Basis of genetic adaptation to heavy metal stress in the acidophilic green alga *Chlamydomonas acidophila*

Fernando Puente-Sánchez[a], Silvia Díaz[b], Vanessa Penacho[c], Angeles Aguilera[d] and Sanna Olsson[e,f*]

Running title: Genetic adaptation of extremophilic *C. acidophila*

[a]Systems Biology Program, Centro Nacional de Biotecnología (CNB-CSIC), Calle Darwin 3, 28049, Madrid, Spain

[b]Dept. Physiology, genetics and microbiology, Complutense University of Madrid (UCM), Calle José Antonio Novais 12, 28040 Madrid, Spain

[c]Bioarray, S.L. Parque Científico y Empresarial de la UMH. Edificio Quorum III. Avenida de la Universidad s/n, 03202 Elche, Alicante, Spain

[d]Centro de Astrobiología (CSIC-INTA), Carretera de Ajalvir Km 4, 28850 Torrejón de Ardoz, Madrid, Spain

[e]INIA Forest Research Centre (INIA-CIFOR), Dept. Forest Ecology and Genetics, Carretera de la Coruña km 7.5, 28040 Madrid, Spain

[f]Dept. Agricultural Sciences, P.O. Box 27, 00014 University of Helsinki, Finland

*corresponding author : Sanna Olsson, INIA Forest Research Centre (INIA-CIFOR), Dept. Forest Ecology and Genetics, Carretera de A Coruña km 7.5, E-28040 Madrid, Spain, tel. +34 913476773, fax: +34 913476767, e-mail: sanna.olsson@helsinki.fi

Abstract

To better understand heavy metal tolerance in *Chlamydomonas acidophila*, an extremophilic green alga, we assembled its transcriptome and measured transcriptomic expression before and after Cd exposure in this and the neutrophilic model microalga *Chlamydomonas reinhardtii*. Genes possibly related to heavy metal tolerance and detoxification were identified and analyzed as potential key innovations that enable this species to live in an extremely acid habitat with high levels of heavy metals. In addition we provide a data set of single orthologous genes from eight green algal species as a valuable resource for comparative studies including eukaryotic extremophiles.

Our results based on differential gene expression, detection of unique genes and analyses of codon usage all indicate that there are important genetic differences in *C. acidophila* compared to *C. reinhardtii*. Several efflux family proteins were identified as candidate key genes for adaptation to acid environments. This study suggests for the first time that exposure to cadmium strongly increases transposon expression in green algae, and that oil biosynthesis genes are induced in *Chlamydomonas* under heavy metal stress. Finally, the comparison of the transcriptomes of several acidophilic and non-acidophilic algae showed that the *Chlamydomonas* genus is polyphyletic and that acidophilic algae have distinctive aminoacid usage patterns.

Key words: heavy metals, transcriptomics, green algae, Río Tinto, extremophiles, transposons

1 Introduction

Cd is a widespread environmental pollutant which is even at low concentrations extremely toxic to aquatic microorganisms, in particular microalgae (Brayner et al., 2011; Wang et al., 2013). In spite of its harmfulness there exist very few studies on transcriptomic alterations caused by increased levels of this or other heavy metals in green algae, (Hutchins et al., 2010; Jamers et al., 2013; Zhang et al., 2014). Cd binds to organic

molecules by forming bonds with sulfur and nitrogen, thereby inactivating proteins causing a broad range of adverse effects. It is easily absorbed and bio-accumulated by lower organisms and transferred to higher trophic levels in food chain. It has been shown to inhibit growth (Okamoto et al., 1996), chlorophyll and chloroplast synthesis (Lamai et al., 2005), cause disintegration of the cell wall as well as induce a large increase in superoxide dismutase (SOD) activity, indicative of oxidative stress (Okamoto et al., 1996). Additionally, Cd replaces zinc and selenium at the active sites of enzymes, competes with other ions in membrane transport, andecreases RNA and DNA synthesis as well as photosynthetic pigments and proteins (Prasad et al., 1999; Wang et al., 2013).

