Evolution in Chrysophyceae regarding the nutritional mode and intraspecific variation based on comparative genomics

INAUGURAL-DISSERTATION

zur Erlangung
des Doktorgrades
Dr. rer. nat.

der Fakultät für

Biologie

an der

Universität Duisburg-Essen

vorgelegt von

Stephan Majda

aus Aschaffenburg
16. Juni 2020

Die der vorliegenden Arbeit zugrunde liegenden Experimente wurden in der Abteilung für Biodiversität der Universität Duisburg-Essen durchgeführt.

Gutachter: Prof. Dr. Jens Boenigk

Gutachter: Prof. Dr. Bánk Beszteri

Vorsitzender des Prüfungsausschusses: Prof. Dr. Florian Leese

Tag der mündlichen Prüfung: 05.05.2020

DOI: 10.17185/duepublico/71907
URN: urn:nbn:de:hbz:464-20200616-103042-2

Preface & Acknowledgements

Dear reader,

welcome to the first and only version of my dissertation. Three publications in three years don't seem much, but the workload behind the bioinformatics was extensive. I created over 240 computer scripts (although including some replicates due to different version variations). This would not have been possible without the support and freedom of employment that I have received. Therefore, I would like to thank my superiors Jens and Daniela for their backup and time they have spent for me. I would also like to thank the rest of the team for the good working atmosphere. Further, I would like to thank Bank for taking time as second supervisor. Finally, I would like to thank my wife for accepting my "I'll do that as soon as I have time.", I will now need a new excuse.

Summary

Many protists are capable of photosynthesis. In several cases, species developed from phototrophy via mixotrophy to heterotrophy. Due to this development, in Chrysophyceae several species evolved independently of each other, making them an ideal model to study this change. In order to understand the evolution and effects of nutritional shift, the main experiments carried out so far had been feeding experiments and few transcriptome experiments.
In the present thesis, whole genome sequencing of several Chrysophyceae strains provide far-reaching insights into microevolution and evolution of nutritional change. Genome characteristics such as ploidy, GC content, gene density, functional diversity of genes, metabolic pathways and shared genes were analyzed supported by bioinformatics methods. Additionally, genome and cell size were determined by flow cytometry.
Three *Poteriospumella lacustris* strains were investigated for intraspecific variation: about 70 % of the genes are shared and genetic variation is mainly present in the secondary metabolism. The main difference between the strains is the different ploidy. Considering the different feeding modes, the following can be stated:

- The cell and genome size depends on the mode of nutrition. Both decrease from phototrophic to heterotrophic organisms. Mixotrophs lie in between.
- The GC content in heterotrophs is higher, due to different carbon and nutrient limitations.
- Reduction of the genome was subject to neutral evolution. Gene loss was predominantly accidental.

My work generates extensive insights into differentiation, changes of the nutritional mode and their effects, which contributes to the general understanding of evolution.

Zusammenfassung

Viele Protisten sind zur Photosynthese fähig. In mehreren Fällen entwickelten sich Arten aus einer zuvor phototrophen Ernährung über die Mixotrophie zur Heterotrophie. In Chrysophyceae evolvierten diesbezüglich mehrere Arten unabhängig voneinander, was sie zu einem idealen Modell zur Untersuchung dieser Veränderung macht. Um die Evolution und Auswirkung des Ernährungswechsels zu verstehen, wurden bislang hauptsächlich Fraßversuche und wenige Transkriptom Versuche durchgeführt.
Die vorliegende Arbeit liefert durch Genomsequenzierung mehrerer Chrysophyceae Stämme weitreichende Erkenntnisse zu Mikroevolution und Evolution des Ernährungswechsels. Mit Hilfe bioinformatischer Methoden wurden Ploidy, GC Gehalt, Gendichte, funktionelle Vielfalt der Gene, Stoffwechselwege und gemeinsame Gene analysiert. Zusätzlich wurden Genom- und Zellgröße mittels Durchflusszytometrie bestimmt.
Anhand dreier *Poteriospumella lacustris* Stämme konnte die innerartliche Variation gezeigt werden: ca. 70 % der Gene werden geteilt und genetische Variationen liegen überwiegend im sekundären Metabolismus vor. Vor allem unterscheiden sich die Stämme durch unterschiedliche Ploidy-Grade. Betrachtet man die unterschiedlichen Ernährungsmodi, lässt sich folgendes erkennen:

- Die Zell- und Genomgröße hängt vom Ernährungsmodus ab. Beide Parameter nehmen von phototrophen zu heterotrophen Organismen ab. Mixotrophe liegen dazwischen.
- Der GC Gehalt in Heterotrophen ist höher, begründet durch unterschiedliche Kohlenstoff- und Nährstofflimitierungen.
- Reduktion des Genoms unterlag neutraler Evolution. Genverlust war überwiegend zufällig.

Meine Arbeit liefert umfangreiche Erkenntnisse zu Differenzierung, Wechsel des Ernährungsmodus und deren Auswirkungen auf das Genom, welche zum generellen Verständnis der Evolution beitragen.

Contents

Chapter 1

Introduction

1.1 Modes of nutrition

Every organism needs energy and components such as carbon and nutrients to survive, grow and reproduce. There are different approaches of nutrient uptake. Autotrophy or phototrophy enables an organism to gain energy from light. On the contrary, heterotrophic species obtain energy by degrading organic molecules. An organism capable of both is called a mixotroph. Between phototrophy and heterotrophy there are numerous intermediate gradations. Based on the resource availability a mixotrophic organism switches between the modes of nutrition, whereby it prefers a primary nutrionional type (Jones, 2000).

The chloroplast in protists and plants was obtained by an endosymbiotic event (Mereschkowsky, 1905). Consequently, phototrophic protists possessed the ability of phagocytosis at least in their evolutionary history (Raven, 1997). In contrast to the acquisition of the phototrophy, the return to the heterotrophy occurred frequently and in several phyla independently (Krause, 2012; Kamikawa et al., 2015; Suzuki et al., 2018). Additionally, different feeding types can appear within one lineage, for example in Chrysophyceae (Hausmann et al., 2004; Graham et al., 2009).

1.2 Chrysophytes

The chrysophytes are a group of flagellated protists with predominantly freshwater habitation (Andersen, 2007). They were originally subdivided into the closely related classes Chrysophyceae (Pascher, 1914) and Synurophyceae (Andersen, 1987), but both are currently combined together (Lavau et al., 1997; Andersen, 1987). Chrysophytes were also called golden algae, because of their brownish colour based on the pigment fucoxanthin (Jeffrey et al., 2011). However, heterotrophic species are colourless, since they reduced their plastid. The cell envelope of chrysophytes can be covered with and without silica scales (Škaloud et al., 2013). Besides, some species are capable of building cysts (called stomatocyst) (Pascher, 1912; Sandgren, 1991).

Chrysophytes are globally distributed (Kristiansen, 2000) and abundant in freshwater (del Campo and Massana, 2011; Kammerlander et al., 2015; Weitere and Arndt, 2003). They carry out ecological functions of high relevance. For example, heterotrophic chryso-

phytes are among the most important grazers of bacteria-sized microorganisms (Finlay and Esteban, 1998; Bjorbækmo et al., 2019). Further, photosynthetic algae generate around 50% of the global oxygen production (Field et al., 1998). Especially, photosynthetic chrysophytes are predominat primary producer in oligotrophic waters (Wolfe and Siver, 2013).

Chrysophytes cover any nuance of nutrition between phototrophy, mixotrophy and phagotrophy. There are even gradations within one species, e.g. different prey behaviour and different expression of photosynthesis related genes in strains of *Ochromonas* sp. (Lie et al., 2018). This nutritional flexibility led to several independent shifts of nutrional modes within the chrysophytes (Grossmann et al., 2016; Dorrell et al., 2019). The change of nutrition was accompanied by plastid reduction and disabling of related pathways (Graupner et al., 2018).

1.3 Species boundary and intraspecific variation

Like May (May, 2010) imagined, would the first question of an alien visitor be: "How many distinct life forms–species–does your planet have?". Indeed, it is hard to tell since we have discovered only a small fraction of species on Earth ($\sim$14%) and in the ocean ($\sim$9%) (Mora et al., 2011) as well as there are inconsistent species concepts (there exist more than 30 concepts; Zachos 2016). Two initial aspects were morphological distinction (Ernst, 1963) and the predefinition that sexual reproduction can only take place within one species (Barr and Pollard, 1962; Linné and Stearn, 1957). The awareness of cryptic species, which were indistinguishable by morphology, led to a need for molecular methods to recognize accurate species richness (Hillis, 1987). Thereby, the question arises : How similar are species on a molecular level and how flexible is their intraspecific variation? A study based on 90,000 prokaryotic genomes determined clear species boundaries (Jain et al., 2018). However, other phyla likely have different extents of intraspecific variation.

1.4 Next Generation Sequencing & Bioinformatics

Next Generation Sequencing triggered a boost in molecular data. However, despite the advantages of assembly and binning methods the processing of sequencing data is still a huge challenge. In principle, a sequencing project always takes place in the following order: sequencing, filtering undesired reads, assembly, data analysis.

Sequencing

Sequencing generates short (e.g. Illumina or IonTorrent technique) or long (e.g. PacBio or MinION technique) fragments of the DNA called *reads*. Both variants have their advantages and disadvantages. In general, short reads have a lower error rate and are cheaper (Quail et al., 2008), while long reads could span over repeat regions and are easier to handle in computer applications (Carneiro et al., 2012). If you want to sequence a genome for which no reference exists, the *de novo* whole-genome shotgun sequencing is suitable. In this case the DNA is fragmented randomly by restriction enzymes or shearing action and synthesized multiple times (Sanger et al., 1980).

Genome Assembly

The sequenced reads have to be pieced together to reconstruct the genome. One possibility would be to join overlapping reads (Staden, 1979). Here, the reads are placed on top of each other in such a way that as few mismatches as possible occur. This could be determined by algorithms like the Smith-Waterman algorithm (Smith and Waterman, 1981) or the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). The aligning and merging of the optimal reads build so called *contigs*. Their base composition is defined by the majority of the present nucleotides at a given position. This is called *consensus* sequence. Several contigs with known order and distance to each other are defined as *scaffolds*.

Nowadays, reads are merged within graphs. A graph consist of a root and its outgoing branches. A special case are De Bruijn graphs (de Bruijn, 1946). Each sequence can be divided in fragments of a certain length. This fragments are called *k*-mers, whereby k represents the length of the fragment (e.g. "ATGC" -> 4-mer). A De Bruijn graph is based on all possible *k*-mers of a sequence. Each node is a unique *k*-mer and is connected by overlapping parts in the order of the original sequence. This can be used in bioinformatics (Pevzner et al., 1989; Idury and Waterman, 1995), since after sequencing DNA fragments exist, which were divided in *k*-mers and ordered by their overlapping parts. The genome is determined by walking the most likely (highest read coverage) path through the graph and connecting passed nodes (Pevzner et al., 1989; Idury and Waterman, 1995), whereby there is an attempt to use each node only once. The main advantage is that the algorithm behind De Bruijn graphs is much faster than the traditional overlap approach and needs less memory (Flicek and Birney, 2009; Li et al., 2011). One disadvantage is that repeats build loops in the graph, which are usually not resolved, since each node should only be used once. This disadvantage can be reduced e.g. by variable *k*-mer sizes (Lima et al., 2017) or appropriate statistics (Novák et al., 2010).

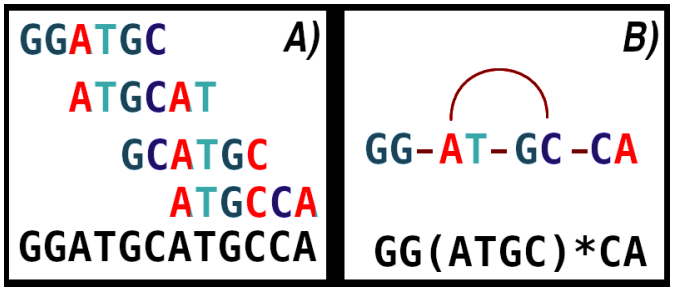


Figure 1-1: Concept of assembly methods

There exists two common methods of sequence assembly:
A: Reads (coloured) were overlapped to build consensus sequence (black).
B: Out of the reads *k*-mers (here 2-mers, coloured) were extracted and connected with its adjacent *k*-mers. In the the example sequence: GGAT**GC**AT**GC**CA the GC occurs two times followed by different *k*-mers. This ambiguity causes a branching. The repetition of ATGC provokes a loop, whereby their frequency in the consensus sequence (black) is vague.

A further assembly improvement is the combination of long and short sequencing techniques. These, so called hybrid-assemblies combine the advantage of the low error rate

from short reads and the structure terminability of long-read-methods (Sovic et al., 2016; Antipov et al., 2016). Here, out of the short reads a De Bruijn graph is built. Afterwards, the long reads overlay the pathways through the graph, which enables connecting edges and resolving repeat regions (Antipov et al., 2016).

Before the assembly process, the reads were filtered to reduce the error rate of the assembly. Thereby, reads of bad quality are removed or partly trimmed.

Binning

Another reason for filtering reads could be to select a specific organelle or species in a mixed sample. *Binning* is the process of separating reads or contigs to groups (also called *bins*), which have similar features (ideally belonging to the same species). From this point of view bins are comparable to *OTU*s (operational taxonomic units) containing one or dozens of species. The binning can be based on two techniques:

- Taxonomic or phylogenetic assignment
- Compositional features

The taxonomic binning uses sequence similarity to known taxa (Dröge and McHardy, 2012). Therefore, either the complete sequence is aligned against a known sequence (Tringe et al., 2005) or marker genes, mostly 16S rRNA, were used (Langille et al., 2013). Compositional features are for example abundance or GC content (Tyson et al., 2004). Instead of GC content often tetranucleotide frequencies are used, which is more accurate (Teeling et al., 2004). Both, phylogenetic and compositional methods could be combined to increase accuracy (Wu et al., 2016). After the binning process, each bin can be assembled separately. This second assembly consists of far less chimeric sequences than the first assembly with all reads.

Data analysis

The assembled genome can be used for numerous types of analysis. One way is the prediction of genes. This can be done in two ways:

- The *extrinsic content sensors* method is based on the comparison with homologous genes, which either works fine in well-known prokaryotes or find about 50% of the genes (Mathè et al., 2002).
- *Ab initio* methods look for gene patterns.

These gene patterns were evaluated by specific functional sites (or signals) in the genomic sequence like transcription factor binding sites, TATA boxes, poly(A) sites, generally ATG with exceptions, and stop codons (Mathè et al., 2002). Further, compositional information as k-mer occurrences, frame length and composition (Saeys et al., 2007) or Z-curve (a mathematical method to reduce multidimensional data) parameters (Gao and Zhang, 2004)

can also be used. Usually, Markov chain models are applied to differentiate coding and non-coding regions (Borodovsky and Lomsadze, 2011; Stanke et al., 2006).

After gene prediction, the genes are normally annotated. The annotation is an alignment of the predicted genes with a gene or protein database, which have additional biological information to these genes. Typical databases are the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa and Goto, 2000), Uniprot (UniProt Consortium, 2018) or *Gene Ontology* (GO) (The Gene Ontology Consortium, 2019).

Problems, problems, problems... or the challenges of bioinformatics

During these bioinformatical steps several difficulties may occur:

- Binning:
  - Taxonomic assignment is crucial in non-axenic cultures. However, the assignments quality of current binning tools is good up to the family rank, but lacks on genus or species level and low abundant strains (Sczyrba et al., 2017).
  - The rRNA marker genes are only present at about 0.1% of the reads (McHardy et al., 2007).
  - Plasmids or mitochondria would possibly bin separated from their host, because their compositional features differ (Ingman et al., 2000; Davis and Olsen, 2009). Further, some species have multiple organelles or plasmids (Macrina et al., 1978; Bendich, 1987) leading to higher abundances than the nuclear genome.
- Assembly:
  - Errors in the binning process directly effect the assembly quality by increased chimeric sequences.
  - *De novo* assemblies of new species are more complex than reference based assemblies.
  - Larger (eukaryotic) genomes are more complex and harder to assemble (Chu et al., 2013; Bradnam et al., 2013). Especially, repeats are error-prone in Illumina sequencing (Heydari et al., 2019).
  - Hybrid assemblies are not always better and can come along with a higher error rate (De Maio et al., 2019).
  - Assembly and binning tools had been developed and tested on predominantly prokaryotic datasets (Sczyrba et al., 2017).
- Gene prediction and annotation:
  - Gene prediction is even more difficult in large genomes (Guigo et al., 2000).

- Since Markov chains are based on recurrent patterns in DNA sequences, short or fragmented sequences in draft genomes complicate the gene prediction (Stanke et al., 2006).
- Large genes (for example, the human dystrophin gene with 79 exons and a size of 2.3 Mb; Nobile et al. 1997), large introns, overlapping genes from different strands, frameshifts, intron in non-coding regions etc. are often misleading (Mathè et al., 2002).
- Gene databases contain information about gene functions. However, this informations are mostly based on model organisms. The same also applies to annotated pathways, which can vary considerably between model and non-model species.
- A found gene does not necessarily mean that it is also functional.

Knowing these error sources enables to develop the best possible solution according to the circumstances.

1.5 The scope and objectives of this work

In protists a shift from phototrophy to heterotrophy occurred many times independently in the course of evolution, i.e. this switch is essential for understanding the evolution of today's biodiversity. Comparative genomics of species with different nutrition modes reveal how this transformation took place and indicates what frequently triggered this evolutionary pattern as well as plastid loss. Knowledge about nutrition acquisition development supports the understanding of global biodiversity and ecology.

Until now, only one genome of Chrysophyceae has been publicly released (Bråte et al., 2019) besides my publications. The genomic data is a fundament for future studies and reference to related (and also not well genomically researched) stramenopiles. Besides, the evolutionary aspects mentioned above, my data enables understanding genomic structure and build a basis for functional gene analyses (for example, by facilitate knock-out experiments).

Comparative genomics relies on a large amount of high-throughput data, which is why bioinformatics is necessary. My work is one of the first investigating the intra- and interspecific genomic variation in free-living flagellates. Furthermore, it is among the first investigating the pan/core genome of protists. Hence, I developed bioinformatics methods, scripts and pipelines to fulfil the species specificity and the experimental setup. These pipelines have been designed so that future sequencing projects can base on their easy automation. I determined the intraspecific variation in Chrysophyceae (chapter II), before I explored the interspecific variation based on different nutritional modes (chapter III) with the overall aim to gain knowledge about the impact of the nutritional shift.

Objective: Determine intraspecific variation within Chrysophyceae (chapter II)

- There's a lot of disagreement in species delimitation and the flexibility of genetic variation (Zachos, 2016). Therefore, I determined the intraspecific variation in three strains of *Poteriospumella lacustris* regarding genomic features as gene content, gene density, ploidy and mutation rate. Finally, the intraspecific variation of these strains should be transferable to represent the intraspecific variation in Chrysophyceae.

Objective: Explore impact and evolution of nutritional shift (chapter III)

- The first noticeable effect of the different nutritional modes is the cell size. I investigated the correlation between nutritional mode and cell size respectively genome size based on flow cytometry. Further, the genomes between different trophies were compared concerning GC content, gene density, gene content and functions. The transcriptome gives reasons to expect an evolutionary gene loss pattern (Graupner et al., 2018). It should be examined whether this can be verified.

Automation (included in chapter II & III)

- All analyses should be repeatable with further strains. Therefore, I created a bioinformatic pipeline automating the working steps genome assembly, binning, gene prediction, all analyses steps and visualisation without the necessity of manual parameter adjustment.

Chapter 2

Intraspecific variation

Publication 1:

Intraspecific Variation in Protists: Clues for Microevolution from Poteriospumella lacustris (Chrysophyceae)

Stephan Majda, Jens Boenigk, Daniela Beisser

Published in: Genome Biology and Evolution (2019), 11(9): 2492-2504.
https://doi.org/10.1093/gbe/evz171

Contribution to this publication:

conception & planning:	30 %	planing bioinformatical and laboratory methods
experimental work:	95 %	cultivation, DNA isolation, testing methods
data analysis:	100 %	testing methods, programming, debugging, collect web data, data analysis, statistic analysis, interpretation of results
writing the manuscript:	75 %	writing the draft, creation of tables and figures, several rounds of internal revision
revising the manuscript:	70 %	text revision and additional analyses

....
Stephan Majda Prof. Dr. Jens Boenigk

Intraspecific Variation in Protists: Clues for Microevolution from *Poteriospumella lacustris* (Chrysophyceae)

Stephan Majda *, Jens Boenigk†, and Daniela Beisser†

Department of Biodiversity, Duisburg-Essen, Germany

†These authors contributed equally to this work.

*Corresponding author: E-mail: stephan.majda@uni-due.de.

Accepted: August 2, 2019

Data deposition: This project has been deposited at NCBI under the accessions PRJNA504602, PRJNA504603, PRJNA504604.

Abstract

Species delimitation in protists is still a challenge, attributable to the fact that protists are small, difficult to observe and many taxa are poor in morphological characters, whereas most current phylogenetic approaches only use few marker genes to measure genetic diversity. To address this problem, we assess genome-level divergence and microevolution in strains of the protist *Poteriospumella lacustris*, one of the first free-living, nonmodel organisms to study genome-wide intraspecific variation.

Poteriospumella lacustris is a freshwater protist belonging to the Chrysophyceae with an assumed worldwide distribution. We examined three strains from different geographic regions (New Zealand, China, and Austria) by sequencing their genomes with the Illumina and PacBio platforms.

The assembled genomes were small with 49–55 Mb but gene-rich with 16,000–19,000 genes, of which ~8,000 genes could be assigned to functional categories. At least 68% of these genes were shared by all three species. Genetic variation occurred predominantly in genes presumably involved in ecological niche adaptation. Most surprisingly, we detected differences in genome ploidy between the strains (diploidy, triploidy, and tetraploidy).

In analyzing intraspecific variation, several mechanisms of diversification were identified including SNPs, change of ploidy and genome size reduction.

Key words: chrysomonad flagellates, genome comparison, whole genome sequencing, ploidy.

Introduction

Genetic variation permits flexibility and survival of a population under changing environmental conditions (Reed and Frankham 2003) and leads over time to genetic differences between strains or populations (diversification) from different geographic regions or environments. Genetic variation therefore provides insights into the evolutionary history of a species (Knoll 1994; Darling et al. 2004).

Eukaryotes can have large intraspecific variation in DNA content (Parfrey et al. 2008). For example, recent studies of intraspecific genetic variation using DNA fingerprinting techniques in aquatic phytoplankton have identified high levels of diversity. Average gene diversity, which gives the probability that two alleles chosen at random from a population will be different from each other, range from 39% to 88% (Godhe and Rynearson 2017). Logares et al. (2009) found among five dinoflagellates a genetic diversity between 20% and 90%. Several further studies demonstrated a surprisingly high intraspecific genetic diversity in both marine and limnic species (John et al. 2004; Evans et al. 2005; Rynearson and Armbrust 2005; Hayhome et al. 2007). However, these experiments were all based on microsatellites or DNA fingerprinting and reflect only a small part of the entire intraspecific variation in microeukaryotes.

A high intraspecific variation may not be considered surprising as under a neutral model, a population's genetic diversity depends (besides the gene's mutation rate) on the effective population size (Kimura 1979). Because of their short generation time and huge population size (Watts et al. 2013), the genetic variation in protists may thus be enormous. Despite a potentially high intraspecific variation, the mechanisms of microevolution and with that of speciation may not

necessarily correspond to those discussed for multicellular organisms: Many protist taxa are sexually reproducing (Raikov 1995; Heywood and Magee 1976) but a separation of gene pools between subpopulations does not necessarily occur (Fenchel and Finlay 2004) as protists have a high potential for long-range and persistent dispersal and many taxa show a cosmopolitan distribution (e.g., Finlay 2002).

On the other hand, protist populations, in particular in freshwater lakes, are separated as lakes are considered to act like islands (as in the theory of island biogeography; MacArthur and Wilson 1967); that is, despite the huge population size and the ease of dispersal, populations may be (at least temporarily) separated. Corresponding to these considerations, limited distribution and geographic separation have been shown to apply also to protist populations (Fernandez et al. 2017; Bestová et al. 2018; Boenigk et al. 2018). However, the geographic differences on the community level, in particular the low contribution of geography as demonstrated for protists, do not necessarily indicate evolutionary separation of subpopulations but may result from regional extinctions of species or temporal fluctuations between plankton and seed bank. Despite recent indications for geographic isolation of protist communities, it remains uncertain to what extent this separation applies to the level of subpopulation, that is, to intraspecific variation, which would provide a basis for speciation by geographic separation. The huge population sizes of protists and the relative ease of dispersal may, hinder or even prevent speciation by geographic separation as known for many multicellular organisms.

The extent of intraspecific genomic variation in protists is still obscure. Here, we examine the genomic molecular variation between three clonal strains of the heterotrophic chrysophyte species *Poteriospumella lacustris*. The strains were collected in different geographical regions (JBC07 in China, JBM10 in Austria, and JBNZ41 in New Zealand; Boenigk et al. 2005). Despite the geographic remote sampling sites they have identical SSU and ITS sequences except for one base deviating in the ITS region of strain JBNZ41 (from 835 bp) and show similar ecophysiological characteristics (Boenigk et al. 2005, 2007). Despite this high similarity backing up the affiliation with the same species, transcriptome studies indicated a high intraspecific variation (Beisser et al. 2017). It remains unclear to what extent the reported gene content variation is artificial due to differential gene expression or rather reflects true genomic intraspecific variation. Further, the transcriptome studies suggested a different degree of genetic variation for different pathways but again conclusions remained vague due to incomplete gene coverage.

In order to resolve the puzzle of genetic variation within this free-living protist species we here study the intraspecific genome-level variation. We test the hypothesis that genes coding for primary and secondary pathways differently reflect the accumulation of intraspecific molecular variation. In particular, we expect genes affiliated with the basal metabolism to be conserved, whereas genes affiliated with the secondary metabolism and with pathways directly interacting with the environment should be more diverse due to adaptations to changing environmental conditions; that is, the gene variation in primary metabolism is assumed to be higher than in secondary metabolism. We further test whether intraspecific variation with respect to the accumulation of point mutations, changes of ploidy and genome reduction is weak as would be expected based on the close relatedness. Or alternatively, whether the variation is high as would be expected due to the large population sizes and the global distribution of this species. We further analyze intraspecific genomic variation for indications of evolutionary differentiation and recent population bottlenecks.

In order to address the above hypotheses we sequenced the three strains using Illumina and PacBio sequencing platforms and the assembled genomes were examined by comparing, for example, gene content, gene density, SNPs, proportion of repetitive regions, ploidy, and GC content. We further identified gene functions and pathways using the *Kyoto Encyclopedia of Genes and Genomes* (KEGG) database (Kanehisa and Goto 2000).

Materials and Methods

Cultivation and Sequencing

Three clonal strains of *P. lacustris* (JBM10, JBNZ41, JBC07; Findenig et al. 2010) from the culture collection of the working group were cultivated according to Hahn et al. (2003) in NSY medium under a light:dark cycle of 12:12 h at 16 C. Before harvesting, the axenic cultures were tested for contaminations using light microscopy. The DNA was isolated (BioBudget, my-Budget DNA Mini Kit, Krefeld, Germany) and sequenced by sequencing provider (BGI, Hong Kong) using the Illumina Hiseq XTen technology (insert size 300 bp, BGI in-house reagents) for the strains JBC07, JBM10, and JBNZ41 and PacBio RSII for strain JBM10. PacBio processing was done with gTUBE (Covaris, USA) for shearing genome DNA to 20 kb, DNA Template Prep Kit 3.0, DNA/Polymerase Binding Kit and DNA sequencing Reagent 4.0 (Pacific Biosciences, USA).

Genome Assembly

Unless otherwise stated, the default settings were used for the following programs. A Snakemake (Koster and Rahmann 2012) automated workflow was created to process the sequencing data. We benchmarked the N50 statistic of different assemblers (SPAdes, v3.10.1 [Nurk et al. 2013], Celera, v8.3 [Myers et al. 2000], ABySS, v2.0.2 [Simpson et al. 2009], CANU, v1.5 [Koren et al. 2017]) and chose supported by literature, for example Forouzan et al. (2017), Sovic et al. (2016) and the implementation of hybrid approaches, the following assembler and procedure: First CANU (with parameters:

genomeSize = 96m correctedErrorRate = 0.105; Koren et al. 2017) was used to assemble the PacBio reads of strain JBM10. The genome size parameter was chosen based on estimates from Olefeld et al. (2018). Subsequently, with SPAdes (v3.10.1, with parameters: –untrustedcontigs; Nurk et al. 2013) the Illumina reads of each strain were assembled using the PacBio reads of JBM10 as scaffolding template. By this hybrid assembly approach a simplified de Bruijn graph was constructed and overlaid with an assembly graph of long contigs at graph edges to close gaps and resolve repeats (Antipov et al. 2016) combining low error rates with long scaffolds. The Illumina reads were decisive for the assembled sequence allowing comparisons between strain sequences later on. After assembly, contigs smaller than 500 bp were discarded. About 23,000 18S DNA sequences of Chromista (from NCBI) were blasted (BlastN, v2.5.0, with parameters: -percidentity 99; Boratyn et al. 2012) against the scaffolds to validate the correctness of the strains and to exclude the possibility of contamination. Genome size was first estimated by kmers ($k = 21$) with KMC (Dlugosz et al. 2017) and GenomeScope (Underwood et al. 2017). However, we changed the method due to discrepancies (see supplementary table S1, Supplementary Material online) with the estimation from nuclear staining and flow cytometry (Olefeld et al. 2018). Hence, the length of all contigs was summed up for each strain. Genome characteristics were compared with all available stramenopile genomes from NCBI (last accessed February 2019). Genomes smaller than 50 Mb were excluded, since these are usually parasites or organelles, which are not suitable for comparison.

The Benchmarking Universal Single-Copy Orthologs (BUSCO) software (v3.0.2; Simao et al. 2015) was used to verify the existence of all essential orthologous genes and to measure genome completeness. For BUSCO data sets of protists and eukaryota were used. During data submission to NCBI, the genomes were aligned to publicly available organelle genomes and classified mitochondrial contigs.

Gene Prediction

Different approaches of gene prediction were tested (details see supplementary file, Supplementary Material online). In the final approach, the gene pattern of *Arabidopsis thaliana* was chosen as model for gene prediction with AUGUSTUS (v3.3 with parameters: –species = arabidopsis –gff3 = on –singlestrand = true –UTR = off; Stanke et al. 2006). Mapping the RNA sequences of the strains (Beisser et al. 2017) to the predicted genes with Bowtie (v.2.2.8 with parameters: –very-sensitive-local; Langmead et al. 2009) functioned as validation of the prediction procedure.

Mitochondrial genes were predicted and annotated by aligning the genes of the *Ochromonas danica* mitochondrial genome (from NCBI, ACCESSION number: NC_002571) with the contigs, that were identified as mitochondrial sequences, using Minimap2 (2.16-r922; with parameters: -c -G 80 K; Li 2018).

Gene Clustering

Since the genome assembly yielded many contigs, the gene prediction could likely miss several genes between or at the edge of the contigs. Mapping reads from one strain against exclusive genes from another strain showed which genes could possibly be build with the read set. We merged the predicted genes of all strains and clustered them with CD-HIT (v4.6.8 with parameters: cd-hit-est -n 8 -M 20000 -T 18 -s 0.8 -aL 0.8 -aS 0.8 -G 0; Li and Godzik 2006). Gene clusters were merged by their consensus sequence. In order to prevent overlooking of genes, the reads of each strain were mapped with BWA (Li and Durbin 2009) against this pool of predicted genes. Including this information, all genes could be covered completely (coverage $> 5\times$) by the sequencing reads of each strain (see fig. 1). To exclude that genes were missing due to incorrect assemblies, the strain specific genes of JBC07 and JBNZ41 were further aligned to the PacBio-sequenced contigs of JBM10. We aligned the pool of predicted genes against each genome to detect correct genes with Minimap2 (v.2.9-r720; with parameters: -x splice -G 80K). The alignment of predicted genes of JBM10 against its genome was used as a validation of whether spliced genes were correctly aligned. Multiple genes aligning on the same strand with an overlap of at least 10% of the gene length were declared as one gene or a splice-variant of the gene. Genes aligning on several genomes were declared as shared genes whereas genes aligning to only one genome were specific.

Gene Annotation

Using Diamond (v0.9.10.111; Buchfink et al. 2015), we aligned the predicted genes to the KEGG (Release 2014-06-23; Kanehisa and Goto 2000) and UniProt database (Release 2017-09-18; Pundir et al. 2017) to obtain KEGG Orthology (KO) identifiers. Both databases complemented each other. In case of inconsistencies the lower *e*-value defined the assignment. KO identifiers operate as unique flags for a functionally orthologous group of genes. Thereby, a species independent gene annotation and comparison is made possible. Furthermore, the "KEGG Mapper—Reconstruct Module" tool was used to reveal the metabolic pathways. Thereby, a module describes a pathway that is necessary for a defined function. A functional module consisting of only two genes was confirmed if it was completely, whereas modules with more than two genes were allowed to miss a maximum of one gene. Inspection of the essential primary metabolic pathways (relevant for energy production, metabolism of carbohydrates, lipids, amino acid and genetic information processing) was used as completeness check for the genome.

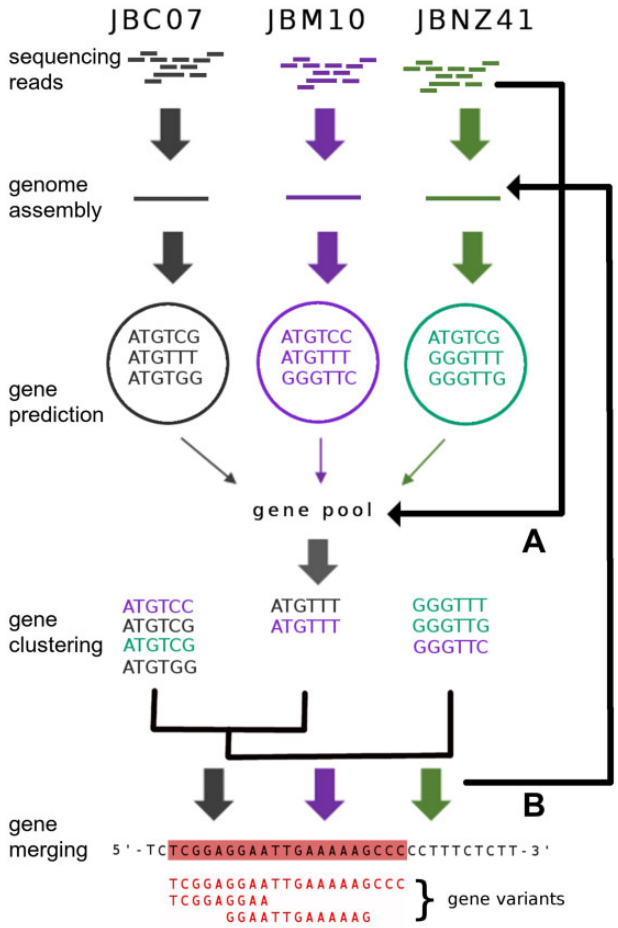


Fig. 1.—Flow chart of gene prediction and analysis. AUGUSTUS (Stanke et al. 2006) predicted genes based on assemblies of each strain genome. Predicted genes were pooled and clustered with CD-HIT (Li and Godzik 2006). (*A*) Sequenced reads of each strain were mapped against the gene pool. (*B*) Clustered genes were aligned against each genome. Genes overlapping on the same strand were merged. This procedure prevents overlooking of genes and combines gene duplicates and variants. Step *A*: All reads of each strain could be map back. Step *B*: Genes aligned either specific or to multiple genomes.

We searched for signal peptides in all predicted genes of each strain with SignalP (v5.0; with parameters: -batch 100000, Almagro Armenteros et al. 2019) and HECTAR (Gschloessl et al. 2008) to identify proteins that are translocated across membranes. 25% of the genes with the lowest validation score identified by SignalP were filtered out to increase specificity. These genes were grouped in the category *genes encoding organelle targeted proteins* and analyzed like the other KEGG functional groups.

Genome Similarity, Ploidy Estimation, and Analysis of Repetitive Regions

Genome similarity based on the average nucleotide identity (ANI) was calculated with FastAni (v1.1 Jain et al. 2017) for the whole genome. We aligned the mitochondrial sequences pairwise between the strains with Minimap2 (v2.16-r922; with parameters: -cx asm5 -Y) to determine similarity between them. Since they only differ in a few bases, the mitochondrial sequences were additionally mapped (Minimap2 with parameters: -c –splice) against the PacBio contigs of JBM10 to exclude all mitochondrial sequences originate from the same PacBio template.

RepeatScout (v1.0.5; Price et al. 2005) was used to construct a de novo repetitive sequences library for each strain. Including this libraries RepeatMasker (v4.0.7; http://www.repeatmasker.org; last accessed August 2019) identified repetitive sequences and masked them. Genome ploidy was estimated by nQuire (version from April 5, 2018; with parameters: nQuire create -c 20 -q 15 with following denoise step; Weiß et al. 2018). Additionally, ploidy was determined for the repeat masked genomes as well as separately for each contig longer than 10,000 bp. For the contig-wise ploidy estimation, it was evaluated whether the probability for a certain ploidy level was at least 10% higher compared with the others. In this case ploidy counts were normalized by contig length. The tool nQuire uses different frequencies of SNP mutations on heterozygous alleles to determine the ploidy. Additionally, ploidy was estimated based on *k*-mer frequencies with *smudgeplot* (v0.1.3 with parameters: -k21 -m300 -ci1 -cs10000; https://github.com/tbenavi1/smudgeplot; last accessed August 2019). This package determines the number of *k*-mer pairs differing by one SNP and compares them to their relative coverage.

Gene Analysis

GATK haplotypecaller (v4.0.6.0 with parameters: ploidy 4 –emit-ref-confidence GVCF –output-mode EMIT ALL CONFIDENT SITES –annotate-with-num-discovered-alleles true –max-reads-per-alignment-start 0; McKenna et al. 2010) was used for variant calling on the whole genome. Here, all strains were consistently evaluated with parameter *-ploidy 4*, which served as upper boundary, but had no influence on the lower ploidy. Variants were excluded if read depth was <8 or the variant occurred in $<10\%$ of the reads. Next, the genes with their variants were extracted and subdivided by KEGG functional categories. For each kind of variation (SNP, insertion, deletion) the occurrence per gene was determined. Additionally, we compared the gene variation pairwise for each strain. If at any position there was a mismatch between two genes and no allelic variation could induce a match we listed a mutation. To normalize the data, the number of mutations were divided by the length of the gene to obtain the mutation rate of the allele. Additionally, identical genes between strains were counted. Afterwards a one-way ANOVA test verified significant differences (P-value < 0.01) within a functional group for each allelic variation (SNP, insertion, deletion) between the three strains. A following

Wilcoxon signed-rank tested which of the functional groups differed significantly. For the mutation analysis between strains, we repeated these steps, expect Kruskal–Wallis test was performed instead of ANOVA, because the data were not normally distributed.

Genes were divided into shared (if they could be aligned by Minimap2 (v.2.9-r720; with parameters: -x splice -G 80 K to more than one genome)) or exclusive genes and counted. The R package eulerr (v4.1.0; Larsson 2018) was used to visualize the amount of shared genes. The gene density (d) was calculated for the number of genes (n) by the formula:

$$d = \frac{n}{\sum bp_{contigs>500\text{bp}}} * 10^6 \quad (1)$$

Gene densities were compared with 197 protists from the Ensembl Protists database (accessed December 2017, https://protists.ensembl.org).

Results

Genome Assembly and Gene Prediction

Illumina sequencing generated 80.8 (JBC07), 79.6 (JBM10), and 109 (JBNZ41) million 150 bp-long paired end reads and 207,000 PacBio reads (JBM10) with a total length of 1,505 Mbps (reads lengths: mean = 7,243 bp, median = 7,130 bp). The sequence data were assembled into draft genomes for each strain (see table 1). The assembly of the PacBio reads for JBM10 produced 695 contigs (N50 = 143,709 bp). The combination of short Illumina reads and the assembly of PacBio reads as template resulted in a high number of contigs (9,122–13,826, see table 2). The obtained contigs had a coverage between 116× and 153×. The proportion of repeat regions amounted to 12–16%. The mitochondrial genome of each strain could be assembled almost completely (JBC07: 38,850 bp, 2 contigs; JBM10: 38,860 bp, 2 contigs; JBNZ41: 38,814 bp, 1 contig). About 760 genes with signal peptides were identified. Mapping against the PacBio contigs of JBM10 showed all mitochondrial sequences were based on the strain specific Illumina reads. From the BUSCO reference gene set for eukaryotes 81.5–83.8% genes were recovered. However, in the gene set for protists 54.5–55.8% genes were recovered (see supplementary table S2, Supplementary Material online). Taking into account the genome reduction in *P. lacustris* and analyzing the annotated genes using KEGG, verified that essential metabolic pathways were covered (see supplementary table S14, Supplementary Material online).

The gene density was roughly 310–370 genes/Mb. In comparison to other protists, *P. lacustris* is among the species with the lowest gene densities, but with a comparable density to related species in the group of stramenopiles (see fig. 2).

The gene prediction was evaluated with the respective transcriptome, but the RNA sequences could not be used to train and create a specific model for *P. lacustris*. Despite this, the data could be used to improve an existing model and for validation and accuracy checking. Applying the prediction model of *A. thaliana* 81–84% of RNA reads could be mapped back to the genome. Improving the model with transcriptomic data enhanced the mapping by 0.16%. In total, 55,941 (JBC07), 60,147 (JBM10), and 62,248 (JBNZ41) genes were predicted. By pooling the predicted genes and clustering we obtained a total amount of 47,500 unique genes, of which 25,693 genes could be annotated. This procedure merged duplicates and gene variants. The pooled genes were aligned back to each genome (sensitivity > 99%, see fig. 1), resulting in the following numbers of genes: The genomes of JBC07, JBM10, and JBNZ41 contained 17,315, 16,915 and 19,494 genes, respectively, of which 7,453, 7,960, and 8,525 were annotated with KEGG identifiers.

All three strains combined contained 21,551 genes. 14,756 genes (68.5%) were present in every strain (see fig. 3). 80.8% of the genes were shared with at least one other strain. The proportions of genes found in a single strain were 3.5% (JBC07), 3.1% (JBM10), and 12.5% (JBNZ41). Shared genes had a sequence similarity of 97.3% and 0.0018 deletions/inserts per gene on average.

Table 1

Genome Size Estimations

P. lacustris strain	JBC07	JBM10	JBNZ41
Total genome size (Flow cytometry) [Mb]	157.2	96.5	177.5
Haploid genome size (Flow cytometry) [Mb]	52.4	48.3	44.4
(Haploid) assembly size [Mb]	49.4	54.7	52.8
Estimated ploidy level	Triploid	Diploid	Tetraploid

Note.—The genome size was estimated in Olefeld et al. (2018) by nuclear staining and flow cytometry. The haploid genome size of the flow cytometry was recalculated by the total size and the ploidy estimation. Assembly size reflects the sum over the length of all contigs longer than 500 bp.

Ploidy Estimation

Ploidy estimation indicates triploidy in JBC07, diploidy in JBM10 and tetraploidy (or diploidy) in JBNZ41 based on the SNPs of the denoised data set as well as on the data set with excluded repeats (details see supplementary tables S4, S5 and fig. S2, Supplementary Material online). For the ploidy estimation on single contigs each data set consisted of 850–934 contigs. The criteria of a >10% higher probability of a certain ploidy was fulfilled for 78–79% (JBM10, JBNZ41) and 65% (JBC07) of the contigs. Estimating the ploidy for individual contigs of JBC07, the majority of the contigs (45%) was triploid, but also a high proportion of contigs (39%) were rated diploid (see supplementary S4, Supplementary Material online). The *k*-mer based approach approved the estimated ploidy levels (JBC07: triploid, JBM10: diploid, and JBNZ41: tetraploid; see fig. 4). The variation in ploidy level was

Table 2
Overview of Sequencing and Genome Characteristics

Species	JBC07	JBM10	JBNZ41	*Ectocarpus siliculosus*	*Nannochloropsis oceanica*	average stramenopiles
# contigs	9,122	9,400	13,826	13,533	32	—
N50	40,792	52,370	24,662	32,613	—	—
Coverage	116	128	153	—	—	—
GC %	53.1	53.1	52.9	53.5	54.1	49.7
Repetitive regions %	13.2	16.6	12.4	—	—	—
Predicted genes	55,941	60,147	62,248	—	7k–10k	—
Final genes	17,315	16,915	19,494	16,269	7k–10k	8,368
Annotated genes	7,453	7,960	8,525	—	222	—
Gene length (median)	1,612	1,887	1,331	243	455	—
Gene length (mean)	3,169	3,582	2,871	226	687	—
Gene density [Mb^{-1}]	350	309	369	83	—	269
Ratio coding DNA to total %	55.5	55.3	53.0	13.2	—	—
Average nucleotide identity (ANI) [%]	97.5	98.0	97.4	—	—	—

NOTE.—Predicted genes were aligned to the genome. Overlapping genes were merged to final genes. The gene density describes the number of genes per MB, with the gene lengths varying widely, whereas the ratio of coding DNA to total DNA is independent of the gene length. The genome similarity was expressed by the ANI. The stramenopiles data are based on the average of the statistics data of all available stramenopiles genomes ($n = 97$) from NCBI (https://www.ncbi.nlm.nih.gov/genome/? term=stramenopiles; last accessed August 2019). *Ectocarpus siliculosus* data also originate from NCBI. *Nannochloropsis oceanica* data are based on NCBI and Wang et al. (2014).

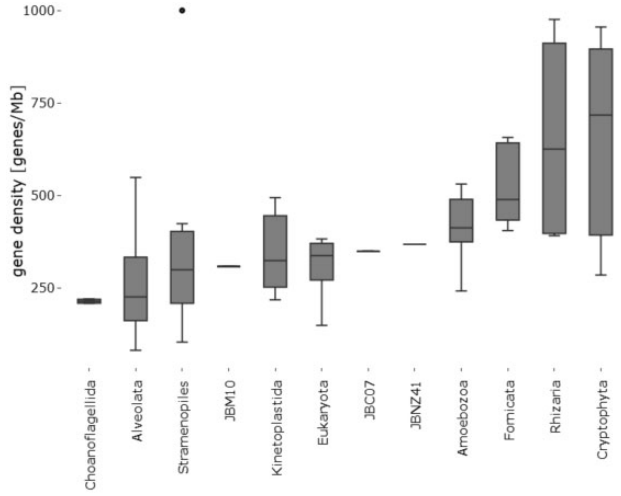


FIG. 2.—Gene density. Comparison of *Poteriospumella lacustris* against 197 protists from the Ensembl Protists database.

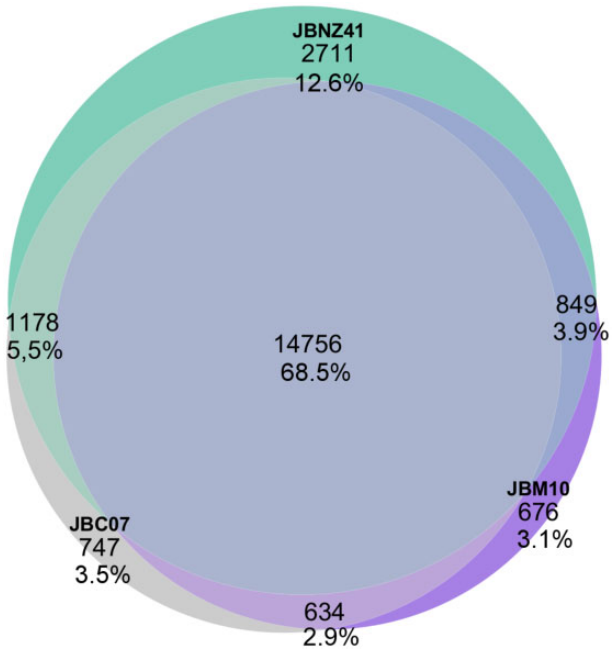


FIG. 3.—Venn diagram. Proportion of shared genes. Percentages refer to the total number of 21,551 genes.

unexpected and shows that the intraspecific molecular differences were higher than initially assumed.

