).

NoChimera: Number of sequences remaining after chimera removal. (not abundance-weighted; unique sequences).

Consensus_filtered: Number of sequences included in consensus clusters. (not abundance-weighted; unique sequences).

Final_reads: Final number of reads per sample after reapplying the abundance of each sequence (*i.e.*, after expanding dereplicated sequences back to their original read counts). This represents the effective number of reads used in downstream ecological and diversity analyses (abundance-weighted; true read counts).

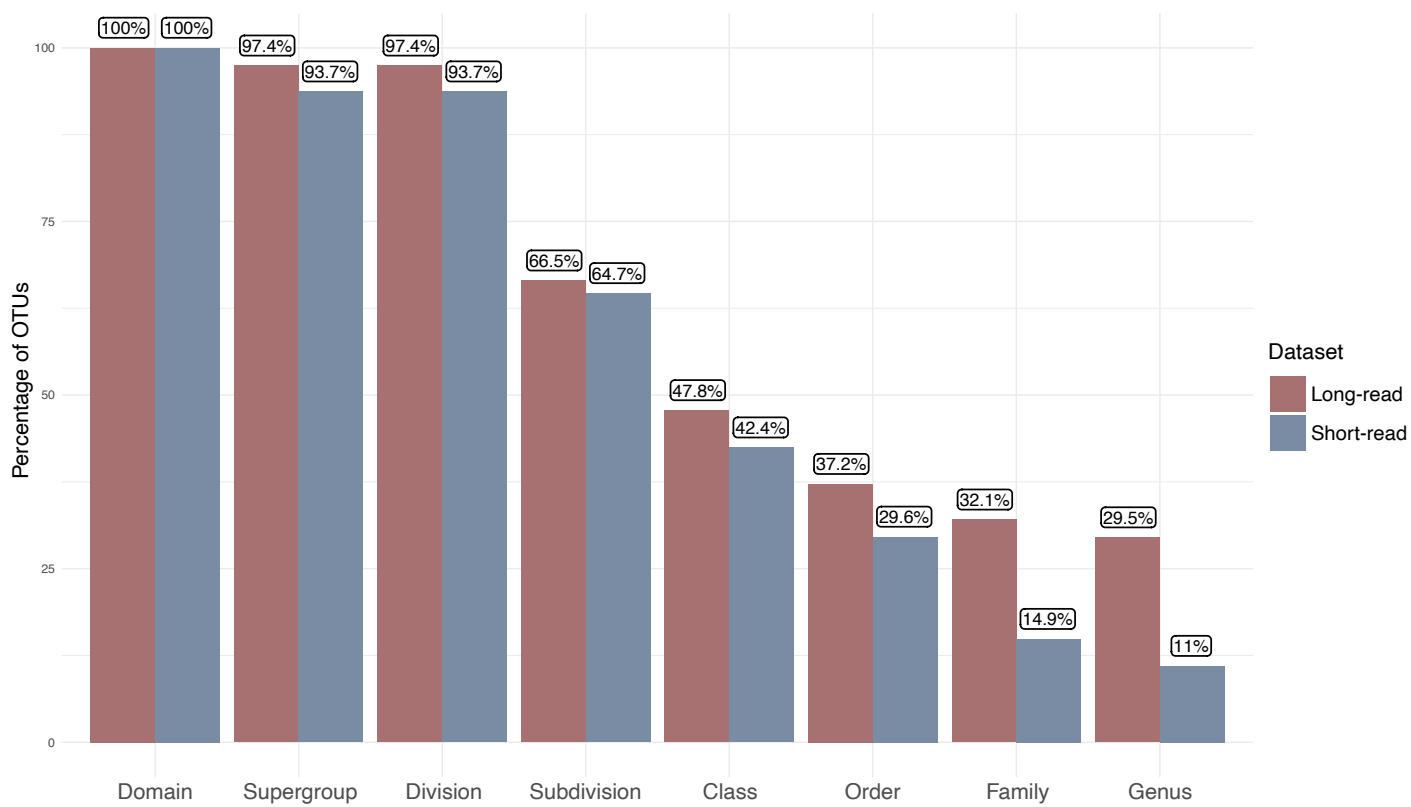


Figure S1. Percentage of OTUs assigned up to each taxonomic level in both datasets.