The extremophilic green alga *Chlamydomonas acidophila* grows in very acidic environments (pH 2.3-3.4). Metal sequestration in vacuoles seems to be an important mechanism in cadmium tolerance and detoxification in *C. acidophila* (Aguilera and Amils, 2005) but there is evidence that also unique genetic features in *C. acidophila* contribute to its high heavy metal tolerance (Olsson *et al*., 2015; Olsson et al., 2017). The strain analyzed in this work was isolated from Río Tinto (SW Spain), one of the most extensive examples of natural extreme acidic environments (Fernández-Remolar et al., 2003). The river has a low pH (ranging from 0.8 to 2.5) buffered by ferric iron and with high concentrations of heavy metals (Aguilera et al., 2006). These extreme conditions are produced by the metabolic activity of chemolithotrophic prokaryotes that are found in high numbers in its waters (González-Toril et al., 2003). Despite these extreme environmental conditions, Río Tinto shows an unexpected degree of eukaryotic diversity (Amaral-Zettler et al., 2011). Cd was chosen for this study due to its toxicity and also because it is found in very high concentrations in Río Tinto, with local average amounts that can reach ca. 40 mg/L (Aguilera et al., 2007).

Research on extremophilic organisms significantly contribute to our understanding of the evolution of heavy metal tolerance in plants and algae. The results enable detection of novel genes potentially useful for

biotechnology and phytoremediation of contaminated water resources. In spite of this, there is very limited genetic data available for *C. acidophila* while the genome of *C. reinhardtii* has been sequenced and annotated (Merchant et al., 2007; Manichaikul et al., 2009). For *C. reinhardtii*, there also exist several physiological, molecular, and genetic studies including experimental verification of the functionality of the predicted ORFs (Ghamsari et al., 2011). To increase genomic resources in *C. acidophila* we assembled an improved transcriptome for this non-model species. We compared it to the transcriptomes of the model microalga *Chlamydomonas reinhardtii* from the same genus and other publicly available green algal transcriptomes. To explain how *C. acidophila* is able to survive extreme environments we used transcriptomic sequencing and qRT-PCR to detect transcriptional changes caused by high Cd concentrations in *C. reinhardtii* and *C. acidophila* and identified possible adaptive key genes. The high level of genes with unknown function as well as lack of an annotated genome assembly makes the identification of important genes involved in heavy metal detoxification in *C. acidophila* challenging. In spite of these difficulties we provide new information on heavy metal tolerance in this organism, extremophiles and green algae in general.

2 Material and Methods

2.1 Sample collection, cultivation and exposure to Cadmium

Chlamydomonas acidophila strain RT46 was collected from water samples taken in 2010 at the CEM sampling station of Río Tinto (SW Spain) (Aguilera et al., 2006), and isolated to grow in the presence of antibiotics, vancomycin 50 µg/mL, cefotaxime 100 µg/mL and chloramphenicol 15 µg/mL (Sigma), on agar plates made with 0.22 µm-filtered river water. Individual colonies were transferred into K medium (Keller et al., 1987), pH 2. A strain of *Chlamydomonas reinhardtii* (CC-1374, SAG 77.81) was purchased from the Chlamydomonas Resource Center (University of Minnesota) and grown in K medium, pH 7. The K medium was supplied with the same antibiotics as the ones used for cell isolation (vancomycin 50 µg/mL, cefotaxime 100 µg/mL and chloramphenicol 15 µg/mL).

The algae were grown under ca. 70 µE s^{-1} m^{-2} irradiance provided by day-light fluorescent tubes, 16:8 h LD cycle and 22 °C of temperature. The cultivations were refreshed every two weeks in corresponding growth media and cells undergoing exponential growth were grown to be treated with metalloid solutions. To reach exponential growth 5 ml of *Chlamydomonas* cultivate was transferred into an 1 L Erlenmeyer bottle with 500 ml medium. After 10 days of growth 15 ml of cultivate was transferred into three 2 L Erlenmeyer bottles with 980 ml medium in each.

For the transcriptomic sequencing a Cd solution ($CdCl_2$ X 2 ½ H_2O) with a final concentration of 245 µM was used. Earlier studies on *Chlamydomonas* showed a peak of gene expression between three and four hours in genes involved in cadmium tolerance (Hanikenne et al., 2005; Olsson et al. 2017). Therefore time points for cell collection were set before the treatment, at 3h and 6h after Cd exposure. The cells were collected in 50 ml Falcon tubes, centrifuged for 10 min in 5000 rpm, the supernatant was discarded and the pellets frozen at -80 °C until RNA extraction. For qRT-PCR cultures were treated with following solutions: 1 µM Cd solution ($CdCl_2$ X 2 ½ H_2O), 1mM Cu ($CuSO_4$ X $5H_2O$), 10 mM Fe ($FeSO_4$ X 7H2O), 1mM As (III) ($AsNaO_2$) or 5 mM As(V) (Na_2HAsO_4) and cells were collected at 1, 3 or 24 h after exposure.