Variant Analysis

Separating the SNPs, inserts and deletions according to metabolic pathway categories showed differences in their mutation rate. Likewise, the strains differ in the mutation rate within a metabolic pathway category. We examined if mutation rates of a specific functional group differed between strains (ANOVA or Kruskal–Wallis test, see supplementary tables S12 and S13, Supplementary Material online) and if mutation rates of a specific strain differed between functional groups (Wilcoxon signed-rank test, see supplementary tables S6–S11, Supplementary Material online). We first enumerated the allelic variation of each strain (see fig. 5*A*, see supplementary tables S6–S8, Supplementary Material online). The strain JBM10 had except for the category *nucleotide metabolism* a significantly ($P < 0.01$) lower variation rate than JBC07 or JBNZ41 (see fig. 5*A*, ANOVA see supplementary table

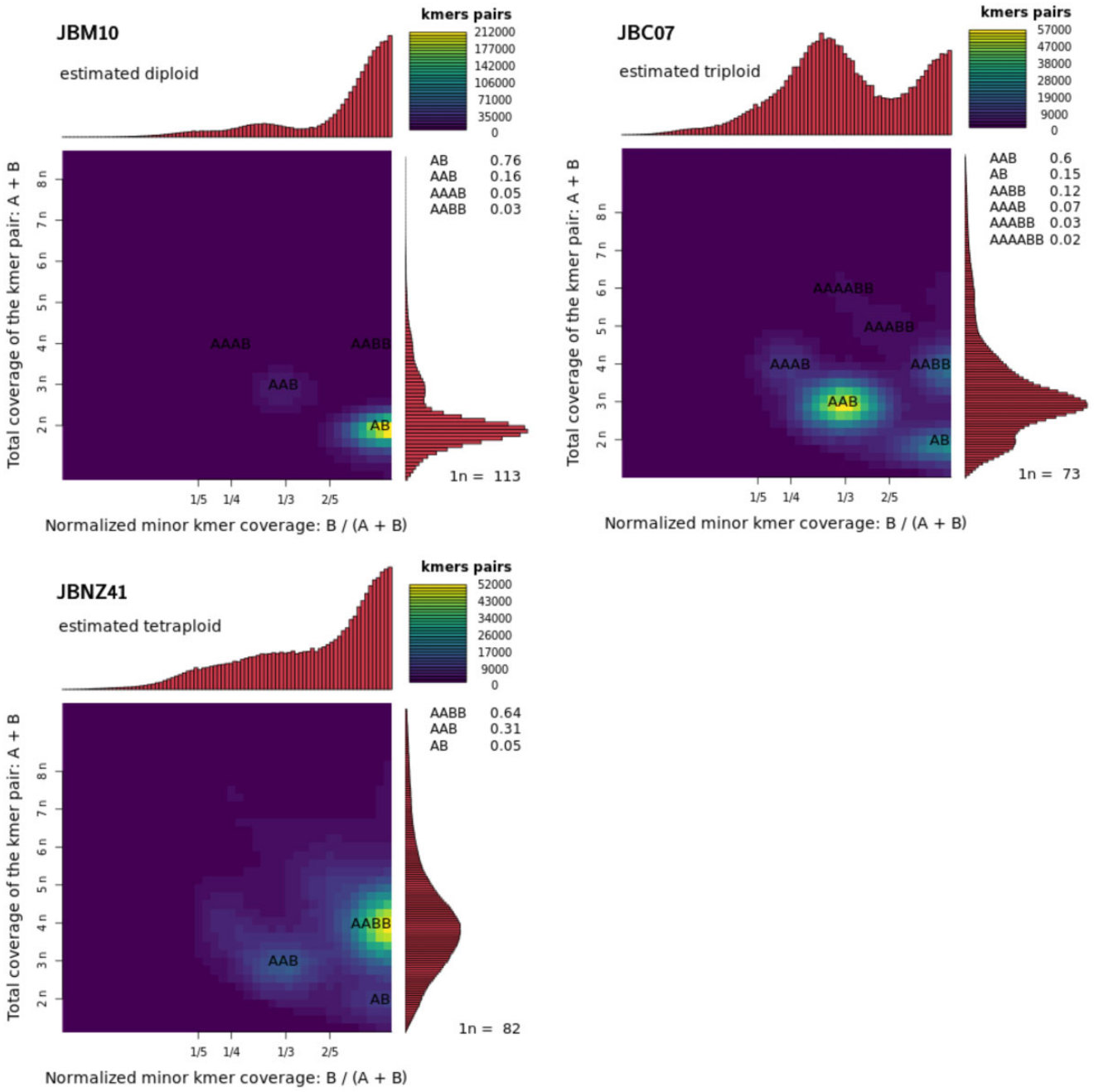


Fig. 4.—*K*-mer based ploidy estimation. The heatmap reflects coverage of *k*-mer pairs differing by one base. The ratio of the characters A to B represents the ratio of these *k*-mer pairs (e.g., 67% ATGTC and 33% ATGTT conforms AAB). The coverage distribution on the right side indicates ploidy levels (left scale). The distribution on the top side is based on the coverage normalized by the ratio. The determined ploidy levels of the strains were diploidy (JBM10), triploidy (JBC07), and tetraploidy (JBNZ41). n = average *k*-mer coverage, *k*-mer size = 21.

S12, Supplementary Material online). Molecular variation between the strains differed significantly in the categories *unannotated genes, genetic information processing, organelle targeting genes* and *signaling and cellular processes* (see supplementary table S12, Supplementary Material online). Molecular variation between the functional groups of each strain was significant ($P < 0.01$) for: *genetic information processing* (all strains), energy metabolism (JBM10, JBNZ41), *unannotated genes* (JBM10, JBNZ41), *signaling and cellular processes* (JBM10), and *carbohydrate metabolism* (JBNZ41) (see supplementary tables S6–S8, Supplementary Material online).

A pairwise comparison of genes between different strains demonstrated genetic divergence (see fig. 5*B*, statistic see

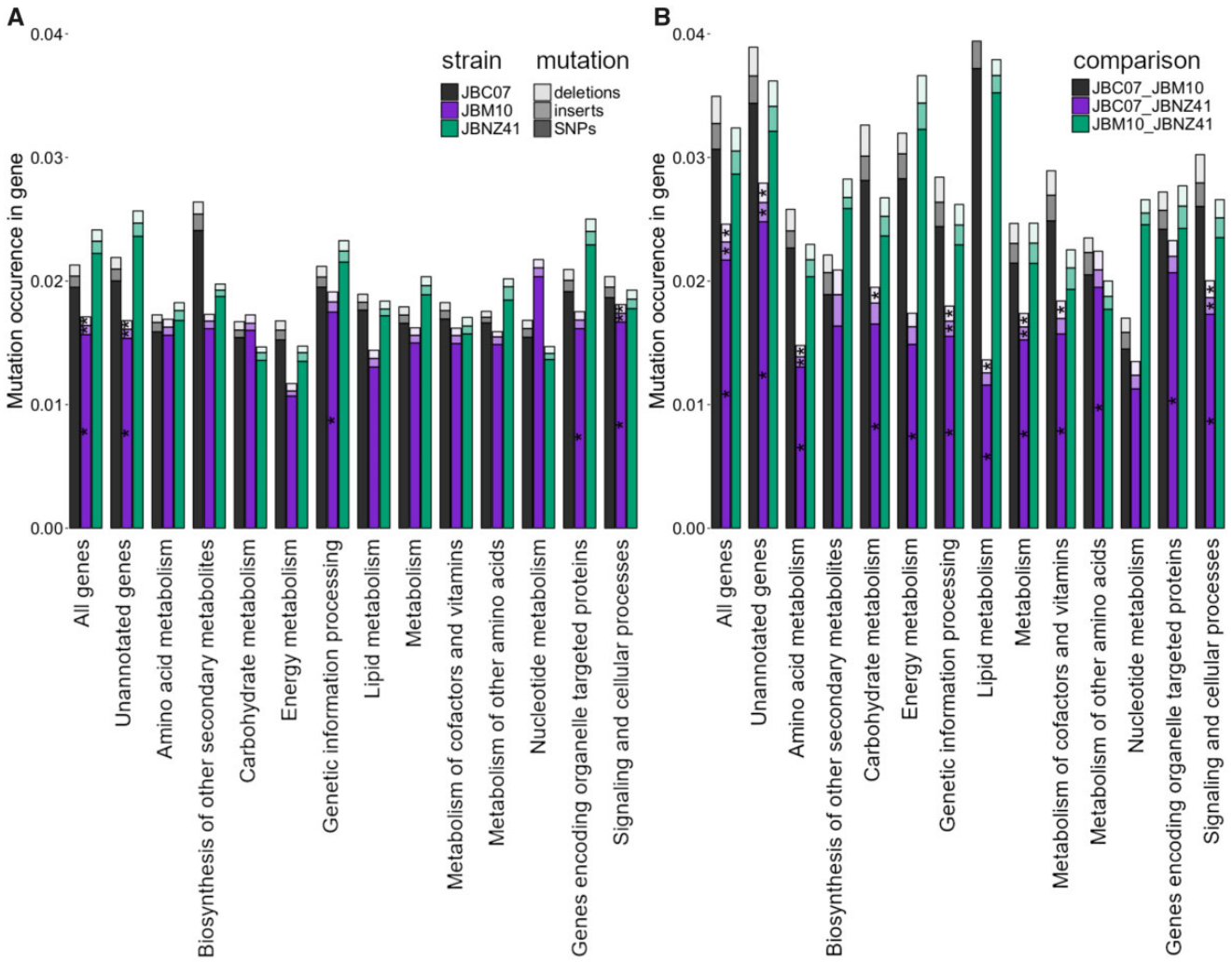


Fig. 5.—Allelic variation and mutation distribution. Occurrence of SNPs (dark), insertions (semitransparent), and deletions (bright) in the strains JBC07 (gray), JBM10 (purple), and JBNZ41 (green). The significance (P-value < 0.01) within one functional group was calculated by ANOVA. The asterisk was placed in the middle column (JBM10) and refers to one mutation type within the three strains. (*A*) The distribution of mutations displays the allelic variation within one strain. There the energy metabolism is most conserved. (*B*) Pairwise comparison of genes demonstrates which groups evolved apart and which stay conserved. Especially JBM10 deviates from the other two strains.

supplementary tables S9–S11, Supplementary Material online). Across nearly all categories JBM10 deviated from the other two strains. Exceptional categories were *biosynthesis of other secondary metabolism* and *nucleotide metabolism*, which showed no significant difference (see supplementary table S13, Supplementary Material online). In the pairwise comparison between strains the group of *unannotated genes* differed the most (see supplementary tables S9–S11, Supplementary Material online). Moreover, ~1,000 identical genes were shared between JBC07 and JBNZ41, whereas JBM10 had 584 (JBC07) and 430 (JBNZ41) identical genes in common (see supplementary table S3, Supplementary Material online).

Counting genes based on the KEGG functional assignment showed small differences between the strains (see fig. 6). Notably, JBM10 contains more genes assigned to *proteasome* and JBNZ41 comprises more genes of *cofactor and vitamin biosynthesis* compared with the other two strains.

The *genes encoding organelle targeted proteins* identified by HECTAR were subdivided in the groups signal peptide, signal anchor, mitochondrion and chloroplast. The associated mutation rate between these classes, however, did not differ significantly ($P > 0.05$), which is why we kept the combined group *genes encoding organelle targeted proteins.*

The ANI was between 97.4% and 98.0% (see table 2). Despite the high similarity of the entire genome, we could identify distinctions between primary and secondary metabolism. The number of SNPs in pathways of the secondary metabolism was generally higher (see fig. 5*B* and supplementary tables S9–S11, Supplementary Material online). However,

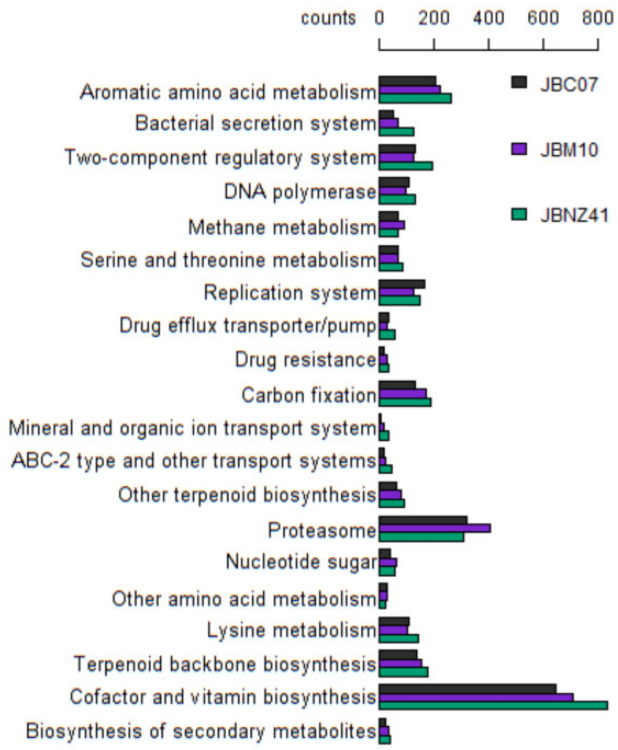


Fig. 6.—Gene count based on the KEGG hierarchy functional assignment. Only functional categories with at least 20 counts and >10% standard deviation between the strains are shown.

the SNPs gave us insufficient information about recent population development due to higher variation caused by different ploidies. The mitochondrial DNA was almost identical (one insertion in JBM10 and 1–5 SNPs between each strain, despite ~38,000 bp length).

Discussion

Genome Assembly and Gene Prediction

Genome comparison of three *P. lacustris* strains (JBM10, JBNZ41, JBC07) revealed similarities in genes and their structural arrangement. Although the strains were assembled based on Illumina and PacBio sequencing, the assembly size of the genomes deviated from the estimated sizes based on nuclear staining (Olefeld et al. 2018) (see table 1). This discrepancy may indicate different ploidy levels which were not considered by Olefeld et al. (2018) (see discussion on ploidy below). Additionally, the high number of contigs in the genome assemblies reflected the extent of fragmentation. It is clearly a challenge to assemble large and possibly repetitive eukaryotic genomes and new technologies such as MinION sequencing with an even larger read length could possibly enhance the genome assemblies. Nevertheless, the amount of contigs was comparable to other eukaryotic hybrid assemblies (e.g., *Aegilops tauschii*: >24,000 contigs [Zimin et al. 2017], *Melopsittacus undulatus*: >15,000 contigs [Ganapathy et al. 2014]).

The BUSCO analysis could not answer the question of genome completeness adequately since the comparative gene set was unsuitable for *P. lacustris*. In nonmodel microbial eukaryotes the model gene sets used in BUSCO or CEGMA could be limited due to distantly related model organisms as was shown for example in dinoflagellates of the genus *Symbiodinium* (Aranda et al. 2016; Liu et al. 2018).

The completeness of essential primary metabolic pathways (citrate cycle, pentose phosphate pathway, nucleotide and protein biosynthesis, etc.) in our analyses, assessed with the KEGG Mapper—Reconstruct Module tool, affirmed genome integrity. Furthermore, we found genes affiliated with phototrophic pathways as well as pathways related to autotrophy (e.g., sulfur assimilation), which are presumably remainings from the phototrophic ancestor (Beisser et al. 2017; Graupner et al. 2018). More evolved heterotrophs have lost these pathways, providing evidence that *P. lacustris* is in an early stage of heterotrophy (Graupner et al. 2018). The reduction of phototropic pathways confirms the natural selection and adaptation of *P. lacustris*.

Since our gene analysis is based on an unclosed genome it could naturally lead to errors, especially with respect to gene counts (Denton et al. 2014). Therefore, we reduced the number of predicted genes by clustering and marking of duplicated genes if they overlapped on the same strand in the alignment between genes and genome. As a result, the total amount of 178,000 genes decreased to around 18,000 genes in each strain (see fig. 1). The approach of merging overlapping genes may, lead to a loss of nested genes. This loss is neglect able since the number of nested genes of different strands in eukaryotic genomes is very small (0.7–0.8% of the total amount of genes, Sanna et al. 2008). The gene level is slightly lower than the estimated ~20,000 genes found in the transcriptome study of *P. lacustris* (Graupner et al. 2017). However, the gene number in the study of Graupner et al. (2017) was based on gene components generated by the Trinity software, which could therefore be higher because of isoforms and variants of genes.

Gene Density

Even though the genome sizes of heterotrophic chrysomonads is smaller than in mixotrophic or phototrophic relatives (Olefeld et al. 2018), the gene density remains at a comparable level to other protist species, even compared with predominantly phototrophic relatives (see fig. 2). It is assumed that heterotrophic chrysophytes have a selection pressure toward small cell sizes, which enable more effective preying on ultra small bacteria (Hansen 1992; Hansen et al. 1994; Olefeld et al. 2018). Deletions of noncoding DNA sequences are likely mechanisms to decrease the genome size. The gene density varies between the strains due to differences in gene

length. However, the average ratio of coding DNA in proportion to total DNA is consistent, although there are small differences in the proportion of repeat regions (see table 1).

Ploidy

Surprisingly, ploidy levels differed between the strains despite their close relatedness. Earlier attempts to stain the chromosomes for microscopic ploidy assessment failed because of the small nucleus and chromosome size of *P. lacustris*. By using flow cytometry, it has been shown that the genome size of JBM10 is distinctly smaller than those of the other two strains (Olefeld et al. 2018). This method assumes equal ploidy and can only determine the total amount of DNA. On the basis of the genome assembly the haploid genome size is ~50 Mb for all three strains and from the molecular data, we have indications that the grade of ploidy differs between the three strains. Allelic distribution of SNPs suggests diploidy in JBM10, triploidy in JBC07, and tetraploidy in JBNZ41. This is in accordance with the genome size estimates of (Olefeld et al. 2018)—the genome of JBNZ41 is approximately twice as large as that of JBM10 and the genome size of JBC07 is in-between (see table 1). However, as we found only two allelic variants for numerous genes we assume that JBNZ41 became tetraploid with a recent genome duplication. This would explain the strong peak indicating diploidy for many genes with weaker (even though pronounced) peaks indicating tetraploidy for other genes and at the same time the larger genome size (see supplementary fig. S1, Supplementary Material online).

Tetraploid strains with a characteristic diploid distribution were also reported for *Saccharomyces cerevisiae* (Zhu et al. 2016). The ploidy estimation based on kmers, which is independent of possible assembly biases approved our results (see fig. 4). Nevertheless, all strains showed at least partial triploid-like peaks. This could be induced by a large number of paralogous genes or gene duplications only on one locus. Polyploidy has also been found in some other relatives of the stramenopiles, for example, in diatoms (Parks et al. 2018), oomycetes (Martens and Van de Peer 2010), and brown algae (Cock et al. 2010). Together with the genome size data from (Olefeld et al. 2018) our data provide evidence for different levels of ploidy in closely related strains. Nonetheless, the presumable genome duplication contradicts the hypothesis of a strong selection pressure toward small cell sizes and small genomes (de Castro et al. 2009; Olefeld et al. 2018) indicating that other factors beyond predator–prey interactions may also be significant in the genome evolution of heterotrophic chrysophytes.

Variation between Strains

The larger genome size of related mixotrophic taxa indicate that *P. lacustris* reduced the genome size as it developed a heterotrophic mode of nutrition. The decreased genome size but constant gene density implies a loss of genes. In general, mutations already occur, before a gene gets lost or the gene function is changing. Further, after a gene lost its function, mutations appear more frequently. Consequently, genes with more mutations are potentially less important for an organism or several copies may exist. Organisms with a higher ploidy often show higher mutation frequencies (e.g., Ohnishi et al. 2004; Uauy et al. 2009; Krasileva et al. 2017), because they have multiple genes as back up or several gene variants for special conditions. Therefore, we compared the variation of genes grouped by their function (see fig. 5*A*). Some groups (e.g., *biosynthesis secondary metabolism*) seemed to have a high variation, but differences were nonsignificant due to the high standard deviation of mutation occurrence. In all three strains the category of *genetic information processing* contained the most highly variable genes and, in contradiction, also the highest proportion of identical genes between the strains (see supplementary table S3, Supplementary Material online). Apparently, *genetic information processing* is vital. The high number of mutations in genes affiliates with this pathway may indicate that this system is manifold fail-safe so that mutations presumably can be tolerated without severe consequences. In JBM10, noticeably many mutations occurred in the group of *cellular processes* belonging predominantly to *cGMP signaling, DNA damage-induced cell cycle checkpoints, MAPK signaling* and *Cell cycle—G2/M transition*. The variation in the MAPK pathway may indicate decreased environmental stress including osmotic and thermal changes (Jimenez et al. 2004). Variances in *cell cycle supervision* may have led to a higher proliferation rate or cell death (Hartwell and Weinert 1989; Stark and Taylor 2006). However, the growth rate of *P. lacustris* is comparable to related species like *Poterioochromonas malhamensis* or *Dinobryon divergens* (Rottberger et al. 2013). Like in the category *genetic information processing*, the genes affiliated with *cellular processing* are present in multiple copies. On the other hand, genes assigned to energy metabolism comprise fewer variations possibly indicating that fewer mutations are acceptable. In general JBM10 has a lower allelic variation than JBC07 or JBNZ41. Since JBM10 also has a smaller genome size (Olefeld et al. 2018), but a similar gene density (see table 2), this strain consequently possess fewer gene copies and therefore less allelic variation. Especially, higher ploidy would enable higher recombination rates (Song et al. 1995). In addition, polyploids comprise significantly more allelic variation than diploids (Li et al. 2017), which may explain for higher variation within JBC07 and JBNZ41.

We performed a pairwise comparison for each gene shared between strains to count mismatches between two genes if no allelic variation could induce an identical sequence. This enabled us to detect possibly identical genes within the alleles between strains as well as the genetic variation (see supplementary table S3, Supplementary Material online). The genes of strains JBC07 and JBNZ41 are genetically more

similar to each other than to JBM10 (see fig. 5). This indicates a closer relation between JBC07 and JBNZ41, which also originate from geographically closer sites. It must be noted, that with increasing allelic variation the probability for random matches also increases. In other words, the probability that the same alleles are found between a triploid and a tetraploid organism is higher than when compared with a diploid strain. Thus, allelic variation could not be clearly assigned to either differences in ploidy or phylogenetic relatedness.

Most mutations occurred in the set of genes without a functional group assignment (see fig. 5*B*). One reason could be a false positive gene interpretation during the prediction procedure (noncoding sequence interpreted as unknown gene). It can be assumed that noncoding DNA presumably has a weaker selection pressure and therefore more mutations (Andolfatto 2005). The KEGG database relies on genes assigned with a function. These information are gained from well-known organisms which are not closely affiliated with our target strains but with other supergoups. Possibly the annotated *P. lacustris* genes could have only the essential genes in common with these model organisms. Hence, the other (unassigned) genes might be necessary for ecological niche adaptation or species-specific functions, but not for key primary metabolic pathways. Our data suggest that the secondary metabolism in *P. lacustris* is subject to stronger genetic changes than the primary metabolism.

The number of shared genes (see fig. 3) reflects the close relationship between the three strains. In contrast to the indications from sequence variation discussed above, the strains JBC07 and JBM10 have more genes in common, whereas JBNZ41 has the highest number of strain-specific genes indicating a more distant relation. This constellation of the relationships was also described in Beisser et al. (2017) and Stoeck et al. (2008), but contradicts the findings of Graupner et al. (2017) based on orthologous genes. The three strains share 68.5% of all genes. However, the overlap of shared genes could rise when the genome sequences are completed. On the other hand, Graupner et al. (2017) determined around 92% annotated shared genes ($n = 2{,}000$) and 50% sequences variations in general ($n = 20{,}000$), which confirms our results. *Poteriospumella lacustris* has a slightly less gene diversity as the phytoplankton *Emiliania huxleyi* (75% shared genes; Read et al. 2013) and more than the fungus *Zymoseptoria tritici* (58% shared genes; Plissonneau et al. 2018). Because of the high number of shared genes, it is not surprising that the gene count varies little by function (see fig. 6).

The extent of variation in the mitochondrial sequences should be similar or even larger than the variation between the genomes (Smith 2015). However, the mitochondrial DNA remains conserved. This findings accord to phylogenetic analyses of COI genes within chrysophytes, where 18 strains of *P. lacustris* clustered together (Bock et al. 2017). However, the conserved mDNA could not be generalized within chrysophytes since other species varied in the COI gene (Bock et al. 2017). Furthermore, there are some other species showing very low intraspecifc mitochondrial variation (like the coral Octocorallia, McFadden et al. 2010). In comparison *genes encoding organelle targeted proteins* have similar mutation rates to all other genes (see supplementary tables S6–S11, Supplementary Material online).

Subsequent studies should include further species to cover the whole class of Chrysophyceae and especially include representatives with phototrophic and mixotrophic nutrition in order to shed light on the genome evolution in the course of the multiple parallel nutritional adaptation from mixotrophy to heterotrophy as well as species of different ploidy levels to consider the ploidy as an influencing variable. Further, a more extensive analysis of transposons could extend the analysis of genome evolution in Chrysophyceae.

Conclusions

Our study provides a comprehensive genome analysis and created one of the first reference genomes within the Chrysophyceae. The intraspecific genome variation of *P. lacustris* is high, especially the level of ploidy. Most mutations occurred in unannotated genes, which are likely related to secondary metabolism and to the adaption to a particular niche. We thus can reject the hypothesis that mutations are randomly distributed across pathways and metabolic categories. Since all strains differ in the degree of ploidy, it was not possible to deduce past population bottlenecks based on the allelic variation.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Thanks to Micah Dunthorn for proofreading and Sabina Marks for lab assistance. We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen. This work was supported by the DFG Project BO 3245/19 and DFG Projekt BO 3245/17.

Author's Contributions

J.B. conceived the study; J.B., S.M. designed the lab experiments; S.M. and D.B. designed computational procedure; S.M. performed the experiment, analyzed the data, and drafted the manuscript; S.M., J.B., and D.B. interpreted the data; J.B. and D.B. revised and edited the manuscript; all authors read and approved the final manuscript.

Literature Cited

Almagro Armenteros JJ, et al. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 37(4):420–423.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437(7062):1149–1152.

Antipov D, et al. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 32(7):1009–1015.

Aranda M, et al. 2016. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. Sci Rep. 6:39734.

Beisser D, et al. 2017. Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. PeerJ 5:e2832.

Bestová H, et al. 2018. Ecological and biogeographical drivers of freshwater green algae biodiversity: from local communities to large-scale species pools of desmids. Oecologia 186(4):1017–1030.

Bock CA, et al. 2017. Genetic diversity in chrysophytes: comparison of different gene markers. Fottea 17(2):209–221.

Boenigk J, et al. 2005. High diversity of the 'Spumella-like' flagellates: an investigation based on the SSU rRNA gene sequences of isolates from habitats located in six different geographic regions. Environ Microbiol. 7(5):685–697.

Boenigk J, et al. 2007. Differential thermal adaptation of clonal strains of a protist morphospecies originating from different climatic zones. Environ Microbiol. 9(3):593–602.

Boenigk J, et al. 2018. Geographic distance and mountain range structure freshwater protist communities on a European scale. Metabarcod Metagenomics. 2:e21519.

Boratyn GM, et al. 2012. Domain enhanced lookup time accelerated BLAST. Biol Direct. 7:12.

Buchfink B, et al. 2015. Fast and sensitive protein alignment using diamond. Nat Methods. 12(1):59–60.

Cock JM, et al. 2010. The Ectocarpus genome sequence: insights into brown algal biology and the evolutionary diversity of the eukaryotes. New Phytol. 188(1):1–4.

Darling KF, et al. 2004. Molecular evidence links cryptic diversification in polar planktonic protists to quaternary climate dynamics. Proc Natl Acad Sci USA. 101(20):7657–7662.

de Castro F, et al. 2009. Reverse evolution: driving forces behind the loss of acquired photosynthetic traits. PLoS One 4(12):e8465.

Denton JF, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 10(12): 1003998.

Dlugosz M, et al. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics 33(17):2759–2761.

Evans KM, et al. 2005. High levels of genetic diversity and low levels of genetic differentiation in north sea pseudo-nitzschia pungens (bacillariophyceae) populations1. J Phycol. 41(3):506–514.

Fenchel T, Finlay BJ. 2004. The ubiquity of small species: patterns of local and global diversity. BioScience 54(8):777–784.

Fernandez LD, et al. 2017. Geographical distance and local environmental conditions drive the genetic population structure of a freshwater microalga (Bathycoccaceae; Chlorophyta) in Patagonian lakes. FEMS Microbiol Ecol. 93:10.

Findenig BM, et al. 2010. Taxonomic and ecological characterization of stomatocysts of spumella-like flagellates (chrysophyceae)1. J Phycol. 46(5):868–881.

Finlay BJ. 2002. Global dispersal of free-living microbial eukaryote species. Science 296(5570):1061–1063.

Forouzan E, et al. 2017. Evaluation of nine popular de novo assemblers in microbial genome assembly. J Microbiol Methods. 143:32–37.

Ganapathy G, et al. 2014. High-coverage sequencing and annotated assemblies of the budgerigar genome. GigaScience 3(1):11.

Gschloessl B, et al. 2008. HECTAR: a method to predict subcellular targeting in heterokonts. BMC Bioinformatics. 9:393.

Godhe A, Rynearson T. 2017. The role of intraspecific variation in the ecological and evolutionary success of diatoms in changing environments. Philos Trans R Soc Lond B Biol Sci. 372:1728.

Graupner N, et al. 2018. Evolution of heterotrophy in chrysophytes as reflected by comparative transcriptomics. FEMS Microbiol Ecol. 94(4):1–11.

Graupner N, et al. 2017. Functional and phylogenetic analysis of the core transcriptome of ochromonadales. Metabarcod Metagenomics. 1:e19862.

Hahn MW, et al. 2003. Isolation of novel ultramicrobacteria classified as actinobacteria from five freshwater habitats in Europe and Asia. Appl Environ Microbiol. 69(3):1442–1451.

Hansen B, et al. 1994. Prey size selection, feeding rates and growth dynamics of heterotrophic dinoflagellates with special emphasis on gyrodinium spirale. Limnol Oceanogr. 39(2):395–403.

Hansen PJ. 1992. Prey size selection, feeding rates and growth dynamics of heterotrophic dinoflagellates with special emphasis on gyrodinium spirale. Mar Biol. 114(2):327–334.

Hartwell LH, Weinert TA. 1989. Checkpoints: controls that ensure the order of cell cycle events. Science 246(4930):629–634.

Hayhome BA, et al. 2007. Intraspecific variation in the dinoflagellate peridinium volzip1. J Phycol. 23(4):573–580.

Heywood P, Magee PT. 1976. Meiosis in protists. Some structural and physiological aspects of meiosis in algae, fungi, and protozoa. Bacteriol Rev. 40(1):190–240.

Jain C, et al. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* 9 (1):5114.

Jimenez C, et al. 2004. Phosphorylation of MAP kinase-like proteins mediate the response of the halotolerant alga *Dunaliella viridis* to hypertonic shock. Biochim Biophys Acta. 1644(1):61–69.

John U, et al. 2004. Utility of Amplified Fragment Length Polymorphisms (AFLP) to analyse genetic structures within the Alexandrium tamarense species complex. Protist 155(2):169–179.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1):27–30.

Kimura M. 1979. The neutral theory of molecular evolution. Sci Am. 241(5):98–100.

Knoll AH. 1994. Proterozoic and early cambrian protists: evidence for accelerating evolutionary tempo. Proc Natl Acad Sci USA. 91(15):6743–6750.

Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.

Koster J, Rahmann S. 2012. Snakemake – a scalable bioinformatics workflow engine. Bioinformatics 28(19):2520–2522.

Krasileva KV, et al. 2017. Uncovering hidden variation in polyploid wheat. Proc Natl Acad Sci USA. 114(6):E913–E921.

Langmead B, et al. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3):R25.

Larsson J. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. 2018. R package version 4.1.0. https://cran.r-project.org/package=eulerr

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.
Li Y, et al. 2017. Changing ploidy as a strategy: the Irish potato famine pathogen shifts ploidy in relation to its sexuality. MPMI 30(1):45–52.
Liu H, et al. 2018. Symbiodinium genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. Commun Biol. 1(1):95.
Logares R, et al. 2009. Genetic diversity patterns in five protist species occurring in lakes. Protist 160(2):301–317.
MacArthur R, Wilson E. 1967. The theory of Island biogeography. Princeton: Princeton University Press.
Martens C, Van de Peer Y. 2010. The hidden duplication past of the plant pathogen phytophthora and its consequences for infection. BMC Genomics. 11(1):353.
McFadden CS, et al. 2010. Insights into the evolution of octocorallia: a review. Integr Compar Biol. 50(3):389–410.
McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
Myers EW, et al. 2000. A whole-genome assembly of drosophila. Science 287(5461):2196–2204.
Nurk S, et al. 2013. Assembling single-cell genomes and mini-metagenomes from chimeric mda products. J Comput Biol. 20(10):714–737.
Ohnishi G, et al. 2004. Spontaneous mutagenesis in haploid and diploid *Saccharomyces cerevisiae*. Biochem Biophys Res Commun. 325(3):928–933.
Olefeld JL, et al. 2018. Genome size of chrysophytes varies with cell size and nutritional mode. Org Divers Evol. 18:163.
Parfrey LW, et al. 2008. The dynamic nature of eukaryotic genomes. Mol Biol Evol. 25(4):787–794.
Parks MB, et al. 2018. Phylogenomics reveals an extensive history of genome duplication in diatoms (bacillariophyta). Am J Bot. 105(3):330–347.
Plissonneau C, et al. 2018. Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol. 16(1):5.
Price AL, et al. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):351–358.
Pundir S, et al. 2017. UniProt protein knowledgebase. Methods Mol Biol. 1558:41–55.
Raikov IB. 1995. Meiosis in protists: recent advances and persisting problems. Eur J Protistol. 31(1):1–7.
Read BA, et al. 2013. Pan genome of the phytoplankton Emiliania underpins its global distribution. Nature 499(7457):209–213.
Reed DH, Frankham R. 2003. Correlation between fitness and genetic diversity. Conserv Biol. 17(1):230–237.
Rottberger J, et al. 2013. Influence of nutrients and light on autotrophic, mixotrophic and heterotrophic freshwater chrysophytes. Aquat Microb Ecol. 71(2):179–191.
Rynearson TA, Armbrust EV. 2005. Maintenance of clonal diversity during a spring bloom of the centric diatom *Ditylum brightwellii*. Mol Ecol. 14(6):1631–1640.
Sanna CR, et al. 2008. Overlapping genes in the human and mouse genomes. BMC Genomics. 9:169.
Simao FA, et al. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Simpson JT, et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):1117–1123.
Smith DR. 2015. Mutation rates in plastid genomes: they are lower than you might think. Genome Biol Evol. 7(5):1227–1234.
Song K, et al. 1995. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. Proc Natl Acad Sci USA. 92(17):7719–7723.:
Sovic I, et al. 2016. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. Bioinformatics 32(17):2582–2589.
Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34(Web Server issue):435–439.
Stark GR, Taylor WR. 2006. Control of the g2/m transition. Mol Biotechnol. 32(3):227–248.
Stoeck T, et al. 2008. Multigene phylogenies of clonal Spumella-like strains, a cryptic heterotrophic nanoflagellate, isolated from different geographical regions. Int J Syst Evol Microbiol. 58(3):716–724.
Uauy C, et al. 2009. A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. BMC Plant Biol. 9(1):115.
Underwood CJ, et al. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33(14):2202–2204.
Wang D, et al. 2014. *Nannochloropsis* genomes reveal evolution of microalgal oleaginous traits. PLoS Genet. 10(1):e1004094.
Watts PC, et al. 2013. A century-long genetic record reveals that protist effective population sizes are comparable to those of macroscopic species. Biol Lett. 9(6):20130849.
Weiß CL, et al. 2018. nQuire: a statistical framework for ploidy estimation using next generation sequencing. BMC Bioinformatics. 19(1):122.
Zhu YO, et al. 2016. Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. G3 (Bethesda) 6(8):2421–2434.
Zimin AV, et al. 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 27(5):787–792.

Associate editor: John Archibald

Chapter 3

Evolution of the nutritional mode

Publication 1:

Genome size of chrysophytes varies with cell size and nutritional mode

Jana L. Olefeld, Stephan Majda, Dirk C. Albach,
Sabina Marks, Jens Boenigk

Published in: Org Divers Evol (2018) 18: 163.
https://doi.org/10.1007/s13127-018-0365-7

Contribution to this publication:

conception & planning:	0 %	
experimental work:	0 %	
data analysis:	15 %	Genome size comparison for several phyla
writing the manuscript:	35 %	discussion, proofreading
revising the manuscript:	25 %	discussion

....
Stephan Majda Prof. Dr. Jens Boenigk

Organisms Diversity & Evolution (2018) 18:163–173
https://doi.org/10.1007/s13127-018-0365-7

ORIGINAL ARTICLE

Genome size of chrysophytes varies with cell size and nutritional mode

Jana L. Olefeld[1] · Stephan Majda[1] · Dirk C. Albach[2] · Sabina Marks[1] · Jens Boenigk[1]

Received: 12 October 2017 / Accepted: 22 April 2018 / Published online: 9 May 2018

Abstract

The cellular content of nuclear DNA varies up to 200,000-fold between eukaryotes. These differences can arise via different mechanisms. In particular, cell size and nutritional mode may influence evolution of the nuclear DNA content. Chrysophytes comprise organisms with different cell organizations and nutritional modes. Heterotrophic clades evolved independently several times from phototrophic or mixotrophic ancestors. Thus, chrysophytes are an ideal model taxon for investigating the effect of the nutritional mode on cellular DNA content. We investigated the genome size of heterotrophic, mixotrophic, and phototrophic chrysophytes. We demonstrate that cell sizes and genome sizes differ significantly between taxa with different nutritional modes. Phototrophic strains tend to have larger cell volumes and larger genomes than heterotrophic strains do. The investigated mixotrophic strains had intermediate cell volumes and small to intermediate genome sizes. Heterotrophic chrysophytes had the smallest genomes and cell volumes compared to other chrysophytes. In general, genome size increased with cell volume, but cell volume only partially explained the variation in genome size. In particular, genome sizes of mixotrophic strains were smaller than expected based on cell sizes.

Keywords Genome evolution · C-value · Heterotrophic nanoflagellates · Algae · Microbial eukaryotes · Stramenopiles

Introduction

The evolution of genome sizes is of special interest because huge variation in the amount of DNA (up to 200,000-fold for eukaryotes) are the result of complex interactions between various evolutionary forces (Kapraun 2005; Cavalier-Smith 2005; Gupta et al. 2016). Studies on the amount of nuclear DNA and on genome size (C-value) can have practical as well as predictive uses, as these parameters are important traits of a species (Bennett and Leitch 2005). Often, the amount of nuclear DNA is much higher than necessary for coding or regulatory sequences, a fact which is known as the C-value paradox (Dolezel et al. 1998; Gupta et al. 2016). However, genome size is generally correlated with cell volume (Dolezel et al. 1998); the widespread view that genome size will increase with the complexity of an organism seems therefore to hold only in the general sense that most eukaryotes have more DNA than prokaryotes or viruses (Cavalier-Smith 1982). Genome size is furthermore important for our understanding of its effect on phenotypic traits (nucleotype) in response to different environmental conditions (Kapraun 2005; Cavalier-Smith 1978; Dolezel et al. 1998).

During the life cycle of a species, changes in the genome size of certain cells (referring to gamets sensu *lato*) are possible, which is especially important if dealing with single-celled organisms. Haploid vegetative cells (e.g., in dinophytes) as well as diploid vegetative cells (e.g., in diatoms) are known. In general, asexual reproduction occurs, but also sexual reproduction (e.g., by producing haploid eggs and sperms or fusion of haploid vegetative cells) can be triggered (Koester et al. 2007; Kremp 2013).

Protists are particularly suited for studying mechanisms of genome size evolution in eukaryotes, because they are unicellular organisms which respond quickly to environmental changes. Further, the number of cell types (even in colonial species) is limited, and we face many different kinds of life cycles which have mostly evolved independently several times in phyla with different modes of nutrition and different

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13127-018-0365-7) contains supplementary material, which is available to authorized users.

✉ Jana L. Olefeld
jana.olefeld@uni–due.de

[1] Department of Biodiversity, University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, Germany

[2] Department of Biodiversity and Evolution of Plants, Carl von Ossietzky University Oldenburg, Carl-von-Ossietzky-Straße 9-11, 26129 Oldenburg, Germany

cell structures (Cavalier-Smith 1980). Since some of the unicellular protists have considerably larger genomes than multicellular eukaryotes, studying single-celled protists allows the investigation of the full range of eukaryote C-values (Cavalier-Smith 1982).

Here, we focus on the chrysophytes, which group in the Ochrophyta (Stramenopiles). They comprise a range of organisms with different life cycle stadia (e.g., silicate cysts or flagellated cells), cell organizations (e.g., colonial, branched, or single-celled morphologies), and nutrition modes such as phototrophy, heterotrophy, and mixotrophy (Boenigk et al. 2006). The Synurales are considered to be phototrophic (Kristiansen and Škaloud 2017) even though it is uncertain whether they can occasionally take up particles. However, as particulate food uptake is uncertain in the Sunyrales and of minor importance anyway, the Synurales are considerd to be phototrophic in the context of this study.

Heterotrophic chrysophytes evolved through parallel evolution several times independently within the Chrysophyceae (Grossmann et al. 2016). The plastids of these heterotrophic chrysophytes are in different stages of reduction and may be involved in several metabolic pathways as well as contain numerous photopigments (Graupner et al. 2018). However, the heterotrophic chrysophytes have in common the loss of chlorophyll and photosynthetic pathways as it had been demonstrated for several lineages so far (Graupner et al. 2018). They rely on particulate food uptake and prey on bacteria and other bacteria-sized microorganisms to gain carbon (Boenigk et al. 2005). Most chrysophyte species are mixotrophic, i.e., capable of photosynthesis and phagotrophy. In mixotrophs, particle uptake may be an additional source of nutrient uptake as, for instance, in nutrient-poor waters, or be the major route of carbon uptake. All different gradations from primarily phototrophic to primarily heterotrophic are realized in mixotrophic chrysophytes (Bock et al. submitted 2018).

Due to this diverse nutritional strategies, chrysophytes are important primary producers but also among the dominant bacterivores (Boenigk et al. 2005; Bock et al. 2014; Grossmann et al. 2016; Bock et al. submitted 2018).

Since chrysophytes are abundant in most temperate, freshwater habitats and occupy different niches, they play an important role in the regulation of microbial food webs (Bock et al. 2014).

Sexual reproduction occurs but is believed to be rare in chrysophytes. The dominant form of chrysophytes is a, probably haploid, vegetative swarmer, which can generally reproduce asexually, undergo encystment, or fuse to form a zygote. It has been suggested that sexuality can be induced, e.g., at high cell densities (e.g., in *Synura*) or by cyst formation (e.g., in *Dinobryon*) (Sandgren 1991; Kristiansen und Škaloud 2017).

In our study, we want to investigate the relationship between different nutrition modes and genome size in different chrysophyte strains with respect to their predominant nutrition mode. The basis of the genome size measurement is the correlation between DNA content and fluorescent staining. Linear regression of reference genome peaks enables calculation of an unknown sample DNA content. This has been shown highly reproducible for certain staining methods (Dolezel et al. 1998). Further, numerous authors stated and proved a correlation of cell volume and genome size (e.g., Bennett 1972). Additionally, heterotrophic chrysophytes are generally smaller than phototrophic ones, since they do only host remnants of a plastid (Škaloud et al. 2014; Grossmann et al. 2016; Graupner et al. 2018). Following these, we hypothesize phototrophic taxa to have larger genomes than heterotrophic ones, due to the differences in cell sizes. We further hypothesize that mixotrophic strains have a genome size intermediate between the other two groups, depending on the degree of phagocytosis and photosynthesis.

Methods

Chrysophyte cultures were obtained from the culture collection at University Duisburg-Essen and originate from sampling sites distributed worldwide (Table 1). They comprise various chrysophytes belonging to different orders and genera. Phototrophic chrysophyte strains are assigned to the Synurales and comprise the genera *Mallomonas* and *Synura*. Mixotrophic strains (mostly affiliated with the genera *Dinobryon*) are assigned to the Ochromonadales, as are most of the investigated heterotrophic chrysophyte strains. The investigated heterotrophic chrysophyte strains comprise representatives of various genera, e.g., *Spumella*, *Pedospumella*, or *Segregatospumella* (Findenig et al. 2010; Grossmann et al. 2016).

Further, this study will evaluate 1C-values, referring to the overall DNA content of a haploid nucleus of a cell, which is usually constant in any species (Greilhuber et al. 2005). This ensures that differences, caused by the ploidy level, are evaded, even if we could not observe any sexual reproduction in our cultures, during the measurements. Therefore, the terms "genome size" and "C-value" have the same meaning.