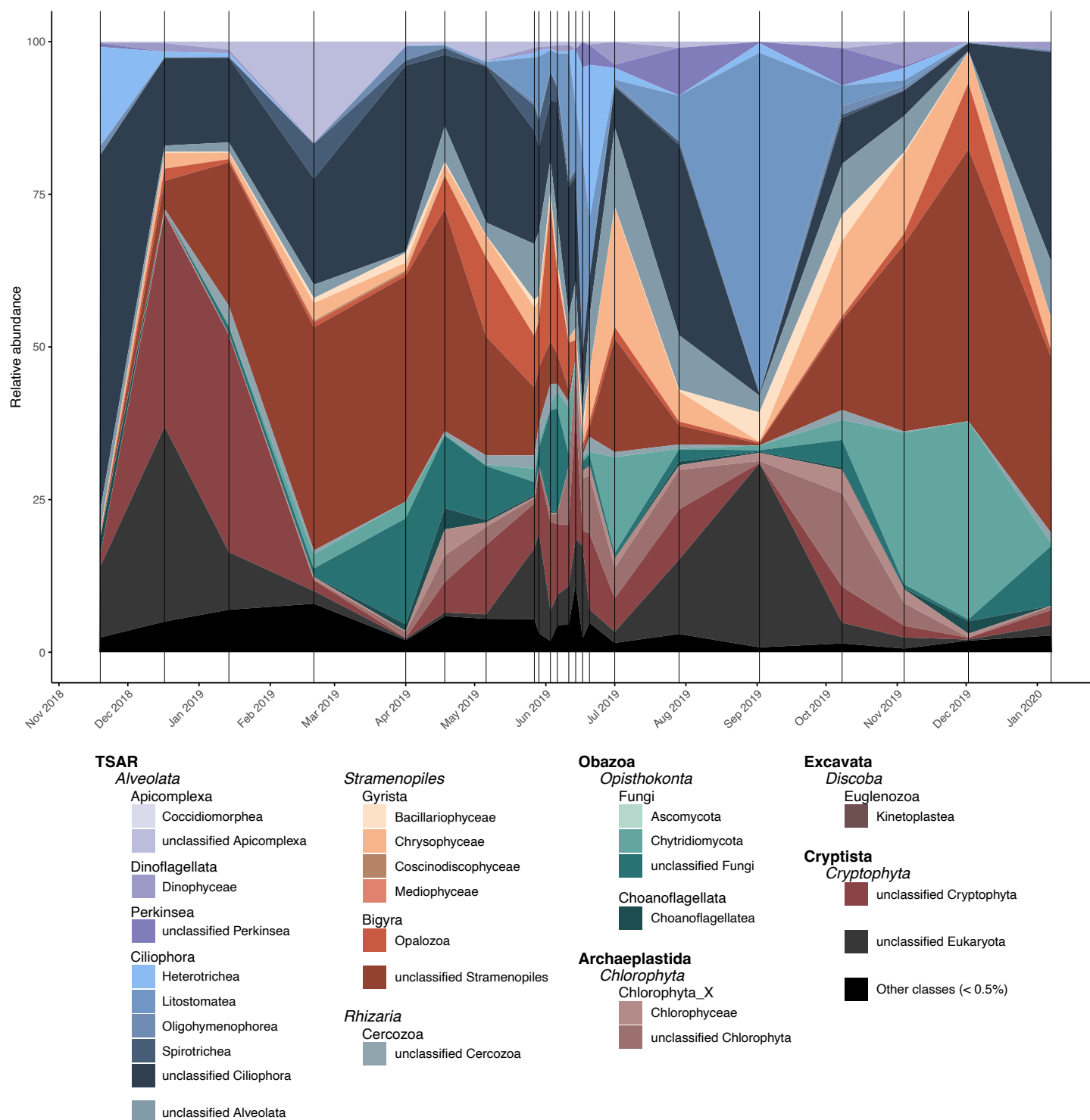


Figure S2. Temporal evolution of the relative abundances of the most abundant eukaryotic classes throughout the studied period using short reads.

The black zone on the bottom represents the cumulative proportion of other classes with low abundances (< 0.5% of the total reads). Vertical bars represent sampling dates.