Culture conditions

Cultures were grown in different media (IB (Hahn et al. 2003), NSY (Hahn et al. 2003), DY-V (Keller and Andersen, unpublished), or WC (Guillard and Lorenzen 1972)) depending on their growth requirements (Table S1) under a light/dark cycle of 12:12 h at 16 °C. Culture flasks were only opened under sterile conditions. Phototrophic strains were grown in 500 mL TC-flasks, heterotrophic strains in 30-mL TC-flasks with bacteria supplied (*Limnohabitans planktonicus*; strain IID5^{T}). Bacteria cultures were raised on 3 g L^{-1} NSY medium and

Table 1 Strain description including reference of species delimination

Strain	Species	Culture collection	Other designations	Origin	Order	Species published in
JBAF 33	*Acrispumella msimbasiensis*	Uni-DUE		Tanzania	Ochromonadales	(Grossmann et al. 2016)
JBC 27	*Chromulinospumella sphaerica*	Uni-DUE		China	Chromulinales	(Grossmann et al. 2016)
A-R4-D6	*Cornospumella fuschlensis*	Uni-DUE		Austria	Ochromonadales	(Grossmann et al. 2016)
9-4-C1	HF	Uni-DUE		Austria		
N1846	HF	Uni-DUE		Japan		
1006	*Pedospumella encystans*	Uni-DUE		Antarctica	Ochromonadales	(Findenig et al. 2010)
JBM|S 11	*Pedospumella encystans*	Uni-DUE /SAG	SAG 2324	Austria	Ochromonadales	(Findenig et al. 2010)
JBC|S 23	*Pedospumella sinomuralis*	Uni-DUE		China	Ochromonadales	(Grossmann et al. 2016)
JBC 07	*Poteriospumella lacustris*	Uni-DUE		China	Ochromonadales	(Findenig et al. 2010)
JBM 10	*Poteriospumella lacustris*	Uni-DUE/NCMA/SAG	CCMP 3167/SAG 2323	Austria	Ochromonadales	(Findenig et al. 2010)
JBNZ 41	*Poteriospumella lacustris*	Uni-DUE		New Zealand	Ochromonadales	(Findenig et al. 2010)
A-R3-A3	*Segregatospumella dracosaxi*	Uni-DUE		Austria	Segregatales	(Grossmann et al. 2016)
JBNZ 39	*Spumella lacusvadosi*	Uni-DUE		New Zealand	Ochromonadales	(Grossmann et al. 2016)
A-R4-A6	*Spumella rivalis*	Uni-DUE			Ochromonadales	(Findenig et al. 2010)
376hm	*Spumella* sp.	Uni-DUE		Antarctica	Ochromonadales	(Cienkowsky 1870)
LO244K-D	*Spumella* sp.	Uni-DUE/CCAC/NCMA	M3950 B/CCMP 3058	Austria	Ochromonadales	(Cienkowsky 1870)
199hm	*Spumella vulgaris*	Uni-DUE/SAG	SAG 2322	Antarctica	Ochromonadales	(Findenig et al. 2010)
933-7	*Chlorochromonas danica*	Uni-DUE/SAG		Denmark	Ochromonadales	(Andersen et al. 2017)
DS	*Poterioochromonas malhamensis*	Uni-DUE		Germany	Ochromonadales	(Péterfi 1969)
M2953	*Bitrichia* sp.	Uni-DUE/CCAC	M2953 B	Austria	Hibberdiales	(Woloszynska 1914)
FU24K-BA	*Dinobryon bavaricum*	Uni-DUE/CCAC	M2950 B	Austria	Ochromonadales	(Imhof 1890)
FU18K-A	*Dinobryon divergens*	Uni-DUE		Austria	Ochromonadales	(Imhof 1887)
FU22K-AK	*Dinobryon divergens*	Uni-DUE/CCAC	M3941 B	Austria	Ochromonadales	(Imhof 1887)
WA20K-H	*Dinobryon divergens*	Uni-DUE/CCAC	M2972 B	Austria	Ochromonadales	(Imhof 1887)
WA26K-D	*Dinobryon divergens*	Uni-DUE		Austria	Ochromonadales	(Imhof 1887)
LO226K-S	*Dinobryon pediforme*	Uni-DUE/CCAC	M2958 B	Austria	Ochromonadales	(Steineck 1916)
OE22K-D	*Dinobryon sociale*	Uni-DUE/NCMA	CCMP 2884	Austria	Ochromonadales	(Ehrenberg 1834)
OE26K-V	*Dinobryon sociale var. americana* cf. *div. schauinslandii*	Uni-DUE		Austria	Ochromonadales	(Ehrenberg 1834)
AU32K-E	*Dinobryon* sp.	Uni-DUE		Austria	Ochromonadales	(Ehrenberg 1834)
FU29K-J	*Dinobryon* sp.	Uni-DUE		Austria	Ochromonadales	(Ehrenberg 1834)
WA32K-W	*Dinobryon* sp.	Uni-DUE		Austria	Ochromonadales	(Ehrenberg 1834)
WI32K-F	*Dinobryon* sp.	Uni-DUE		Austria	Ochromonadales	(Ehrenberg 1834)
PR26K-G	*Epipyxis* sp.	Uni-DUE/CCAC	M2966 B	Austria	Ochromonadales	(Ehrenberg 1838)
FU36K-N	*Kephyrion* sp.	Uni-DUE		Austria	Ochromonadales	(Pascher 1911)

Table 1 (continued)

Strain	Species	Culture collection	Other designations	Origin	Order	Species published in
WA34K-E	*Uroglena* sp.	Uni-DUE/CCAC	M2977 B	Austria	Ochromonadales	(Woloszynska 1914)
WA18K-M	*Mallomonas annulata*	Uni-DUE		Austria	Synurales	(Harris 1967)
PR26K-H	*Mallomonas caudata*	Uni-DUE/NCMA	CCMP 2893	Austria	Synurales	(Iwanoff 1899)
WA40K-F	*Mallomonas caudata*	Uni-DUE		Austria	Synurales	(Iwanoff 1899)
OE26K-M	*Mallomonas* cf. *tonsurata*	Uni-DUE/NCMA	CCMP 2895	Austria	Synurales	(Teiling 1912)
B 601	*Mallomonas kalinae*	Uni –DUE/CAUP		Czech Republic	Synurales	(Rezácova 2006)
OE40K-J	*Mallomonas* sp.	Uni-DUE/NCMA	CCMP 3271	Austria	Synurales	(Perty 1852)
WI26K-B	*Mallomonas* sp.	Uni-DUE		Austria	Synurales	(Perty 1852)
S 20.45	*Synura heteropora*	Uni-DUE/CAUP	B 709	Scotland	Synurales	(Škaloud et al. 2014)
WA18K-A	*Synura petersenii*	Uni-DUE/NCMA	CCMP 2898	Austria	Synurales	(Korshikov 1929)
WA18K-U	*Synura* sp.	Uni-DUE		Austria	Synurales	(Ehrenberg 1834)
LO234K-E	*Synura sphagnicola*	Uni-DUE/CCAC	M2959 B	Austria	Synurales	(Korshikov 1929)

Culture collections: *Uni-DUE* Culture Collection of the working group Jens Boenigk, Essen, *SAG* Sammlung von Algen Kulturen, Göttingen, *CCAC* culture collection of algae, cologne, *NCMA* National Center for Marine Algae and Microbiota, East Boothbay, *CAUP* culture collection of algae, Prague), *HF* heterotrophic flagellate

transferred to the correspondent chrysophyte culture medium. The chrysophyte strains *Ochromonas danica* (strain SAG933-7), *Poterioochromonas malhamensis* (strain DS), and *Poteriospumella laucustris* (strains JBC07, JBM10, and JBNZ41) were grown axenically in 3 g L^{-1} NSY medium. The cultures were regularly checked using light microscopy.

Sample preparation

Nuclei extraction for plant reference standards (Table 2) was performed separately from chrysophyte samples as according to Galbraith et al. (1983) and Albach and Greilhuber (2004), using OTTO-1 isolation buffer and fresh plant leaf tissue: a piece of plant leaf (1 cm^2) was placed into a Petri dish and covered with 550 μL OTTO-1 isolation buffer. The leaf piece was chopped with a sharp razor blade. Afterwards, the resulting suspension was filtered through a 30 μm nylon mesh (Partec®). Afterwards, 50 μL of RNAse-A solution (3 g L^{-1}) was added, and the suspension was incubated in a preheated water bath at 37 °C for 30 min.

For chrysophyte sample preparation, we used cultures with a cell density of 1–3 × 10^5 cells mL^{-1}. Samples were centrifuged to separate the cells from the medium (about 12,000 rpm; 2–4 min). The pellets were diluted in 450 μL OTTO-1 isolation buffer and incubated at 50 °C for 5–10 min to achieve cell breakdown and make the nucleus accessible. Afterwards, 50 μL of RNAse-A solution was added to the chrysophyte sample, and the cell suspension was incubated for 30 min at room temperature.

In cases, in which the chrysophyte *Synura sphagnicola* (strain LO234K-E) was used as standard for the measurement, both chrysophyte cultures were also prepared separately.

After RNAse incubation, the chrysophyte sample suspension was added to the reference standard suspension and mixed gently. This mixture of 450 μL was stained with 2 mL of propidium iodide (PI) (Dolezel et al. 1992) and incubated at 4 °C for 60 min.

Genome sizes were measured in four replicates containing a sample and the standard. The fluorescence of stained nuclei was quantified using a CyFlowSL flow cytometer (Partec GmbH, Münster, Germany) with excitation by a Green NdYAG laser tuned at 532 nm. Genome size was estimated as according to Dolezel et al. (1992). The conversion of mass values into base-pair numbers was achieved using the factor 1 pg = 978 Mbp (Dolezel et al. 2003).

Cell sizes of the chosen chrysophytes (living samples) were measured using a Nikon Eclipse Ti-S inverted microscope with 40× magnification. Cell length and width were obtained using the "NIS-Elements Basic Research" software. The cell volume was calculated using the formula for calculation of ellipsoid volumes. Average values were calculated from at least 35 cells.

Table 2 Reference genome sizes of all used standards for flow cytometry

Standard	1C value (reference GS)	Order	Published in
Raphanus sativus	0.55 pg	Brassicales	(Doležel et al. 1998)
Solanum pseudocapsicum	1.29 pg	Solanales	(Temsch et al. 2010)
Hedychium gardnerianum	2.01 pg	Zingiberales	(Meudt et al. 2015)
Synura spagnicola (strain LO234K-E)	0.20 pg	Synurales	Newly established

Statistical analyses

Statistical analyses were performed using the program SigmaPlot 12.5 (Systat Software GmbH, Erkrath, Germany) and the software R (®The R Foundation). To evaluate if there were significant differences in genome sizes or cell sizes of phototrophic, heterotrophic, and mixotrophic chrysophyte strains, we used Kruskal-Wallis one-way ANOVAs on ranks (Kruskal and Wallis 1952), as the data did not show a normal distribution (Shapiro and Wilk 1965). ANOVAs were followed by Dunn's pairwise multiple comparison procedure (Dunn 1964). The family-wise error rate (FWER) was corrected using Šidák's adjustment (Šidák 1967). We determined the relationship between genome size and cell size within one nutrition mode, as well as between all chrysophyte strains, using Bayesian regression models, determining the Bayesian R_s^2 following the solution of Gelman et al. (2017).

Results

Average cell volumes of investigated strains ranged between 111.85 ± 65.67 µm^3 (strain N1846) and 8356.78 ± 2961.43 µm^3 (strain WA40K-F; *M.caudata*) (Fig. 1). The interspecific variation of cell size was huge over all examined strains. M. caudata (strain WA40K-F) had a cell volume about 10 times higher than any other examined chrysophyte strain. Within one strain, cell volume variation increased with increasing cell volume. We detected significant differences between cell volumes of heterotrophic and phototrophic strains ($p = 0.0002$). The cell volume of mixotrophic strains was intermediate, significantly differing from phototrophic cell volumes ($p = 0.0437$) but not from heterotrophic cell volumes ($p = 0.0902$).

Our genome size measurements demonstrate that chrysophytes cover a wide range of eukaryotic genome sizes, ranging from 0.045 ± 0.001 pg (44.2 Mbp; strain 9-4-C1) to 12.426 ± 0.191 pg (12,177.4 Mbp; strain WA40K-F; *M. caudata*) (Fig. 2). Mean genome sizes of heterotrophic chrysophytes ($\tilde{x}^H = 0.077$) varied significantly from mean genome sizes of phototrophic chrysophytes ($\tilde{x}^P = 0.837$; $p \leq 0.0001$), as well as from mean genome sizes of mixotrophic chrysophytes ($\tilde{x}^M = 0.154$; $p = 0.0060$). The difference between mean genome sizes of phototrophic and mixotrophic chrysophytes also varied significantly ($p = 0.0015$). Standard deviation was comparable small within one strain, among all

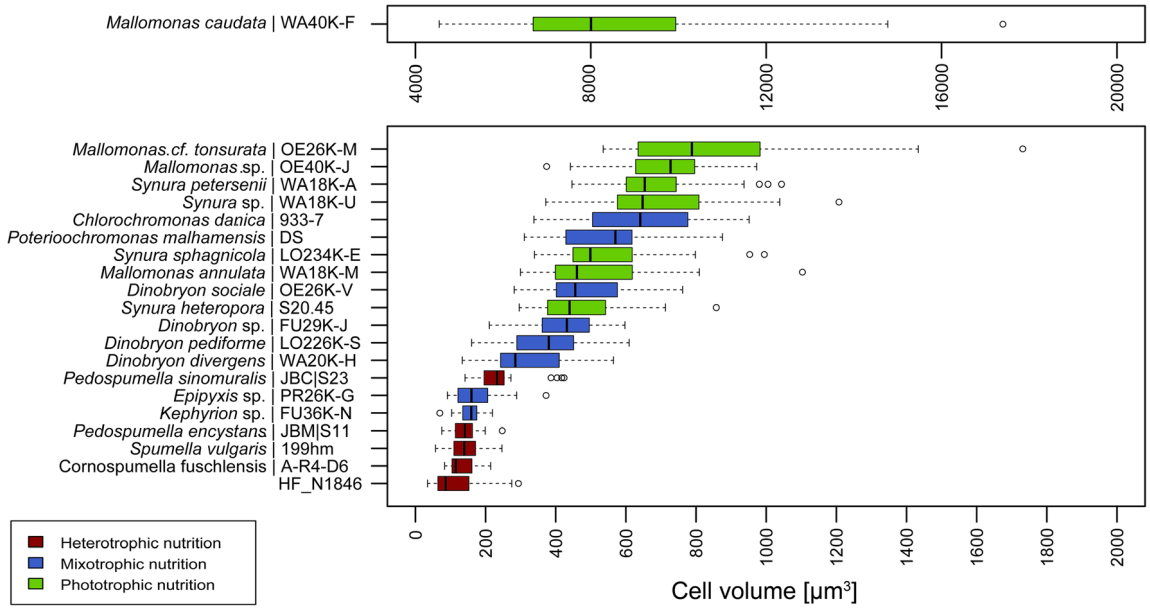


Fig. 1 Cell volumes [µm^3] of different chrysophytes. Different colors represent the different nutritional modes present. Phototrophic chrysophytes (light green) do have highest cell volumes compared to heterotrophic (dark red) and mixotrophic (blue) chrysophytes. Heterotrophic chrysophytes had the smallest cell volumes. HF = Heterotrophic flagellate

stains great variations occurred, especially between heterotrophic and phototrophic chrysophytes (Table S1).

Our results confirm that genome size generally increases with cell size in chrysophytes ($R_s^2 = 0.992$). However, within the different trophic modes, this trend could only be confirmed for phototrophic ($R_s^2 = 0.997$) and weakly for heterotrophic strains ($R_s^2 = 0.557$) (Fig. 3). By contrast, for the investigated mixotrophic strains, genome size did not increase with cell size ($R_s^2 = 0.077$). Nevertheless, results of partial correlations should be interpreted with caution since they are based on a small number of samples.

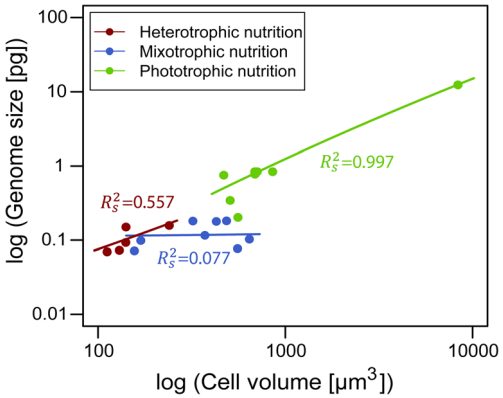


Fig. 3 Regression analysis of log (genome size [pg]) and log (cell volume [µm^3]). Different colors represent the different nutritional modes present (light green = phototrophic; drak red = heterotrophic; blue = mixotrophic). The analysis displays a positive relationship between genome size and cell volume for heterotrophic and phototrophic chrysophyte strains. There is no positive relationship within mixotrophic chrysophyte strains

Discussion

Our study reveals differences in genome size and cell volume regarding the different nutritional modes observable in chrysophytes. Genome sizes of the investigated chrysophytes were generally within the range of reported genome sizes for eukaryotes (Fig. 4). In particular, the genome sizes of phototrophic strains were comparable to those reported for other taxonomic groups, whereas genome sizes of heterotrophic and mixotrophic genomes were among the smallest

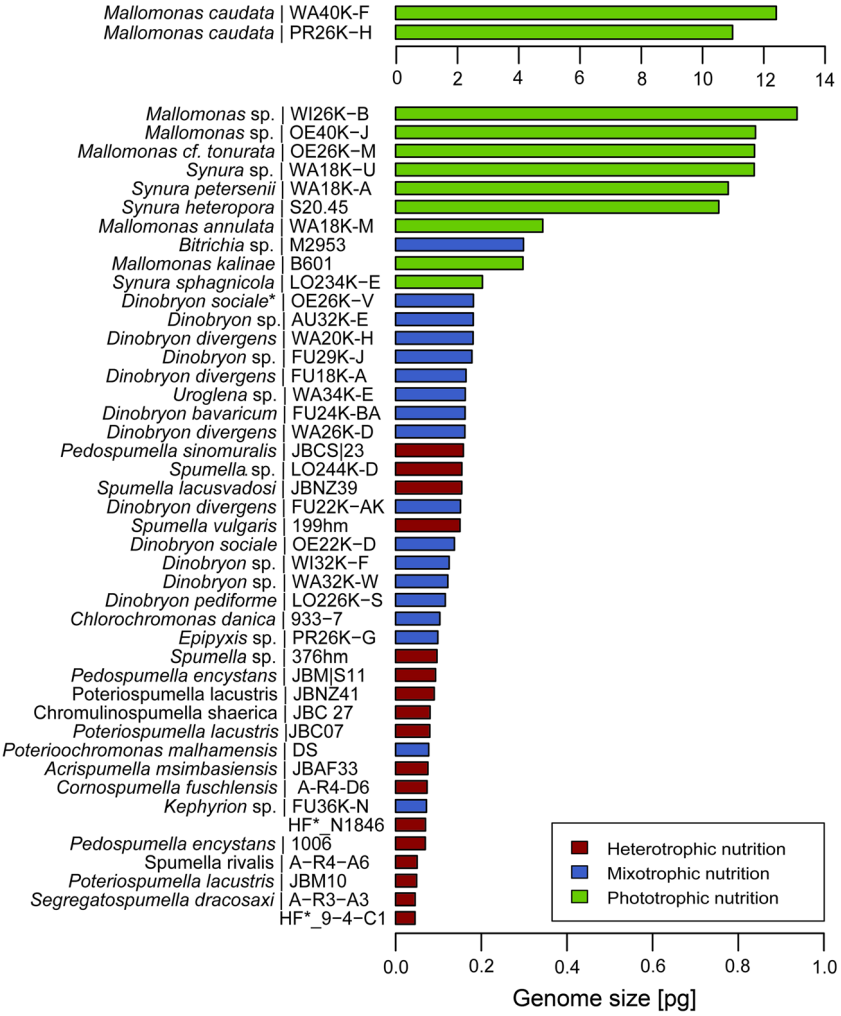


Fig. 2 Genome size [pg] of investigated chrysophytes. Different colors represent the different nutritional modes present. Heterotrophic chrysophytes (dark red) tend to have smaller genome sizes, compared to phototrophic chrysophytes (light green), while mixotrophic chrysophytes (blue) show intermediate genome sizes. **Dinobryon sociale var. americana* cf. *div.* schauinslandii; HF = Heterotrophic flagellate

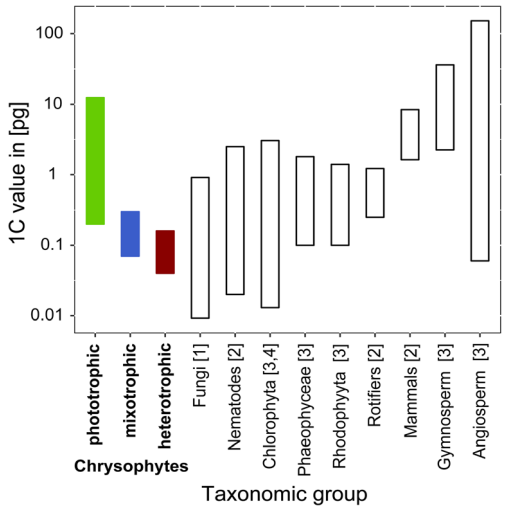


Fig. 4 Comparison of genome size within different taxonomic groups. Mixotrophic (blue) and heterotrophic (dark red) chrysophytes rank among the smallest eukaryotic genomes (values obtained from [1] Mohanta and Bae 2015; Egertová and Sochor 2017; [2] Gregory 2017; [3] Bennett 2012; [4] Courties et al. 1994)

genomes for compared eukaryotes (Fig. 4) (values obtained from Courties et al. 1994; Bennett 2012; Mohanta and Bae 2015; Egertová and Sochor 2017; Gregory 2017).

Consistent with studies on the Baccillariophyta (Conolly et al. 2008) and various plant species (Price et al. 1973; Cavalier-Smith 1985b; Dolezel et al. 1998), genome size increases with the cell volume, at least for the investigated phototrophic and heterotrophic chrysophyte strains (Fig. 3).

While generally in accordance with our initial hypotheses, results for mixotrophic organisms should be interpreted with care due to the bias introduced by the over-representation of the genus *Dinobryon*, among mixotrophic chrysophytes, which does not represent the whole diversity of genome sizes and cell volume in this trophic mode, since chrysophytes comprise numerous other mixotrophic genera (Beisser et al. 2017; Bock et al. 2017). In addition, the relative share of phototrophy and heterotrophy could affect genome size in mixotrophic taxa as well, since both nutritional strategies are used by different mixotrophic species to different extents. Furthermore, plastid size varies in these taxa, ranging from large plastids in genera like *Dinobryon*, with a photosynthetically active plastid, to genera like *Poterioochromonas*, in which the plastid is greatly reduced. In addition, one can detect great variation in cell size within the investigated mixotrophic chrysophytes: the predominant phototrophic strains of *Dinobryon* are larger and show a higher variability in cell volume than the predominantly heterotrophic strains, in particular *Poterioochromonas*. Therefore, we hypothesize that not only the genome size but also cell volume is associated with the degree of use of either phagocytosis or photosynthesis.

Assuming that cell volume is the dominant factor in determining genome sizes, all factors that influence cell size, including, e.g., grazing pressure, should affect genome size evolution as well (Conolly et al. 2008). Indeed, Bennett (1972) suggests that the genome has a function beyond the coding of genetic information: it also determines, e.g., the size of the nuclear envelope, which surrounds the genome and selects for a karyoplasmatic ratio that allows the optimal transfer of RNA (Cavalier-Smith 1985a).

However, since the investigated taxa affiliated with different trophic modes differed systematically in cell size, the effect of cell size and trophic mode could not clearly be separated.

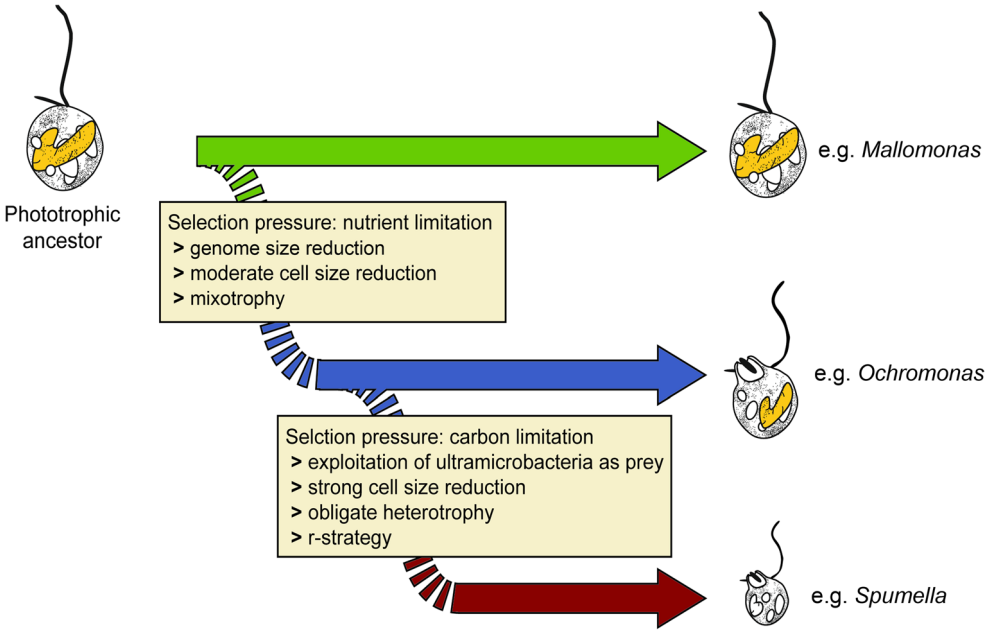


Fig. 5 Model of evolution of genome size, cell volume, and nutritional mode of chrysophytes: nutrient limitations may have driven genome size reduction in the ancestors of mixotrophic (and heterotrophic) chrysophytes, as well as the evolution of phagotrophic mechanisms to attain additional nutrients. Cell size reduction is supposedly a more gradual process, coming into play in taxa which were already able to obtain nutrients by phagotrophy, which optimized food uptake by the optimization of the predator-prey size ratio. This may have triggered the evolution of obligate heterotrophs in many chrysophyte lineages independently

Systematic differences in cell volume and genome size between phototrophic, heterotrophic, and mixotrophic chrysophyte strains are likely to be associated with different life strategies. Phototrophic strains are associated with larger body size, a longer lifespan, and lower growth rates (Kapraun 2005; Cavalier-Smith 1980). In contrast, smaller genomes are associated with small body size and short generation time (Kapraun 2005; Cavalier-Smith 1980). Consistent with these observations, genome downsizing could result in increasing growth rates, because the time needed for mitosis and meiosis would be reduced (Hessen et al. 2010). This theory is supported by findings of Raven (1997), who discovered that pure phototrophs have lower maximum-specific growth rates than mixotrophs, which, in turn, have lower rates than pure heterotrophs (Raven 1997).

This trend is consistent for the investigated taxa and seems to be independent of the phylogenetic affiliation at least for mixotrophic and heterotrophic chrysophytes. As the Synurales are the only phototrophic taxon within Chrysophyceae, independent control groups within Chrysophyceae are missing with respect to phototrophs. Therefore, functional and phylogenetic effects cannot completely be resolved with respect to phototrophs.

Downsizing of genomes can occur in environments in which nitrogen (N) and phosphate (P) are limited. By contrast, genome size expansion is more likely in environments with high N and P supply, which suggests a possible role of N and P availability in the evolution of genomes (Sardans et al. 2012). The potential routes of nutrient uptake and nutrient availability in heterotrophic, mixotrophic, and phototrophic taxa could therefore systematically affect cell and genome size evolution. This is in accordance with our observation that the mode of nutrition appears to affect cell volume and genome size evolution.

As chrysophytes branch within the Ochrophyta, which comprise the phototrophic Stramenopiles, it is likely that mixotrophic and heterotrophic chrysophytes evolved from a phototrophic ancestor (Cavalier-Smith 1999; Aleoshin et al. 2016). A further evidence for this hypothesis is that the Synurophytes, which consist of only phototrophic representatives, branch relatively basal in a phylogenetic tree. However, the heterotrophic Paraphysomonadida also branch basal indicating that the loss of phototrophy occurred early in some branches of Chrysophyceae (Bock et al. 2017). As the basal branching pattern is not yet sufficiently resolved, we cannot fully exclude the possibility of early heterotrophic ancestors within Chrysophyceae even though such a scenario appears highly unlikely as it would require secondary "revival" of the plastid and plastid metabolism (Graupner et al. 2018).

However, if we assume a phototrophic ancestor of the chrysophytes to be limited in terms of nutrient availability, the acquisition of phagocytosis and invention of mixotrophy to evade such limitations would be a plausible scenario (Burkholder et al. 2008) (Fig. 5). Especially the competition in allocation of P between DNA and RNA owing to P limitations could be a factor responsible for genome downsizing in evolutionary time (Hessen et al. 2008; Hessen et al. 2010). With invention of mixotrophy, the ratio of C/N would rise, which is consistent with increased grazing activity (Sardans et al. 2012).

Abandonment of photosynthesis could also be due to energetical factors. In a purely phototrophic organism, the photosynthetic apparatus can account for 50% of the energetic costs of cell synthesis and comprises a huge amount of the cellular biomass (Raven 1997). For pure heterotrophs (with only remains of plastids), the corresponding fraction is usually below 10%. In mixotrophic strains, some predominantly heterotrophic species have a strongly reduced photosynthetic apparatus, such as *Poterioochromonas*, whereas predominantly phototrophic taxa usually have a large photosynthetic apparatus. Therefore, the fraction of the energy and nutrients required by the photosynthetic apparatus may account for the difference between predominantly phototrophic or heterotrophic life strategies (Raven 1997).

Since mixotrophic organisms can obtain nutrients via phagotrophy, further selection towards downsizing of genomes due to a lack of nutrients is relaxed. Continuing, we would find genome sizes similar to those of heterotrophic chrysophytes if the process only depended on P limitations. Similarly, mixotrophic genome sizes would be comparable to those of phototrophic chrysophytes if the maintenance of the photosynthetic apparatus decided.

Therefore, the evolution towards smaller genomes seems to be driven by carbon (C) limitations (Fig. 5): Chrysophytes generally lack uptake mechanisms for carbon in the form of bicarbonate; they use carbon dioxide (CO_2), instead, which is primarily available in slightly acidic environments (Maberly et al. 2009). Predominantly, phototrophic chrysophytes are thus more restricted to acidic environments than predominantly heterotrophic ones. Heterotrophic chrysophytes are able to cover their requirements by preying upon bacteria. Due to this, a reduction of the cell size to optimize predator-prey interactions is reasonable. A small size allows for grazing on ultramicrobacteria (Castro et al. 2009) and therefore expands the spectrum of potential bacterial prey. This would be particularly beneficial in nutrient-limited environments (Nygaard and Tobiesen 1993), but if cell size decreases to a certain point, the necessity and ability of holding a functional plastid are questionable. As a consequence, purely heterotrophic nutrition modes can develop. Furthermore, optimized preying will result in the necessity to survive in environments with limited light, because bacterial abundances are higher in profundal regions resulting from decomposition processes.

By contrast, phototrophs are consistently large. The greater cell size of phototrophs allows for hosting a big functional plastid and may further be beneficial for swimming towards

the sunlit upper water layer (Waite et al. 1997). However, for very large cells, the efficiency of light harvesting seems to decrease with increasing size (Geider et al. 1986). For diatoms, it was shown that large phototrophic cells have smaller relative chlorophyll concentrations (Agustí 1991) and that abundant chlorophyll interferes mutually, the so-called "package effect" (Finkel and Irwin 2000).

Surprisingly, for *Mallomonas caudata*, we could detect very high genome sizes and cell volumes. Higher cell volumes could serve as preying protection (Lampert and Sommer 2007). Generally, we can find diverse kinds of protection against preying, especially in Synurophytes, e.g., colony forming of *Synura* or cell size and spikes of different *Mallomonas* species. One mechanism leading to larger cell volumes is polyploidy. However, chrysophytes and their chromosomes are too small to determine the chromosome number by light microscopy. This restriction has also been observed in other species (Alberts et al. 2002; McIntyre 2012).

In general, a downsizing of genomes caused by nutrient limitations seems to be plausible in the case of chrysophytes because of the various nutritional modes observable. Although variations in nuclear-DNA content can arise via many other mechanisms as well, for instance by chromosome polymorphisms as well as polyploidy or duplications, there is still no generally accepted theory of what determines genome size (Cavalier-Smith 1982, 2005; Bennett and Leitch 2005). Particularly, polyploidy, which is common in plants (Soltis and Soltis 1999), could be another factor responsible for genome size variation and an important process in the genome evolution of chrysophytes (Conolly et al. 2008). Whole genome sequencing of selected chrysophyte strains is planned; this will presumably give us a greater insight into genome size evolution in chrysophytes.

Our findings suggest that

(i) the transition from phototrophy to mixotrophy, which occurred presumably early in chrysophyte evolution, is the force behind genome size reduction.

(ii) This may indicate that nutrient limitations on the ancestors of the mixotrophic (and heterotrophic) chrysophytes may have driven both genome size reduction as well as the evolution of phagotrophic mechanisms as a means of attaining additional nutrients in the presumably strongly nutrient-limited ancestors of these chrysophyte lineages.

(iii) Cell volume reduction is presumably a more gradual process which comes into play largely in taxa that were already able to obtain nutrients via phagotrophy as an adaptation to specific prey. Once the chrysophyte ancestors evolved the ability to take in food phagotrophically, optimizing food uptake via the optimization of the predator-prey size ratio presumably influenced cell size, e.g., by increasing the ability to prey upon ultramicrobacteria which consequently may have triggered the evolution of obligate heterotrophs in many chrysophyte lineages independently.

Acknowledgements We are grateful to Silvia Kempen and the working group Biodiversity and Evolution of Plants (Carl von Ossietzky University Oldenburg) who made it possible to measure genome sizes by flow cytometry and provided helpful commentaries to the results. We also would like to thank the German Research Foundation for financial support (projects BO 3245/17 and BO 3245/19).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Agustí, S. (1991). Allometric scaling of light absorption and scattering by phytoplankton cells. *Canadian Journal of Fisheries and Aquatic Sciences, 48*(5), 763–767. https://doi.org/10.1139/f91-091.

Albach, D. C., & Greilhuber, J. (2004). Genome size variation and evolution in Veronica. *Annals of botany., 94*(6), 897–911. https://doi.org/10.1093/aob/mch219.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. (Hg.) (2002). Molecular biology of the cell. *Chapter 4: the global structure of chromosomes. 4th edition.* New York: Garland Science.

Aleoshin, V. V., Mylnikov, A. P., Mirzaeva, G. S., Mikhailov, K. V., & Karpov, S. A. (2016). Heterokont Predator Develorapax marinus gen. et sp. nov. - A Model of the Ochrophyte Ancestor. *Frontiers in microbiology, 7*, 1194. https://doi.org/10.3389/fmicb.2016.01194.

Andersen, R. A., Graf, L., Malakhov, Y., & Su Yoon, H. (2017). Rediscovery of the Ochromonas type species Ochromonas triangulata (Chrysophyceae) from its type locality (Lake Veysove, Donetsk region, Ukraine). *Phycologia, 56*(6), 591–604. https://doi.org/10.2216/17-15.1.

Beisser, D., Graupner, N., Bock, C., Wodniok, S., Grossmann, L., Vos, M., Sures, B., Rahmann, S., & Boenigk, J. (2017). Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. *PeerJ, 5*, e2832. https://doi.org/10.7717/peerj.2832.

Bennett, M. D. (1972). Nuclear DNA content and minimum generation time in herbaceous plants. *Proceedings of the Royal Society B: Biological Sciences, 181*(1063), 109–135. https://doi.org/10.1098/rspb.1972.0042.

Bennett, M. D. (2012). Plant DNA C-value Database (release 6.0). *www.kew.org/cvalues*.

Bennett, M. D., & Leitch, I. J. (2005). Plant genome size research: A field in focus. *Annals of botany, 95*(1), 1–6. https://doi.org/10.1093/aob/mci001.

Bock, C., Medinger, R., Jost, S., Psenner, R., Boenigk, J. (2014). Seasonal variation of planktonic chrysophytes with special focus on Dinobryon. *Fottea*, 179–190. https://doi.org/10.5507/fot.2014.014.

Bock, C., Chatzinotas, A., & Boenigk, J. (2017). Genetic diversity in chrysophytes. Comparison of different gene markers. *Fottea, 17*(2), 209–221. https://doi.org/10.5507/fot.2017.005.

Bock, C., Wessel, C., Wu, W., Marks, S., Olefeld, J., Jensen, M., Boenigk, J. (submitted 2018). Cool and shady: Ecophysiological preferences of chrysophytes. *Aquatic Microbial Ecology.*

Boenigk, J., Pfandl, K., & Hansen, P. J. (2006). Exploring strategies for nanoflagellates living in a ´wet desert´. *Aquatic Microbial Ecology, 44*, 71–83. https://doi.org/10.3354/ame044071.

Boenigk, J., Pfandl, K., Stadler, P., & Chatzinotas, A. (2005). High diversity of the 'Spumella-like' flagellates. An investigation based on the SSU rRNA gene sequences of isolates from habitats located in six different geographic regions. *Environmental microbiology, 7*(5), 685–697. https://doi.org/10.1111/j.1462-2920.2005.00743.x.

Burkholder, J. M., Glibert, P. M., & Skelton, H. M. (2008). Mixotrophy, a major mode of nutrition for harmful algal species in eutrophic waters. *Harmful Algae, 8*(1), 77–93. https://doi.org/10.1016/j.hal.2008.08.010.

de Castro, F., Gaedke, U., & Boenigk, J. (2009). Reverse evolution: Driving forces behind the loss of acquired photosynthetic traits. *PLoS One, 4*(12), e8465. https://doi.org/10.1371/journal.pone.0008465.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. *Journal of Cell Science, 1978*(34), 247–278.

Cavalier-Smith, T. (1980). r- ad K-tactics in the evolution of protists developmental systems: cell and genome size, phenotype diversifying selection and cell cycle patterns. *BioSystems (12),* 43–59.

Cavalier-Smith, T. (1982). Skeletal DNA and the evolution of genome size. *Annual review of biophysics and bioengineering, 11*, 273–302. https://doi.org/10.1146/annurev.bb.11.060182.001421.

Cavalier-Smith, T. (1985a). The evolution of genome size. In T. Cavalier-Smith (Ed.), *Eukaryote gene numbers, non-coding DNA and genome size*. London, UK: Wiley.

Cavalier-Smith, T. (1985b). The evolution of genome size. In T. Cavalier-Smith (Ed.), *Eukaryote gene numbers, non-coding DNA and genome size.* London, UK: Wiley.

Cavalier-Smith, T. (2005). Economy, speed and size matter: Evolutionary forces driving nuclear genome miniaturization and expansion. *Annals of botany, 95*(1), 147–175. https://doi.org/10.1093/aob/mci010.

Cavalier-Smith, T. (1999). Principles of protein and lipid targeting in secondary Symbiogenesis. Euglenoid, dinoflagellate, and Sporozoan plastid origins and the eukaryote family tree, 2. *J Eukaryotic Microbiology, 46*(4), 347–366. https://doi.org/10.1111/j.1550-7408.1999.tb04614.x.

Cienkowsky, L. (1870). Über Palmellaceen und einige Flagellaten. *Archiv für Mikroskopische Anatomie, 6*, 421–438.

Conolly, J. A., Oliver, M. J., Beaulieu, J. M., Knight, C. A., Tomanek, L., & Moline, M. A. (2008). Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). *Journal of Phycology, 44*, 124–131. https://doi.org/10.1111/j.1529-8817.2007.00452.x.

Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chrétiennot-Dinet, M. J., Neveux, J., et al. (1994). Smallest eukaryotic organism. *Nature, 370*(6487), 255. https://doi.org/10.1038/370255a0.

Dolezel, J., Bartos, J., Voglmayr, H., & Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry Part A : the journal of the International Society for Analytical Cytology, 51*(2), 127–128; author reply 129. https://doi.org/10.1002/cyto.a.10013.

Dolezel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysák, M. A., Nardi, L., & Obermayer, R. (1998). Plant genome size estimation by flow cytometry. Inter-laboratory comparison. *Annals of botany, 82* (suppl_1), 17–26. https://doi.org/10.1093/oxfordjournals.aob.a010312.

Dolezel, J., Sgorbati, S., & Lucretti, S. (1992). Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol Plant, 85*(4), 625–631. https://doi.org/10.1034/j.1399-3054.1992.850410.x.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics, (6)*. 241–252.

Egertová, Z., & Sochor, M. (2017). The largest fungal genome discovered in Jafnea semitosta. *Plant Syst Evol, 303*(7), 981–986. https://doi.org/10.1007/s00606-017-1424-9.

Ehrenberg, C. G. (1834). Dritter Beitrag zur Erkenntniss grosser Organisation in der Richtung des kleinsten Raumes. *Abhandlungen der Königlichen Akademie der Wissenschaften zu Berlin, 1833*, 145–336.

Ehrenberg, C. G. (1838). Die Infusionsthierchen als vollkommene Organismen: Ein Blick in das tiefere organische Leben der Natur (pp. i–xviii, [1-4], 1–547, [1]. Leipzig: Verlag von Leopold Voss.

Findenig, B. M., Chatzinotas, A., & Boenigk, J. (2010). Taxonomic and ecological characterization of stomatocysts ofspumella-like flagellates (Chrysophyceae). *Journal of Phycology, 46*(5), 868–881. https://doi.org/10.1111/j.1529-8817.2010.00892.x.

Finkel, Z. V., & Irwin, A. J. (2000). Modeling size-dependent photosynthesis: Light absorption and the allometric rule. *Journal of theoretical biology, 204*(3), 361–369. https://doi.org/10.1006/jtbi.2000.2020.

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., Firoozabady, E. (1983) Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science (220),* 1049–1051. https://doi.org/10.1126/science.220.4601.1049.

Geider, R. J., Platt, T., Raven, J. A. (1986). Size dependence of growth and photosynthesis in diatoms a synthesis. *Marine Ecology Programme Series (30),* 93–104.

Gelman, A., Goodrich, B., Gabry, J., Ali, I. (2017), R-squared for Bayesian regression models. *Unpublished via http://www.stat.columbia.edu/~gelman/research/unpublished/.*

Graupner, N., Jensen, M., Bock, C., Marks, S., Rahmann, S., Beisser, D., Boenigk, J. (2018). Evolution of heterotrophy in chrysophytes as reflected by comparative transcriptomics. *FEMS microbiology ecology, 94*. https://doi.org/10.1093/femsec/fiy039

Gregory, T. R. (2017). Animal Genome Size Database. *www.genomesize.com*.

Greilhuber, J., Dolezel, J., Lysák, M. A., & Bennett, M. D. (2005). The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. *Annals of botany, 95*(1), 255–260. https://doi.org/10.1093/aob/mci019.

Grossmann, L., Bock, C., Schweikert, M., & Boenigk, J. (2016). Small but Manifold - Hidden Diversity in "Spumella-like Flagellates". *The Journal of eukaryotic microbiology, 63*(4), 419–439. https://doi.org/10.1111/jeu.12287.

Guillard, R. R. L., & Lorenzen, C. J. (1972). Yellow-geen algae with Chloropyllide C. *Journal of Phycology, 8*(1), 10–14. https://doi.org/10.1111/j.1529-8817.1972.tb03995.x.

Gupta, A., LaBar, T., Miyagi, M., & Adami, C. (2016). Evolution of genome size in asexual digital organisms. *Scientific reports, 6*, 25786. https://doi.org/10.1038/srep25786.

Hahn, M. W., Lunsdorf, H., Wu, Q., Schauer, M., Hofle, M. G., Boenigk, J., & Stadler, P. (2003). Isolation of novel Ultramicrobacteria classified as Actinobacteria from five freshwater habitats in Europe and Asia. *Applied and Environmental Microbiology, 69*(3), 1442–1451. https://doi.org/10.1128/AEM.69.3.1442-1451.2003.

Harris, K. (1967). Variability in Mallomonas. *Journal of General Microbiology, 46*, 185–191.

Hessen, D. O., Jeyasingh, P. D., Neiman, M., & Weider, L. J. (2010). Genome streamlining and the elemental costs of growth. *Trends in ecology & evolution, 25*(2), 75–80. https://doi.org/10.1016/j.tree.2009.08.004.

Hessen, D. O., Ventura, M., & Elser, J. J. (2008). Do phosphorus requirements for RNA limit genome size in crustacean zooplankton? *Genome, 51*(9), 685–691. https://doi.org/10.1139/G08-053.

Imhof, O. E. (1887). Studien über die Fauna hochalpiner Seen, insbesondere des Cantons Graubünden. *Jahresbericht der Naturforschenden Gesellschaft Graubündens, 30,* 45–164.

Imhof, O. E. (1890). Das Flagellatengenus Dinobryon. *Zoologischer Anzeiger, 13*, 483–488.

Iwanoff, L. A. (1899). Beitrag zur Kenntnis der Morphologie und Systematik der Chrysomonadinen. *Bulletin de l'Académie Impériale des Sciences de Saint Pétersbourg, 5*, 247–262.

Kapraun, D. F. (2005). Nuclear DNA content estimates in multicellular green, red and brown algae: Phylogenetic considerations. *Annals of botany, 95*(1), 7–44. https://doi.org/10.1093/aob/mci002.

Koester, J. A., Brawley, S. H., Karp-Boss, L., & Mann, D. G. (2007). Sexual reproduction in the marine centric diatom Ditylum brightwellii (Bacillariophyta). *European Journal of Phycology, 42*(4), 351–366. https://doi.org/10.1080/09670260701562100.

Korshikov, A. A. (1929). Studies on the Chrysomonads I. *Archiv für Protistenkunde, 67*, 253–290.

Kremp, A. (2013). Diversity of dinoflagellate life cycles: Facets and implications of complex strategies. *Biological and Geological Perspectives of Dinoflagellates*, S. 189–198. https://doi.org/10.1144/TMS5.18.

Kristiansen, J., Škaloud, P. (2017). Chrysophyta. In *Handbook of the Protists, Chrysophyta.* https://doi.org/10.1007/978-3-319-28149-0_43

Kruskal, W. H., Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*, 583–621.

Lampert, W., & Sommer, U. (2007). *Limnoecology: The ecology of lakes and streams* (2nd ed.). Oxford: Oxford Univ. Press.

Maberly, S. C., Ball, L. A., Raven, J. A., & Sultemeyer, D. (2009). Inorganic carbon Aquisition by Chrysophytes. *Journal of Phycology, 45*(5), 1052–1061. https://doi.org/10.1111/j.1529-8817.2009.00734.x.

McIntyre, P. J. (2012). Cytogeography and genome size variation in the Claytonia perfoliata (Portulacaceae) polyploid complex. *Annals of botany, 110*(6), 1195–1203. https://doi.org/10.1093/aob/mcs187.

Meudt, H. M., Rojas-Andrés, B. M., Prebble, J. M., Low, E., Garnock-Jones, P. J., & Albach, D. C. (2015). Is genome downsizing associated with diversification in polyploid lineages of Veronica? *Botanical Journal of the Linnean Society, 178*(2), 243–266. https://doi.org/10.1111/boj.12276.

Mohanta, T. K., & Bae, H. (2015). The diversity of fungal genome. *Biological procedures online, 17*, 8. https://doi.org/10.1186/s12575-015-0020-z.

Nygaard, K., Tobiesen, A. (1993). Bacterivory in algae a survival strategy during nutrient limitation. *Limnol Oceanog (38),* 273–279.

Pascher, A. (1911). Über Nannoplanktonten des Süsswassers. *Berichte der deutsche botanischen Gesellschaft, 29*, 523–533.

Perty, M. (1852). Zur Kenntniss kleinster Lebensformen: nach Bau, Funktionen, Systematik, mit Specialverzeichniss der in der Schweiz beobachteten. *Verlag von Jent & Reinert*, 1–228.

Péterfi, L. S. (1969). The fine structure of Poterioochromonas malhamensis (Pringsheim) comb. nov. with special reference to the lorica. *Nova Hedwigia, 17*, 93–103.

Price, H. J., Sparrow, A. G., & Naumann, A. F. (1973). Correlations between nuclear volume, cell volume, and DNA content in meristematic cells of herbaceous angiosperms. *Experentia, 29*, 1028–1029.

Raven, J. A. (1997). Phagotrophy in phototrophs. *Limnol. Oceanogr., 42*(1), 198–205. https://doi.org/10.4319/lo.1997.42.1.0198.

Rezácova, M. (2006). Mallomonas kalinae (Synurophyceae), a new species of alga from northern Bohemia, Czech Republic. *Preslia, 78*, 353–358.

Sandgren, C. D. (1991). Chrysophyte reproduction an resting cysts: A paleolimnologist's primer. *Journal of Paleolimnology, 5*, 1–9.

Sardans, J., Rivas-Ubach, A., & Peñuelas, J. (2012). The elemental stoichiometry of aquatic and terrestrial ecosystems and its relationships with organismic lifestyle and ecosystem structure and function. A review and perspectives. *Biogeochemistry, 111*(1–3), 1–39. https://doi.org/10.1007/s10533-011-9640-9.

Shapiro, S. S., Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrica (52),* 591–611. https://doi.org/10.2307/2333709.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association (62),* 626–633.

Škaloud, P., Škaloudová, M., Procházková, A., & Němcová, Y. (2014). Morphological delineation and distribution patterns of four newly described species within the Synura petersenii species complex (Chrysophyceae, Stramenopiles). *European Journal of Phycology, 49*(2), 213–229. https://doi.org/10.1080/09670262.2014.905710.

Soltis, D. E., Soltis, P. S. (1999). Polyploidy: recurrent formation and genome evolution. *Trends Ecology and Evolution, 14* 348–352.

Steinecke, F. (1916). Die Algen des Zehlaubruches in systematischer und biologischer Hinsicht. *Schriften der koniglichen physikalisch-okonomischen Gesellschaft zu Konigsberg, 56*, 1–138.

Teiling, E. (1912). Schwedische Planktonalgen. I. *Phytoplankton aus dem Råstasjön bei Stockholm. Svensk Botanisk Tidskrift, 6*(2), 267–281.

Temsch, E. M., Greilhuber, J., & Krisai, R. (2010). Genome size in liverworts. *Preslia, 82*, 63–80.

Waite, A., Fischer, A., Thompson, P. A., Harrison, P. J. (1997). Sinking rate versus cell volume relationships illuminate sinking rate control mechanisms in marine diatoms. *MEPS* (157), 97–108. https://doi.org/10.3354/meps157097.

Woloszynska, J. (1914). Zapiski algologiczne/Algologische Notizen. Sprawozdania z Posiedzen Towarzystwa Naukowego Warszawskiego. *Wydzial II, 7*, 22–26.

Publication 3:

Nutrient-driven genome evolution revealed by comparative genomics of chrysomonad flagellates

Stephan Majda, Daniela Beisser, Jens Boenigk

Submitted to: Nature Microbiology

Contribution to this publication:

conception & planning:	30 %	planing bioinformatical and laboratory methods
experimental work:	100 %	cultivation, DNA isolation, testing methods
data analysis:	100 %	testing methods, programming, debugging, collect web data, data analysis, statistic analysis, interpretation of results
writing the manuscript:	75 %	writing the draft, creation of tables and figures

....
Stephan Majda Prof. Dr. Jens Boenigk

RESEARCH

Nutrient-driven genome evolution revealed by comparative genomics of chrysomonad flagellates

Stephan Majda*, Daniela Beisser† and Jens Boenigk†

*Correspondence:
stephan.majda@uni-due.de
Department of Biodiversity, Duisburg-Essen, Essen, Germany
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Phototrophic eukaryotes have evolved mainly by the primary or secondary uptake of photosynthetic organisms. A return to heterotrophy occurred multiple times in various protistan groups such as Chrysophyceae, despite the expected advantage of autotrophy. There is the assumption that the evolutionary shift to mixotrophy and further to heterotrophy is triggered by a differential importance of nutrient and carbon limitation. We sequenced the genomes of 16 chrysophyte strains and compared their genomes in terms of size, function and sequence characteristics in relation to photo-, mixo- and heterotrophic nutrition. All strains were sequenced with Illumina and partly with PacBio resulting in an assembly size between 40 to 120 Mb. Heterotrophic taxa have reduced genomes and a higher GC content of up to 59% as compared to phototrophic taxa. In addition, we identified several complete facultative pathways in their genomes, which were largely not detected by transcriptome sequencing. Heterotrophs have a large pan genome, but a small core genome, indicating a differential specialization of the distinct lineages. The pan genome of mixotrophs and heterotrophs taken together but not the pan genome of the mixotrophs alone, covers the complete functionality of the phototrophic strains indicating a random reduction of genes. The observed ploidy ranges from di- to tetraploidy and was found to be independent of taxonomy or trophic mode. Our results substantiate an evolution driven by nutrient and carbon limitation.

Keywords: heterokonts; stramenopiles; whole genome sequencing; gold algae; autotrophic nutrition

Introduction

Competition for nutrients and organic carbon are among the major ecological selective forces in the evolution of eukaryotes. The presence of phototrophic, heterotrophic and mixotrophic taxa and evidence for multiple gains and losses of photosynthesis across nearly all major eukaryotic supergroups reflects the significance of nutritional constraints in the evolution of life [1–3]. Taxa which have experienced such an evolutionary modification of their basic nutritional strategy must bear hallmarks of this fundamental switch in their genomes. Here we track down the genomic fingerprints of the eco-evolutionary constraints linked to the varying significance of nutrient and carbon shortage.

Recent hypotheses suggest that the shift to mixotrophy was one way to overcome nutrient limitations while the shift towards heterotrophy was caused by carbon limitations [4, 5]. In contrast to former individual case studies, here we use the parallel

evolution of heterotrophic Chrysophyceae from phototrophic and mixotrophic ancestors in order to separate general directions and constraints in genome evolution from random modifications.

For phototrophic organisms, nutrients usually are a limiting factor. Phagotrophic uptake of bacteria and with that of nutrients and organic compounds could overcome limitations of essential nutrients such as nitrogen and phosphorus [6]. Alleviating the nutrient limitation may drive taxa into another challenge as nutrient shortage may simply be replaced by prey/carbon shortage. Towards the smaller size spectrum of eukaryotic life certain free-living species reduced their cell size to prey more efficiently on ultramicrobacteria [4, 5].

The changing relevance of either nutrient or carbon limitation alters constraints in genome evolution which should be reflected by the incorporation of nucleotides with different costs, the loss of obsolete genes and the evolution of new gene functions.

The Chrysophyceae within the Ochrophyta (Stramenopiles) are especially suited to investigate the evolutionary significance of this shift of the nutritional mode since the loss of photosynthesis occurred several times independently within this group [7–10]. Using transcriptome sequencing Graupner et al. [11] analysed plastid-targeting and -encoded genes of heterotrophic chrysophytes compared to photo- and mixotrophs in which they identified different stages of plastid pathway reduction and degradation of accompanying structures. Dorrell et al. [10] extended the analysis to include plastid genomes and identified shared losses of function across chrysophytes. To our knowledge, all existing studies so far focused on the plastid and plastid-related functions but none yet on changes in the nuclear genome related to changes in trophy. Findings from transcriptome sequencing of 18 chrysophyte strains provide first insights into molecular changes between trophic modes showing that heterotrophs possess a reduced repertoire of genes related to photosynthesis but an increased or up-regulated repertoire of pathways associated with food uptake and motility [12]. Apart from this, changes in the nuclear gene content, genome size, GC content, ploidy and further genomic features connected to trophy have not been analysed. Therefore, in the presented study we examine genomes of 16 chrysophytes including phototrophic, mixotrophic and heterotrophic lineages to investigate the impact of the nutritional shift and its drivers.

We hypothesize that:
(1) The nuclear genome is reduced in size from photo- over mixo- to heterotrophic species, either as a result of nutrient limitations or as a size adaptation to feed on small bacteria. Likewise, the ploidy in heterotrophs is lower.
(2) Accompanying the genome reduction and specialization of heterotrophs, we will find a loss of functional genes, especially no longer required genes for photosynthesis, biosynthesis of certain amino acids and plastid-targeting genes, as well as a decrease in intergenic regions, reflected by an increased gene density.
(3) The GC content of photo- and mixotrophic species is lower than in heterotrophic species if nutrients were the limiting factor during evolution.

Results

Genome assembly and gene content

Sequencing generated at least 60 million 150 bp-long paired end Illumina reads for each strain (see Table 1) and a total of 700,000 PacBio reads (details see Table S1). The comparison of binning strategies has shown that MetaBAT removed much more reads than MaxBin2 resulting in one-third smaller assembly sizes (details see Table S2). Since the steps for filtering of prokaryotic reads were identical resulting in similar qualities, we chose MaxBin2 for binning. On average 61.2% (median)

Table 1 Sequencing and assembly statistics.

strain	reads [M]	GC [%]	coverage	used reads∗ [%]	N50	number of contigs [10^3]
JBC27	70	51.3	115.7	90.6	4039	35
A-R4-D6	77.9	51.7	38.1	22.8	2506	46
1006	83.7	54.5	13.9	6.2	1019	54
JBMS11	59.7	51.4	52.2	11.9	3876	29
JBC07⋆	80.8⋆	53.1	116⋆	100	40,792⋆	9,122⋆
JBM10⋆	79.6⋆	53.1	128⋆	100	52,370⋆	9,400⋆
JBNZ41⋆	109.0⋆	52.9	153⋆	100	24,662⋆	13,826⋆
199hm	79.8	47.9	19.9	14.9	2675	54
CCAC 4401B	66.5	43.9	45	50	1694	79
FU18K-A	64.3	45.1	60.1	70.7	1836	77
LO226K-S	81.6	51.6	111.4	79.6	2530	50
PR26K-G	67.1	34.1	88.9	52.2	11429	17
933-7	88	45.4	298.6	100	97456	14
DS	90.4	40.4	235.6	100	22680	17
WA18K-M	67.3	40.1	57.5	61.8	7384	43
LO234KE	112.5	46.9	88.9	61.2	2024	77

∗ The percentage of reads after binning and filtering out prokaryotic classified reads ⋆ values from Majda [13]

of the sequencing reads were used for the assembly (see Table 1). Despite similar assembly sizes and N50 values between axenically and non-axenically cultured strains (see Table 1,2), the BUSCO analysis revealed partly incomplete genomes. The completeness of recovered genes of the draft genomes ranged from 4 to 61% (median 29.8%, see Table 4), whereas about 61% recovery of the BUSCO protist dataset seems to represent the largest almost complete chrysophyte genome (based on axenic culture assembled with Illumina and PacBio reads). Assembly size and with that detection of genes and completeness of the genomes seems to be affected by the presence of bacteria; i.e., the quality of the assemblies was higher for axenic strains. However, the following analyses of ploidy, GC content and gene density are independent of genome completeness.

We predicted 23,000 to 120,000 genes for the analysed chrysophyte species (see Table 5). In assembled genomes, parts of long genes can possibly lie on several contigs and be counted multiple times thus overestimating the number of genes. We therefore clustered similar or duplicated genes to orthologous groups (OG) obtaining numbers in the range from 17,000 to 60,000, comparable to gene numbers obtained from transcriptome sequences of Chrysophyceae (range of 8,275 to 72,269; [12], *Ochromonas* sp. 19,692 genes; [14]). Between 14 and 25 % of these orthologous groups could be annotated with a KEGG Orthology (KO) ID for the assignment

to metabolic pathway. It is not surprising that additional use of PacBio sequencing and axenic strains lead to better assembly quality (see supp. Table S3, Fig. S3). However, the omission of axenic and PacBio sequenced species leads to similar results concerning gene and pathway composition. The pathway compositions regarding the KEGG functional groups differed mostly between single strains, instead of nutritional modes (see Fig. 1). Investigated pathways were almost complete and more extensive in the genome compared to the transcriptome (see Fig. 2, also Fig. S7 to S12).

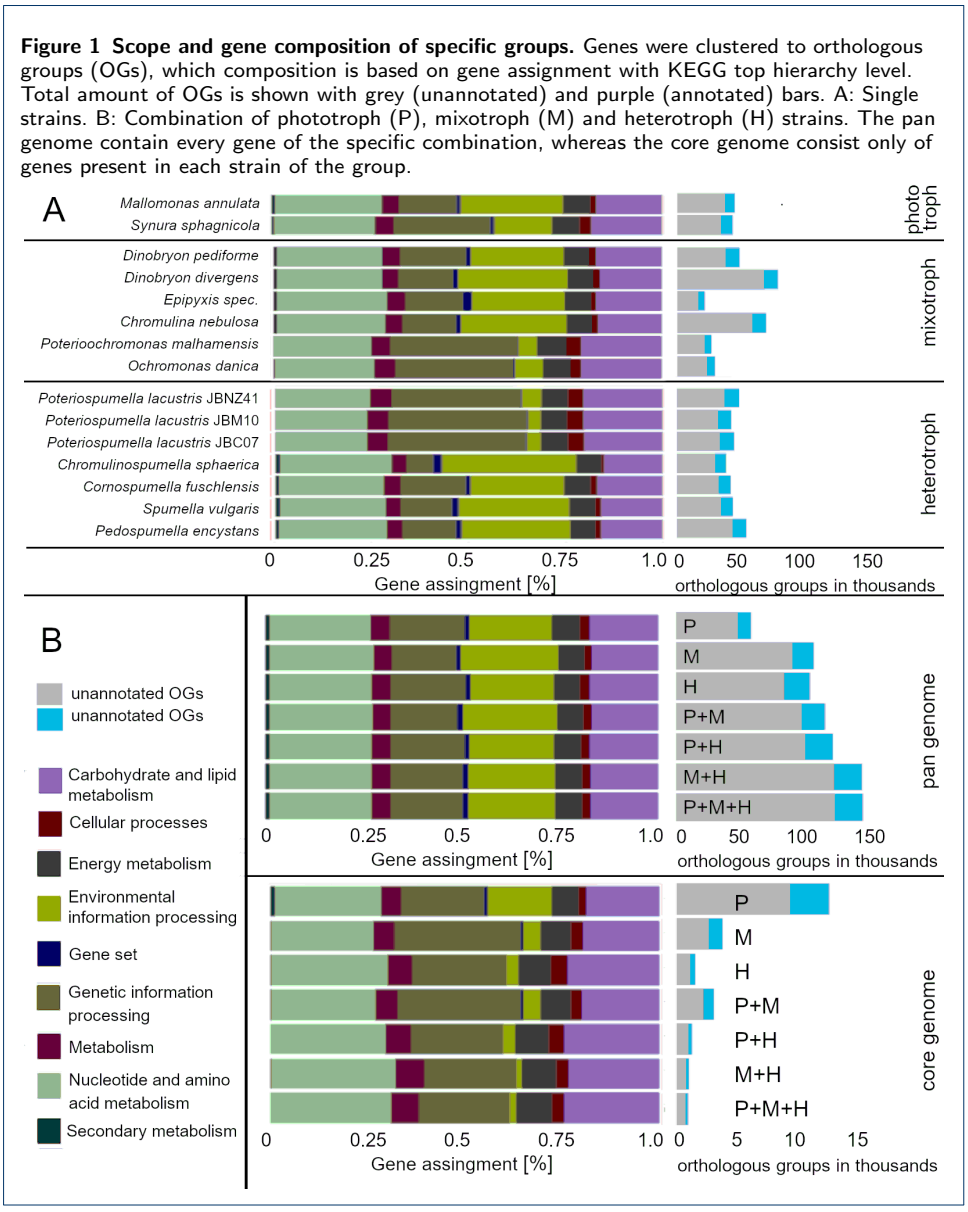


Figure 1 Scope and gene composition of specific groups. Genes were clustered to orthologous groups (OGs), which composition is based on gene assignment with KEGG top hierarchy level. Total amount of OGs is shown with grey (unannotated) and purple (annotated) bars. A: Single strains. B: Combination of phototroph (P), mixotroph (M) and heterotroph (H) strains. The pan genome contain every gene of the specific combination, whereas the core genome consist only of genes present in each strain of the group.

Pan genome and core genome

To expand the analysis of genes and pathways in single species, we compared the sets of orthologous groups (OGs) between the trophic mode. The pan genome, the entire gene set of a group of species, was contrasted to the core genome, the intersection of

genes within the group of species. The pan genome of all investigated phototrophic species contained 47,750 orthologous groups, while the pan genome of mixotrophic taxa contained 90,463 OGs and that of heterotrophic taxa 83,503 OGs. Combining the pan genomes of two trophic modes increased the total number of OGs (Fig. 1), while the combination of all trophic modes resulted in a similar number of OGs as the combination of mixotrophs and heterotrophs only. Each combination of orthologous groups of different nutritional modes resulted in similar compositions of KEGG functional groups at the top hierarchy level (see Fig. 1) and the next lower level (see Fig. S2). In contrast to the pan genome, the intersection between the OGs of phototrophic species resulted in the largest core genome (phototroph OGs: 8,934; mixotroph OGs: 2,561; heterotroph OGs: 1,084), as a consequence of the small group size, while the core genome of the heterotrophic species was the smallest.

Table 2 Genome size estimations. The genome size was estimated in [5] by nuclear staining and flow cytometry.

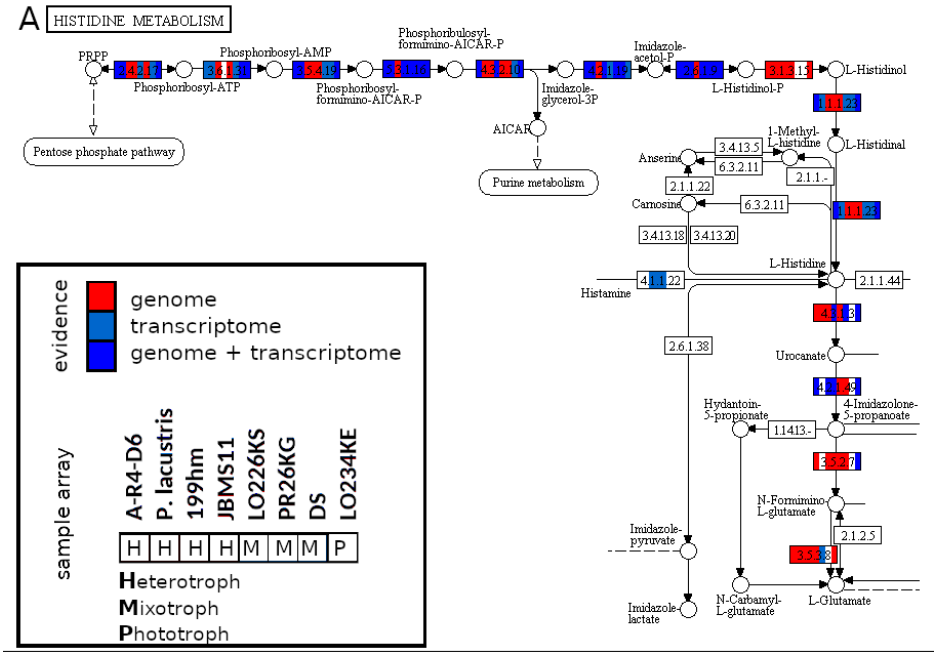
strain	assembly size (haploid) [Mbp]	estimated total size (flow cytometry) [Mbp]	ploidy
JBC27	82	157	di
A-R4-D6	70	143.2	di
1006	56	135	-
JBMS11	61	182.2	tetra
JBC07⋆	49.4	314.4	tri
JBM10⋆	54.7	193	di
JBNZ41⋆	52.8	355	tetra
199hm	90	293.6	di
CCAC 4401B	111	-	tri
FU18K-A	113	321.4	di or tri
LO226K-S	88	226.6	tri
PR26K-G	59	192.6	di or tetra
933-7	44	201.4	tetra
DS	58	150.8	tri
WA18K-M	109	671.8	(di)
LO234KE	116	396.2	di

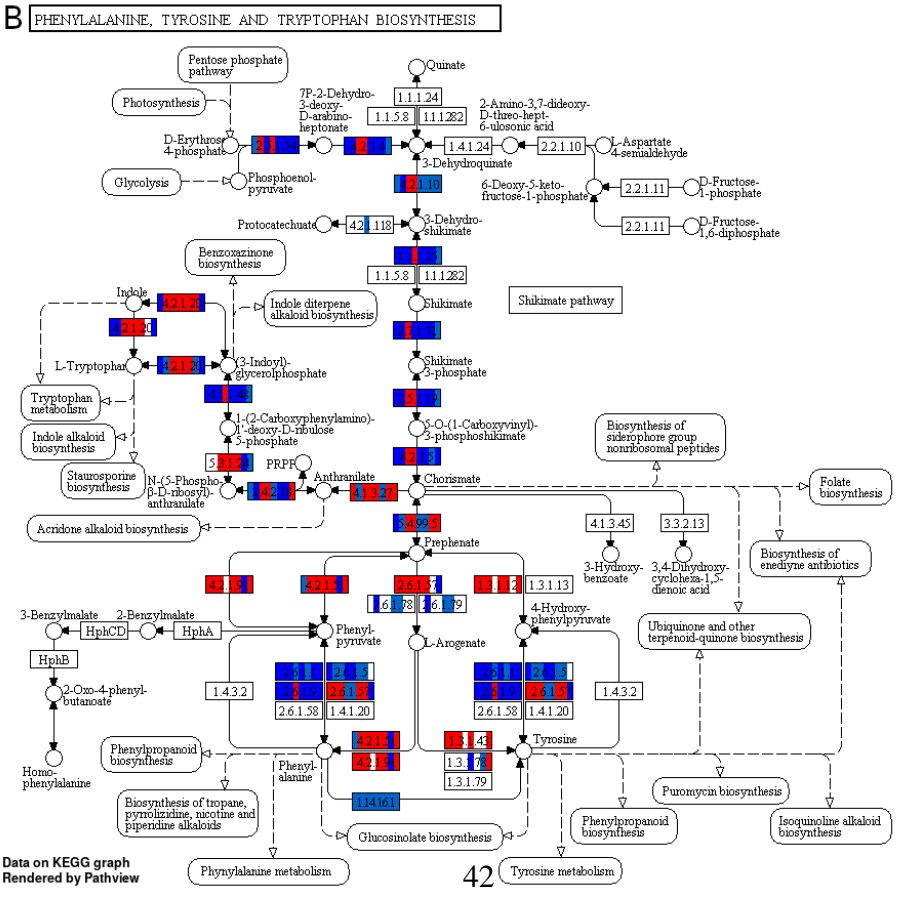
⋆ values from Majda [13]

Genome size

Considering only genomes with high completeness, hetero- and mixotrophic species did not differ largely in assembly size, ranging at around 67 Mbp. We observed significant differences in assembly size between all three groups (anova p-value $<$ 0.05), but these differences were mainly due to the substantially larger genomes of the phototrophic species (see Table 2). Due to few species in this groups, we refrain from drawing a clear conclusion here.

Figure 2 Genomic data complete gene presence. Each box represents a gene and is divided into eight segments. The segments represent from left to right the strains: A-R4-D6, *P. lacustris* (pool of JBC07, JBM10 and JBNZ41), 199hm, JBMS11 (heterotroph, red marked in the legend), LO226KS, PR26KG , DS (mixotroph, blue marked) and LO234KE (phototroph, green marked). The segment is coloured based on the evidence (red: genome, bright blue: transcriptome, dark blue: both). **A:** The path from PRPP (phosphoribosyl pyrophosphate) to histidine were incomplete for 199hm and JBMS11 based on transcriptomic data, but could be completed by genomic data. Whereas from histidine to glutamate was neither evidence for strain DS. **B:** Genomic data complete pathway between 3-Dehydroquinate and Chorismate for strains 199hm and JBMS11.

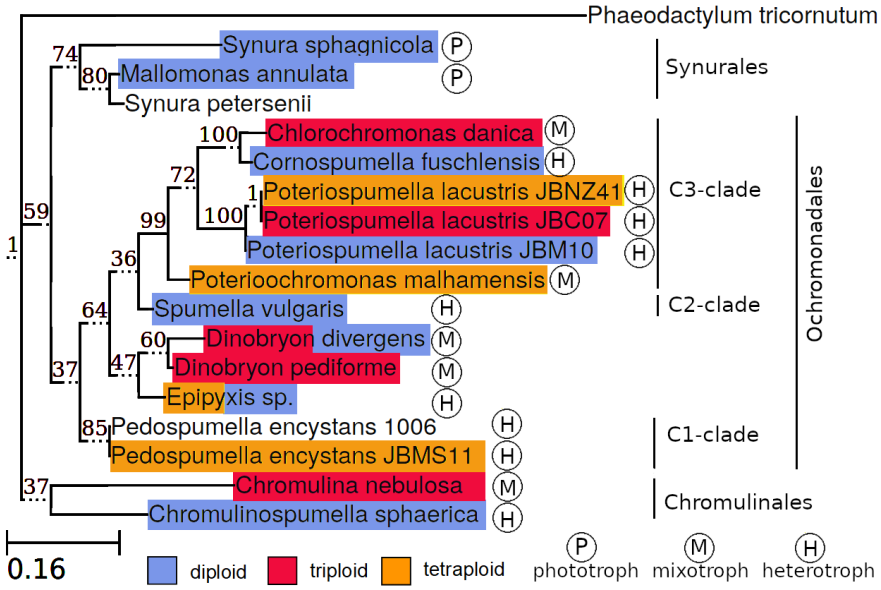




Ploidy

The ploidy of the investigated species ranged from diploidy to tetraploidy and seemed to be independent of nutritional mode and taxonomy (see Fig. 3). *Chromulinospumella sphaerica*, *Cornospumella fuschlensis*, *Poteriospumella lacustris* strain JBM10, *Spumella vulgaris* and *Synura sphagnicola* were diploid, *Poteriospumella lacustris* strain JBC07, *Chromulina nebulosa*, *Dinobryon pediforme* and *Poterioochromonas malhamensis* were triploid and *Pedospumella encystans* strain JBMS11, *Poteriospumella lacustris* strain JBNZ41 and *Chlorochromonas danica* were tetraploid (see Table 2) For *Mallomonas annulata* we had only poor indications of diploidy and the ploidy level was ambiguous in *Dinobryon divergens* (di- or tetraploid) and *Epipyxis sp.* (di- or triploid) (see Fig. S5). For *Pedospumella encystans* 1006 data was not sufficient to obtain a clear result.

Figure 3 Taxonomic distribution of ploidy. Ploidy (blue = diploid, red = triploid and yellow = tetraploid) is independent of nutritional mode(**p**hototroph, **m**ixotroph and **h**eterotroph) or taxonomy. Phylogenetic tree is based on 18S rDNA.



Gene density

In general, there is a correlation between genome size and gene density in eukaryotes [15]. We expected to find a correlation between nutritional mode and gene density, especially in heterotrophs because of the strong selection towards smaller cells. The average gene density ranged from 1,031 (phototroph), over 1,049 (mixotroph) to 1,520 (heterotroph) genes/Mb. A trend towards higher gene density with increased heterotrophy was visible, but these differences were not significant (p-value = 0.32, see Fig. 4).

GC content

The mean GC content of the phototrophic (43.5 %) and mixotrophic group (43.4 %) was similar, whereas heterotrophs (51.6 %) had a strongly increased GC content

(anova: p-value < 0.01, posthoc test mixo- versus heterotrophs p-value < 0.01, see Fig. 4).

The strains *P. encystans* 1006 and *S. vulgaris* 199hm inhabit arctic regions. Both strains had the highest GC content (see Table 1). Excluding both arctic organism from the heterotrophic group in the ANOVA test resulted in a still higher GC content compared to phototrophs and mixotrophs (p-values < 0.01 in anova and posthoc test). We separated the total GC content based on coding sequences, introns and non-coding sequences (see Table S4). The mean GC content of non-coding regions was smallest (35.5 %), intron GC content was in between (41.4 %) and coding sequences had the highest GC content (54.8 %) (pairwise t-test, each p-value ≤ 0.001), whereas nutritional mode had no significant impact (anova: p-value > 0.05). Besides, we did neither obtain correlations between assembly size nor the total genome size and the GC content (p-values > 0.5).

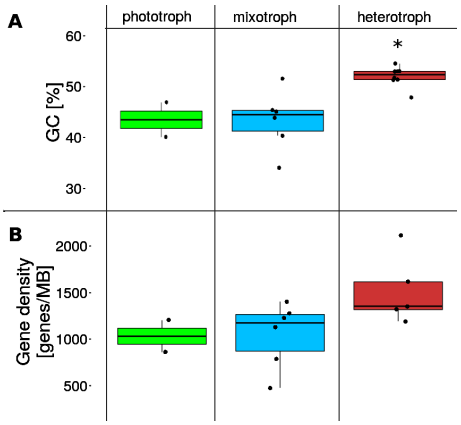


Figure 4 GC content and gene density in relation to nutritional mode. A: The GC content depend on the nutritional mode ($p < 0.001$). B: There is no correlation between trophic mode and gene density.

Discussion

The comparative genome analysis of chrysophytes revealed genome reduction accompanied by an increase in the genomic GC content and an increase in gene density as concomitant phenomenon of the evolutionary switch from phototrophy to heterotrophy. Gene reduction occurred in all heterotrophic lineages but different genes were lost in the various phylogenetic lineages presumably reflecting differential evolutionary selective pressures and adaptations to different environments in the distinct lineages. Even though genome reductions evolved independently, gene losses were predominately observed for non-annotated genes; i.e., genes of unknown functions, while genes and pathways of the primary metabolism were kept in all lineages.

Genome size and ploidy variation

We found considerable differences of genome size and ploidy between strains. While genome size reduction seems to be related to the change of trophy but independent of phylogeny, a variation of ploidy was found in different lineages and seemingly independent of the nutritional mode. From the assembly sizes we could not definitely assess differences attributable to the different trophic modes, but the two larger phototrophic genomes suggest it (see Table 2). Nucleic acids are amongst the most phosphorus and nitrogen-demanding biomolecules and large genomes are costly to build and maintain, thus under nutrient limitation a reduction in genome size is expected. Further, the genome size has been shown to correlate with cell volume (e.g., [16]), which his reduced in phagotrophic species preying on ultramicrobacteria. This

is supported by a study from Olefeld et al. [5] analysing 46 chrysophyte strains, in which a reduction in cell volume and genome size from photo- over mixo- to heterotrophs could be clearly shown with flow cytometry. The latter study calculated haploid genome size based on the assumption of diploid genomes in chrysophytes. As we show here, ploidy varies between strains and thus variation in ploidy overlays the general trend of genome size reduction. Taking the different ploidies into account the genome size estimates based on flow cytometry [5] and based on assembled genomes (this study) largely correspond except for few strains for which the ploidy level could not unambiguously determined (Table 2). In particular, the genomes of phototrophic strains were considerably larger as compared to those of heterotrophic and mixotrophic strains.

Ploidy varied between diploidy and tetraploidy and was independent of phylogeny and nutritional mode. The independence from phylogeny was not surprising as different ploidy levels have already been demonstrated within one species [13, 17]. Apparently, a change in ploidy occurs frequently and is often associated with broader ecological niches and/or invasiveness due to increased heterozygosity and flexibility [18]. In particular, the extra gene copies in polyploid genomes may be beneficial for evolving new functions, novel gene combinations and modified gene expression. For instance, flowering plant lineages polyploidize at a rate which is about 2-10% of the speciation rate [19, 20]. In protists, which replicate predominantly asexual, the rate could even be higher, since the disadvantages of vulnerability to infertility or pairing difficulties in meiosis [21] are of little relevance. Furthermore, polyploidy is advantageous with respect to gene redundancy [21], especially in predominantly asexual species in order to prevent the accumulation of mutations (Muller's ratchet) [22]. However, not only increases in ploidy but also a decrease of the ploidy level may frequently occur as it has been demonstrated for tetraploid or triploid yeast within in a short time period of only 186 generations [23]. Irrespective of the potential advantages of polyploid genomes, an increase in ploidy level counteracts the attempt of reducing genome size, since cell size and DNA content usually correlate [24, 25]. Polyploid organisms form larger cells [26, 27], but the reason for this correlation is not yet clearly understood. Both, the higher amount of DNA and the higher expression of proteins have been suggested to be causative for this correlation [28, 29]. However, deviations from this correlation are known within the same genus [30]. These deviations imply that more factors are decisive for the cell size. Likewise, even though plants often increase their cell size by increasing their ploidy level [27], preventing the ploidy enhancement did not affect cell size [31]. Summarising, our data point to a trade-off between the advantages of genome size reduction and polyploidization: Cell size reduction – and with that genome size reduction – is discussed as an adaptation to increase the efficiency in preying on small bacteria, in particular on ultramicrobacteria. At the same time genome size reduction lowers the costs of DNA buildup and reproduction. In contrast, polyploidization may allow for a higher flexibility enabling the taxa to populate broader niches and may prevent or at least slow down the accumulation of mutations. When nutrients are not limiting this trade-off should level off as it has also been shown for cultures with repeated transfer [32], since selection pressure decrease under laboratory conditions. However, due to selective advantages of both, genome size reduction and polyploidy, it seems reasonable that polyploidy may also evolve under nutrient limitations.

Pan genome and core genome

To analyse the gene composition and gene loss we compiled for each nutritional group and the combinations the pan genome and the core genome. The pan genome comprises genes present in at least one representative, while the core genome comprises genes present in all representatives. One major finding of our study is the deviation between core and pan genome of the heterotrophic taxa. The small core genome in contrast to the large pan genome of the heterotrophs implies that the strong genome reduction in the evolution of the heterotrophs was accompanied by a strong niche specialization. Even though numerous genes have been lost in each heterotrophic lineage the distinct lineages lost different genes. Further, our data demonstrate that in all trophic groups some of the genes are not found in at least one of the other trophic groups. However, taken together the group of mixotrophic and heterotrophic species, their combined genes cover almost all genes and pathways found in the phototrophic species. We expected to see a gradual reduction of pathways related to photoautotrophy depending on the stage of nutritional mode. This gene loss pattern has been shown for the plastome, even if exceptions such as genes related to Rubisco could still persist [33]. The study of Graupner et al. [11] suggests that for pathways directly linked to photosynthesis and genes encoded in the plastid a certain sequence of gene losses occurs, accompanying the shift from mixotrophy to heterotrophy. Here we show that this is not the case for the majority of nucleus-encoded genes but gene loss seems to be different for distinct lineages indicating that different strains presumably were exposed to different evolutionary forces. This is supported by an increased fraction of genes related to environmental information processing and secondary metabolism in the photo- and mixotrophic group (see Fig. 1) while the heterotrophic species have a more specialized and unique gene inventory in these functional groups.

We further demonstrate that primary pathways and annotated pathways of the secondary metabolism are almost complete in the genome. Gene reductions concern mostly genes which could not be annotated; i.e., genes of unknown function which presumably are important for niche specialization but not for the basic requirements of the cells. For instance, the pathways for the biosynthesis of amino acids were complete in all investigated strains even though amino acids can, in principle, also be taken up with food in the phagotrophic taxa (see Fig. 5, also Fig. S7 to S12). This completeness of pathways seems to be different for plastid-encoded genes as suggested by transcriptome data and supported by the structural reduction as observed in microscopical analyses [9]. The maintenance of the pathways of the primary (and to a large part also of the secondary) metabolism could be advantageous for backing-up nutritional requirements in the case of food shortage. As we analyzed the occurrence of genes, our results do not necessarily imply that the genes are functional. But the fact that pathways were mostly found to be complete supports that the genes are still functional. Still, undetected mutations could have rendered genes non-functional or modified their function, for certainty, additional verifications would be needed.

Gene density

We could detect a trend towards higher gene density with increased heterotrophy, but possibly due to the small sample size the difference between groups was not

significant. The gene densities were generally high as compared to gene densities of other organisms. Stramenopiles have on average a gene density of 200-400 genes/Mb [13], whereas prokaryotes or archaea typically have around 1,000 genes/Mb. As eukaryotes with small genomes are known to have gene densities similar to bacteria [15] and constant gene densities independent of genome size [34] our estimates still seem reliable. However, this high gene density could partly be due to a certain overestimation in the gene prediction, in particular comprise a certain bias due to merging of repeat regions during assembly and a potential presence of genes on both strands.

GC content

The GC content is known to be correlated with genome size in bacteria but the relationship for other kingdoms is less clear. Studies analyzing genome size and base composition in available sequenced genomes in various kingdoms found a correlation between average GC content and genome size which was positive in bacteria (specifically Proteobacteria and Actinobacteria), weakly positive in Ascomycota fungi and some plants, but negative in animals and indifferent in two analysed protistan phyla [35–37]. We show that the GC content of chrysophytes did not correlate with genome size. In contrast, the GC content was significantly different between nutritional modes. While the mean GC content of the phototrophic and mixotrophic group was similar, the heterotrophs showed a strongly increased GC content. In general, an increased GC content is associated with higher needs regarding nutrients and energy [38, 39]. Adenine and guanine require 5, cytosine 3 and thymine/uracil only 2 nitrogen atoms, thus the difference in costs is observable in the GC content of a genome. Species living in or adapted to nitrogen-limited environments therefore often use nucleotides that require fewer nitrogen such as A and T [40]. Additionally, at least in bacteria, organisms show distinct GC patterns that are not explained by phylogeny but by similar environments [41]. It has been shown that plants that require more nitrogen to conduct photosynthesis experience stronger selection to minimize nitrogen biosynthesis costs [42]. As the evolution towards heterotrophy; i.e., the establishment of an alternative nutrient source, is typically related to adaptations to low nutrient availability, a lower GC content should be expected in phototrophic chrysophytes. The evolutionary constraint of nutrient limitation should be already relaxed in mixotrophic species, but as the photosynthetic apparatus is costly and accounts for about half of the cell protein [43], mixotrophs may be subject to similar constraints as the phototrophs. Heterotrophic species should predominantly be limited by carbon not by nutrients and thus can afford a higher GC content. The increased energy demand of high GC content led to a low GC content in non-coding regions and introns, whereas coding sequences remained GC-rich encoding cheaper amino acids [38]. Aside from nutritional contraints the GC content may also be affected by temperature. Within the heterotrophic taxa the two strains having the highest GC content originate from Antarctica. Increased GC content was documented in plant [36] and fish [44] species adapted to cold climate and could thus likewise be increased in these two species. Since GC-rich DNA is more stable it possibly indicates a mechanism during cell freezing.

Conclusions

We demonstrate that nutritional constraints drive genome evolution and that limitations in nutrient and carbon acquisition leave footprints in genome evolution. Based on comparative genome analysis of lineages which independently and in parallel evolved heterotrophy from mixotrophic ancestors we separated genomic footprints linked to a nutritional switch from random shifts. In particular, genome size reduction and a shift in GC content reflect changes in nutrient limitation during the evolution of obligate heterotrophy from phototrophic and mixotrophic ancestors. Gene losses accompany these changes but vary between lineages indicating a presumably increasing niche separation and differentiation between the distinct heterotrophic lineages. In the broader context of eukaryotic megaevolution the discovery of disparate gene losses in different lineages and its implication of an accelerated differentiation is intriguing as it adds a new facette on the evolution of eukaryotic diversity. Nutrient and carbon limitation may not only be crucial for the transformation of feeding strategies but furthermore speed-up the evolutionary diversification and thus contribute to rapid radiations on short evolutionary time scales.

Materials and methods

Cultivation and sequencing

We cultivated and sequenced 16 strains (see Table 3) according to Hahn and Majda et al. [13, 45]. Non-axenic heterotrophic and mixotrophic cultures grew with bacterial food supply (*Limnohabitans planktonicus*; strain IID5T). Two days before DNA harvesting in these cultures the feeding with bacteria was omitted.

Table 3 General information of examined species.

species name	strain	nutrient mode∗	reference	sequenced†	medium
Chromulinospumella sphaerica	JBC27	H	[9]	I	IB
Cornospumella fuschlensis	A-R4-D6	H	[9]	I	IB
Pedospumella encystans	1006	H	[46]	I	IB
Pedospumella encystans	JBMS11	H	[46]	I / P	IB
Poteriospumella lacustris⋆	JBC07	H	[46]	I	NSY
Poteriospumella lacustris⋆	JBM10	H	[46]	I / P	NSY
Poteriospumella lacustris⋆	JBNZ41	H	[46]	I	NSY
Spumella vulgaris	199hm	H	[47]	I	IB
Chromulina nebulosa	CCAC 4401B	M	[47]	I	WC
Dinobryon divergens	FU18K-A	M	[48]	I	WC
Dinobryon pediforme	LO226K-S	M	[49]	I	IB
Epipyxis sp.	PR26K-G	M	[50]	I	WC
Chlorochromonas danica⋆	933-7	M	[51, 52]	I / P	NSY
Poterioochromonas malhamensis⋆	DS	M	[53]	I / P	NSY
Mallomonas annulata	WA18K-M	P	[54]	I	WC
Synura sphagnicola	LO234KE	P	[55]	I / P	WC

∗ H = heterotroph, M = mixotroph, P = phototroph, † I = Illumina Hiseq XTen, P = PacBio RSII
⋆ axenic cultures, processed according to Majda [13]

Genome assembly and binning

Unless otherwise stated, the default settings were used for the following programs. An automated workflow with Snakemake [56] processed the sequencing data. Figure S1 gives an overview of the assembly and binning procedure. SPAdes (v3.13.0; with

parameters: –meta [57]) assembled the Illumina reads. If PacBio reads were available they were incorporated in the SPAdes assembly and additionally assembled with CANU (v1.8; with parameters: genomeSize=100m, correctedErrorRate=0.105 [58]), whereby the genome size estimations from Olefeld et al. [5] were used if possible. We renounced genome size estimation based on *k*-mers, since it often lead to size bias in chrysophytes [13], especially through read filtering in non-axenic strains. To classify the reads we combined a tool using compositional features and another applying taxonomical methods. The tool MaxBin2 (v2.2.5;59) binned the contigs created by SPAdes. Subsequently, Kraken2 (v2.0.7 60) classified the bins taxonomically. Therefore, the NCBI database (Release 2018-11-01) and the assemblies of the axenic cultures *Chlorochromonas danica*, *Poterioochromonas malhamensis* and *Poteriospumella lacustris* were used as reference. Including the axenic cultures enabled to classify chloroplast bins as eukaryotic instead of prokaryotic and prevented their exclusion. Bins consist out of several contigs. In some cases the bins contain contigs classified as eukaryotic as well as prokaryotic or unclassified. Depending on the bin composition, we classified and processed unclassified contigs as follows:
1. If the number of eukaryotes is at least half the number of the bacteria, the unclassified are eukaryotic.
2. If the number of unclassified contigs is twice as large as the number of bacteria, the unclassified are eukaryotic.
We have determined these points because the number of known bacteria in the NCBI database is much higher than that of the related protists.
Additionaly, the binning tool MetaBAT(v2.12.1 61) was tested followed by the same classification and filtering steps. The bins were merged into two files containing either eukaryotic or bacterial sequences. The Illumina reads were aligned with Bowtie2 (v.2.3.0; [62]) against the bacterial contigs. Subsequently, the hits were mapped again with Bowtie2 against the eukaryotic contigs. Unaligned reads were marked as bacterial and excluded in further processing. Normally, reads mapping to both the bacterial and the eukaryotic contigs were randomly assigned. To keep the eukaryotic reads in this case we use an incremental mapping.
The contigs of the CANU assembly were only binned by Kraken2, as the long contig length caused sufficient classification accuracy. In a second step, we filtered out all contigs that were not eukaryotic. We repeated the SPAdes assembly (without *–meta* parameter) with filtered Illumina reads and filtered PacBio reads if available (parameter: *–untrusted-contigs*). Contigs smaller than 500 bp were discarded.
Finally, the Benchmarking Universal Single-Copy Orthologs (BUSCO) software (v3.0.1; protists dataset, 63) was used to verify the existence of essential genes.

Ploidy estimation

Genome ploidy was estimated based on *k*-mer frequencies with *smudgeplot* (v0.1.3 with parameters: -k21 -m300 -ci1 -cs10000; `https://github.com/tbenavi1/smudgeplot`).

Gene prediction

The tool AUGUSTUS (v3.3; with parameters: –gff3=on –progress=true –singlestrand=true –UTR=off –species=arabidopsis; [64]) was used for gene prediction. For the strains

Table 4 BUSCO analysis results(protists set)

strain	complete [%]	singelton [%]	duplicate [%]	fragment [%]	missing [%]
JBC27	3.7	2.3	1.4	0	96.3
A-R4-D6	23.7	14.9	8.8	0.5	75.8
1006	9.8	9.3	0.5	0.9	89.3
JBMS11	29.8	22.8	7	0.9	69.3
JBC07⋆	55.8	45.6	10.2	0.0	44.2
JBM10⋆	56.3	51.2	5.1	0.0	43.7
JBNZ41⋆	53.5	51.2	2.3	0.9	35.8
199hm	11.2	9.9	1.3	1.8	87.0
CCAC 4401B	53.4	46	7.4	0.9	45.7
FU18K-A	46.5	43.7	2.8	0.5	53
LO226K-S	11.6	9.3	2.3	0.9	87.5
PR26K-G	60.9	59.5	1.4	1.4	37.7
933-7	60.9	54.4	6.5	0.5	38.6
DS	60.5	52.6	7.9	0.9	38.6
WA18K-M	57.7	56.3	1.4	0.9	41.4
LO234KE	27	22.8	4.2	0.9	72.1

⋆ values from Majda [13]

199hm, A-R4-D6, JBMS11, LO226K-S, LO234KE, PR26KG and *DS* transcriptomic data supported the gene prediction. In this case RepeatScout (v1.0.5; [65]) was used to extract repetitive sequences for each strain. Subsequently, the genomes were masked by BEDTools (v2.28, with parameters: maskfasta -soft; [66]). This prevented aligning RNA reads or predicting genes in repeat regions. Finally, AUGUSTUS parameter were modified to: –softmasking=1 –species=arabidopsis –gff3=on –singlestrand=true –UTR=off –alternatives-from-evidence=false –extrinsicCfgFile –hintsfile. In the extrinsicCfgFile the following relevant weighting were changed to weight more strongly the supporting transcriptomic sequences:
exon hit bonus: 1e+10 (instead 1), nonexonpart malus: 2 (instead 1)
Aligning the transcriptome data with the genome by Minimap2 (v.2.9-r720; with parameters: -c -L -x splice -G 80K -t 16; 67) created the hintsfile.

Gene annotation and clustering

Genes were annotated according to [13] by aligning the predicted genes to the KEGG (Release 2014-06-23, [68]) and UniProt database (Release 2017-09-18, [69]) with Diamond (v0.9.10.111; [70]).
OrthoFinder (v2.2.6; with parameters: -S diamond) clustered the protein sequences of all strains to *orthologous groups* (OG). The majority of gene annotations within a orthologous group determined the annotation of the whole group. We built the union (all OGs) and intersection (OGs shared among all) between species of one nutritional mode (phototroph, mixotroph and heterotroph) and between nutritional modes. Additionally, for these sets the composition of KEGG functional groups were determined.

GC content and gene density

The GC content was determined for the whole genome and for intergenic regions. Significance of gene density and GC content was calculated by one-way ANOVA test for each nutritional mode. The gene density (d) was calculated for the number of genes (n) by the formula:

$$d = \frac{n}{\sum bp_{contigs>500bp}} \cdot 10^6 \quad (1)$$

Table 5 Gene prediction and orthologous groups.

strain	#OrthoGroups	#genes	annotated [%]
JBC27	30,998	62,845	23.1
A-R4-D6	33,640	78,807	23.1
1006	43,528	79,427	20.3
JBMS11	30,479	54,910	19.6
JBM10	34,008	42,266	24.8
JBC07	35,720	41,883	24.7
JBNZ41	38,877	45,773	24.6
199hm	35,008	77,543	21.2
CCAC 4401B	55,495	91,636	15.8
FU18K-A	62,902	93,394	13.7
LO226K-S	39,124	80,221	23.2
PR26K-G	17,019	23,363	23.5
933-7	23,539	36,312	22.5
DS	21,158	33,769	21.8
WA18K-M	35,854	65,203	17.5
LO234KE	34,689	119,235	21.8

Pathway analysis

Present annotated genes were visualized in the KEGG pathway maps by the R package Pathview (v1.24.0; [71]). We focused on the following pathways:
biotin metabolism, histidin metabolism, tryptophan metabolism, nitrogen metabolism, lysine biosynthesis, pyhenylalanine, tyrosine and tryptohan biosynthesis, thiamine metabolism and valine, leucine and isoleucine biosynthesis.
In addition, a comparison was made to determine whether genes from a transcriptome dataset [12] were also present.

Phylogenetic tree

The 18S sequence of each strain, *Phaeodactylum tricornutum* (strain: KMMCC B-386, accession number: GQ452860.1, as outgroup) and *Synura petersenii* (strain: CCMP857, accession number: EF165117.1) were aligned with the Clustal Omega algorithm [72] and processed with EMBLs Simple Pyhlogeny tool [73]. We calculated maximum likelihood with W-IQ-TREE (v1.6.11; with parameters: -st DNA -bb 1300 [74]) and visualized the tree with ETE [75].

Competing interests
The authors declare that they have no competing interests.

Author's contributions
JB conceived the study; JB, SM designed the lab experiments; SM and DB designed computational procedure; SM performed the experiment and analyzed the data.; SM, JB and DB interpreted the data; SM, JB and DB wrote the manuscript; all authors read and approved the final manuscript.

Acknowledgements
We thank Micah Dunthorn for proof-reading and reviewing.

Data availability
DNA-sequencing data that support the findings of this study are available from NCBI BioProject IDs: PRJNA546545 and PRJNA548251.

References

1. Krause K. Plastid Genomes of Parasitic Plants: A Trail of Reductions and Losses. Bullerwell CE, editor. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012.
2. Kamikawa R, Yubuki N, Yoshida M, Taira M, Nakamura N, Ishida Ki, et al. Multiple losses of photosynthesis in Nitzschia (Bacillariophyceae). Phycological Research. 2015;63(1):19–28. Available from: `https://onlinelibrary.wiley.com/and/abs/10.1111/pre.12072`.
3. Suzuki S, Endoh R, Manabe RI, Ohkuma M, Hirakawa Y. Multiple losses of photosynthesis and convergent reductive genome evolution in the colourless green algae Prototheca. Sci Rep. 2018 01;8(1):940.

4. de Castro F, Gaedke U, Boenigk J. Reverse evolution: driving forces behind the loss of acquired photosynthetic traits. PloS one. 2009 December;4(12):e8465. Available from: http://europepmc.org/articles/PMC2794545.
5. Olefeld JL, Majda S, Albach DC, Marks S, Boenigk J. Genome size of chrysophytes varies with cell size and nutritional mode. Organisms Diversity & Evolution. 2018 May;Available from: https://and.org/10.1007/s13127-018-0365-7.
6. Oborník M. Endosymbiotic Evolution of Algae, Secondary Heterotrophy and Parasitism. Biomolecules. 2019;9(7):266.
7. Boenigk J, Pfandl K, Stadler P, Chatzinotas A. High diversity of the 'Spumella-like' flagellates: an investigation based on the SSU rRNA gene sequences of isolates from habitats located in six different geographic regions. Environ Microbiol. 2005 May;7(5):685–697.
8. Cavalier-Smith T, Chao EE. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J Mol Evol. 2006 Apr;62(4):388–420.
9. Grossmann L, Bock C, Schweikert M, Boenigk J. Small but Manifold - Hidden Diversity in "Spumella-like Flagellates" [Journal Article]. J Eukaryot Microbiol. 2016;63(4):419–39. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26662881.
10. Dorrell RG, Azuma T, Nomura M, Audren de Kerdrel G, Paoli L, Yang S, et al. Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. Proc Natl Acad Sci USA. 2019 04;116(14):6914–6923.
11. Graupner N, Jensen M, Bock C, Marks S, Rahmann S, Beisser D, et al. Evolution of heterotrophy in chrysophytes as reflected by comparative transcriptomics. FEMS Microbiol Ecol. 2018 04;94(4).
12. Beisser D, Graupner N, Bock C, Wodniok S, Grossmann L, Vos M, et al. Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. PeerJ. 2017;5:e2832.
13. Majda S, Boenigk J, Beisser D. Intraspecific variation in protists: clues for microevolution from Poteriospumella lacustris (Chrysophyceae). Genome Biol Evol. 2019 Aug;.
14. Lie AAY, Liu Z, Terrado R, Tatters AO, Heidelberg KB, Caron DA. A tale of two mixotrophic chrysophytes: Insights into the metabolisms of two Ochromonas species (Chrysophyceae) through a comparison of gene expression. PLoS ONE. 2018;13(2):e0192439.
15. Hou Y, Lin S. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. PLoS ONE. 2009 Sep;4(9):e6978.
16. Bennett MD, Leitch IJ. Plant genome size research: a field in focus. Annals of botany. 2005;95(1):1–6.
17. Lewis WHe. Polyploidy: Biological Relevance. New York; 1980.
18. Baatout S. Molecular basis to understand polyploidy. Hematology and Cell Therapy. 1999 Aug;41(4):169–170. Available from: https://doi.org/10.1007/s00282-999-0169-5.
19. Otto SP, Whitton J. POLYPLOID INCIDENCE AND EVOLUTION. Annual Review of Genetics. 2000;34(1):401–437. PMID: 11092833. Available from: https://doi.org/10.1146/annurev.genet.34.1.401.
20. Meyers LA, Levin DA. On the abundance of polyploids in flowering plants. Evolution. 2006 Jun;60(6):1198–1206.
21. Comai L. The advantages and disadvantages of being polyploid. Nat Rev Genet. 2005 Nov;6(11):836–846.
22. Muller HJ. Some Genetic Aspects of Sex. The American Naturalist. 1932;66(703):118–138. Available from: http://www.jstor.org/stable/2456922.
23. Gerstein AC, McBride RM, Otto SP. Ploidy reduction in *Saccharomyces cerevisiae*. Biology Letters. 2008;4(1):91–94. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rsbl.2007.0476.
24. Mirsky A, Ris H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. The Journal of general physiology. 1951;34(4):451.
25. Cavalier-Smith T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 2005 Jan;95(1):147–175.
26. Sherman F. [1] Getting started with yeast. In: Guide to Yeast Genetics and Molecular Biology. vol. 194 of Methods in Enzymology. Academic Press; 1991. p. 3 – 21.
27. Melaragno JE, Mehrotra B, Coleman AW. Relationship between endopolyploidy and cell size in epidermal tissue of Arabidopsis. The Plant Cell. 1993;5(11):1661–1668.
28. Galitski T, Saldanha AJ, Styles CA, Lander ES, Fink GR. Ploidy Regulation of Gene Expression. Science. 1999;285(5425):251–254. Available from: https://science.sciencemag.org/content/285/5425/251.
29. Tsukaya H. Does Ploidy Level Directly Control Cell Size? Counterevidence from Arabidopsis Genetics. PLOS ONE. 2013 12;8(12). Available from: https://doi.org/10.1371/journal.pone.0083729.
30. Hoang PTN, Schubert V, Meister A, Fuchs J, Schubert I. Variation in genome size, cell and nucleus volume, chromosome number and rDNA loci among duckweeds. Sci Rep. 2019 Mar;9(1):3234.
31. Leiva-Neto JT, Grafi G, Sabelli PA, Dante RA, Woo Ym, Maddock S, et al. A Dominant Negative Mutant of Cyclin-Dependent Kinase A Reduces Endoreduplication but Not Cell Size or Gene Expression in Maize Endosperm. The Plant Cell. 2004;16(7):1854–1869. Available from: http://www.plantcell.org/content/16/7/1854.
32. Lewis WH. In: Polyploidy: Biological Relevance. Basic Life Sciences. Springer US; 2012. Available from: https://books.google.de/books?id=5K3eBwAAQBAJ.
33. Graham SW, Lam VKY, Merckx VSFT. Plastomes on the edge: the evolutionary breakdown of mycoheterotroph plastid genomes. New Phytologist. 2017;214(1):48–55. Available from: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.14398.
34. Xu L, Chen H, Hu X, Zhang R, Zhang Z, Luo ZW. Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms. Molecular Biology and Evolution. 2006 04;23(6):1107–1108. Available from: https://and.org/10.1093/molbev/msk019.
35. Li XQ, Du D. Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. PLoS ONE. 2014;9(2):e88339.

36. Šmarda P, Bureš P, Horovà L, Leitch IJ, Mucina L, Pacini E, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc Natl Acad Sci USA. 2014 Sep;111(39):E4096–4102.
37. Romiguier J, Ranwez V, Douzery EJ, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 2010 Aug;20(8):1001–1009.
38. Chen WH, Lu G, Bork P, Hu S, Lercher MJ. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. Nat Commun. 2016 Apr;7:11334.
39. Rocha EP, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002 Jun;18(6):291–294.
40. Seward EA, Kelly S. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. Genome Biol. 2016 11;17(1):226.
41. Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. EMBO reports. 2005;6(12):1208–1213. Available from: https://www.embopress.org/doi/abs/10.1038/sj.embor.7400538.
42. Kelly S. The Amount of Nitrogen Used for Photosynthesis Modulates Molecular Evolution in Plants. Mol Biol Evol. 2018 07;35(7):1616–1625.
43. Raven JA, Beardall J, Larkum AW, Sanchez-Baracaldo P. Interactions of photosynthesis with genome size and function. Philos Trans R Soc Lond, B, Biol Sci. 2013 Jul;368(1622):20120264.
44. Zhang D, Hu P, Liu T, Wang J, Jiang S, Xu Q, et al. GC bias lead to increased small amino acids and random coils of proteins in cold-water fishes. BMC Genomics. 2018;19(1):315. Available from: https://doi.org/10.1186/s12864-018-4684-z.
45. Hahn MW, Lunsdorf H, Wu Q, Schauer M, Hofle MG, Boenigk J, et al. Isolation of novel ultramicrobacteria classified as actinobacteria from five freshwater habitats in Europe and Asia. Appl Environ Microbiol. 2003 Mar;69(3):1442–1451.
46. Findenig BM, Chatzinotas A, Boenigk J. TAXONOMIC AND ECOLOGICAL CHARACTERIZATION OF STOMATOCYSTS OF SPUMELLA-LIKE FLAGELLATES (CHRYSOPHYCEAE)1. Journal of Phycology. 2010;46(5):868–881. Available from: http://dx.and.org/10.1111/j.1529-8817.2010.00892.x.
47. Cienkowski L. Uber Palmellaceen unde einige Flagellaten [Journal Article]. Arch Mikrosk Anat. 1870;6:421–438.
48. Imhof OE. Studien über die Fauna hochalpiner Seen, insbesondere des Cantons Graubünden. Jahresbericht der Naturforschenden Gesellschaft Graubündens. 1887;30:45–164.
49. Steinecke F. Die Algen des Zehlaubruches in systematischer und biologischer Hinsicht. Schriften der koniglichen physikalisch-okonomischen Gesellschaft zu Konigsberg. 1916;56:1–138. Available from: http://diatombase.org/aphia.php?p=sourcedetails&id=264372.
50. Ehrenberg CG. Die Infusionsthierchen als vollkommene Organismen: Ein Blick in das tiefere organische Leben der Natur. Leipzig; 1838.
51. Pringsheim EG. Uber Ochromonas danica n. sp. und andere Arten der Gattung. Archives of Microbiology. 1955;23:181–194.
52. Andersen RA, Graf L, Malakhov Y, Yoon HS. Rediscovery of the Ochromonas type species Ochromonas triangulata (Chrysophyceae) from its type locality (Lake Veysove, Donetsk region, Ukraine). Phycologia. 2017;56(6):591–604.
53. S PL. The fine structure of Poterioochromonas malhamensis (Pringsheim) comb. nov. with special reference to the lorica. Nova Hedwigia. 1969;17:93–103.
54. Harris K. Variability in Mallomonas. Journal of General Microbiology. 1967;46:185–191.
55. A KA. Studies on the Chrysomonads I. Arch Protistenk. 1929;67:253–290.
56. Koster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine [Journal Article]. Bioinformatics. 2012;28(19):2520–2. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22908215.
57. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, et al. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products [Journal Article]. J Comput Biol. 2013;20(10):714–37. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24093227.
58. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation [Journal Article]. Genome Res. 2017;27(5):722–736. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28298431.
59. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016 Feb;32(4):605–607.
60. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014 Mar;15(3):R46.
61. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
62. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009 Mar;10(3):R25. Available from: https://and.org/10.1186/gb-2009-10-3-r25.
63. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct;31(19):3210–3212.
64. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006 Jul;34(Web Server issue):W435–439.
65. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005 Jun;21 Suppl 1:i351–358.
66. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar;26(6):841–842.
67. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 09;34(18):3094–3100.
68. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes [Journal Article]. Nucleic Acids Res. 2000 Jan;28(1):27–30.
69. Pundir S, Martin MJ, O'Donovan C. UniProt Protein Knowledgebase. Methods Mol Biol. 2017;1558:41–55.

70. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND [Journal Article]. Nat Methods. 2015;12(1):59–60. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25402007.
71. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013 Jul;29(14):1830–1831.
72. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011 Oct;7:539.
73. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research. 2019 July;47(W1):W636—W641. Available from: http://europepmc.org/articles/PMC6602479.
74. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Research. 2016 04;44(W1):W232–W235. Available from: https://and.org/10.1093/nar/gkw256.
75. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution. 2016 02;33(6):1635–1638. Available from: https://and.org/10.1093/molbev/msw046.

Additional Files

supplement.pdf

The supplement contains among others information about sequencing and binning statistics, ploidy figures, additional examined pathways.

Chapter 4

Discussion

4.1 Determination of intraspecific variation within Chrysophyceae

Determination of intraspecific variation is an essential step before comparing species to get the scope of molecular flexibility. The three *P. lacustris* strains showed differences in ploidy (JBC07 = triploid, JBM10 = diploid, JBNZ41 = tetraploid), mutation rate (higher ploidy correlates with increased mutation rate) and shared genes. On the other hand, there are some stable characteristics like gene density, mitochondrial genome and proportion of repetitive sequences.

The ploidy varies frequently within the Chrysophyceae. This indicates that genome duplication or ploidy reduction occur often. Polyploidization is a mechanism to adapt to niches or invade and colonize niches (Baduel et al., 2018). Some algae use it even seasonally for environmental adaptation (Thornber, 2006). Thereby, the range of ploidy reach from diploid to tetraploid within the investigated chrysophytes. This range must be the optimal trade-off between cost effectiveness (regarding number of nucleotides) and the three characteristics: gene redundancy (advantages e.g. Muller's ratchet (Muller, 1932), transcription and protein synthesis rate; Priest and Priest 1969; Rumeur et al. 1981; Baduel et al. 2018). Nevertheless, species with much higher gene copies e.g. ciliates (Coyne et al., 2011), or predominantly haploid species e.g. dinoflagellates (Taylor, 1987) occupy similar habitats. It can be concluded from this that ploidy is strain specific and at least not directly affected by habitat.

The question arises: What are the characteristics of intraspecific variation? The three *P. lacustris* strains shared 70% of their genes. The core genome of heterotrophic Chrysophyceae is smaller compared to mixotrophic or phototrophic chrysophytes. Therefore, the amount of shared genes cannot be generalized for the whole taxa and is likely higher in mixo- and phototrophs. However, genes related to the primary metabolism are stronger conserved (regarding mutations as well as occurrence in the genome) than unannotated genes. These unannotated genes are probably specific to ecological niche adaptations, which is why they are under stronger selection. Graupner et al. (2017) found 92-93% shared annotated genes within the transcriptome of *P. lacustris* confirming a conserved primary metabolism. Consequently, mostly other factors determine intraspecific variation. Compar-

ing transcriptomic data with genetic data revealed distinctions regarding presence/absence of genes within several chrysophytes. Accordingly, the gene expression differs, which has been shown intraspecifically by two *Ochromonas* spp. strains (Lie et al., 2018). In fact, a pyhlogenetic tree based on transcriptome data is predominantly determined by relation and almost independent of environmental origin (Yang et al., 2017). In other words, the genome defines the transcriptome, which is a more accurate indicator of genetic variation within a species. In conclusion, the main difference of intraspecific variation in Chrysophyceae is ploidy and the variation of unannotated genes and the gene expression level.

There are several mechanisms by which intraspecific variation influences ecology (Bolnick et al., 2011): For example, variation in size leads to a larger range of interactions regarding predation and prey defence, as smaller sizes favour the escape of predators. Hence, a large exemplar of a predator is able to feed from large prey, respectively a small predator feeds from small prey. This size variation results in decreased food competition within a population. Further, high variation enables positive selection without a need for new mutants (Barrett and Schluter, 2008) and increases functional diversity (Bolnick et al., 2011). Large intraspecific variation might even exceed interspecific variation (Weisse, 2002), wherefore it is so important to ecology.

Some parameters that influence genetic variation cannot be clarified. Most microbial eukaryotes are globally distributed (Finlay, 2002). Even if the *P. lacustris* strains were geographically separated (Austria, China and New Zealand), it is not distinguishable whether an exchange or an invasion took place. A gene flow from one strain to another would decrease the variation substantial. On the other hand, large and adapted populations persist small amounts of invaders and lakes behave like isolated islands (Ventura et al., 2014; MacArthur and Wilson, 1967) preserving gene variants. Additionally, the effective population size or population bottlenecks (caused by drastic events like ice age) affect the genetic variation, whereby large populations have increased genetic variation (Kimura, 1979; Lebret et al., 2012). However, there is no record of population history. Hence, estimating the effect of these factors is difficult up to impossible.

4.2 Impact and evolution of nutritional shift

Charles Darwins theory of evolution analogously consists of two processes (Nei, 2013; Darwin, 1859):

- generation of new variations
- natural selection of advantageous variations

One of these variations is the shift of nutrition. While endosymbiosis occurred extremely rarely during the entire evolution (Palmer, 2003; Reyes-Prieto et al., 2007), evolvement towards heterotrophy took place several times independently of each other (Krause, 2012; Kamikawa et al., 2015; Suzuki et al., 2018). Additionally,endosymbiosis is a single event and evolving from phototrophy via mixotrophy to heterotrophy is a slow process. This process is characterized predominantly by neutral mutations (Kimura, 1979) and under selective environmental conditions adaptations emerge (Muller, 1950; Haldane, 1957). A

change of nutrition is likely under food shortage. Also, switching the niche avoids competition for nutrients. The higher GC level in heterotrophic species indicated that nutrients were the limiting factor in phototrophs and mixotrophs, whereas carbon limitation triggered the evolution towards obligate heterotrophy. A decrease of GC content can be achieved by neutral substitutions of nucleotides resulting in synonymous codons. I assume that the common chrysophyte ancestor had a GC content similar to the heterotrophs, since mixotrophs and phototrophs were selected towards lower GC content, while heterotrophic species kept the GC level consistent based on other selection criteria.

The genome and cell size is big in phototrophs and small in heterotrophs presumably due to grazing effectiveness (de Castro et al., 2009). Since larger genomes correlate with greater cell sizes (Vialli, 1957; Baetcke et al., 1967), a heterotrophic strain should decrease its genome content to obtain the most beneficial size. However, mutations are random and there is no intentionally controlled evolution (besides transposons to a lesser extent)(Kimura, 1979; Nei, 2013). Although, the primary metabolism is stronger conserved in *P. lacustris*, the mutations are not directed. This becomes clear from the pan and core genome of the heterotrophic strains: genes were randomly lost, whereas optional genes and pathways were retained. Additionally, the polyploidy increases nuclear genome content. Hence, different aspects underlie size selection such as the plastid, which is strongly reduced in heterotrophs (Grossmann et al., 2016; Dorrell et al., 2019). Additionally, the carbon storage (in the form of amylose or amylopectin) in heterotrophs is low, leading to small starvation resistance (Neilson and Lewin, 1974; Weithoff and Wacker, 2007; Deschamps et al., 2008).

Even if the transcriptome is subject to random evolution (Yang et al., 2017; Khaitovich et al., 2004), changes seem easier to implement. Accordingly, in chrysophytes optional genes were switched off instead of lost. This coincides with the view, that species differences (especially morphology) are mainly caused by development of regulatory elements and secondarily by gene mutations (Ohno, 1972; King and Wilson, 1975). This is analogous to differentiation in metazoa, where cell fate is determined by the transcriptome. Although, the development after endosymbiosis points from photosynthesis towards heterotrophy, the reverse direction would be possible by reactivation of certain genes. On the other hand, the reduction of the plastid seems irreversible.

The evolution from phototrophy to heterotrophy increased biodiversity in protists. This is beneficial for ecosystems by stabilizing the system (Grilli et al., 2017) and very likely increasing the functionality (Schwartz et al., 2000). Furthermore, mixotrophy increases the transfer of biomass to larger size classes of the food chain increasing mean organism size and vertical carbon flux (Ward and Follows, 2016).

4.3 Future studys and recommendations

I have shown that the shift of nutritional mode is related to nutrient and carbon limitation. However, the question remains, why the loss of phototrophy is an advantage at all? Even though obligate heterotrophy is advantageous for several chrysophte lineages, there are original heterotrophic microorganism, which could adapt to this niche much longer. Hence, it would be interesting, which beneficial genes descend from the plastid enabling competitive fitness. Additionally, a comparison with other taxa could support persistence and usefulness of genes. With regard to intraspecific variation, mixo- or phototrophic species could also be investigated to determinate variation. This would create a contrast to the here presented intraspecific variation, which is based on extensively reduced heterotrophs.

Polyploidization could directly lead to speciation (Vries, 1903; Nei, 2013). Therefore, a cross-breeding experiment for *P. lacustris* could clarify if offspring is possible. This can be realised by marking each strain DNA with different fluorescent dyes.

As already described in the introduction, there are many difficulties in bioinformatic processing. Future genome sequencing should be done with axenic strains. Filtering is associated with a loss of many reads, especially as a result of strict precautions to avoid contaminations. This affects also the gene annotation. Additionally, only around 20 % of the genes could be annotated. The unannotated genes cause a big gap in our understanding of organisms regarding functional diversity, ecology and evolution.

4.4 Conclusion

I have shown that the shift of nutrition is related to characteristics as GC content and genome size. This supports the hypothesis about the evolutionary scenario, that mixotrophy is based on nutrient limitation and heterotrophy is based on carbon limitation (de Castro et al., 2009).
Based on the genome comparisons in *P. lacustris* I could reveal that the intraspecific variation is mainly based on ploidy and unannotated genes, which were predominantly associated with secondary metabolism and niche adaptation. Further, I demonstrated random gene reduction between species with different nutritional modes, implying neutral evolution. Accordingly, genes of facultative pathways were switched off instead of lost. These findings indicate a specific selection in microevolution, but mainly neutral evolution in the overall and long-term context of Macroevolution. Hereby, I contribute to the knowledge of how the nutritional switch progressed evolutionarily, which is relevant not only to Chrysophyceae but also for many other phyla in terms of biodiversity and ecology.

Chapter 5

Appendix

Acronyms

μE microeinstein

μL microlitre

18S 18S ribosomal RNA

ANI average nucleotide identity

ANOVA one-way analysis of variance

bp base pairs

BUSCO Benchmarking Universal Single-Copy Orthologs

C Celsius or carbon

CO_2 carbon dioxide

COI cytochrome c oxidase I

DFG Deutsche Forschungsgemeinschaft

dip diploid

DNA deoxyribonucleic acid

FWER family-wise error rate

g gram

GC guanine-cytosine

GO Gene Ontology

h hour

H heterotroph

I Illumina

ID identifier

ITS internal transcribed spacer

K kilo

KB kilo bases

KEGG Kyoto Encyclopedia of Genes and Genomes

KO KEGG Orthology

L litre

M mixotroph

MB or Mb mega bases

Mbp mega base pairs

min minute

mL millilitre

mRNA messanger RNA

N nitrogen

NCBI National Center for Biotechnology Information

nm nanometer

OG orthologous group

OTU operational taxonomic unit

P PacBio, p-value, probability, phototroph or phosphate

PacBio Pacific Biosciences

pg picogram

PI propidium iodide

rRNA ribosomal ribonucleic acid

SNP single nucleotide polymorphism

sp. or spp. species

SSU small subunit

tet tetraploid

tri triploid

Uniprot Universal Protein Ressource

5.1 List of Figures

5.2 List of Tables

5.3 Supplementary files

Supplement to the paper: intraspecific variation in protists: clues for microevolution from *Poteriospumella lacustris* (Chrysophyceae)

Detailed Methods

Gene Prediction

The gene prediction was tested in different ways:
(1) At the first place, Tophat2 (Kim et al., 2013) combined with GeneMark-ET (Borodovsky and Lomsadze, 2011) was applied to train the gene prediction with RNA-Seq data (Beisser et al., 2017). Subsequently, the tool AUGUSTUS (Stanke et al., 2006) was used for gene prediction. To verify the prediction results, we mapped the RNA reads to the predicted genes with Bowtie (v.2.2.8 with parameters: –very-sensitive-local; Langmead et al., 2009). This attempt was discarded, because of the low back mapping rate (65-70%)
(2) To create a species-specific prediction model, we followed the instructions on http://augustus.gobics.de/binaries/retraining.html with RNA-Seq data (Beisser et al., 2017). It was not achievable to create a sufficient *P. lacustris* specific model with the current data. Therefore, the gene pattern of *Arabidopsis thaliana* was chosen as model instead.
(3) AUGUSTUS can improve the prediction model with additional information on expressed sequence tag (EST). Because the transcriptomic reads (Beisser et al., 2017) were not satisfactory, we used known algal sequences. Therefore, algae sequences of Alga-PrAS database (`http://alga-pras.riken.jp/menta.cgi/static/algapras/Alga-PrAS\_Resource.zip;retrieved12/2017`) were aligned with the assembled genome of strain JBM10 by Exonerate (v2.2.0 with parameters: –model protein2genome; Slater and Birney, 2005). The alignment of about 500,000 possibly present protein sequences determined potential introns, which where masked with BEDTools (v2.27, with parameters: maskfasta -bed GFF; Quinlan and Hall, 2010). RNA reads of the

strains (Beisser et al., 2017) were mapped to the masked genome. Thereby, gene occurrence was confirmed and consensus sequences of the alignment were declared as EST. In a trial run a file with the ESTs was used to support the gene prediction of AUGUSTUS.
(4) Finally, we used AUGUSTUS (v3.3 with parameters: –species=arabidopsis –gff3=on –singlestrand=true –UTR=off; Stanke et al., 2006) for gene prediction. Because of the minimal benefit (e.g. 0.16% enhancement of the back mapping rate) of attempt (3) to (4) and the expensive computational calculations we favored (4).
(5) Retrospectively, we compared (4) with an additional approach. The RNA Seq data (Beisser et al., 2017) was aligned with the genome by Minimap2 (2.16-r922; with parameters: -c -L -x splice -G 80K -t 16). The tool RepeatScout (v1.0.5; Price et al., 2005) provided repeat sequences, which were used to mask the genome with BEDTools (v2.28, with parameters: maskfasta -soft). Afterwards, AUGUSTUS (v3.3 with parameters: –softmasking=1 –species=arabidopsis –gff3=on –singlestrand=true –UTR=off –alternatives-from-evidence=false –extrinsicCfgFile –hintsfile) was used to predict the genes. In the extrinsicCfgFile the following relevant weightings were set:
exon hit bonus: 1e+10 (instead 1), nonexonpart malus: 2 (instead 1)
Genes of attempt (4) and (5) were clustered with CD-HIT (v4.7 with parameters: cd-hit-est -c 0.85 -s 0.7; Li and Godzik, 2006). Both methods share 94% of predicted genes (mean identity= 98.9, standard deviation 3.0). Approach (5) predicted 2% less genes. Methods (4) and (5) generated comparable results within the usual variation between gene prediction approaches.

Literature Cited

Beisser D, et al. 2017. Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. PeerJ 5:e2832.

Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using genemark.hmm-E and genemark-ES. Curr Protoc Bioinformatics. 35: 4.6.1–10.

Kim D, et al. 2013. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14(4):R36.

Langmead B, et al. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3):R25.

LiW, Godzik A. 2006. Cd-hit: a fast programfor clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.

Price AL, et al. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):351–358.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6(1):31.

Stanke M, et al. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 34(Web Server issue):435–439.

Table S1: ***K*-mer based genome size estimation**
GenomeScope (version 1.0; with parameters: k = 21, read = 150, max cov. 10,000; http://qb.cshl.edu/genomescope/)

JBC07

property	min	max
Heterozygosity	0.526297%	0.558878%
Genome Haploid Length	73,203,643 bp	73,946,828 bp
Genome Repeat Length	61,182,149 bp	61,803,288 bp
Genome Unique Length	12,021,494 bp	12,143,539 bp
Model Fit	74.7753%	76.031%
Read Error Rate	0.458844%	0.458844%

JBM10

property	min	max
Heterozygosity	1.84751%	1.86612%
Genome Haploid Length	38,945,278 bp	38,994,499 bp
Genome Repeat Length	11,944,582 bp	11,959,679 bp
Genome Unique Length	27,000,695 bp	27,034,821 bp
Model Fit	90.1891%	91.2805%
Read Error Rate	0.487495%	0.487495%

JBNZ41

property	min	max
Heterozygosity	0.395286%	0.406731%
Genome Haploid Length	69,026,950 bp	69,267,407 bp
Genome Repeat Length	51,677,816 bp	51,857,836 bp
Genome Unique Length	17,349,135 bp	17,409,571 bp
Model Fit	85.7327%	90.24%
Read Error Rate	0.450912%	0.450912%

Table S2: **BUSCO genome completeness check.** The data sets for eukaryotes and protist were used to check the genome integrity. Additionaly, the genomes of *Nannochloropsis oceanica* and *Ectocarpus siliculosus* (from NCBI) were used as comparison.

BUSCO set	*P. lacustris* (JBC07)	*P. lacustris* (JBM10)	*P. lacustris* (JBNZ41)	*Nannochloropsis oceanica*	*Ectocarpus siliculosus*
EUK set (303 BUSCO groups)					
Complete [%]	79.2	79.9	75.2	78.9	77.6
Complete and single-copy [%]	68.6	70.0	70.6	76.9	76.6
Complete and duplicated [%]	10.6	9.9	4.6	2.0	1.0
Fragmented [%]	4.6	3.6	6.3	5.5	6.9
Missing [%]	16.2	16.5	18.5	15.8	15.5
protist set (215 BUSCO groups)					
Complete [%]	55.8	56.3	53.5	63.3	68.9
Complete and single-copy [%]	45.6	51.2	51.2	63.3	68.4
Complete and duplicated [%]	10.2	5.1	2.3	0.0	0.5
Fragmented [%]	0.0	0.0	0.9	0.9	0.9
Missing [%]	44.2	43.7	45.6	35.8	30.2

Table S3: **Identical genes.** Number of genes that are 100% identical in at least one allele between two strains. Only functional groups containing more than 40 genes are listed. JBC07 and JBNZ41 share about twice as many identical genes as they share with JBM10. The group of genetic information processing and metabolism, mainly enzymes, are highly conserved between all strains.

KEGG functional hierarchy L1	JBC07 JBM10	JBC07 JBNZ41	JBM10 JBNZ41
All genes	584	1041	430
Unannotated genes	356	717	242
Amino acid metabolism	6	8	13
Biosynthesis of other secondary metabolites	0	1	0
Carbohydrate metabolism	6	3	3
Energy metabolism	2	10	6
Genetic information processing	92	126	53
Lipid metabolism	2	1	1
Metabolism	52	52	39
Metabolism of cofactors and vitamins	8	7	7
Metabolism of other amino acids	4	9	4
Nucleotide metabolism	0	4	1
Signaling and cellular processes	38	67	33

Table S4: **Ploidy estimation for each contig.** Quantity of contigs assigned with ploidy level and normalized by contig length. Only contigs longer than 10 kbp were counted and used, if ploidy assessment value differ at least 10% from the other ploidy estimations. Recent tetraploidy could also be assigned to diploidy in contigs with insufficient diverging SNPs.

Ploidy	JBC07	JBM10	JBNZ41
diploid	84.0	**184.1**	**172.3**
triploid	**97.2**	41.1	81.3
tetraploid	33.6	48.7	48.1
percentage used contigs	65	79	78

Table S5: **Ploidy estimation of the nQuire tool.** Columns free, dip (diploid), tri (triploid) and tet (tetraploid) show log values of model probabilities. Rows starting with "d_" indicate the divergence to the real model. The model with the smallest deviation determine the ploidy level."

file	JBC07_denoised.bin	JBM10_denoised.bin	JBNZ41_denoised.bin
free	253,438.79	567,363.69	308,681.32
dip	73,869.62	537,303.73	292,502.65
tri	204,201.79	167,745.84	123,628.44
tet	78,299.84	330,984.04	164,817.24
d_dip	179,569.18	**30,059.95**	**16,178.66**
d_tri	**49,237.01**	399,617.85	185,052.88
d_tet	175,138.95	236,379.65	143,864.08

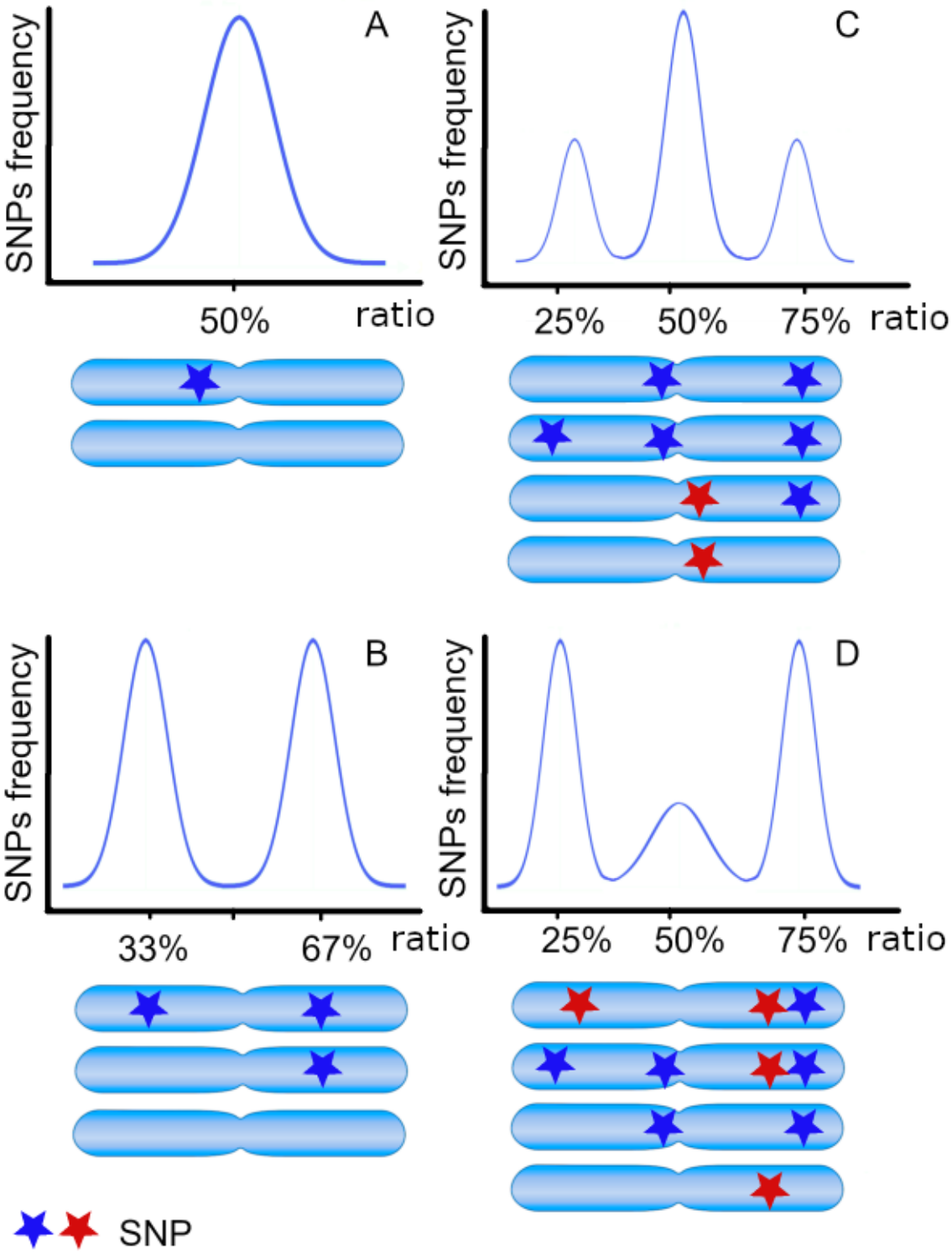


Figure S1 Schematic of SNPs based ploidy estimation. The ratio of alleles with the SNPs (red and blue stars) to alleles without indicates the ploidy level. The *P. lacustris* strains showed distributions associated with diploidy (**A**, JBM10), triploidy (**B**, JBC07) and recent tetraploidy (**C**, JBNZ41), whereas ancient tetraploidy (**D**) was not found. A recent tetraploidy organism has a shift of the majority in SNP frequencies to 50% as it arises from the doubling of a diploid genome. The red stars indicate the majority ratios.

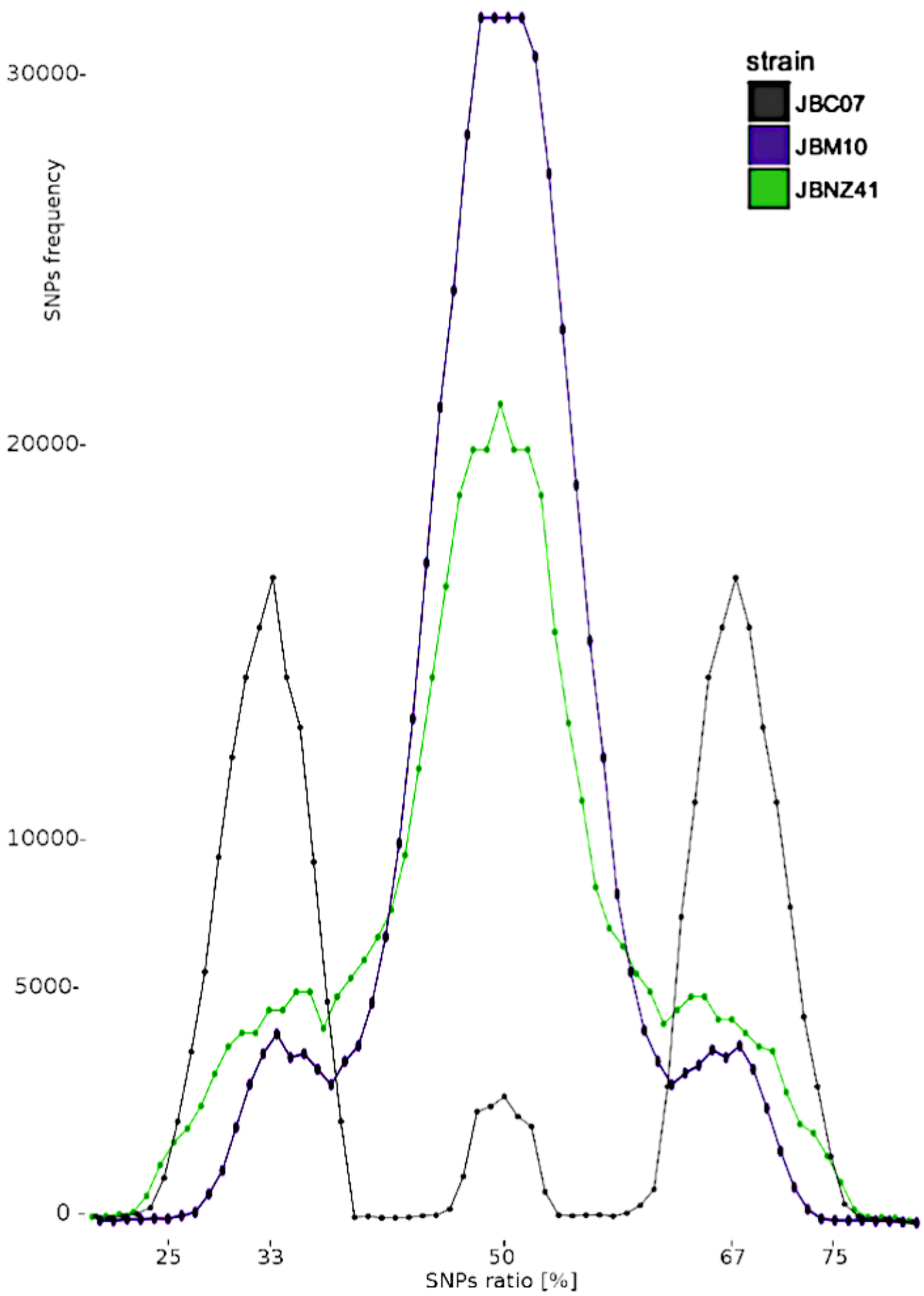


Figure S2 Ploidy estimation based on SNPs. The ratio of SNPs correlates with the ploidy. The determined ploidy levels of the strains were diploidy (JBM10), triploidy (JBC07) and diploidy or tetraploidy (JBNZ41).

Table S6, S7 and S8: Strain specific Wilcoxon signed-rank tests on gene variation within functional groups validate pairwise differences. P-values (< 0.01) were bold.

JBC07

	All genes	Unannotated genes	Amino acid metabolism	Biosynthesis of other secondary metabolites	Carbohydrate metabolism	Energy metabolism	Genetic information processing	Lipid metabolism	Metabolism	Metabolism of cofactors and vitamins	Metabolism of other amino acids	Nucleotide metabolism	Organelle targeting genes
Unannotated genes	8.8E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Amino acid metabolism	5.1E-01	5.4E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Biosynthesis of other secondary metabolites	4.6E-01	4.6E-01	2.3E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Carbohydrate metabolism	5.1E-01	5.4E-01	9.5E-01	2.9E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA
Energy metabolism	4.9E-01	5.1E-01	9.4E-01	2.3E-01	9.5E-01	NA	NA	NA	NA	NA	NA	NA	NA
Genetic information processing	**1.1E-04**	**7.7E-05**	1.6E-01	5.1E-01	2.3E-01	2.3E-01	NA	NA	NA	NA	NA	NA	NA
Lipid metabolism	8.7E-01	8.7E-01	4.9E-01	5.4E-01	5.1E-01	4.6E-01	9.5E-01	NA	NA	NA	NA	NA	NA
Metabolism	4.8E-01	5.1E-01	9.1E-01	2.3E-01	8.8E-01	8.4E-01	**3.6E-03**	5.1E-01	NA	NA	NA	NA	NA
Metabolism of cofactors and vitamins	6.8E-01	7.0E-01	9.5E-01	3.0E-01	9.3E-01	8.8E-01	2.4E-01	5.5E-01	9.5E-01	NA	NA	NA	NA
Metabolism of other amino acids	9.3E-01	9.3E-01	9.5E-01	4.6E-01	9.3E-01	8.8E-01	5.4E-01	8.4E-01	9.9E-01	9.5E-01	NA	NA	NA
Nucleotide metabolism	8.8E-01	8.8E-01	9.5E-01	4.4E-01	9.3E-01	9.3E-01	5.1E-01	7.0E-01	9.5E-01	9.7E-01	9.5E-01	NA	NA
Organelle targeting genes	2.3E-01	2.4E-01	9.5E-01	2.3E-01	9.3E-01	8.8E-01	**7.7E-05**	4.9E-01	8.8E-01	9.5E-01	9.5E-01	9.5E-01	NA
Signaling and cellular processes	9.1E-01	8.8E-01	4.9E-01	4.6E-01	5.1E-01	4.9E-01	2.3E-01	8.8E-01	4.6E-01	6.2E-01	8.8E-01	8.8E-01	2.3E-01

Table S6

JBM10

	All genes	Unannotated genes	Amino acid metabolism	Biosynthesis of other secondary metabolites	Carbohydrate metabolism	Energy metabolism	Genetic information processing	Lipid metabolism	Metabolism	Metabolism of cofactors and vitamins	Metabolism of other amino acids	Nucleotide metabolism	Organelle targeting genes
Unannotated genes	**8.9E-04**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Amino acid metabolism	3.8E-01	1.7E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Biosynthesis of other secondary metabolites	6.2E-01	4.6E-01	8.7E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Carbohydrate metabolism	8.7E-01	6.2E-01	7.3E-01	7.3E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA
Energy metabolism	4.1E-02	1.1E-01	2.3E-02	1.0E-01	1.1E-01	NA	NA	NA	NA	NA	NA	NA	NA
Genetic information processing	**1.3E-14**	**9.8E-22**	3.3E-01	7.8E-01	2.2E-01	**5.5E-05**	NA	NA	NA	NA	NA	NA	NA
Lipid metabolism	7.3E-01	9.9E-01	3.1E-01	3.3E-01	6.2E-01	2.9E-01	7.5E-02	NA	NA	NA	NA	NA	NA
Metabolism	7.0E-01	2.1E-01	6.5E-01	7.2E-01	9.9E-01	3.0E-02	**3.1E-03**	5.4E-01	NA	NA	NA	NA	NA
Metabolism of cofactors and vitamins	4.9E-01	2.2E-01	8.9E-01	8.1E-01	8.1E-01	3.0E-02	2.7E-01	4.0E-01	7.3E-01	NA	NA	NA	NA
Metabolism of other amino acids	4.6E-01	6.2E-01	2.9E-01	3.2E-01	4.6E-01	8.0E-01	1.0E-01	6.5E-01	4.0E-01	2.9E-01	NA	NA	NA
Nucleotide metabolism	5.2E-02	3.0E-02	1.1E-01	2.5E-01	1.1E-01	**2.1E-03**	2.5E-01	3.0E-02	7.6E-02	1.0E-01	6.4E-02	NA	NA
Organelle targeting genes	2.3E-01	9.2E-01	1.6E-01	3.8E-01	5.5E-01	1.6E-01	**1.7E-08**	9.7E-01	2.2E-01	2.2E-01	6.3E-01	2.3E-02	NA
Signaling and cellular processes	5.7E-03	**4.6E-05**	8.0E-01	9.9E-01	5.5E-01	**2.1E-03**	1.7E-01	2.2E-01	2.2E-01	7.2E-01	2.2E-01	1.6E-01	**2.1E-03**

Table S7

JBNZ41

	All genes	Unannotated genes	Amino acid metabolism	Biosynthesis of other secondary metabolites	Carbohydrate metabolism	Energy metabolism	Genetic information processing	Lipid metabolism	Metabolism	Metabolism of cofactors and vitamins	Metabolism of other amino acids	Nucleotide metabolism	Organelle targeting genes
Unannotated genes	**3.2E-03**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Amino acid metabolism	3.5E-02	9.5E-03	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Biosynthesis of other secondary metabolites	4.8E-01	3.4E-01	8.1E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Carbohydrate metabolism	**9.0E-06**	**1.2E-06**	5.5E-02	1.1E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA
Energy metabolism	**4.2E-04**	**1.0E-04**	1.6E-01	2.2E-01	7.0E-01	NA	NA	NA	NA	NA	NA	NA	NA
Genetic information processing	1.0E-02	3.8E-01	**4.0E-03**	2.6E-01	**2.9E-07**	**2.0E-05**	NA	NA	NA	NA	NA	NA	NA
Lipid metabolism	6.8E-01	4.9E-01	6.3E-01	8.2E-01	2.5E-02	8.3E-02	3.4E-01	NA	NA	NA	NA	NA	NA
Metabolism	1.0E-02	**5.3E-04**	5.5E-01	9.4E-01	**3.1E-03**	2.2E-02	**1.5E-04**	8.4E-01	NA	NA	NA	NA	NA
Metabolism of cofactors and vitamins	1.2E-02	**3.4E-03**	7.3E-01	6.3E-01	1.3E-01	3.4E-01	**1.2E-03**	4.1E-01	3.0E-01	NA	NA	NA	NA
Metabolism of other amino acids	2.8E-01	2.0E-01	9.3E-01	7.6E-01	2.6E-01	4.2E-01	1.4E-01	6.3E-01	6.8E-01	9.3E-01	NA	NA	NA
Nucleotide metabolism	1.0E-02	**4.9E-03**	2.6E-01	2.4E-01	8.7E-01	9.3E-01	**3.1E-03**	1.4E-01	9.7E-02	4.1E-01	4.1E-01	NA	NA
Organelle targeting genes	2.5E-01	3.0E-02	2.2E-01	6.9E-01	**1.5E-04**	**3.4E-03**	6.5E-03	9.3E-01	3.6E-01	9.7E-02	4.1E-01	2.6E-02	NA
Signaling and cellular processes	**4.8E-08**	**6.7E-11**	9.1E-01	7.3E-01	1.9E-02	1.1E-01	**8.4E-11**	4.9E-01	2.1E-01	7.6E-01	9.4E-01	2.5E-01	1.0E-02

Table S8

Table S9, S10 and S11: Pairwise Wilcoxon signed-rank tests on gene variation within functional groups between two strains validate differences. P-values (< 0.01) were bold.

JBC07 - JBM10

	All genes	Unannotated genes	Amino acid metabolism	Biosynthesis of other secondary metabolites	Carbohydrate metabolism	Energy metabolism	Genetic information processing	Lipid metabolism	Metabolism	Metabolism of cofactors and vitamins	Metabolism of other amino acids	Nucleotide metabolism	Organelle targeting genes
Unannotated genes	**2.0E-29**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Amino acid metabolism	**8.0E-06**	**4.4E-11**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Biosynthesis of other secondary metabolites	3.2E-02	**2.1E-03**	6.8E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Carbohydrate metabolism	8.3E-01	1.4E-01	3.4E-02	7.4E-02	NA	NA	NA	NA	NA	NA	NA	NA	NA
Energy metabolism	3.6E-01	3.6E-02	1.3E-01	9.8E-02	6.8E-01	NA	NA	NA	NA	NA	NA	NA	NA
Genetic information processing	**1.4E-34**	**2.2E-75**	4.2E-01	5.1E-01	4.7E-02	2.1E-01	NA	NA	NA	NA	NA	NA	NA
Lipid metabolism	5.6E-01	1.2E-01	9.8E-02	1.3E-01	8.3E-01	8.9E-01	1.8E-01	NA	NA	NA	NA	NA	NA
Metabolism	**6.5E-14**	**1.3E-30**	4.3E-01	5.3E-01	3.9E-02	2.1E-01	8.3E-01	1.8E-01	NA	NA	NA	NA	NA
Metabolism of cofactors and vitamins	2.0E-01	**2.2E-03**	7.1E-02	1.3E-01	5.6E-01	1.0E+00	1.3E-01	8.3E-01	1.3E-01	NA	NA	NA	NA
Metabolism of other amino acids	1.4E-02	**6.1E-04**	7.1E-01	8.3E-01	7.1E-02	1.6E-01	4.5E-01	1.3E-01	5.0E-01	1.3E-01	NA	NA	NA
Nucleotide metabolism	**8.0E-06**	**3.9E-08**	3.8E-02	2.8E-01	**8.0E-04**	**2.1E-03**	1.4E-02	**2.1E-03**	1.1E-02	**1.0E-03**	2.4E-01	NA	NA
Organelle targeting genes	2.6E-01	1.8E-02	1.4E-01	2.8E-01	5.1E-01	8.9E-01	2.8E-01	7.8E-01	2.3E-01	8.9E-01	2.2E-01	6.7E-03	NA
Signaling and cellular processes	**8.2E-11**	**6.7E-29**	1.3E-01	2.6E-01	1.4E-01	4.4E-01	1.7E-01	3.6E-01	2.2E-01	3.4E-01	2.3E-01	**2.6E-03**	5.6E-01

Table S9

JBC07 - JBNZ41

	All genes	Unannotated genes	Amino acid metabolism	Biosynthesis of other secondary metabolites	Carbohydrate metabolism	Energy metabolism	Genetic information processing	Lipid metabolism	Metabolism	Metabolism of cofactors and vitamins	Metabolism of other amino acids	Nucleotide metabolism	Organelle targeting genes
Unannotated genes	**5.0E-08**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Amino acid metabolism	7.6E-02	**3.4E-03**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Biosynthesis of other secondary metabolites	9.6E-01	9.1E-01	4.4E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Carbohydrate metabolism	4.2E-01	1.6E-01	9.6E-01	6.8E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA
Energy metabolism	4.2E-01	1.5E-01	9.2E-01	6.8E-01	9.7E-01	NA	NA	NA	NA	NA	NA	NA	NA
Genetic information processing	**4.9E-16**	**1.3E-28**	9.2E-01	4.2E-01	9.1E-01	8.7E-01	NA	NA	NA	NA	NA	NA	NA
Lipid metabolism	8.5E-01	4.6E-01	6.8E-01	8.7E-01	8.5E-01	9.2E-01	5.8E-01	NA	NA	NA	NA	NA	NA
Metabolism	1.7E-02	**5.3E-06**	6.8E-01	6.8E-01	9.2E-01	9.6E-01	1.6E-01	9.1E-01	NA	NA	NA	NA	NA
Metabolism of cofactors and vitamins	3.5E-01	7.5E-02	9.1E-01	6.8E-01	9.6E-01	9.6E-01	6.9E-01	9.1E-01	9.4E-01	NA	NA	NA	NA
Metabolism of other amino acids	9.6E-02	3.0E-02	6.0E-01	3.0E-01	6.0E-01	6.0E-01	6.0E-01	3.9E-01	3.5E-01	4.6E-01	NA	NA	NA
Nucleotide metabolism	4.7E-01	2.4E-01	9.7E-01	6.7E-01	9.6E-01	9.2E-01	9.6E-01	6.8E-01	9.1E-01	9.4E-01	7.6E-01	NA	NA
Organelle targeting genes	2.5E-01	9.6E-01	8.9E-03	9.1E-01	1.6E-01	1.5E-01	**9.4E-07**	4.2E-01	**3.4E-03**	9.6E-02	3.2E-02	2.1E-01	NA
Signaling and cellular processes	**3.4E-03**	**7.9E-08**	7.6E-01	6.8E-01	9.4E-01	9.5E-01	1.6E-01	9.1E-01	9.6E-01	9.6E-01	3.5E-01	9.1E-01	**3.4E-03**

Table S10

JBM10 - JBNZ41

	All genes	Unannotated genes	Amino acid metabolism	Biosynthesis of other secondary metabolites	Carbohydrate metabolism	Energy metabolism	Genetic information processing	Lipid metabolism	Metabolism	Metabolism of cofactors and vitamins	Metabolism of other amino acids	Nucleotide metabolism	Organelle targeting genes
Unannotated genes	**2.7E-16**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Amino acid metabolism	**2.3E-03**	**1.7E-06**	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Biosynthesis of other secondary metabolites	5.5E-01	2.0E-01	7.8E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Carbohydrate metabolism	1.2E-01	7.3E-03	8.7E-01	8.4E-01	NA	NA	NA	NA	NA	NA	NA	NA	NA
Energy metabolism	7.8E-01	2.4E-01	2.4E-01	6.6E-01	4.4E-01	NA	NA	NA	NA	NA	NA	NA	NA
Genetic information processing	**1.9E-19**	**2.3E-41**	7.9E-01	8.4E-01	9.9E-01	2.4E-01	NA	NA	NA	NA	NA	NA	NA
Lipid metabolism	3.1E-01	5.3E-02	8.4E-01	9.9E-01	8.5E-01	6.3E-01	8.5E-01	NA	NA	NA	NA	NA	NA
Metabolism	**3.2E-06**	**2.1E-14**	6.8E-01	8.9E-01	9.0E-01	3.6E-01	7.9E-01	9.8E-01	NA	NA	NA	NA	NA
Metabolism of cofactors and vitamins	5.6E-01	5.0E-02	2.4E-01	7.7E-01	5.3E-01	9.0E-01	2.4E-01	6.3E-01	3.9E-01	NA	NA	NA	NA
Metabolism of other amino acids	1.4E-01	2.6E-02	8.5E-01	6.6E-01	7.2E-01	3.6E-01	6.8E-01	6.8E-01	6.8E-01	3.6E-01	NA	NA	NA
Nucleotide metabolism	8.8E-02	1.4E-02	7.8E-01	6.5E-01	7.0E-01	2.4E-01	6.6E-01	6.6E-01	6.2E-01	2.4E-01	9.9E-01	NA	NA
Organelle targeting genes	9.0E-01	5.3E-02	1.7E-02	6.3E-01	2.0E-01	8.4E-01	**1.0E-03**	3.8E-01	1.4E-02	6.3E-01	2.0E-01	1.3E-01	NA
Signaling and cellular processes	**2.9E-07**	**8.6E-18**	6.2E-01	9.5E-01	8.4E-01	4.1E-01	6.1E-01	9.9E-01	8.7E-01	4.5E-01	6.2E-01	5.7E-01	1.8E-02

Table S11

Table S12: ANOVA test for intraspecific mutation distribution.

P-values < 0.01 are marked bold. The mutation rates of each strain differ in the groups *all genes, unannotated genes* an *signaling and cellular processes* in the amount of SNPs, inserts and deletions per gene. In the groups *genetic information processing* and *organelle targeted genes* only the proportion of SNPs differ signifcantly.

Metabolic_pathway	SNPs	inserts	deletions
All genes	**1.1E-71**	**2.6E-16**	**2.5E-16**
Unannotated genes	**2.9E-88**	**1.5E-20**	**9.0E-19**
Amino acid metabolism	8.5E-01	9.1E-01	8.1E-01
Biosynthesis of other secondary metabolites	2.0E-01	1.4E-01	6.7E-02
Carbohydrate metabolism	1.7E-02	4.5E-01	1.1E-02
Energy metabolism	1.1E-01	2.5E-01	2.4E-01
Genetic information processing	**4.9E-04**	4.4E-01	1.6E-01
Lipid metabolism	8.4E-02	7.2E-01	7.1E-01
Metabolism	3.0E-01	5.7E-01	3.8E-01
Metabolism of cofactors and vitamins	4.8E-01	6.8E-01	3.0E-01
Metabolism of other amino acids	4.9E-01	9.3E-01	9.2E-01
Nucleotide metabolism	4.6E-02	1.8E-01	1.2E-01
Organelle targeted genes	**2.1E-03**	2.4E-02	6.8E-02
Signaling and cellular processes	**2.6E-04**	**4.9E-04**	**4.1E-04**

Table S13:Kruskal-Wallis test for interspecific mutation distribution.

P-values < 0.01 are marked bold. The mutation rates of each strain are consistent in the groups *Biosynthesis of other secondary metabolites, Metabolism of other amino acids* and *Nucleotide metabolism* in the amount of SNPs, inserts and deletions per gene.

Metabolic_pathway	SNPs	inserts	deletions
All genes	**0.0E+00**	**1.7E-104**	**8.9E-134**
Unannotated genes	**3.6E-256**	**1.8E-61**	**2.2E-73**
Amino acid metabolism	**7.9E-07**	**4.7E-03**	**3.3E-04**
Biosynthesis of other secondary metabolites	4.8E-01	3.1E-01	9.6E-01
Carbohydrate metabolism	**9.1E-06**	9.9E-01	**1.2E-03**
Energy metabolism	**1.1E-04**	1.3E-02	5.5E-02
Genetic information processing	**2.8E-85**	**2.1E-22**	**2.9E-30**
Lipid metabolism	**4.3E-03**	6.6E-02	**1.1E-03**
Metabolism	**5.4E-19**	**1.4E-06**	**2.8E-07**
Metabolism of cofactors and vitamins	**1.7E-07**	1.9E-02	**3.2E-05**
Metabolism of other amino acids	**1.0E-03**	5.6E-02	9.3E-01
Nucleotide metabolism	2.8E-01	2.5E-01	3.1E-01
Organelle targeted genes	**9.2E-04**	8.6E-02	6.5E-02
Signaling and cellular processes	**6.5E-32**	**4.5E-05**	**2.5E-11**

Table S14: The KEGG Mapper - Reconstruct Module confirmed metabolic pathways in which max one gene is missing.

ID	Pathway hierarchy	L1	L2	L3	strain		
M00009	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Citrate cycle (TCA cycle, Krebs cycle)	JBC07	JBM10	JBNZ41
M00010	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Citrate cycle, first carbon oxidation, oxaloacetate => 2-oxoglutarate	JBC07	JBM10	JBNZ41
M00011	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Citrate cycle, second carbon oxidation, 2-oxoglutarate => oxaloacetate	JBC07	JBM10	JBNZ41
M00008	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Entner-Doudoroff pathway, glucose-6P => glyceraldehyde-3P + pyruvate			JBNZ41
M00003	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Gluconeogenesis, oxaloacetate => fructose-6P	JBC07	JBM10	JBNZ41
M00001	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate	JBC07	JBM10	JBNZ41
M00002	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Glycolysis, core module involving three-carbon compounds	JBC07	JBM10	JBNZ41
M00004	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Pentose phosphate pathway (Pentose phosphate cycle)			JBNZ41
M00580	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Pentose phosphate pathway, archaea, fructose 6P => ribose 5P	JBC07	JBM10	JBNZ41
M00007	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Pentose phosphate pathway, non-oxidative phase, fructose 6P => ribose 5P	JBC07	JBM10	JBNZ41
M00006	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Pentose phosphate pathway, oxidative phase, glucose 6P => ribulose 5P			JBNZ41
M00005	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	PRPP biosynthesis, ribose 5P => PRPP	JBC07	JBM10	JBNZ41
M00307	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Pyruvate oxidation, pyruvate => acetyl-CoA	JBC07	JBM10	JBNZ41
M00308	Carbohydrate and lipid metabolism		Central carbohydrate metabolism	Semi-phosphorylative Entner-Doudoroff pathway, gluconate => glycerate-3P	JBC07		
M00087	Carbohydrate and lipid metabolism		Fatty acid metabolism	beta-Oxidation	JBC07	JBM10	JBNZ41
M00086	Carbohydrate and lipid metabolism		Fatty acid metabolism	beta-Oxidation, acyl-CoA synthesis	JBC07	JBM10	JBNZ41
M00083	Carbohydrate and lipid metabolism		Fatty acid metabolism	Fatty acid biosynthesis, elongation	JBC07	JBM10	JBNZ41
M00415	Carbohydrate and lipid metabolism		Fatty acid metabolism	Fatty acid biosynthesis, elongation, endoplasmic reticulum	JBC07	JBM10	JBNZ41
M00085	Carbohydrate and lipid metabolism		Fatty acid metabolism	Fatty acid biosynthesis, elongation, mitochondria	JBC07	JBM10	JBNZ41
M00082	Carbohydrate and lipid metabolism		Fatty acid metabolism	Fatty acid biosynthesis, initiation	JBC07	JBM10	JBNZ41
M00088	Carbohydrate and lipid metabolism		Fatty acid metabolism	Ketone body biosynthesis, acetyl-CoA => acetoacetate/3-hydroxybutyrate/acetone	JBC07	JBM10	JBNZ41
M00055	Carbohydrate and lipid metabolism		Glycan metabolism	N-glycan precursor biosynthesis	JBC07	JBM10	JBNZ41
M00073	Carbohydrate and lipid metabolism		Glycan metabolism	N-glycan precursor trimming	JBC07	JBM10	JBNZ41
M00072	Carbohydrate and lipid metabolism		Glycan metabolism	N-glycosylation by oligosaccharyltransferase	JBC07	JBM10	JBNZ41
M00098	Carbohydrate and lipid metabolism		Lipid metabolism	Acylglycerol degradation	JBC07	JBM10	JBNZ41
M00094	Carbohydrate and lipid metabolism		Lipid metabolism	Ceramide biosynthesis	JBC07	JBM10	JBNZ41
M00130	Carbohydrate and lipid metabolism		Lipid metabolism	Inositol phosphate metabolism, PI=> PIP2 => Ins(1,4,5)P3 => Ins(1,3,4,5)P4	JBC07	JBM10	JBNZ41
M00090	Carbohydrate and lipid metabolism		Lipid metabolism	Phosphatidylcholine (PC) biosynthesis, choline => PC	JBC07	JBM10	JBNZ41
M00092	Carbohydrate and lipid metabolism		Lipid metabolism	Phosphatidylethanolamine (PE) biosynthesis, ethanolamine => PE	JBC07	JBM10	JBNZ41
M00093	Carbohydrate and lipid metabolism		Lipid metabolism	Phosphatidylethanolamine (PE) biosynthesis, PA => PS => PE			JBNZ41
M00099	Carbohydrate and lipid metabolism		Lipid metabolism	Sphingosine biosynthesis	JBC07	JBM10	JBNZ41
M00100	Carbohydrate and lipid metabolism		Lipid metabolism	Sphingosine degradation	JBC07	JBM10	JBNZ41
M00089	Carbohydrate and lipid metabolism		Lipid metabolism	Triacylglycerol biosynthesis	JBC07	JBM10	JBNZ41
M00064	Carbohydrate and lipid metabolism		Lipopolysaccharide metabolism	ADP-L-glycero-D-manno-heptose biosynthesis			JBNZ41
M00063	Carbohydrate and lipid metabolism		Lipopolysaccharide metabolism	CMP-KDO biosynthesis			JBNZ41
M00793	Carbohydrate and lipid metabolism		Other carbohydrate metabolism	dTDP-L-rhamnose biosynthesis			JBNZ41
M00632	Carbohydrate and lipid metabolism		Other carbohydrate metabolism	Galactose degradation, Leloir pathway, galactose => alpha-D-glucose-1P	JBC07	JBM10	JBNZ41
M00012	Carbohydrate and lipid metabolism		Other carbohydrate metabolism	Glyoxylate cycle	JBC07	JBM10	JBNZ41
M00013	Carbohydrate and lipid metabolism		Other carbohydrate metabolism	Malonate semialdehyde pathway, propanoyl-CoA => acetyl-CoA	JBC07	JBM10	JBNZ41
M00554	Carbohydrate and lipid metabolism		Other carbohydrate metabolism	Nucleotide sugar biosynthesis, galactose => UDP-galactose			JBNZ41
M00549	Carbohydrate and lipid metabolism		Other carbohydrate metabolism	Nucleotide sugar biosynthesis, glucose => UDP-glucose	JBC07	JBM10	JBNZ41
M00741	Carbohydrate and lipid metabolism		Other carbohydrate metabolism	Propanoyl-CoA metabolism, propanoyl-CoA => succinyl-CoA	JBC07	JBM10	JBNZ41
M00097	Carbohydrate and lipid metabolism		Other terpenoid biosynthesis	beta-Carotene biosynthesis, GGAP => beta-carotene	JBC07	JBM10	JBNZ41
M00364	Carbohydrate and lipid metabolism		Terpenoid backbone biosynthesis	C10-C20 isoprenoid biosynthesis, bacteria	JBC07	JBM10	JBNZ41
M00367	Carbohydrate and lipid metabolism		Terpenoid backbone biosynthesis	C10-C20 isoprenoid biosynthesis, non-plant eukaryotes	JBC07	JBM10	JBNZ41
M00366	Carbohydrate and lipid metabolism		Terpenoid backbone biosynthesis	C10-C20 isoprenoid biosynthesis, plants	JBC07	JBM10	JBNZ41
M00095	Carbohydrate and lipid metabolism		Terpenoid backbone biosynthesis	C5 isoprenoid biosynthesis, mevalonate pathway	JBC07	JBM10	JBNZ41
M00695	Cellular processes		Cell signaling	cAMP signaling	JBC07	JBM10	JBNZ41
M00692	Cellular processes		Cell signaling	Cell cycle - G1/S transition	JBC07	JBM10	JBNZ41
M00693	Cellular processes		Cell signaling	Cell cycle - G2/M transition	JBC07	JBM10	JBNZ41
M00694	Cellular processes		Cell signaling	cGMP signaling	JBC07	JBM10	JBNZ41
M00691	Cellular processes		Cell signaling	DNA damage-induced cell cycle checkpoints	JBC07	JBM10	JBNZ41
M00687	Cellular processes		Cell signaling	MAPK (ERK1/2) signaling	JBC07	JBM10	JBNZ41
M00689	Cellular processes		Cell signaling	MAPK (p38) signaling	JBC07	JBM10	JBNZ41
M00152	Energy metabolism		ATP synthesis	Cytochrome bc1 complex	JBC07	JBM10	JBNZ41
M00151	Energy metabolism		ATP synthesis	Cytochrome bc1 complex respiratory unit	JBC07	JBM10	JBNZ41
M00153	Energy metabolism		ATP synthesis	Cytochrome bd ubiquinol oxidase		JBM10	
M00154	Energy metabolism		ATP synthesis	Cytochrome c oxidase	JBC07	JBM10	JBNZ41
M00156	Energy metabolism		ATP synthesis	Cytochrome c oxidase, cbb3-type	JBC07	JBM10	JBNZ41
M00155	Energy metabolism		ATP synthesis	Cytochrome c oxidase, prokaryotes	JBC07		JBNZ41
M00158	Energy metabolism		ATP synthesis	F-type ATPase, eukaryotes	JBC07	JBM10	JBNZ41
M00157	Energy metabolism		ATP synthesis	F-type ATPase, prokaryotes and chloroplasts	JBC07	JBM10	JBNZ41
M00145	Energy metabolism		ATP synthesis	NAD(P)H:quinone oxidoreductase, chloroplasts and cyanobacteria	JBC07		JBNZ41
M00146	Energy metabolism		ATP synthesis	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex	JBC07	JBM10	JBNZ41
M00147	Energy metabolism		ATP synthesis	NADH dehydrogenase (ubiquinone) 1 beta subcomplex	JBC07	JBM10	JBNZ41

M00143	Energy metabolism	ATP synthesis	NADH dehydrogenase (ubiquinone) Fe-S protein/flavoprotein complex, mitochondria	JBC07	JBM10	JBNZ41
M00144	Energy metabolism	ATP synthesis	NADH:quinone oxidoreductase, prokaryotes	JBC07	JBM10	JBNZ41
M00148	Energy metabolism	ATP synthesis	Succinate dehydrogenase (ubiquinone)	JBC07	JBM10	JBNZ41
M00149	Energy metabolism	ATP synthesis	Succinate dehydrogenase, prokaryotes	JBC07		JBNZ41
M00160	Energy metabolism	ATP synthesis	V-type ATPase, eukaryotes	JBC07	JBM10	JBNZ41
M00171	Energy metabolism	Carbon fixation	C4-dicarboxylic acid cycle, NAD - malic enzyme type	JBC07	JBM10	JBNZ41
M00172	Energy metabolism	Carbon fixation	C4-dicarboxylic acid cycle, NADP - malic enzyme type	JBC07	JBM10	JBNZ41
M00170	Energy metabolism	Carbon fixation	C4-dicarboxylic acid cycle, phosphoenolpyruvate carboxykinase type	JBC07	JBM10	JBNZ41
M00168	Energy metabolism	Carbon fixation	CAM (Crassulacean acid metabolism), dark	JBC07	JBM10	JBNZ41
M00169	Energy metabolism	Carbon fixation	CAM (Crassulacean acid metabolism), light	JBC07	JBM10	JBNZ41
M00579	Energy metabolism	Carbon fixation	Phosphate acetyltransferase-acetate kinase pathway, acetyl-CoA => acetate	JBC07	JBM10	JBNZ41
M00165	Energy metabolism	Carbon fixation	Reductive pentose phosphate cycle (Calvin cycle)	JBC07	JBM10	JBNZ41
M00167	Energy metabolism	Carbon fixation	Reductive pentose phosphate cycle, glyceraldehyde-3P => ribulose-5P	JBC07	JBM10	JBNZ41
M00166	Energy metabolism	Carbon fixation	Reductive pentose phosphate cycle, ribulose-5P => glyceraldehyde-3P	JBC07	JBM10	JBNZ41
M00597	Energy metabolism	Photosynthesis	Anoxygenic photosystem II		JBM10	JBNZ41
M00161	Energy metabolism	Photosynthesis	Photosystem II	JBC07	JBM10	JBNZ41
M00176	Energy metabolism	Sulfur metabolism	Assimilatory sulfate reduction, sulfate => H2S	JBC07	JBM10	JBNZ41
M00254	Environmental information processing	ABC-2 type and other transport systems	ABC-2 type transport system	JBC07	JBM10	JBNZ41
M00256	Environmental information processing	ABC-2 type and other transport systems	Cell division transport system			JBNZ41
M00259	Environmental information processing	ABC-2 type and other transport systems	Heme transport system			JBNZ41
M00252	Environmental information processing	ABC-2 type and other transport systems	Lipooligosaccharide transport system	JBC07		JBNZ41
M00320	Environmental information processing	ABC-2 type and other transport systems	Lipopolysaccharide export system	JBC07		JBNZ41
M00250	Environmental information processing	ABC-2 type and other transport systems	Lipopolysaccharide transport system		JBM10	JBNZ41
M00255	Environmental information processing	ABC-2 type and other transport systems	Lipoprotein-releasing system		JBM10	JBNZ41
M00258	Environmental information processing	ABC-2 type and other transport systems	Putative ABC transport system	JBC07	JBM10	JBNZ41
M00330	Environmental information processing	Bacterial secretion system	Adhesin protein transport system	JBC07	JBM10	JBNZ41
M00571	Environmental information processing	Bacterial secretion system	AlgE-type Mannuronan C-5-Epimerase transport system	JBC07	JBM10	JBNZ41
M00325	Environmental information processing	Bacterial secretion system	alpha-Hemolysin/cyclolysin transport system	JBC07	JBM10	JBNZ41
M00429	Environmental information processing	Bacterial secretion system	Competence-related DNA transformation transporter	JBC07	JBM10	JBNZ41
M00339	Environmental information processing	Bacterial secretion system	RaxAB-RaxC type I secretion system	JBC07	JBM10	JBNZ41
M00326	Environmental information processing	Bacterial secretion system	RTX toxin transport system	JBC07	JBM10	JBNZ41
M00335	Environmental information processing	Bacterial secretion system	Sec (secretion) system	JBC07	JBM10	JBNZ41
M00336	Environmental information processing	Bacterial secretion system	Twin-arginine translocation (Tat) system			JBNZ41
M00331	Environmental information processing	Bacterial secretion system	Type II general secretion system	JBC07	JBM10	JBNZ41
M00333	Environmental information processing	Bacterial secretion system	Type IV secretion system	JBC07	JBM10	JBNZ41
M00334	Environmental information processing	Bacterial secretion system	Type VI secretion system			JBNZ41
M00646	Environmental information processing	Drug efflux transporter/pump	Multidrug resistance, efflux pump AcrAD-TolC			JBNZ41
M00648	Environmental information processing	Drug efflux transporter/pump	Multidrug resistance, efflux pump MdtABC			JBNZ41
M00717	Environmental information processing	Drug efflux transporter/pump	Multidrug resistance, efflux pump NorA	JBC07	JBM10	JBNZ41
M00821	Environmental information processing	Drug efflux transporter/pump	Multidrug resistance, efflux pump TriABC-TolC	JBC07	JBM10	JBNZ41
M00720	Environmental information processing	Drug efflux transporter/pump	Multidrug resistance, efflux pump VexEF-TolC	JBC07	JBM10	JBNZ41
M00742	Environmental information processing	Drug resistance	Aminoglycoside resistance, protease FtsH	JBC07	JBM10	JBNZ41
M00743	Environmental information processing	Drug resistance	Aminoglycoside resistance, protease HtpX			JBNZ41
M00729	Environmental information processing	Drug resistance	Fluoroquinolone resistance, gyrase-protecting protein Qnr	JBC07		
M00247	Environmental information processing	Metallic cation, iron-siderophore and vitamin B12 transport system	Putative ABC transport system			JBNZ41
M00208	Environmental information processing	Mineral and organic ion transport system	Glycine betaine/proline transport system			JBNZ41
M00190	Environmental information processing	Mineral and organic ion transport system	Iron(III) transport system			JBNZ41
M00189	Environmental information processing	Mineral and organic ion transport system	Molybdate transport system	JBC07		JBNZ41
M00188	Environmental information processing	Mineral and organic ion transport system	NitT/TauT family transport system		JBM10	JBNZ41
M00193	Environmental information processing	Mineral and organic ion transport system	Putative spermidine/putrescine transport system	JBC07	JBM10	JBNZ41
M00192	Environmental information processing	Mineral and organic ion transport system	Putative thiamine transport system	JBC07	JBM10	JBNZ41
M00299	Environmental information processing	Mineral and organic ion transport system	Spermidine/putrescine transport system	JBM10		JBNZ41
M00185	Environmental information processing	Mineral and organic ion transport system	Sulfate transport system	JBC07		JBNZ41
M00436	Environmental information processing	Mineral and organic ion transport system	Sulfonate transport system	JBM10		JBNZ41
M00349	Environmental information processing	Peptide and nickel transport system	Microcin C transport system			JBNZ41
M00239	Environmental information processing	Peptide and nickel transport system	Peptides/nickel transport system	JBC07		JBNZ41
M00237	Environmental information processing	Phosphate and amino acid transport system	Branched-chain amino acid transport system		JBM10	JBNZ41
M00222	Environmental information processing	Phosphate and amino acid transport system	Phosphate transport system			JBNZ41
M00236	Environmental information processing	Phosphate and amino acid transport system	Putative polar amino acid transport system		JBM10	JBNZ41

M00219	Environmental information processing	Saccharide, polyol, and lipid transport system	AI-2 transport system		JBC07	JBNZ41
M00201	Environmental information processing	Saccharide, polyol, and lipid transport system	alpha-Glucoside transport system			JBNZ41
M00491	Environmental information processing	Saccharide, polyol, and lipid transport system	arabinogalactan oligomer/maltooligosaccharide transport system			JBNZ41
M00602	Environmental information processing	Saccharide, polyol, and lipid transport system	Arabinosaccharide transport system			JBNZ41
M00206	Environmental information processing	Saccharide, polyol, and lipid transport system	Cellobiose transport system			JBNZ41
M00669	Environmental information processing	Saccharide, polyol, and lipid transport system	gamma-Hexachlorocyclohexane transport system	JBC07		JBNZ41
M00605	Environmental information processing	Saccharide, polyol, and lipid transport system	Glucose/mannose transport system			JBNZ41
M00194	Environmental information processing	Saccharide, polyol, and lipid transport system	Maltose/maltodextrin transport system			JBNZ41
M00670	Environmental information processing	Saccharide, polyol, and lipid transport system	Mce transport system	JBC07		JBNZ41
M00196	Environmental information processing	Saccharide, polyol, and lipid transport system	Multiple sugar transport system			JBNZ41
M00606	Environmental information processing	Saccharide, polyol, and lipid transport system	N,'N''-Diacetylchitobiose transport system			JBNZ41
M00210	Environmental information processing	Saccharide, polyol, and lipid transport system	Phospholipid transport system	JBC07		JBNZ41
M00211	Environmental information processing	Saccharide, polyol, and lipid transport system	Putative ABC transport system	JBC07	JBM10	JBNZ41
M00197	Environmental information processing	Saccharide, polyol, and lipid transport system	Putative fructooligosaccharide transport system			JBNZ41
M00207	Environmental information processing	Saccharide, polyol, and lipid transport system	Putative multiple sugar transport system			JBNZ41
M00221	Environmental information processing	Saccharide, polyol, and lipid transport system	Putative simple sugar transport system	JBC07	JBM10	JBNZ41
M00200	Environmental information processing	Saccharide, polyol, and lipid transport system	Putative sorbitol/mannitol transport system			JBNZ41
M00475	Environmental information processing	Two-component regulatory system	BarA-UvrY (central carbon metabolism) two-component regulatory system			JBNZ41
M00512	Environmental information processing	Two-component regulatory system	CckA-CtrA/CpdR (cell cycle control) two-component regulatory system	JBC07		JBNZ41
M00506	Environmental information processing	Two-component regulatory system	CheA-CheYBV (chemotaxis) two-component regulatory system			JBNZ41
M00452	Environmental information processing	Two-component regulatory system	CusS-CusR (copper tolerance) two-component regulatory system			JBM10
M00445	Environmental information processing	Two-component regulatory system	EnvZ-OmpR (osmotic stress response) two-component regulatory system			JBNZ41
M00524	Environmental information processing	Two-component regulatory system	FixL-FixJ (nitrogen fixation) two-component regulatory system		JBM10	JBNZ41
M00497	Environmental information processing	Two-component regulatory system	GlnL-GlnG (nitrogen regulation) two-component regulatory system			JBNZ41
M00454	Environmental information processing	Two-component regulatory system	KdpD-KdpE (potassium transport) two-component regulatory system			JBNZ41
M00498	Environmental information processing	Two-component regulatory system	NtrY-NtrX (nitrogen regulation) two-component regulatory system	JBC07		JBNZ41
M00434	Environmental information processing	Two-component regulatory system	PhoR-PhoB (phosphate starvation response) two-component regulatory system			JBNZ41
M00511	Environmental information processing	Two-component regulatory system	PleC-PleD (cell fate control) two-component regulatory system	JBC07	JBM10	JBNZ41
M00523	Environmental information processing	Two-component regulatory system	RegB-RegA (redox response) two-component regulatory system			JBNZ41
M00516	Environmental information processing	Two-component regulatory system	SLN1-YPD1-SSK1/SKN7 (osmosensing) two-component regulatory system	JBC07	JBM10	JBNZ41
M00745	Gene set	Drug resistance	Imipenem resistance, repression of porin OprD		JBM10	
M00261	Genetic information processing	DNA polymerase	DNA polymerase alpha / primase complex	JBC07	JBM10	JBNZ41
M00262	Genetic information processing	DNA polymerase	DNA polymerase delta complex	JBC07	JBM10	JBNZ41
M00263	Genetic information processing	DNA polymerase	DNA polymerase epsilon complex	JBC07	JBM10	JBNZ41
M00260	Genetic information processing	DNA polymerase	DNA polymerase III complex, bacteria	JBC07	JBM10	JBNZ41
M00293	Genetic information processing	DNA polymerase	DNA polymerase zeta complex	JBC07	JBM10	JBNZ41
M00343	Genetic information processing	Proteasome	Archaeal proteasome	JBC07	JBM10	JBNZ41
M00342	Genetic information processing	Proteasome	Bacterial proteasome	JBC07	JBM10	JBNZ41
M00337	Genetic information processing	Proteasome	Immunoproteasome	JBC07	JBM10	JBNZ41
M00341	Genetic information processing	Proteasome	Proteasome, 19S regulatory particle (PA700)	JBC07	JBM10	JBNZ41
M00340	Genetic information processing	Proteasome	Proteasome, 20S core particle	JBC07	JBM10	JBNZ41
M00404	Genetic information processing	Protein processing	COPII complex	JBC07	JBM10	JBNZ41
M00408	Genetic information processing	Protein processing	ESCRT-0 complex	JBC07	JBM10	JBNZ41
M00409	Genetic information processing	Protein processing	ESCRT-I complex	JBC07	JBM10	JBNZ41
M00410	Genetic information processing	Protein processing	ESCRT-II complex	JBC07	JBM10	JBNZ41
M00412	Genetic information processing	Protein processing	ESCRT-III complex	JBC07	JBM10	JBNZ41
M00403	Genetic information processing	Protein processing	HRD1/SEL1 ERAD complex	JBC07	JBM10	JBNZ41
M00400	Genetic information processing	Protein processing	p97-Ufd1-Npl4 complex	JBC07	JBM10	JBNZ41
M00401	Genetic information processing	Protein processing	Sec61 complex	JBC07	JBM10	JBNZ41
M00402	Genetic information processing	Protein processing	Translocon-associated protein (TRAP) complex	JBC07	JBM10	JBNZ41
M00296	Genetic information processing	Repair system	BER complex	JBC07	JBM10	JBNZ41
'M00414	Genetic information processing	Repair system	Bloom''s syndrome complex	JBC07	JBM10	JBNZ41'
M00295	Genetic information processing	Repair system	BRCA1-associated genome surveillance complex (BASC)	JBC07	JBM10	JBNZ41
M00297	Genetic information processing	Repair system	DNA-PK complex	JBC07	JBM10	JBNZ41
M00413	Genetic information processing	Repair system	FA core complex	JBC07	JBM10	JBNZ41
M00290	Genetic information processing	Repair system	Holo-TFIIH complex	JBC07	JBM10	JBNZ41

M00291	Genetic information processing	Repair system	MRN complex	JBC07	JBM10	JBNZ41
M00292	Genetic information processing	Repair system	MRX complex	JBC07	JBM10	JBNZ41
M00286	Genetic information processing	Replication system	GINS complex	JBC07	JBM10	JBNZ41
M00285	Genetic information processing	Replication system	MCM complex	JBC07	JBM10	JBNZ41
M00284	Genetic information processing	Replication system	Origin recognition complex	JBC07	JBM10	JBNZ41
M00289	Genetic information processing	Replication system	RF-C complex	JBC07	JBM10	JBNZ41
M00288	Genetic information processing	Replication system	RPA complex	JBC07	JBM10	JBNZ41
M00179	Genetic information processing	Ribosome	Ribosome, archaea	JBC07	JBM10	JBNZ41
M00178	Genetic information processing	Ribosome	Ribosome, bacteria	JBC07	JBM10	JBNZ41
M00177	Genetic information processing	Ribosome	Ribosome, eukaryotes	JBC07	JBM10	JBNZ41
M00182	Genetic information processing	RNA polymerase	RNA polymerase I, eukaryotes	JBC07	JBM10	JBNZ41
M00180	Genetic information processing	RNA polymerase	RNA polymerase II, eukaryotes	JBC07	JBM10	JBNZ41
M00181	Genetic information processing	RNA polymerase	RNA polymerase III, eukaryotes	JBC07	JBM10	JBNZ41
M00183	Genetic information processing	RNA polymerase	RNA polymerase, bacteria	JBC07	JBM10	JBNZ41
M00395	Genetic information processing	RNA processing	Decapping complex	JBC07	JBM10	JBNZ41
M00428	Genetic information processing	RNA processing	eIF4F complex	JBC07	JBM10	JBNZ41
M00430	Genetic information processing	RNA processing	Exon junction complex (EJC)	JBC07	JBM10	JBNZ41
M00390	Genetic information processing	RNA processing	Exosome, archaea	JBC07	JBM10	JBNZ41
M00391	Genetic information processing	RNA processing	Exosome, eukaryotes	JBC07	JBM10	JBNZ41
M00425	Genetic information processing	RNA processing	H/ACA ribonucleoprotein complex	JBC07	JBM10	JBNZ41
M00427	Genetic information processing	RNA processing	Nuclear pore complex	JBC07	JBM10	JBNZ41
M00394	Genetic information processing	RNA processing	RNA degradosome	JBC07	JBM10	JBNZ41
M00392	Genetic information processing	RNA processing	Ski complex	JBC07	JBM10	JBNZ41
M00426	Genetic information processing	RNA processing	Survival motor neuron (SMN) complex	JBC07	JBM10	JBNZ41
M00405	Genetic information processing	RNA processing	THC complex	JBC07	JBM10	JBNZ41
M00393	Genetic information processing	RNA processing	TRAMP complex	JBC07	JBM10	JBNZ41
M00406	Genetic information processing	RNA processing	TREX complex	JBC07	JBM10	JBNZ41
M00399	Genetic information processing	Spliceosome	Cap binding complex	JBC07	JBM10	JBNZ41
M00397	Genetic information processing	Spliceosome	Lsm 1-7 complex	JBC07	JBM10	JBNZ41
M00396	Genetic information processing	Spliceosome	Lsm 2-8 complex	JBC07	JBM10	JBNZ41
M00398	Genetic information processing	Spliceosome	Sm core complex	JBC07	JBM10	JBNZ41
M00355	Genetic information processing	Spliceosome	Spliceosome, 35S U5-snRNP	JBC07	JBM10	JBNZ41
M00353	Genetic information processing	Spliceosome	Spliceosome, Prp19/CDC5L complex	JBC07	JBM10	JBNZ41
M00351	Genetic information processing	Spliceosome	Spliceosome, U1-snRNP	JBC07	JBM10	JBNZ41
M00352	Genetic information processing	Spliceosome	Spliceosome, U2-snRNP	JBC07	JBM10	JBNZ41
M00354	Genetic information processing	Spliceosome	Spliceosome, U4/U6.U5 tri-snRNP	JBC07	JBM10	JBNZ41
M00389	Genetic information processing	Ubiquitin system	APC/C complex	JBC07	JBM10	JBNZ41
M00384	Genetic information processing	Ubiquitin system	Cul3-SPOP complex	JBC07	JBM10	JBNZ41
M00386	Genetic information processing	Ubiquitin system	Cul4-DDB1-CSA complex	JBC07	JBM10	JBNZ41
M00385	Genetic information processing	Ubiquitin system	Cul4-DDB1-DDB2 complex	JBC07	JBM10	JBNZ41
M00388	Genetic information processing	Ubiquitin system	ECS complex	JBC07	JBM10	JBNZ41
M00383	Genetic information processing	Ubiquitin system	ECV complex	JBC07	JBM10	JBNZ41
M00380	Genetic information processing	Ubiquitin system	SCF-BTRC complex	JBC07	JBM10	JBNZ41
M00407	Genetic information processing	Ubiquitin system	SCF-CDC4 complex	JBC07	JBM10	JBNZ41
M00382	Genetic information processing	Ubiquitin system	SCF-FBS complex	JBC07	JBM10	JBNZ41
M00387	Genetic information processing	Ubiquitin system	SCF-FBW7 complex	JBC07	JBM10	JBNZ41
M00411	Genetic information processing	Ubiquitin system	SCF-GRR1 complex	JBC07	JBM10	JBNZ41
M00379	Genetic information processing	Ubiquitin system	SCF-MET30 complex	JBC07	JBM10	JBNZ41
M00381	Genetic information processing	Ubiquitin system	SCF-SKP2 complex	JBC07	JBM10	JBNZ41
M00359	Metabolism	Aminoacyl tRNA	Aminoacyl-tRNA biosynthesis, eukaryotes	JBC07	JBM10	JBNZ41
M00360	Metabolism	Aminoacyl tRNA	Aminoacyl-tRNA biosynthesis, prokaryotes	JBC07	JBM10	JBNZ41
M00361	Metabolism	Nucleotide sugar	Nucleotide sugar biosynthesis, eukaryotes	JBC07	JBM10	JBNZ41
M00362	Metabolism	Nucleotide sugar	Nucleotide sugar biosynthesis, prokaryotes			JBNZ41
M00844	Nucleotide and amino acid metabolism	Arginine and proline metabolism	Arginine biosynthesis, ornithine => arginine	JBC07	JBM10	JBNZ41
M00028	Nucleotide and amino acid metabolism	Arginine and proline metabolism	Ornithine biosynthesis, glutamate => ornithine	JBC07	JBM10	JBNZ41
M00015	Nucleotide and amino acid metabolism	Arginine and proline metabolism	Proline biosynthesis, glutamate => proline	JBC07	JBM10	JBNZ41
M00029	Nucleotide and amino acid metabolism	Arginine and proline metabolism	Urea cycle		JBM10	JBNZ41
M00024	Nucleotide and amino acid metabolism	Aromatic amino acid metabolism	Phenylalanine biosynthesis, chorismate => phenylalanine	JBC07	JBM10	JBNZ41
M00022	Nucleotide and amino acid metabolism	Aromatic amino acid metabolism	Shikimate pathway, phosphoenolpyruvate + erythrose-4P => chorismate	JBC07	JBM10	JBNZ41
M00038	Nucleotide and amino acid metabolism	Aromatic amino acid metabolism	Tryptophan metabolism, tryptophan => kynurenine => 2-aminomuconate			JBC07
M00025	Nucleotide and amino acid metabolism	Aromatic amino acid metabolism	Tyrosine biosynthesis, chorismate => tyrosine	JBC07	JBM10	JBNZ41
M00040	Nucleotide and amino acid metabolism	Aromatic amino acid metabolism	Tyrosine biosynthesis, prephanate => pretyrosine => tyrosine			JBNZ41
M00044	Nucleotide and amino acid metabolism	Aromatic amino acid metabolism	Tyrosine degradation, tyrosine => homogentisate		JBM10	JBNZ41
M00535	Nucleotide and amino acid metabolism	Branched-chain amino acid metabolism	Isoleucine biosynthesis, pyruvate => 2-oxobutanoate			JBNZ41
M00570	Nucleotide and amino acid metabolism	Branched-chain amino acid metabolism	Isoleucine biosynthesis, threonine => 2-oxobutanoate => isoleucine			JBNZ41
M00432	Nucleotide and amino acid metabolism	Branched-chain amino acid metabolism	Leucine biosynthesis, 2-oxoisovalerate => 2-oxoisocaproate			JBNZ41
M00036	Nucleotide and amino acid metabolism	Branched-chain amino acid metabolism	Leucine degradation, leucine => acetoacetate + acetyl-CoA	JBC07	JBM10	JBNZ41
M00019	Nucleotide and amino acid metabolism	Branched-chain amino acid metabolism	Valine/isoleucine biosynthesis, pyruvate => valine / 2-oxobutanoate => isoleucine			JBNZ41
M00141	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	C1-unit interconversion, eukaryotes	JBC07	JBM10	JBNZ41
M00140	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	C1-unit interconversion, prokaryotes	JBC07	JBM10	JBNZ41
M00120	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Coenzyme A biosynthesis, pantothenate => CoA	JBC07	JBM10	JBNZ41
M00121	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Heme biosynthesis, glutamate => heme	JBC07	JBM10	JBNZ41
M00843	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	L-threo-Tetrahydrobiopterin biosynthesis, GTP => L-threo-BH4	JBC07	JBM10	JBNZ41
M00115	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	NAD biosynthesis, aspartate => NAD	JBC07		JBNZ41
M00119	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Pantothenate biosynthesis, valine/L-aspartate => pantothenate	JBC07		JBNZ41

M00572	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Pimeloyl-ACP biosynthesis, BioC-BioH pathway, malonyl-ACP => pimeloyl-ACP			JBNZ41
M00125	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Riboflavin biosynthesis, GTP => riboflavin/FMN/FAD	JBC07	JBM10	JBNZ41
M00846	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Siroheme biosynthesis, glutamate => siroheme			JBNZ41
M00842	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Tetrahydrobiopterin biosynthesis, GTP => BH4	JBC07	JBM10	JBNZ41
M00126	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Tetrahydrofolate biosynthesis, GTP => THF	JBC07	JBM10	JBNZ41
M00841	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Tetrahydrofolate biosynthesis, mediated by PTPS, GTP => THF	JBC07		
M00112	Nucleotide and amino acid metabolism	Cofactor and vitamin biosynthesis	Tocopherol/tocotorienol biosynthesis	JBC07	JBM10	JBNZ41
M00021	Nucleotide and amino acid metabolism	Cysteine and methionine metabolism	Cysteine biosynthesis, serine => cysteine	JBC07	JBM10	JBNZ41
M00017	Nucleotide and amino acid metabolism	Cysteine and methionine metabolism	Methionine biosynthesis, apartate => homoserine => methionine			JBNZ41
M00034	Nucleotide and amino acid metabolism	Cysteine and methionine metabolism	Methionine salvage pathway	JBC07	JBM10	JBNZ41
M00026	Nucleotide and amino acid metabolism	Histidine metabolism	Histidine biosynthesis, PRPP => histidine	JBC07	JBM10	JBNZ41
M00032	Nucleotide and amino acid metabolism	Lysine metabolism	Lysine degradation, lysine => saccharopine => acetoacetyl-CoA	JBC07	JBM10	JBNZ41
M00134	Nucleotide and amino acid metabolism	Polyamine biosynthesis	Polyamine biosynthesis, arginine => ornithine => putrescine	JBC07	JBM10	JBNZ41
M00049	Nucleotide and amino acid metabolism	Purine metabolism	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	JBC07	JBM10	JBNZ41
M00050	Nucleotide and amino acid metabolism	Purine metabolism	Guanine ribonucleotide biosynthesis IMP => GDP,GTP	JBC07	JBM10	JBNZ41
M00048	Nucleotide and amino acid metabolism	Purine metabolism	Inosine monophosphate biosynthesis, PRPP + glutamine => IMP	JBC07	JBM10	JBNZ41
M00546	Nucleotide and amino acid metabolism	Purine metabolism	Purine degradation, xanthine => urea	JBC07	JBM10	JBNZ41
M00052	Nucleotide and amino acid metabolism	Pyrimidine metabolism	Pyrimidine ribonucleotide biosynthesis, UMP => UDP/UTP,CDP/CTP	JBC07	JBM10	JBNZ41
M00051	Nucleotide and amino acid metabolism	Pyrimidine metabolism	Uridine monophosphate biosynthesis, glutamine (+ PRPP) => UMP	JBC07	JBM10	JBNZ41
M00555	Nucleotide and amino acid metabolism	Serine and threonine metabolism	Betaine biosynthesis, choline => betaine	JBC07	JBM10	JBNZ41
M00020	Nucleotide and amino acid metabolism	Serine and threonine metabolism	Serine biosynthesis, glycerate-3P => serine	JBC07	JBM10	JBNZ41
M00018	Nucleotide and amino acid metabolism	Serine and threonine metabolism	Threonine biosynthesis, aspartate => homoserine => threonine	JBC07	JBM10	JBNZ41

Supplement of:

Genome size of chrysophytes varies with cell size and nutritional mode

Table S 1 Strain description including culture medium, reference standard for genome size measurement (additionally *Solanum lycopersicum* (GS = 0.98 pg) was used for genome size estimation in *Synura sphagnicola*), genome size in [pg] including standard deviation (s), genome size converted to [Mbp] and cell volume [µm^3] including standard deviation (s);
* HF = Heterotrophic Flagellate

				Genome size		
Strain	**Species**	**Medium**	**Reference Standard**	**[pg] ± s**	**[Mbp]**	**Cell volume [µm^3] ± s**
JBAF 33	*Acrispumella msimbasiensis*	IB	*Raphanus sativus*	0.075 ± 0.001	73.4	N/A
JBC 27	*Chromulinospumella sphaerica*	IB	*Raphanus sativus*	0.080 ± 0.001	78.5	239.70 ± 68.52
A-R4-D6	*Cornospumella fuschlensis*	IB	*Synura spagnicola*	0.073 ± 0.004	71.6	130.03 ± 35.20
9-4-C1	HF*	IB	*Synura spagnicola*	0.045 ± 0.001	44.1	N/A
N1846	HF*	IB	*Synura spagnicola*	0.070 ± 0.001	68.0	111.85 ± 65.67
1006	*Pedospumella encystans*	IB	*Synura spagnicola*	0.069 ± 0.002	67.5	N/A
JBM|S 11	*Pedospumella encystans*	IB	*Raphanus sativus*	0.093 ± 0.001	91.1	140.68 ± 37.49
JBC|S 23	*Pedospumella sinomuralis*	IB	*Raphanus sativus*	0.158 ± 0.004	154.4	N/A
JBC 07	*Poteriospumella lacustris*	NSY	*Synura spagnicola*	0.080 ± 0.001	77.9	N/A
JBM 10	*Poteriospumella lacustris*	NSY	*Synura spagnicola*	0.049 ± 0.001	47.8	N/A
JBNZ 41	*Poteriospumella lacustris*	NSY	*Synura spagnicola*	0.090 ± 0.001	87.9	N/A
A-R3-A3	*Segregatospumella dracosaxi*	IB	*Synura spagnicola*	0.045 ± 0.001	44.4	N/A
JBNZ 39	*Spumella lacusvadosi*	IB	*Raphanus sativus*	0.154 ± 0.002	150.8	N/A
A-R4-A6	*Spumella rivalis*	IB	*Synura spagnicola*	0.050 ± 0.003	49.0	N/A
376hm	*Spumella* sp.	IB	*Synura spagnicola*	0.096 ± 0.004	94.3	N/A
LO244K-D	*Spumella* sp.	DY-V	*Synura spagnicola*	0.154 ± 0.003	151.1	N/A
199hm	*Spumella vulgaris*	IB	*Raphanus sativus*	0.150 ± 0.001	146.8	141.02 ± 42.86
933-7	*Chlorochromonas danica*	NSY	*Raphanus sativus*	0.103 ± 0.005	100.7	642.35 ± 176.63
DS	*Poterioochromonas malhamensis*	NSY	*Raphanus sativus*	0.077 ± 0.001	75.4	555.33 ± 149.12
M2953	*Bitrichia* sp.	WC	*Solanum pseudocapsicum*	0.298 ± 0.006	291.9	N/A
FU24K-BA	*Dinobryon bavaricum*	WC	*Raphanus sativus*	0.162 ± 0.003	158.5	N/A
FU18K-A	*Dinobryon divergens*	WC	*Raphanus sativus*	0.164 ± 0.002	160.7	N/A
FU22K-AK	*Dinobryon divergens*	WC	*Raphanus sativus*	0.151 ± 0.002	147.9	N/A
WA20K-H	*Dinobryon divergens*	WC	*Solanum pseudocapsicum*	0.181 ± 0.002	176.7	321.09 ± 112.69

WA26K-D	*Dinobryon divergens*	WC	*Solanum pseudocapsicum*	0.162 ± 0.001	158.0	N/A
LO226K-S	*Dinobryon pediforme*	DY-V	*Raphanus sativus*	0.116 ± 0.001	113.3	371.70 ± 107.73
OE22K-D	*Dinobryon sociale*	WC	*Solanum pseudocapsicum*	0.137 ± 0.001	134.1	N/A
OE26K-V	*Dinobryon sociale var. americana cf. div. schauinslandii*	WC	*Solanum pseudocapsicum*	0.181 ± 0.001	177.5	485.87 ± 125.03
AU32K-E	*Dinobryon* sp.	WC	*Solanum pseudocapsicum*	0.181 ± 0.001	177.2	N/A
FU29K-J	*Dinobryon* sp.	WC	*Solanum pseudocapsicum*	0.178 ± 0.001	174.2	428.15 ± 88.64
WA32K-W	*Dinobryon* sp.	WC	*Raphanus sativus*	0.122 ± 0.001	119.1	N/A
WI32K-F	*Dinobryon* sp.	WC	*Raphanus sativus*	0.125 ± 0.002	122.2	N/A
PR26K-G	*Epipyxis* sp.	WC	*Synura spagnicola*	0.098 ± 0.001	96.3	169.25 ± 61.02
FU36K-N	*Kephyrion* sp.	WC	*Synura spagnicola*	0.072 ± 0.001	70.4	156.24 ± 30.45
WA34K-E	*Uroglena* sp.	WC	*Raphanus sativus*	0.162 ± 0.004	158.8	N/A
WA18K-M	*Mallomonas annulata*	WC	*Hedychium gardnerianum*	0.343 ± 0.004	335.9	507.76 ± 162.52
PR26K-H	*Mallomonas caudata*	WC	*Raphanus sativus*	10.999 ± 0.195	10757.0	N/A
WA40K-F	*Mallomonas caudata*	WC	*Hedychium gardnerianum*	12.426 ± 0.191	12152.5	8356.78 ± 2961.43
OE26K-M	*Mallomonas cf. tonsurata*	WC	*Hedychium gardnerianum*	0.838 ± 0.008	819.3	865.03 ± 280.03
B 601	*Mallomonas kalinae*	WC	*Hedychium gardnerianum*	0.297 ± 0.003	290.7	N/A
OE40K-J	*Mallomonas* sp.	WC	*Hedychium gardnerianum*	0.840 ± 0.008	821.7	707.34 ± 143.52
WI26K-B	*Mallomonas* sp.	WC	*Hedychium gardnerianum*	0.937 ± 0.011	916.6	N/A
S 20.45	*Synura heteropora*	DY-V	*Hedychium gardnerianum*	0.754 ± 0.008	737.8	469.41 ± 124.20
WA18K-A	*Synura petersenii*	WC	*Hedychium gardnerianum*	0.777 ± 0.003	759.5	688.95 ± 143.14
WA18K-U	*Synura* sp.	WC	*Hedychium gardnerianum*	0.837 ± 0.005	818.8	686.78 ± 179.05
LO234K-E	*Synura sphagnicola*	DY-V	*Solanum pseudocapsicum; Raphanus sativus; Solanum lycopersicum*	0.203 ± 0.001	198.1	558.77 ± 158.00

Supplement to the paper: Nutrient-driven genome evolution revealed by comparative genomics of chrysomonad flagellates

Table S1: **PacBio sequencing statistics**

strain	reads [k]	bases [billions]	mean length	median length	GC%
JBMS11	106	0.96	9040	7783	53.3
DS	337	2.05	6086	4652	39.0
LO234KE	212	1.79	8439	8031	47.6
933-7	297	2.18	7326	7275	39.9

Table S2: **Binning and classification results**

strain	assigned eukaryotic bins	assigned prokaryotic bins	MaxBin assembly size	MetaBat assembly size
JBC27	7	17	132M	27M
A-R4-D6	6	18	145M	21M
1006	4	30	123M	196K
JBMS11	3	6	94M	61M
199hm	4	14	158M	41M
CCAC 4401B	9	26	226M	34M
FU18K-A	11	33	215M	38M
LO226K-S	6	10	144M	13M
PR26K-G	4	17	101M	40M
WA18K-M	13	28	171M	67M
LO234KE	12	5	245M	67M

Table S3: **Comparison of the pipeline with and without additional PacBio sequencing.** Numbers reflect the amount of predicted genes belonging to different functional groups. The asterisk marks samples excluding PacBio sequences. The quotient indicates by how much the samples differ with and without PacBio. The use of long sequences increase the average contig length improving gene prediction. However, functional groups are affected to different degrees.

strain	DS	LO234KE	DS∗	LO234KE∗	DS/DS∗	LO234KE/LO234KE*
Carbohydrate and lipid metabolism	962	1,392	543	1,295	1.77	1.07
Cellular processes	169	226	66	73	2.56	3.10
Energy metabolism	353	535	243	571	1.45	.94
Environmental information processing	223	1,114	116	1,651	1.92	.67
Gene set	18	73	18	95	1.00	.77
Genetic information processing	1,484	1,845	755	1,146	1.97	1.61
Metabolism	241	388	180	424	1.34	.92
Nucleotide and amino acid metabolism	1,138	1,934	731	2,147	1.56	.90
Secondary metabolism	14	60	6	71	2.33	.85

Table S4: **GC content lower in non-coding regions**

Strain	Total GC [%]	Non-coding GC [%]	Coding GC GC [%]	Intron GC [%]	Trophy
JBC27	51.3	34.4	57.2	45.9	heterotroph
A-R4-D6	51.7	39.1	57.3	44.9	heterotroph
1006	54.5	39.3	58.9	45.8	heterotroph
JBMS11	51.4	39.4	56.3	43.3	heterotroph
199hm	47.9	36.2	57.0	42.4	heterotroph
CCAC 4401B	43.9	30.7	56.8	38.9	mixotroph
FU18K-A	45.1	33.9	55.1	38.6	mixotroph
LO226K-S	51.6	37.1	58.5	44.9	mixotroph
PR26K-G	34.1	27.6	50.1	32.9	mixotroph
933-7	45.4	36.8	49.7	41.6	mixotroph
DS	40.4	34.4	46.4	38.1	mixotroph
WA18K-M	40.1	31.8	53.4	35.5	phototroph
LO234KE	46.9	40.4	54.4	45.0	phototroph

Figure S1: **Flow diagram of assembly, binning and gene prediction.** First, 16 strains were sequenced with Illumina and partly with PacBio. The reads were assembled (blue) to contigs. If the culture were non-axenic a binning step (green) filtered out prokaryotic reads. A second assembly used the filtered data. Subsequently, the genes were predicted (purple). Optional transcriptome data support the prediction.

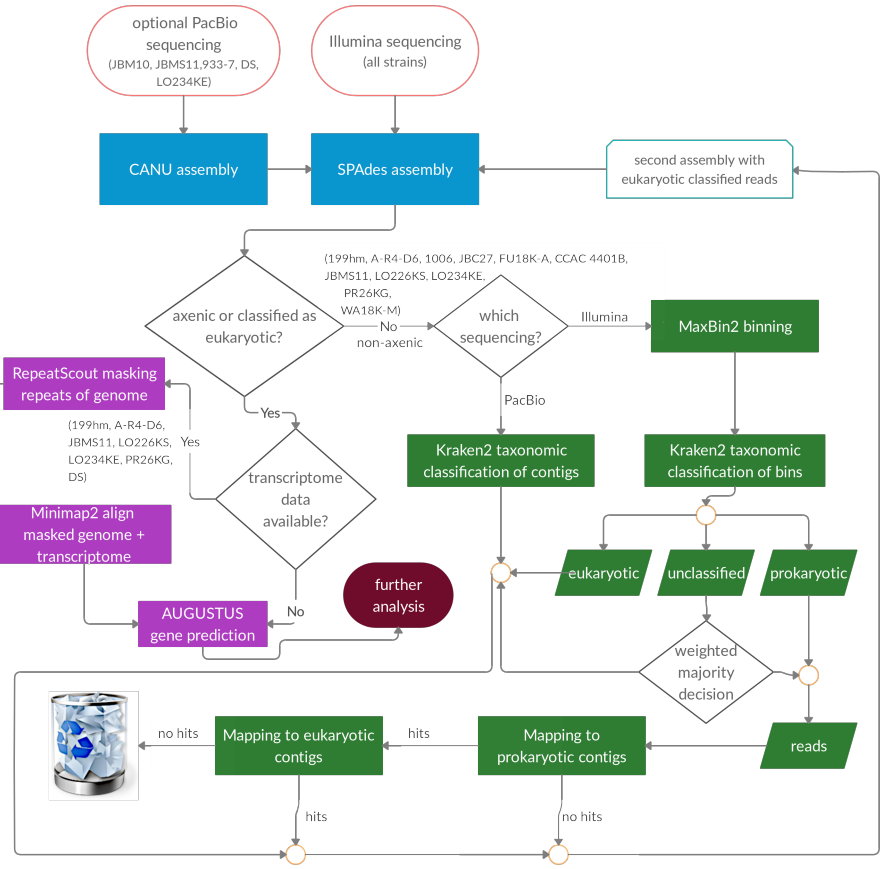


Figure S2: **Proportion of gene assignments for each strain.** Composition of annotated genes according to the second highest hierarchy level of the KEGG functional groups.

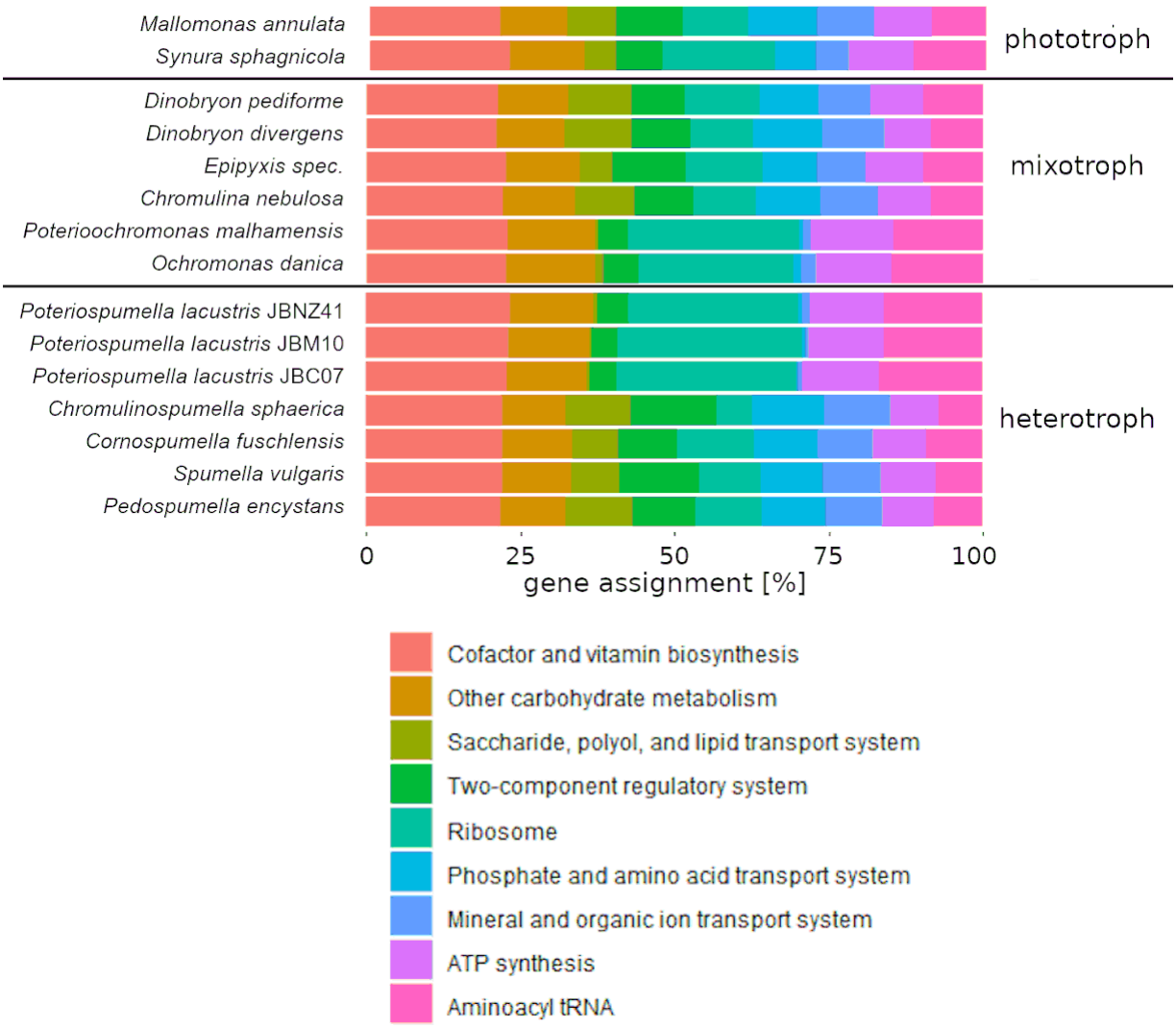
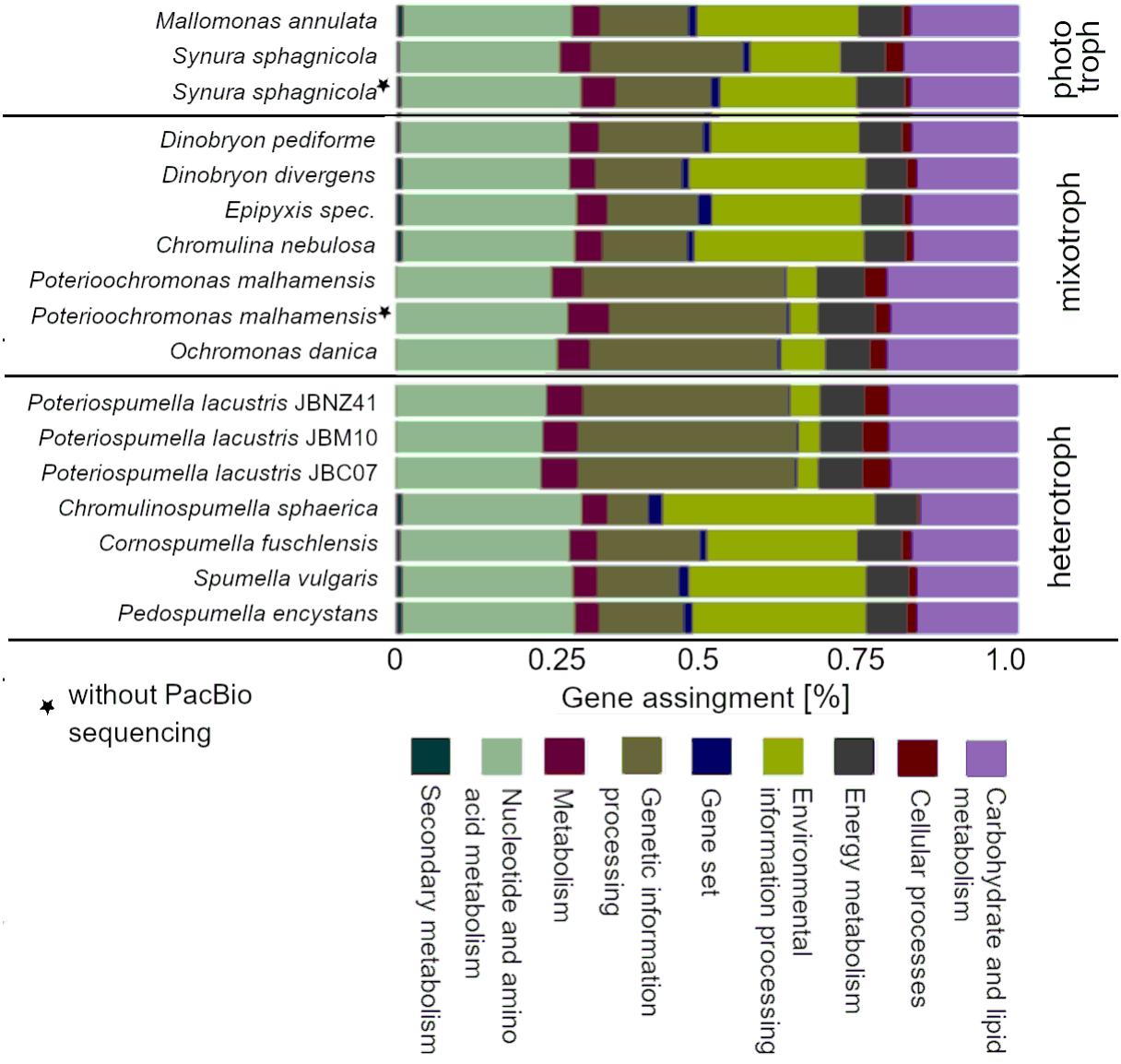


Figure S3: **Additional PacBio sequencing changes proportions of gene assignment.** The strains *P. malhamensis* and *S. sphagnicola* were assembled with and without (marked by star) PacBio sequences. *P. malhamensis* grew under axenic conditions and *S. sphagnicola* did not. Functional groups affected to varying degrees.



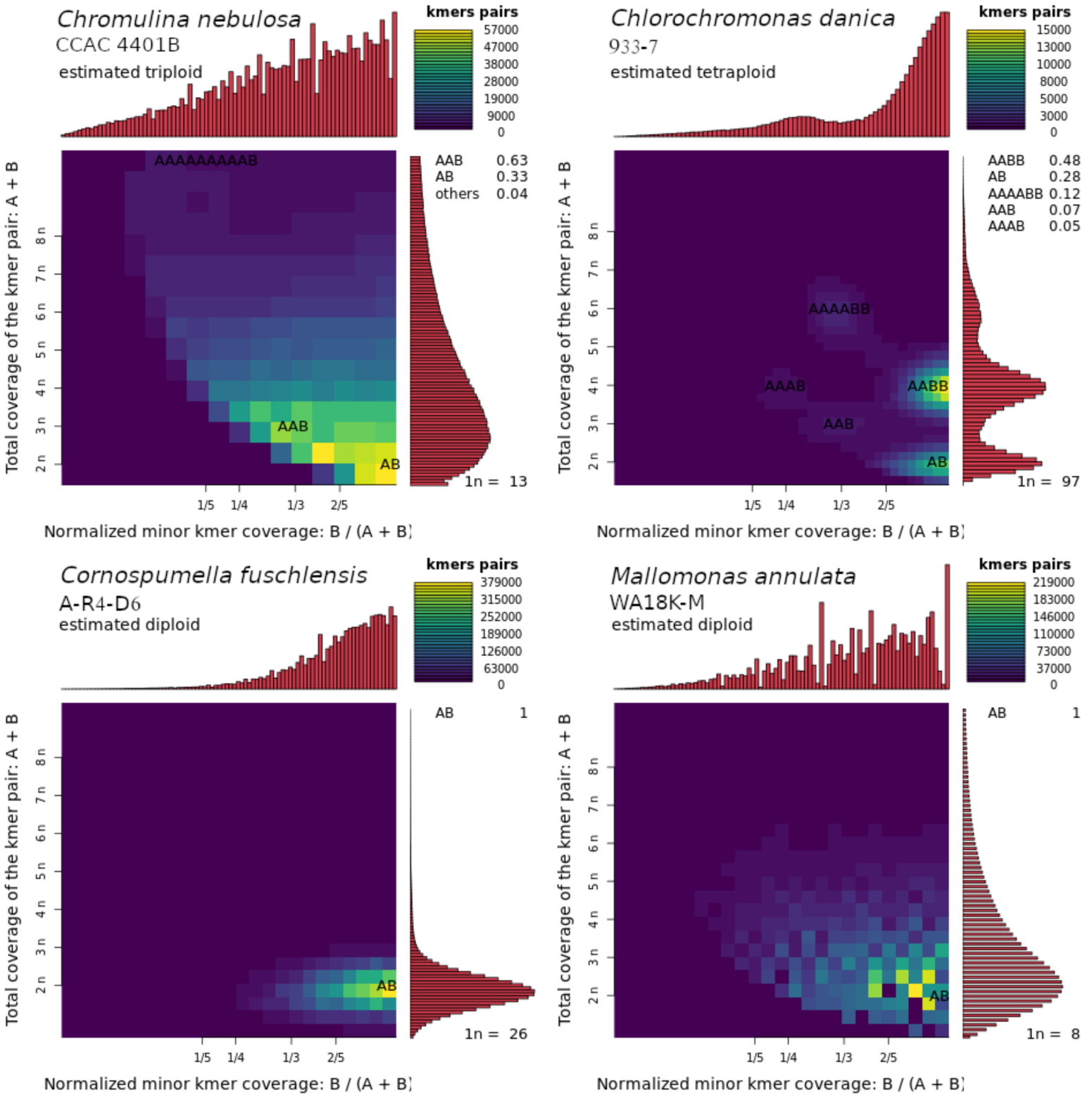


Figure S4: **K-mer based ploidy estimation.** The heatmap reflects coverage of k-mer pairs differing by one base. The ratio of the characters A to B represents the ratio of these k-mer pairs (e.g., 67% ATGTC and 33% ATGTT conforms AAB). The coverage distribution on the right side indicates ploidy levels (left scale). The distribution on the top side is based on the coverage normalized by the ratio. n = average k-mer coverage, k-mer size = 21

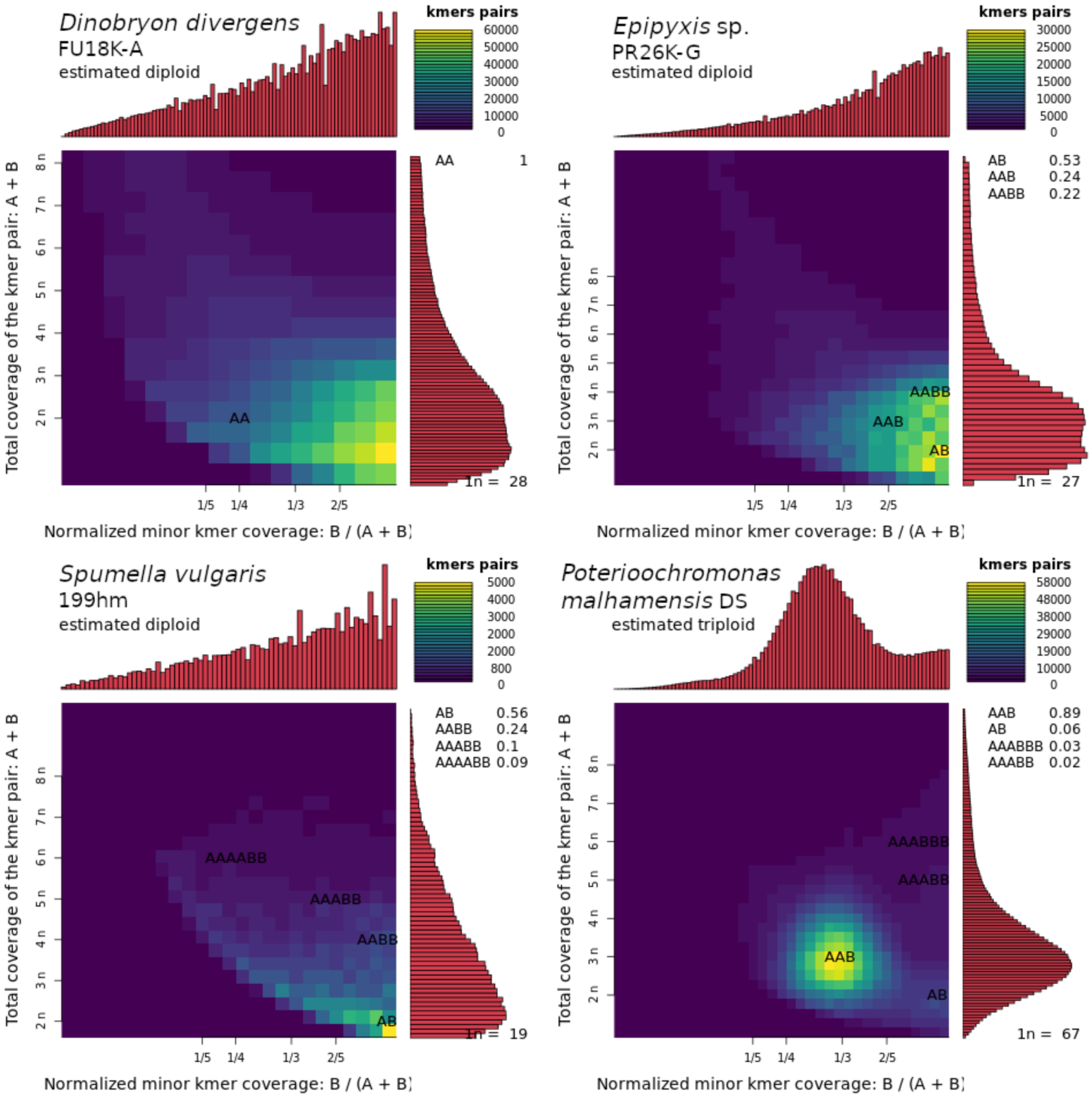


Figure S5: **K-mer based ploidy estimation.** The heatmap reflects coverage of k-mer pairs differing by one base. The ratio of the characters A to B represents the ratio of these k-mer pairs (e.g., 67% ATGTC and 33% ATGTT conforms AAB). The coverage distribution on the right side indicates ploidy levels (left scale). The distribution on the top side is based on the coverage normalized by the ratio. n = average *k*-mer coverage, *k*-mer size = 21

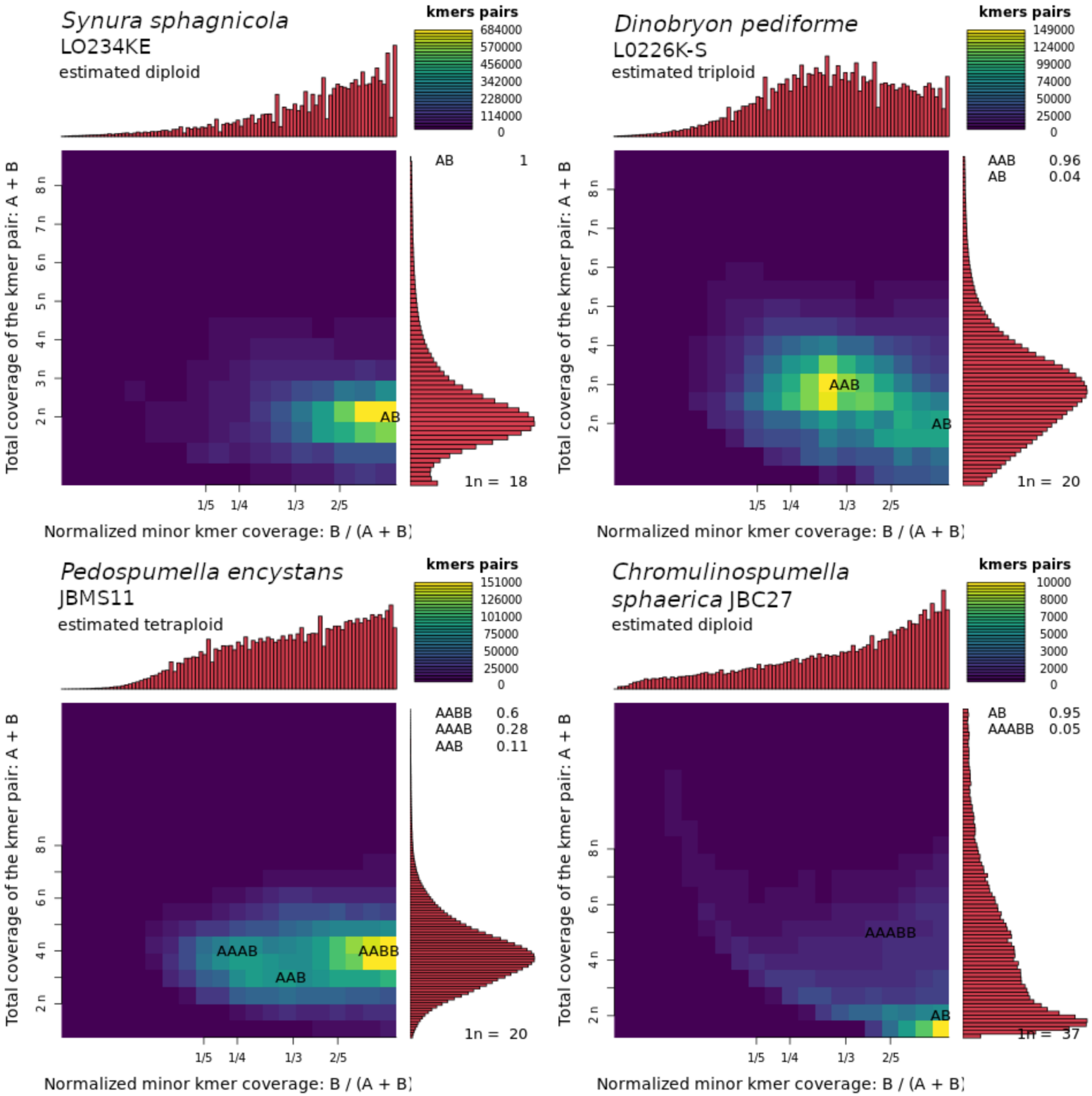


Figure S6: **K-mer based ploidy estimation.** The heatmap reflects coverage of k-mer pairs differing by one base. The ratio of the characters A to B represents the ratio of these k-mer pairs (e.g., 67% ATGTC and 33% ATGTT conforms AAB). The coverage distribution on the right side indicates ploidy levels (left scale). The distribution on the top side is based on the coverage normalized by the ratio. n = average *k*-mer coverage, *k*-mer size = 21

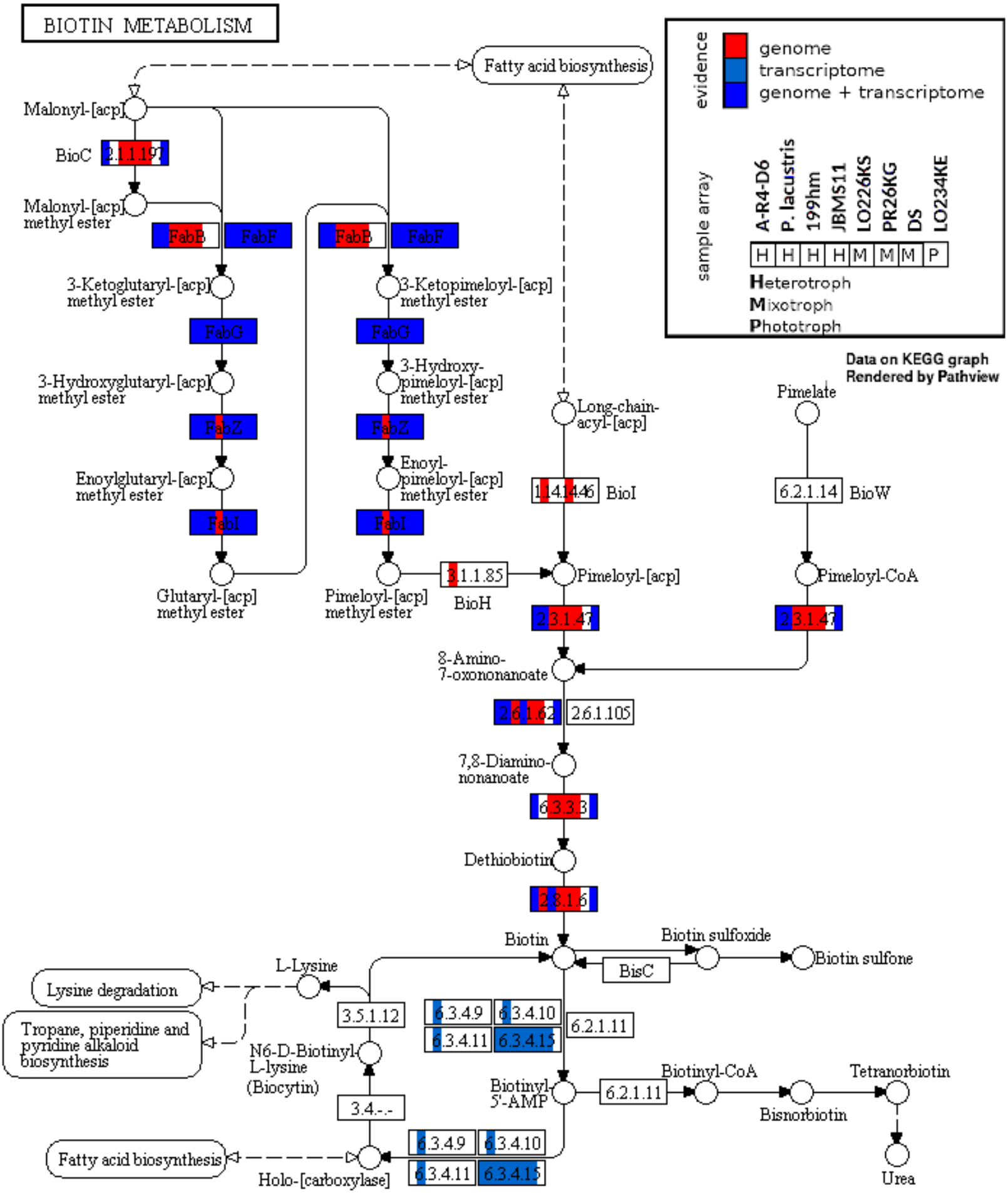


Figure S7: **Biotin meatbolism.** New founded genes are red marked.

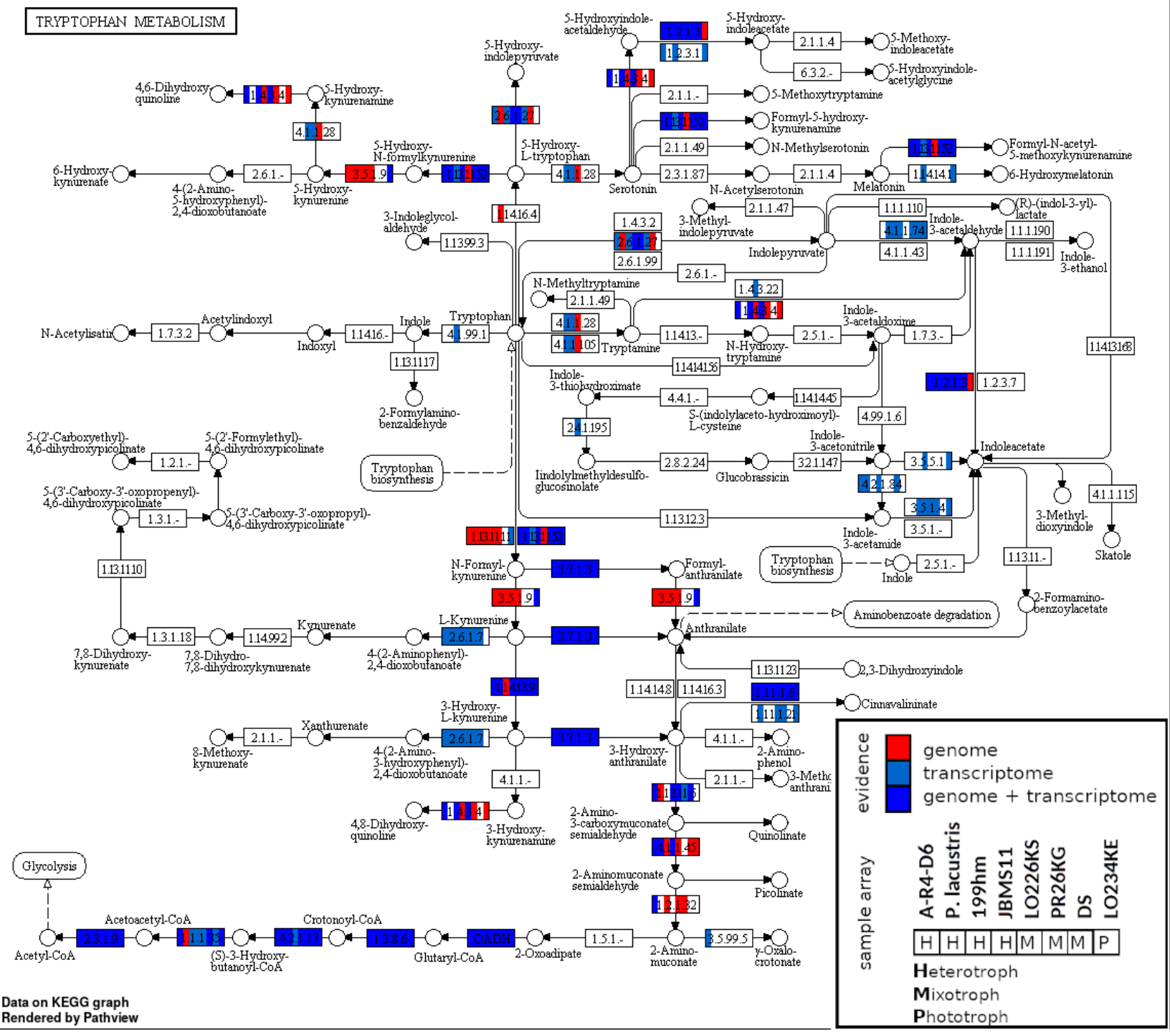


Figure S8: **Tryptophan meatbolism.** New founded genes are red marked.

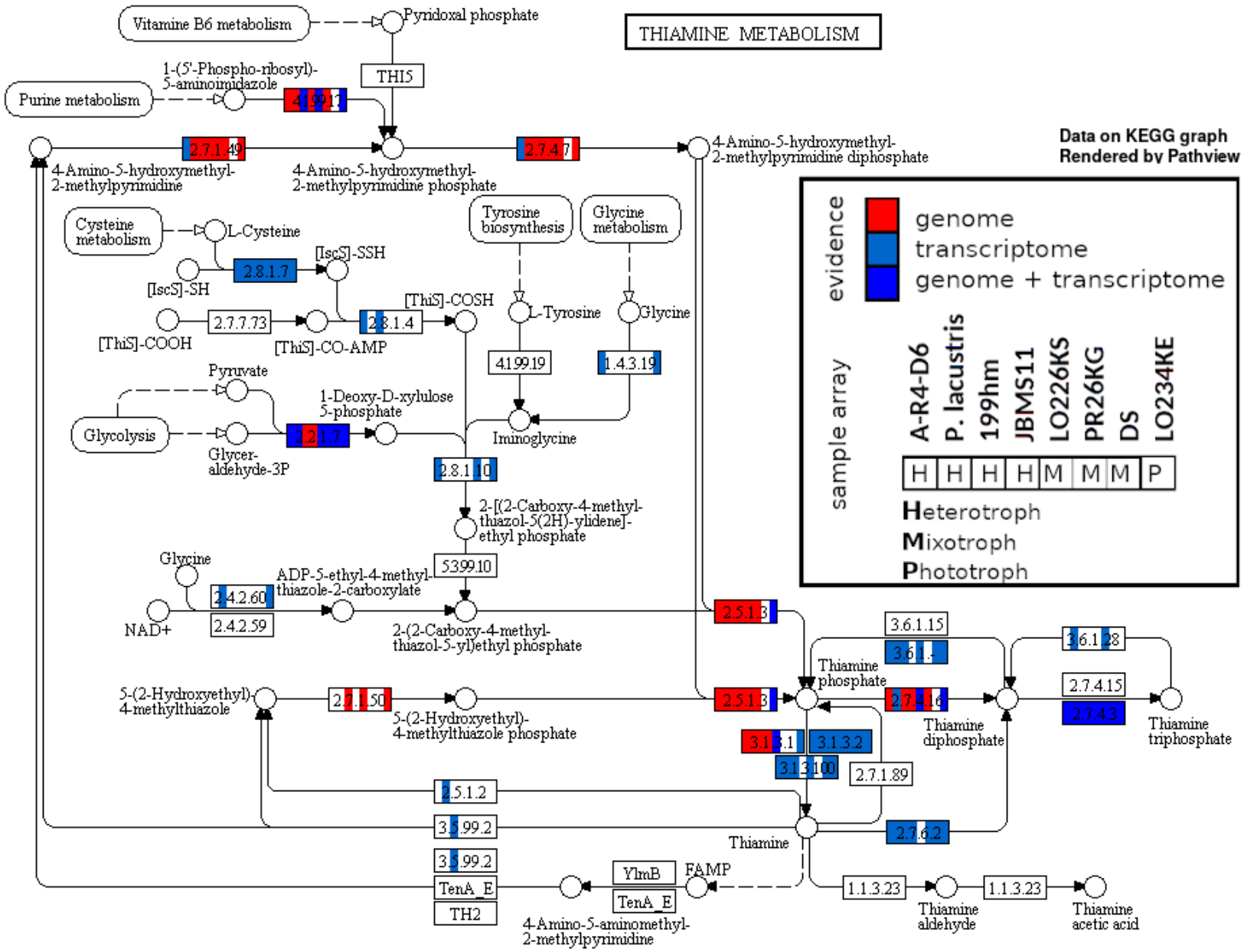


Figure S9: **Thiamine meatbolism.** New founded genes are red marked.

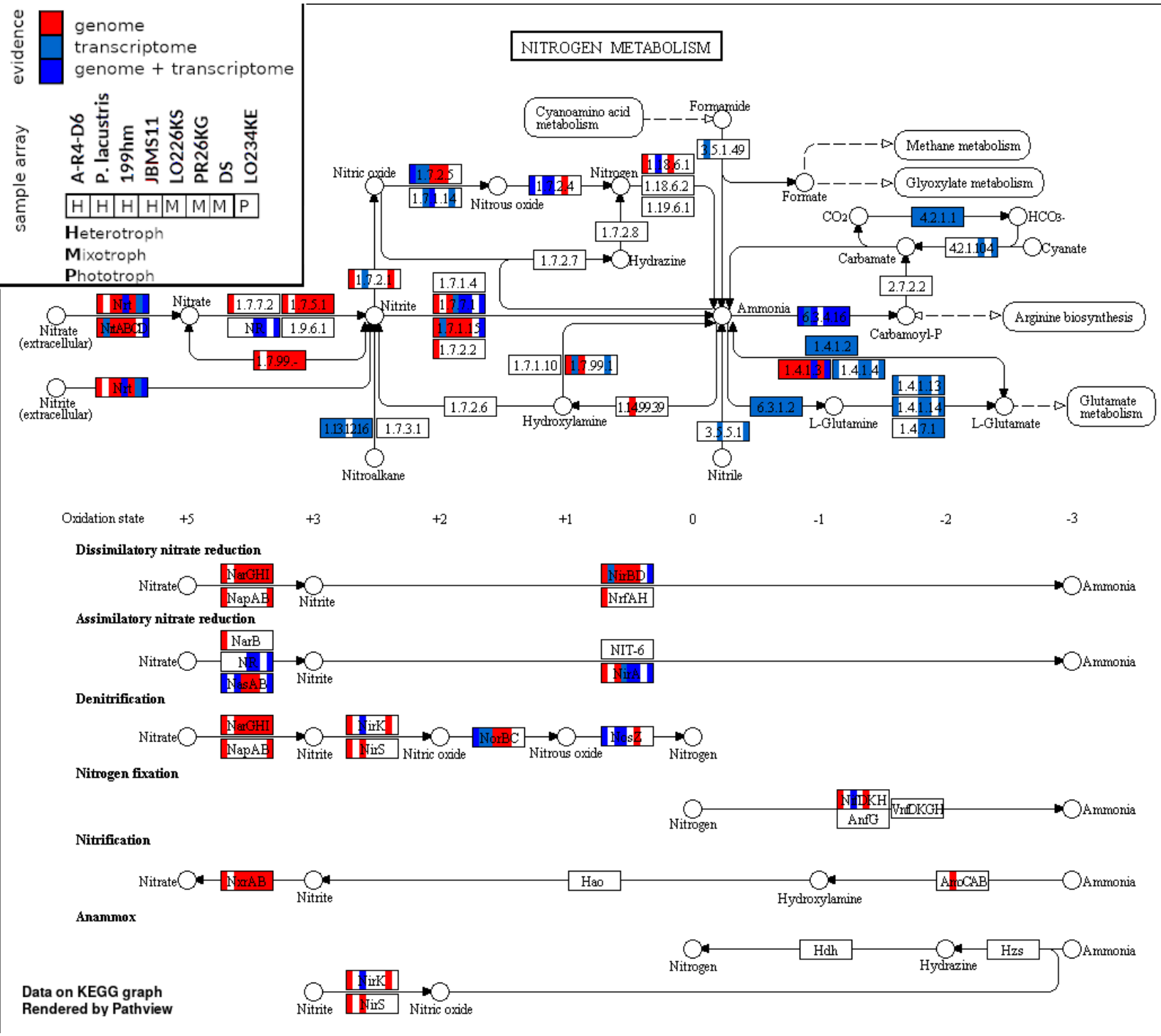


Figure S10: **Nitrogen meatbolism.** New founded genes are red marked.

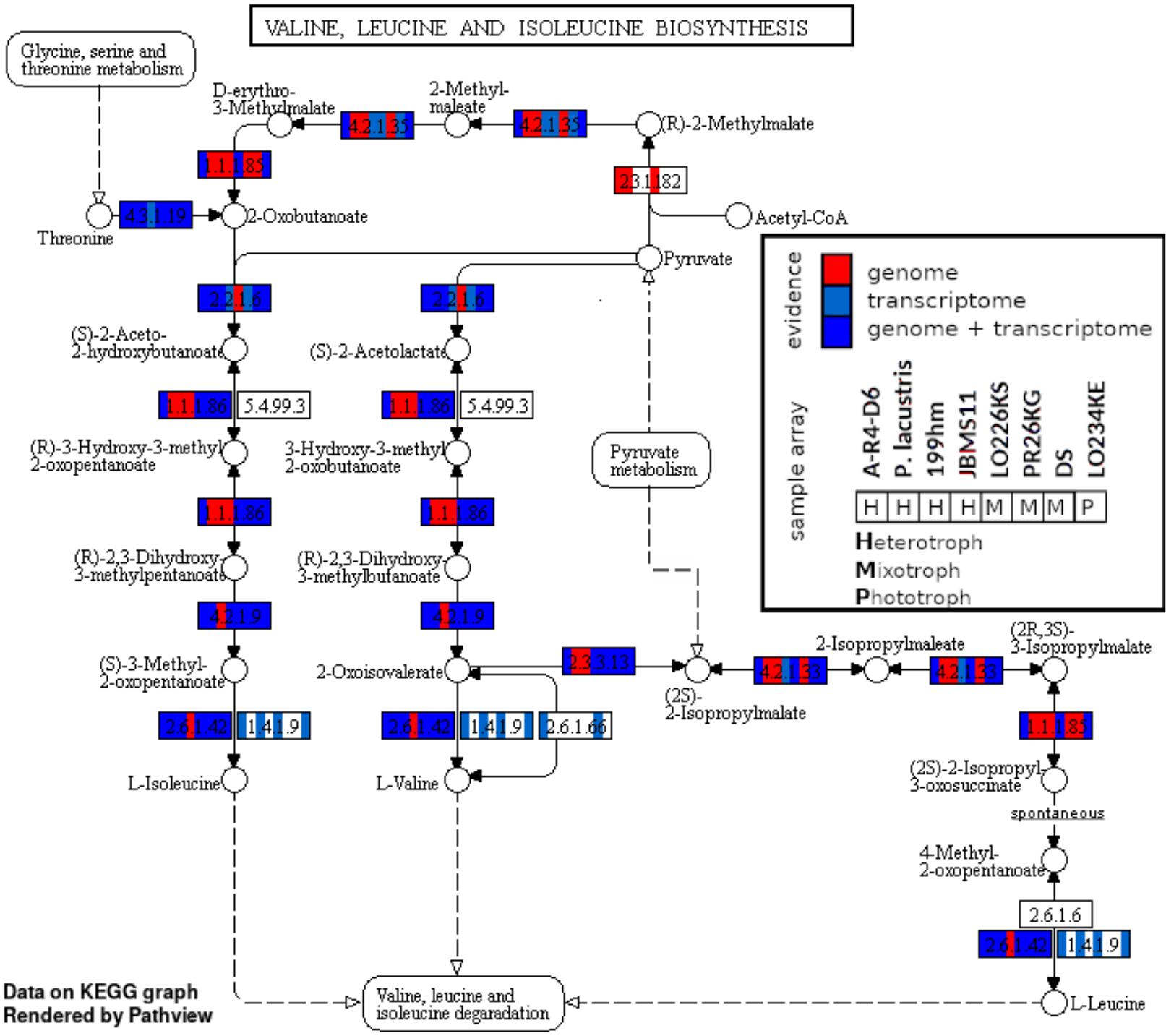


Figure S11: **Valin, leucine and isoleucine biosynthesis.** New founded genes are red marked.

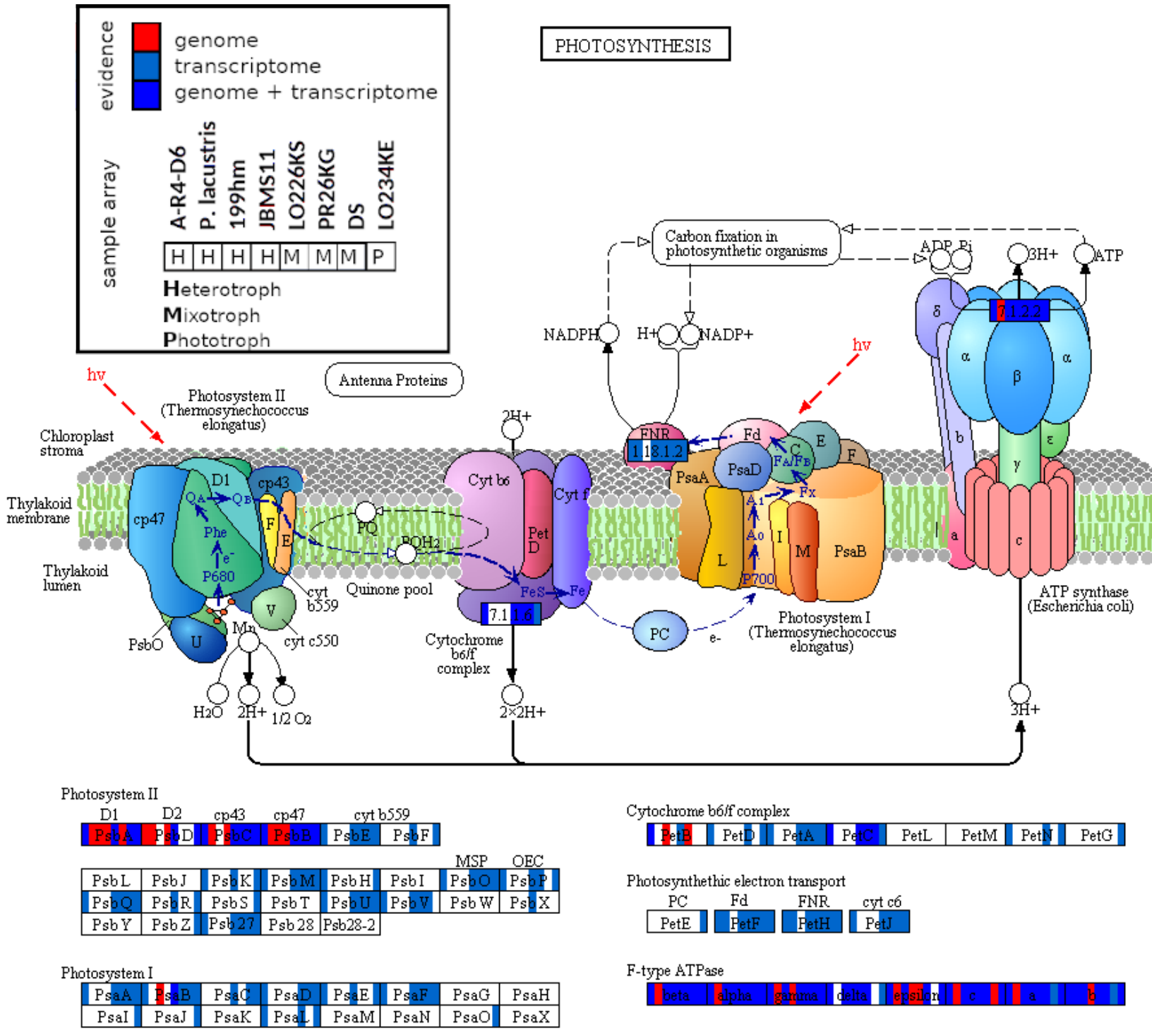


Figure S12: **Photosynthesis.** New founded genes are red marked.

Additional Publications

Contributed publications during my time as a PhD student, which were excluded as part of this thesis. Available under *Creative Commons Attribution 4.0 International License* `https://creativecommons.org/licenses/by-nc/4.0/`.

- Sczyrba, A., Hofmann, P., ... 5 other authors..., Majda, S. et al. Critical Assessment of Metagenome Interpretation–a benchmark of metagenomics software. Nat Methods 14, 1063–1071 (2017) doi:10.1038/nmeth.4458
- Fritz, A., Hofmann, P., Majda, S. et al. CAMISIM: simulating metagenomes and microbial communities. Microbiome 7, 17 (2019) doi:10.1186/s40168-019-0633-6

OPEN

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba[1,2,48], Peter Hofmann[3–5,48], Peter Belmann[1,2,4,5,48], David Koslicki[6], Stefan Janssen[4,7,8], Johannes Dröge[3–5], Ivan Gregor[3–5], Stephan Majda[3,47], Jessika Fiedler[3,4], Eik Dahms[3–5], Andreas Bremges[1,2,4,5,9], Adrian Fritz[4,5], Ruben Garrido-Oter[3–5,10,11], Tue Sparholt Jørgensen[12–14], Nicole Shapiro[15], Philip D Blood[16], Alexey Gurevich[17], Yang Bai[10,47], Dmitrij Turaev[18], Matthew Z DeMaere[19], Rayan Chikhi[20,21], Niranjan Nagarajan[22], Christopher Quince[23], Fernando Meyer[4,5], Monika Balvočiūtė[24], Lars Hestbjerg Hansen[12], Søren J Sørensen[13], Burton K H Chia[22], Bertrand Denis[22], Jeff L Froula[15], Zhong Wang[15], Robert Egan[15], Dongwan Don Kang[15], Jeffrey J Cook[25], Charles Deltel[26,27], Michael Beckstette[28], Claire Lemaitre[26,27], Pierre Peterlongo[26,27], Guillaume Rizk[27,29], Dominique Lavenier[21,27], Yu-Wei Wu[30,31], Steven W Singer[30,32], Chirag Jain[33], Marc Strous[34], Heiner Klingenberg[35], Peter Meinicke[35], Michael D Barton[15], Thomas Lingner[36], Hsin-Hung Lin[37], Yu-Chieh Liao[37], Genivaldo Gueiros Z Silva[38], Daniel A Cuevas[38], Robert A Edwards[38], Surya Saha[39], Vitor C Piro[40,41], Bernhard Y Renard[40], Mihai Pop[42,43], Hans-Peter Klenk[44], Markus Göker[45], Nikos C Kyrpides[15], Tanja Woyke[15], Julia A Vorholt[46], Paul Schulze-Lefert[10,11], Edward M Rubin[15], Aaron E Darling[19], Thomas Rattei[18] & Alice C McHardy[3–5,11]

Methods for assembly, taxonomic profiling and binning are key to interpreting metagenome data, but a lack of consensus about benchmarking complicates performance assessment. The Critical Assessment of Metagenome Interpretation (CAMI) challenge has engaged the global developer community to benchmark their programs on highly complex and realistic data sets, generated from ~700 newly sequenced microorganisms and ~600 novel viruses and plasmids and representing common experimental setups. Assembly and genome binning programs performed well for species represented by individual genomes but were substantially affected by the presence of related strains. Taxonomic profiling and binning programs were proficient at high taxonomic ranks, with a notable performance decrease below family level. Parameter settings markedly affected performance, underscoring their importance for program reproducibility. The CAMI results highlight current challenges but also provide a roadmap for software selection to answer specific research questions.

The biological interpretation of metagenomes relies on sophisticated computational analyses such as read assembly, binning and taxonomic profiling. Tremendous progress has been achieved[1], but there is still much room for improvement. The evaluation of computational methods has been limited largely to publications presenting novel or improved tools. These results are extremely difficult to compare owing to varying evaluation strategies, benchmark data sets and performance criteria. Furthermore, the state of the art in this active field is a moving target, and the assessment of new algorithms by individual researchers consumes substantial time and computational resources and may introduce unintended biases.

We tackle these challenges with a community-driven initiative for the Critical Assessment of Metagenome Interpretation (CAMI). CAMI aims to evaluate methods for metagenome analysis comprehensively and objectively by establishing standards through community involvement in the design of benchmark data sets, evaluation procedures, choice of performance metrics and questions to focus on. To generate a comprehensive overview, we organized a benchmarking challenge on data sets of unprecedented complexity and degree of realism. Although benchmarking has been performed before[2,3], this is the first community-driven effort that we know of. The CAMI portal is also open to submissions, and the benchmarks generated here can be used to assess and develop future work.

We assessed the performance of metagenome assembly, binning and taxonomic profiling programs when encountering major challenges commonly observed in metagenomics. For instance, microbiome research benefits from the recovery of genomes for individual strains from metagenomes[4–7], and many ecosystems have a high degree of strain heterogeneity[8,9]. To date, it is not clear how much assembly, binning and profiling software are influenced by the evolutionary relatedness of organisms, community complexity, presence of poorly categorized taxonomic groups (such as viruses) or varying software parameters.

A full list of affiliations appears at the end of the paper.

RECEIVED 29 DECEMBER 2016; ACCEPTED 25 AUGUST 2017; PUBLISHED ONLINE 2 OCTOBER 2017; DOI:10.1038/NMETH.4458

Fritz *et al. Microbiome* (2019) 7:17
https://doi.org/10.1186/s40168-019-0633-6

Microbiome

SOFTWARE Open Access

CAMISIM: simulating metagenomes and microbial communities

Adrian Fritz[1†], Peter Hofmann[1,2†], Stephan Majda[1,2], Eik Dahms[1,2], Johannes Dröge[1,2], Jessika Fiedler[1,2], Till R. Lesker[1,3], Peter Belmann[1,4], Matthew Z. DeMaere[5], Aaron E. Darling[5], Alexander Sczyrba[4], Andreas Bremges[1,3] and Alice C. McHardy[1,2*]

Abstract

Background: Shotgun metagenome data sets of microbial communities are highly diverse, not only due to the natural variation of the underlying biological systems, but also due to differences in laboratory protocols, replicate numbers, and sequencing technologies. Accordingly, to effectively assess the performance of metagenomic analysis software, a wide range of benchmark data sets are required.

Results: We describe the CAMISIM microbial community and metagenome simulator. The software can model different microbial abundance profiles, multi-sample time series, and differential abundance studies, includes real and simulated strain-level diversity, and generates second- and third-generation sequencing data from taxonomic profiles or de novo. Gold standards are created for sequence assembly, genome binning, taxonomic binning, and taxonomic profiling. CAMSIM generated the benchmark data sets of the first CAMI challenge. For two simulated multi-sample data sets of the human and mouse gut microbiomes, we observed high functional congruence to the real data. As further applications, we investigated the effect of varying evolutionary genome divergence, sequencing depth, and read error profiles on two popular metagenome assemblers, MEGAHIT, and metaSPAdes, on several thousand small data sets generated with CAMISIM.

Conclusions: CAMISIM can simulate a wide variety of microbial communities and metagenome data sets together with standards of truth for method evaluation. All data sets and the software are freely available at https://github.com/CAMI-challenge/CAMISIM

Keywords: Metagenomics software, Microbial community, Benchmarking, Simulation, Metagenome assembly, Genome binning, Taxonomic binning, Taxonomic profiling, CAMI

Introduction

Extensive 16S rRNA gene amplicon and shotgun metagenome sequencing efforts have been and are being undertaken to catalogue the human microbiome in health and disease [1, 2] and to study microbial communities of medical, pharmaceutical, or biotechnological relevance [3–8]. We have since learned that naturally occurring microbial communities cover a wide range of organismal complexities—with populations ranging from half a dozen to likely tens of thousands of members—can include substantial strain level diversity and vary widely in represented taxa [9–12]. Analyzing these diverse communities is challenging.

The problem is exacerbated by use of a wide range of experimental setups in data generation and the rapid evolution of short- and long-read sequencing technologies [13, 14]. Owing to the large diversity of generated data, the possibility to generate realistic benchmark data sets for particular experimental setups is essential for assessing computational metagenomics software.

CAMI, the initiative for the Critical Assessment of Metagenome Interpretation, is a community effort aiming to generate extensive, objective performance overviews of

*Correspondence: alice.mchardy@helmholtz-hzi.de
†Adrian Fritz and Peter Hofmann contributed equally to this work.
[1]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, 38124 Braunschweig, Germany
[2]Formerly Department of Algorithmic Bioinformatics, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany
Full list of author information is available at the end of the article

5.4 Bibliography

Robert A. Andersen. Synurophyceae classis nov., a new class of algae. *American Journal of Botany*, 74(3):337–353, 1987. doi: 10.1002/j.1537-2197.1987.tb08616.x. URL https://bsapubs.onlinelibrary.wiley.com/doi/abs/10.1002/j.1537-2197.1987.tb08616.x.

Robert A Andersen. *15 Molecular systematics of the Chrysophyceae and Synurophyceae*. CRC Press, 2007.

D. Antipov, A. Korobeynikov, J. S. McLean, and P. A. Pevzner. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, 32(7):1009–1015, 04 2016.

Pierre Baduel, Sian Bray, Mario Vallejo-Marin, Filip Kolář, and Levi Yant. The "polyploid hop": Shifting challenges and opportunities over the evolutionary lifespan of genome duplications. *Frontiers in Ecology and Evolution*, 6:117, 2018. ISSN 2296-701X. doi: 10.3389/fevo.2018.00117. URL https://www.frontiersin.org/article/10.3389/fevo.2018.00117.

K. P. Baetcke, A. H. Sparrow, C. H. Nauman, and S. S. Schwemmer. The relationship of dna content to nuclear and chromosome volumes and to radiosensitivity (ld50). *Proceedings of the National Academy of Sciences of the United States of America*, 58(2):533–540, Aug 1967. ISSN 0027-8424. doi: 10.1073/pnas.58.2.533. URL https://www.ncbi.nlm.nih.gov/pubmed/5233456. 5233456[pmid].

C. B. L. Barr and M. Pollard. The "historia plantarum" of john ray. *Transactions of the Cambridge Bibliographical Society*, 3(4):335–338, 1962. ISSN 00686611. URL http://www.jstor.org/stable/41155337.

Rowan D.H. Barrett and Dolph Schluter. Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1):38 – 44, 2008. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2007.09.008. URL http://www.sciencedirect.com/science/article/pii/S0169534707002868.

Arnold J. Bendich. Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays*, 6(6):279–282, 1987. doi: 10.1002/bies.950060608. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.950060608.

Marit F. Markussen Bjorbækmo, Andreas Evenstad, Line Lieblein Røsæg, Anders K. Krabberød, and Ramiro Logares. The planktonic protist interactome: where do we stand after a century of research? *The ISME Journal*, 2019. ISSN 1751-7370. doi: 10.1038/s41396-019-0542-5. URL https://doi.org/10.1038/s41396-019-0542-5.

Daniel I. Bolnick, Priyanga Amarasekare, Márcio S. Araújo, Reinhard Bürger, Jonathan M. Levine, Mark Novak, Volker H.W. Rudolf, Sebastian J. Schreiber, Mark C. Urban, and David A. Vasseur. Why intraspecific trait variation matters in community ecology. *Trends*

in Ecology & Evolution, 26(4):183 – 192, 2011. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2011.01.009. URL http://www.sciencedirect.com/science/article/pii/S0169534711000243.

M. Borodovsky and A. Lomsadze. Eukaryotic gene prediction using genemark.hmm-e and genemark-es. *Curr Protoc Bioinformatics*, Chapter 4:Unit 4 6 1–10, 2011. ISSN 1934-340X (Electronic) 1934-3396 (Linking). doi: 10.1002/0471250953.bi0406s35. URL https://www.ncbi.nlm.nih.gov/pubmed/21901742.

Keith R. Bradnam, Joseph N. Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, Inanç Birol, Sébastien Boisvert, Jarrod A. Chapman, Guillaume Chapuis, Rayan Chikhi, Hamidreza Chitsaz, Wen-Chi Chou, Jacques Corbeil, Cristian Del Fabbro, T. Roderick Docking, Richard Durbin, Dent Earl, Scott Emrich, Pavel Fedotov, Nuno A. Fonseca, Ganeshkumar Ganapathy, Richard A. Gibbs, Sante Gnerre, Élénie Godzaridis, Steve Goldstein, Matthias Haimel, Giles Hall, David Haussler, Joseph B. Hiatt, Isaac Y. Ho, Jason Howard, Martin Hunt, Shaun D. Jackman, David B. Jaffe, Erich D. Jarvis, Huaiyang Jiang, Sergey Kazakov, Paul J. Kersey, Jacob O. Kitzman, James R. Knight, Sergey Koren, Tak-Wah Lam, Dominique Lavenier, François Laviolette, Yingrui Li, Zhenyu Li, Binghang Liu, Yue Liu, Ruibang Luo, Iain MacCallum, Matthew D. MacManes, Nicolas Maillet, Sergey Melnikov, Delphine Naquin, Zemin Ning, Thomas D. Otto, Benedict Paten, Octávio S. Paulo, Adam M. Phillippy, Francisco Pina-Martins, Michael Place, Dariusz Przybylski, Xiang Qin, Carson Qu, Filipe J. Ribeiro, Stephen Richards, Daniel S. Rokhsar, J. Graham Ruby, Simone Scalabrin, Michael C. Schatz, David C. Schwartz, Alexey Sergushichev, Ted Sharpe, Timothy I. Shaw, Jay Shendure, Yujian Shi, Jared T. Simpson, Henry Song, Fedor Tsarev, Francesco Vezzi, Riccardo Vicedomini, Bruno M. Vieira, Jun Wang, Kim C. Worley, Shuangye Yin, Siu-Ming Yiu, Jianying Yuan, Guojie Zhang, Hao Zhang, Shiguo Zhou, and Ian F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, 2013. ISSN 2047-217X. doi: 10.1186/2047-217X-2-10. URL https://doi.org/10.1186/2047-217X-2-10.

J Bråte, J Fuss, S Mehrota, KS Jakobsen, and D Klaveness. Draft genome assembly and transcriptome sequencing of the golden algae hydrurus foetidus (chrysophyceae) [version 3; peer review: 2 approved]. *F1000Research*, 8(401), 2019. doi: 10.12688/f1000research.16734.3.

M. O. Carneiro, C. Russ, M. G. Ross, S. B. Gabriel, C. Nusbaum, and M. A. DePristo. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, 13:375, Aug 2012.

Te-Chin Chu, Chen-Hua Lu, Tsunglin Liu, Greg C. Lee, Wen-Hsiung Li, and Arthur Chun-Chieh Shih. Assembler for de novo assembly of large genomes. *Proceedings of the National Academy of Sciences*, 110(36):E3417–E3424, 2013. ISSN 0027-8424. doi: 10.1073/pnas.1314090110. URL https://www.pnas.org/content/110/36/E3417.

R. S. Coyne, L. Hannick, D. Shanmugam, J. B. Hostetler, D. Brami, V. S. Joardar, J. Johnson, D. Radune, I. Singh, J. H. Badger, U. Kumar, M. Saier, Y. Wang, H. Cai, J. Gu, M. W. Mather, A. B. Vaidya, D. E. Wilkes, V. Rajagopalan, D. J. Asai, C. G. Pearson, R. C. Findly, H. W. Dickerson, M. Wu, C. Martens, Y. Van de Peer, D. S. Roos, D. M. Cassidy-Hanley, and T. G. Clark. Comparative genomics of the pathogenic ciliate Ichthyophthirius multifiliis, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. *Genome Biol.*, 12 (10):R100, Oct 2011.

Charles Darwin. *On the Origin of Species by Means of Natural Selection Or the Preservation of Favoured Races in the Struggle for Life.* H. Milford; Oxford University Press, 1859.

James J Davis and Gary J Olsen. Modal codon usage: assessing the typical codon usage of a genome. *Molecular biology and evolution*, 27(4):800–810, 2009.

N.G. de Bruijn. A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, 49(7):758–764, 1946. ISSN 0370-0348.

Francisco de Castro, Ursula Gaedke, and Jens Boenigk. Reverse evolution: driving forces behind the loss of acquired photosynthetic traits. *PloS one*, 4(12):e8465, December 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0008465. URL `http://europepmc.org/articles/PMC2794545`.

Nicola De Maio, Liam P. Shaw, Alasdair Hubbard, Sophie George, Nicholas D. Sanderson, Jeremy Swann, Ryan Wick, Manal AbuOun, Emma Stubberfield, Sarah J. Hoosdally, Derrick W. Crook, Timothy E. A. Peto, Anna E. Sheppard, Mark J. Bailey, Daniel S. Read, Muna F. Anjum, A. Sarah Walker, Nicole Stoesser, and on behalf of the REHAB consortium. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics*, 5(9):e000294, 2019. doi: https://doi.org/10.1099/mgen.0.000294. URL `https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000294`.

J. del Campo and R. Massana. Emerging diversity within chrysophytes, choanoflagellates and bicosoecids based on molecular surveys. *Protist*, 162(3):435–448, Jul 2011.

Philippe Deschamps, Delphine Guillebeault, Jimi Devassine, David Dauvillée, Sophie Haebel, Martin Steup, Alain Buléon, Jean-Luc Putaux, Marie-Christine Slomianny, Christophe Colleoni, Aline Devin, Charlotte Plancke, Stanislas Tomavo, Evelyne Derelle, Hervé Moreau, and Steven Ball. The heterotrophic dinoflagellate crypthecodinium cohnii defines a model genetic system to investigate cytoplasmic starch synthesis. *Eukaryotic Cell*, 7(5):872–880, 2008. ISSN 1535-9778. doi: 10.1128/EC.00461-07. URL `https://ec.asm.org/content/7/5/872`.

R. G. Dorrell, T. Azuma, M. Nomura, G. Audren de Kerdrel, L. Paoli, S. Yang, C. Bowler, K. I. Ishii, H. Miyashita, G. H. Gile, and R. Kamikawa. Principles of plastid reductive

evolution illuminated by nonphotosynthetic chrysophytes. *Proc. Natl. Acad. Sci. U.S.A.*, 116(14):6914–6923, 04 2019.

Johannes Dröge and Alice C. McHardy. Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Briefings in Bioinformatics*, 13 (6):646–655, 07 2012. ISSN 1467-5463. doi: 10.1093/bib/bbs031. URL https://doi.org/10.1093/bib/bbs031.

Mayr Ernst. *Animal Species and Evolution*. Harvard University Press CY 1963. ISBN 978-0-674-86530-3. URL https://www.degruyter.com/view/product/251693.

C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, 281(5374):237–240, Jul 1998.

B.J. Finlay and G.F. Esteban. Freshwater protozoa: biodiversity and ecological function. *Biodiversity & Conservation*, 7(9):1163–1186, Sep 1998. ISSN 1572-9710. doi: 10.1023/A:1008879616066. URL https://doi.org/10.1023/A:1008879616066.

Bland J. Finlay. Global dispersal of free-living microbial eukaryote species. *Science*, 296(5570):1061–1063, 2002. ISSN 0036-8075. doi: 10.1126/science.1070710. URL http://science.sciencemag.org/content/296/5570/1061.

Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nature Methods*, 6(11):S6–S12, 2009. ISSN 1548-7105. doi: 10.1038/nmeth.1376. URL https://doi.org/10.1038/nmeth.1376.

Feng Gao and Chun-Ting Zhang. Comparison of various algorithms for recognizing short coding sequences of human genes. *Bioinformatics*, 20(5):673–681, 2004.

L.E. Graham, J.M. Graham, and L.W. Wilcox. *Algae*. Benjamin Cummings, 2009. ISBN 9780321559654. URL https://books.google.de/books?id=7NYUAQAAIAAJ.

Nadine Graupner, Jens Boenigk, Christina Bock, Manfred Jensen, Sabina Marks, Sven Rahmann, and Daniela Beisser. Functional and phylogenetic analysis of the core transcriptome of ochromonadales. *Metabarcoding and Metagenomics*, 1:e19862, 2017. doi: 10.3897/mbmg.1.19862. URL https://doi.org/10.3897/mbmg.1.19862.

Nadine Graupner, Manfred Jensen, Christina Bock, Sabina Marks, Sven Rahmann, Daniela Beisser, and Jens Boenigk. Evolution of heterotrophy in chrysophytes as reflected by comparative transcriptomics. *FEMS microbiology ecology*, 2018. ISSN 1574-6941. doi: 10.1093/femsec/fiy039.

Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666): 210, 2017.

L. Grossmann, C. Bock, M. Schweikert, and J. Boenigk. Small but manifold - hidden diversity in "spumella-like flagellates". *J Eukaryot Microbiol*, 63(4):419–39, 2016. ISSN 1550-7408 (Electronic) 1066-5234 (Linking). doi: 10.1111/jeu.12287. URL https://www.ncbi.nlm.nih.gov/pubmed/26662881.

Roderic Guigo, Pankaj Agarwal, Josep F Abril, Moisés Burset, and James W Fickett. An assessment of gene prediction accuracy in large dna sequences. *Genome Research*, 10 (10):1631–1642, 2000.

J. B. S. Haldane. The cost of natural selection. *Journal of Genetics*, 55(3):511, 1957. ISSN 0022-1333. doi: 10.1007/BF02984069. URL https://doi.org/10.1007/BF02984069.

Klaus Hausmann, Norbert Hülsmann, and Renate Radek. Protistology. *Acta Protozoologica*, 43:89–90, 2004.

Mahdi Heydari, Giles Miclotte, Yves Van de Peer, and Jan Fostier. Illumina error correction near highly repetitive dna regions improves de novo genome assembly. *BMC Bioinformatics*, 20(1):298, 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2906-2. URL https://doi.org/10.1186/s12859-019-2906-2.

David M Hillis. Molecular versus morphological approaches to systematics. *Annual review of Ecology and Systematics*, 18(1):23–42, 1987.

Ramana M Idury and Michael S Waterman. A new algorithm for dna sequence assembly. *Journal of computational biology*, 2(2):291–306, 1995.

Max Ingman, Henrik Kaessmann, Svante Pääbo, and Ulf Gyllensten. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813):708, 2000.

Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. High throughput ani analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1):5114, 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07641-9. URL https://doi.org/10.1038/s41467-018-07641-9.

S. W. Jeffrey, Simon W. Wright, and Manuel Zapata. *Microalgal classes and their signature pigments*, pages 3–77. Cambridge Environmental Chemistry Series. Cambridge University Press, 2011. doi: 10.1017/CBO9780511732263.004.

Roger I. Jones. Mixotrophy in planktonic protists: an overview. *Freshwater Biology*, 45(2):219–226, 2000. doi: 10.1046/j.1365-2427.2000.00672.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2427.2000.00672.x.

Ryoma Kamikawa, Naoji Yubuki, Masaki Yoshida, Misaka Taira, Noriaki Nakamura, Kenichiro Ishida, Brian S. Leander, Hideaki Miyashita, Tetsuo Hashimoto, Shigeki Mayama,

and Yuji Inagaki. Multiple losses of photosynthesis in nitzschia (bacillariophyceae). *Phycological Research*, 63(1):19–28, 2015. URL https://onlinelibrary.wiley.com/and/abs/10.1111/pre.12072.

B. Kammerlander, H. W. Breiner, S. Filker, R. Sommaruga, B. Sonntag, and T. Stoeck. High diversity of protistan plankton communities in remote high mountain lakes in the European Alps and the Himalayan mountains. *FEMS Microbiol. Ecol.*, 91(4), Apr 2015.

M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, Jan 2000.

Philipp Khaitovich, Gunter Weiss, Michael Lachmann, Ines Hellmann, Wolfgang Enard, Bjoern Muetzel, Ute Wirkner, Wilhelm Ansorge, and Svante Pääbo. A neutral model of transcriptome evolution. *PLoS biology*, 2(5):e132, 2004.

M. Kimura. The neutral theory of molecular evolution. *Sci. Am.*, 241(5):98–100, Nov 1979.

MC King and AC Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975. ISSN 0036-8075. doi: 10.1126/science.1090005. URL https://science.sciencemag.org/content/188/4184/107.

Kirsten Krause. *Plastid Genomes of Parasitic Plants: A Trail of Reductions and Losses*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-22380-8.

Jørgen Kristiansen. Cosmopolitan chrysophytes. *Systematics and Geography of Plants*, 70(2):291–300, 2000. ISSN 13747886. URL http://www.jstor.org/stable/3668648.

Morgan GI Langille, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, Deron E Burkepile, Rebecca L Vega Thurber, Rob Knight, et al. Predictive functional profiling of microbial communities using 16s rrna marker gene sequences. *Nature biotechnology*, 31(9):814, 2013.

Stephanie Lavau, Gary W. Saunders, and Richard Wetherbee. A phylogenetic analysis of the synurophyceae using molecular data and scale case morphology. *Journal of Phycology*, 33(1):135–151, 1997. doi: 10.1111/j.0022-3646.1997.00135.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0022-3646.1997.00135.x.

K. Lebret, E. S. Kritzberg, R. Figueroa, and K. Rengefors. Genetic diversity within and genetic differentiation between blooms of a microalgal species. *Environ. Microbiol.*, 14 (9):2395–2404, Sep 2012.

Zhenyu Li, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, Nan Li, Yuuki Galaxy, Liu Binghang, Bicheng Yang, and Wei Fan. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph. *Briefings in functional genomics*, 11:25–37, 12 2011. doi: 10.1093/bfgp/elr035.

Alle A. Y. Lie, Zhenfeng Liu, Ramon Terrado, Avery O. Tatters, Karla B. Heidelberg, and David A. Caron. A tale of two mixotrophic chrysophytes: Insights into the metabolisms of two ochromonas species (chrysophyceae) through a comparison of gene expression. *PloS one*, 13(2):e0192439–e0192439, Feb 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0192439. URL https://www.ncbi.nlm.nih.gov/pubmed/29438384.29438384[pmid].

Leandro Lima, Blerina Sinaimeri, Gustavo Sacomoto, Helene Lopez-Maestre, Camille Marchet, Vincent Miele, Marie-France Sagot, and Vincent Lacroix. Playing hide and seek with repeats in local and global de novo transcriptome assembly of short rna-seq reads. *Algorithms for Molecular Biology*, 12(1):2, 2017. ISSN 1748-7188. doi: 10.1186/s13015-017-0091-2. URL https://doi.org/10.1186/s13015-017-0091-2.

Carl von Linné and William T. Stearn. *Species plantarum : a facsimile of the first edition, 1753*. Printed for the Ray Society; sold by B. Quaritch, London, 1957.

RH MacArthur and EO Wilson. *The Theory of Island Biogeography*. Princeton University Press, 1967.

Francis L. Macrina, Dennis J. Kopecko, Kevin R. Jones, Deborah J. Ayers, and Sara M. McCowen. A multiple plasmid-containing escherichia coli strain: Convenient source of size reference plasmid molecules. *Plasmid*, 1(3):417 – 420, 1978. ISSN 0147-619X. doi: https://doi.org/10.1016/0147-619X(78)90056-2. URL http://www.sciencedirect.com/science/article/pii/0147619X78900562.

Catherine Mathè, Marie−France Sagot, Thomas Schiex, and Pierre Rouzé. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30 (19):4103–4117, 10 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf543. URL https://doi.org/10.1093/nar/gkf543.

Robert M. May. Tropical arthropod species, more or less? *Science*, 329(5987):41–42, 2010. ISSN 0036-8075. doi: 10.1126/science.1191058. URL https://science.sciencemag.org/content/329/5987/41.

A. C. McHardy, H. G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, 4(1):63–72, Jan 2007.

Constantin Mereschkowsky. Uber natur und ursprung der chromatophoren im pflanzenreiche. *Biologisches Centralblatt*, 25:293–604, 1905.

C. Mora, D. P. Tittensor, S. Adl, A. G. Simpson, and B. Worm. How many species are there on Earth and in the ocean? *PLoS Biol.*, 9(8):e1001127, Aug 2011.

H. J. Muller. Some genetic aspects of sex. *The American Naturalist*, 66(703):118–138, 1932. ISSN 00030147, 15375323. URL http://www.jstor.org/stable/2456922.

H. J. Muller. Our load of mutations. *American journal of human genetics*, 2(2):111–176, Jun 1950. ISSN 0002-9297. URL https://www.ncbi.nlm.nih.gov/pubmed/14771033. 14771033[pmid].

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, Mar 1970.

Masatoshi Nei. *Mutation-driven evolution*. Oxford Univ. Press, Oxford, 2013. ISBN 0199661731. URL http://digitale-objekte.hbz-nrw.de/storage/2013/07/31/file_11/5208998.pdf.

AH Neilson and RA Lewin. The uptake and utilization of organic carbon by algae: an essay in comparative biochemistry. *Phycologia*, 13(3):227–264, 1974.

Carlo Nobile, Jula Marchi, Vincenzo Nigro, Roland G Roberts, and Gian Antonio Danieli. Exon–intron organization of the human dystrophin gene. *Genomics*, 45(2):421–424, 1997.

Petr Novák, Pavel Neumann, and Jirí Macas. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC bioinformatics*, 11: 378–378, Jul 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-378. URL https://www.ncbi.nlm.nih.gov/pubmed/20633259. 20633259[pmid].

S. Ohno. An argument for the genetic simplicity of man and other mammals. *Journal of Human Evolution*, 1(6):651 – 662, 1972. ISSN 0047-2484. doi: https://doi.org/10.1016/0047-2484(72)90011-5. URL http://www.sciencedirect.com/science/article/pii/0047248472900115.

Jeffrey D Palmer. The symbiotic birth and spread of plastids: how many times and whodunit? *Journal of Phycology*, 39(1):4–12, 2003.

Adolf Pascher. über rhizopoden und palmellastadien bei flagellaten (chrysomonaden), nebst einer übersicht fiber braunen flagellaten. *Arch. Protistenk*, 24:153–200, 1912.

Adolf Pascher. über flagellaten und algen. *Berichte der Deutschen Botanischen Gesellschaft*, 32(2):136–160, 1914. doi: 10.1111/j.1438-8677.1914.tb07573.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1438-8677.1914.tb07573.x.

Pavel A Pevzner, Mark Yu Borodovsky, and Anrey A Mironov. Linguistics of nucleotide sequences ii: stationary words in genetic texts and the zonal structure of dna. *Journal of Biomolecular Structure and Dynamics*, 6(5):1027–1038, 1989.

Robert E. Priest and Jean H. Priest. Diploid and tetraploid clonal cells in culture: Gene ploidy and synthesis of collagen. *Biochemical Genetics*, 3(4):371–382, Aug 1969. ISSN 1573-4927. doi: 10.1007/BF00485721. URL https://doi.org/10.1007/BF00485721.

M. A. Quail, I. Kozarewa, F. Smith, A. Scally, P. J. Stephens, R. Durbin, H. Swerdlow, and D. J. Turner. A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, 5(12):1005–1010, Dec 2008.

J. A. Raven. Phagotrophy in phototrophs. *Limnology and Oceanography*, 42(1):198–205, 1997. URL https://aslopubs.onlinelibrary.wiley.com/and/abs/10.4319/lo.1997.42.1.0198.

Adrian Reyes-Prieto, Andreas P.M. Weber, and Debashish Bhattacharya. The origin and establishment of the plastid in algae and plants. *Annual Review of Genetics*, 41(1): 147–168, 2007. doi: 10.1146/annurev.genet.41.110306.130134. URL https://doi.org/10.1146/annurev.genet.41.110306.130134. PMID: 17600460.

Elisabeth Le Rumeur, Carole Beaumont, Christiane Guillouzo, Maryvonne Rissel, Michel Bourel, and AndrÃl' Guillouzo. All normal rat hepatocytes produce albumin at a rate related to their degree of ploidy. *Biochemical and Biophysical Research Communications*, 101(3):1038 – 1046, 1981. ISSN 0006-291X. doi: https://doi.org/10.1016/0006-291X(81)91853-2. URL http://www.sciencedirect.com/science/article/pii/0006291X81918532.

Yvan Saeys, Pierre Rouzè, and Yves Van de Peer. In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. *Bioinformatics*, 23 (4):414–420, 01 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl639. URL https://doi.org/10.1093/bioinformatics/btl639.

Craig D. Sandgren. Chrysophyte reproduction and resting cysts: A paleolimnologist's primer. *Journal of Paleolimnology*, 5(1):1–9, Jan 1991. ISSN 1573-0417. doi: 10.1007/BF00226555. URL https://doi.org/10.1007/BF00226555.

F. Sanger, A.R. Coulson, B.G. Barrell, A.J.H. Smith, and B.A. Roe. Cloning in single-stranded bacteriophage as an aid to rapid dna sequencing. *Journal of Molecular Biology*, 143(2):161 – 178, 1980. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(80)90196-5. URL http://www.sciencedirect.com/science/article/pii/0022283680901965.

M. W. Schwartz, C. A. Brigham, J. D. Hoeksema, K. G. Lyons, M. H. Mills, and P.J. van Mantgem. Linking biodiversity to ecosystem function: implications for conservation ecology. *Oecologia*, 122(3):297–305, Feb 2000. ISSN 1432-1939. doi: 10.1007/s004420050035. URL https://doi.org/10.1007/s004420050035.

A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Droge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. J?rgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvo?i?t?, L. H. Hansen, S. J. S?rensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y. W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H. H. Lin, Y. C.

Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H. P. Klenk, M. Goker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods*, 14(11):1063–1071, Nov 2017.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147(1):195–197, Mar 1981.

Ivan Sovic, Kresimir Krizanovic, Karolj Skala, and Mile Sikic. Evaluation of hybrid and non-hybrid methods for de novo assembly of nanopore reads. *Bioinformatics*, 32(17): 2582–2589, 2016. doi: 10.1093/bioinformatics/btw237.

R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.*, 6(7):2601–2610, Jun 1979.

M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, 34(Web Server issue): W435–439, Jul 2006.

S. Suzuki, R. Endoh, R. I. Manabe, M. Ohkuma, and Y. Hirakawa. Multiple losses of photosynthesis and convergent reductive genome evolution in the colourless green algae Prototheca. *Sci Rep*, 8(1):940, 01 2018.

FJR Taylor. Ecology of dinoflagellates. *The biology of dinoflagellates*, 1987.

Hanno Teeling, Anke Meyerdierks, Margarete Bauer, Rudolf Amann, and Frank Oliver Glöckner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental microbiology*, 6(9):938–947, 2004.

The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, Jan 2019. ISSN 1362-4962. doi: 10.1093/nar/gky1055. URL `https://www.ncbi.nlm.nih.gov/pubmed/30395331`. 30395331[pmid].

Carol S. Thornber. Functional properties of the isomorphic biphasic algal life cycle. *Integrative and Comparative Biology*, 46(5):605–614, 10 2006. ISSN 1540-7063. doi: 10.1093/icb/icl018. URL `https://doi.org/10.1093/icb/icl018`.

Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A. Salamov, Kevin Chen, Hwai W. Chang, Mircea Podar, Jay M. Short, Eric J. Mathur, John C. Detter, Peer Bork, Philip Hugenholtz, and Edward M. Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, 2005. ISSN 0036-8075. doi: 10.1126/science.1107851. URL `https://science.sciencemag.org/content/308/5721/554`.

Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F

Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37, 2004.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699, 02 2018. ISSN 0305-1048. doi: 10.1093/nar/gky092. URL https://doi.org/10.1093/nar/gky092.

M. Ventura, A. Petrusek, A. Miro, E. Hamrova, D. Bunay, L. De Meester, and J. Mergeay. Local and regional founder effects in lake zooplankton persist after thousands of years despite high dispersal potential. *Mol. Ecol.*, 23(5):1014–1027, Mar 2014.

M Vialli. Volume et contenu en adn par noyau. *Exp. Cell Res. Suppl*, 4:284–293, 1957.

Hugo de Vries. *Die mutationstheorie. Versuche und beobachtungen über die entstehung von arten im pflanzenreich,*, volume Bd.2 (1903). Leipzig,Veit & comp.,, 1903. URL https://www.biodiversitylibrary.org/item/123974. https://www.biodiversitylibrary.org/bibliography/11336 — "Literatur": v. 2, p. [715]-717. — 1. bd. Die entstehung der arten durch mutation.–2. bd. Elementare bastardlehre.

Pavel Škaloud, Jørgen Kristiansen, and Magda Škaloudová. Developments in the taxonomy of silica-scaled chrysophytes – from morphological and ultrastructural to molecular approaches. *Nordic Journal of Botany*, 31(4):385–402, 2013. doi: 10.1111/j.1756-1051.2013.00119.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-1051.2013.00119.x.

Ben A. Ward and Michael J. Follows. Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux. *Proceedings of the National Academy of Sciences*, 113(11):2958–2963, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517118113. URL https://www.pnas.org/content/113/11/2958.

T. Weisse. The significance of inter- and intraspecific variation in bacterivorous and herbivorous protists. *Antonie Van Leeuwenhoek*, 81(1-4):327–341, Aug 2002.

Markus Weitere and Hartmut Arndt. Structure of the heterotrophic flagellate community in the water column of the river rhine (germany). *European Journal of Protistology*, 39(3):287 – 300, 2003. ISSN 0932-4739. doi: https://doi.org/10.1078/0932-4739-00913. URL http://www.sciencedirect.com/science/article/pii/S0932473904701030.

G. Weithoff and A. Wacker. The mode of nutrition of mixotrophic flagellates determines the food quality for their consumers. *Functional Ecology*, 21(6):1092–1098, 2007. doi: 10.1111/j.1365-2435.2007.01333.x. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2435.2007.01333.x.

Alexander P. Wolfe and Peter A. Siver. A hypothesis linking chrysophyte microfossils to lake carbon dynamics on ecological and evolutionary time scales. *Global and Planetary Change*, 111:189 – 198, 2013. ISSN 0921-8181. doi: https://doi.org/10.1016/

j.gloplacha.2013.09.014. URL http://www.sciencedirect.com/science/article/pii/S0921818113002142.

Y. W. Wu, B. A. Simmons, and S. W. Singer. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–7, 2016. ISSN 1367-4811 (Electronic) 1367-4803 (Linking). doi: 10.1093/bioinformatics/btv638. URL https://www.ncbi.nlm.nih.gov/pubmed/26515820.

Jian-Rong Yang, Calum J Maclean, Chungoo Park, Huabin Zhao, and Jianzhi Zhang. Intra and interspecific variations of gene expression levels in yeast are largely neutral:(nei lecture, smbe 2016, gold coast). *Molecular biology and evolution*, 34(9):2125–2139, 2017.

Frank E. Zachos. *An Annotated List of Species Concepts*, pages 77–96. Springer International Publishing, Cham, 2016. ISBN 978-3-319-44966-1.

5.5 Misc

Lebenslauf

Der Lebenslauf ist aufgrund von Datenschutzbestimmungen nicht enthalten.

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) d) + f) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich die vorliegende Dissertation selbständig verfasst und mich keiner anderen als der angegebenen Hilfsmittel bedient, bei der Abfassung der Dissertation nur die angegeben Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen als solche gekennzeichnet habe.

Essen, den 03.02.2020 _____

Unterschrift des/r Doktoranden/in

Erklärung:

Hiermit erkläre ich, gem. § 7 Abs. (2) e) + g) der Promotionsordnung der Fakultät für Biologie zur Erlangung des Dr. rer. nat., dass ich keine anderen Promotionen bzw. Promotionsversuche in der Vergangenheit durchgeführt habe und dass diese Arbeit von keiner anderen Fakultät/Fachbereich abgelehnt worden ist.

Essen, den 03.02.2020 _____

Unterschrift des Doktoranden

Erklärung:

Hiermit erkläre ich, gem. § 6 Abs. (2) g) der Promotionsordnung der Fakultät für Biologie zur Erlangung der Dr. rer. nat., dass ich das Arbeitsgebiet, dem das Thema „Evolution in Chrysophyceae regarding the nutritional mode and intraspecific variation based on comparative genomics“ zuzuordnen ist, in Forschung und Lehre vertrete und den Antrag von Stephan Majda befürworte und die Betreuung auch im Falle eines Weggangs, wenn nicht wichtige Gründe dem entgegenstehen, weiterführen werde.

Essen, den 03.02.2020 _____

Unterschrift eines Mitglieds der Universität Duisburg-Essen