2.2 RNA extraction and sequencing

Total RNA was extracted with TRI Reagent® Solution (Ambion, Life Technologies, CA, USA) following manufacturer's protocol. RNA quality and quantity were estimated using an Agilent 2100 bioanalyzer (Agilent Technologies). RNA library preparation and high-throughput sequencing were carried out in the NGS sequencing Unit (Scientific Park Foundation, Madrid, Spain) using Illumina GAiix sequencing platform. One full lane of 75 basepair long reads for each sample was sequenced to provide sufficient coverage for a representative overview of the expression profile. The generated transcriptome library was non-normalized to allow detection of differences on the gene expression level between the different treatments and untreated cultures.

2.3 Data preprocessing *de novo* hybrid assembly

All raw transcriptomic reads were filtered and trimmed with PRINSEQ lite (version 0.18.3 (Schmieder and Edwards, 2011) in order to remove duplicates and low quality reads (using default parameters except for the following: -min qual mean 25 -derep 12, -ns max p 1 –derep 12 –lc method dust –lc threshold 7 –trim tail left 6 –trim tail right 6 –trim ns left 2 –trim ns right 2 –trim qual left 25 –trim qual right 25).

The single-end Illumina reads from *C. acidophila* obtained in this study were combined with the 454 reads obtained in Olsson et al., (2015). Paired-end Illumina reads were simulated from 454 reads by using the *run_simulate_illuminaPE_from_454ds.sh* script included in the Trinity suite. The resulting reads were subsequently normalized in silico with the *normalize_by_kmer_coverage.pl* included in Trinity. The paired-end normalized reads coming from the 454 dataset were pooled together with the single-end Illumina reads obtained in this study, and assembled with Trinity (release 2013_08_14) (Grabherr et al*.,* 2011) using Jellyfish (Marcais and Kingsford, 2011) for k-mer counting with a maximum memory of 40G, minimum contig length of 200, paired fragment length of 350 and a maximum butterfly heap space of 20G. Contigs with a BLASTn identity of more than 90% to the *E. coli*, *C. reinhardtii* and human transcriptomes were discarded.

2.4 Abundance estimation and transcriptome coverage analysis

The RSEM software package (version 1.1.18.modified) (Li and Dewey, 2011) was used to estimate the gene and isoform expression values. For *C. acidophila*, a reference transcriptome was generated from the Trinity assembly by using the RSEM commands e*xtract-transcript-to-gene-map-from-trinity* and *rsem-prepare-reference* with default parameters. For *C. reinhardtii,* the reference transcriptome v4.0 (Merchant et al., 2007) available from Phytozome (http://www.phytozome.net/) was used as a reference for estimating transcript expression. Reads from the six samples were aligned separately to the reference transcriptomes by using Bowtie (version 0.12.7) (Langmead et al., 2009) and expression values for each sample were obtained with RSEM. The resulting

expression counts were normalized with the trimmed mean of M-values method, as implemented in the edgeR package (version 2.15.0) (Robinson et al., 2010). The transcripts with a log2 fold change higher than 6 and FPKM (Fragments Per Kilobase Million) of more than 20 in at least one sample were selected for further analysis. For *C. acidophila*, only the longest transcript per Trinity subcomponent was reported.

In order to assess the coverage of each sequence in our *C. acidophila* assembly, reads from the three *C. acidophila* samples were pooled and aligned against the reference transcriptome. We used the *align-reads.pl* script included in the Trinity package (release 2013_08_14), resorting to Bowtie (version 0.12.7) to perform the alignment. The script also utilized Samtools (version 0.1.18) (Li et al., 2009) for SAM-format alignment manipulations. The output file bowtie out.coordSorted.bam, which contains both properly mapped pairs and single unpaired fragment reads, was used as input f*or Quali*map (version 0.6) (García-Alcalde et al., 2012) in order to estimate transcript coverage.

2.5 Taxonomic and functional annotation

All transcripts were annotated via BLASTx searches (Altschul et al., 1997). For taxonomic annotation GenBank's non-redundant protein database (nr) was used. For functional annotation two other major databases, Uniprot's Swiss-Prot and TrEMBL protein databases were used in addition to the nr database to get more accurate information on genetic functions. Taxonomic and functional information from the multiple databases for each differentially expressed contig was collected into a table preferring the most accurate functional annotation from Swiss-Prot when available using the methods and scripts modified from de Wit et al., (2012).

2.6 Protein prediction and orthology search with OrthoFinder across green algal transcriptomes

To identify orthologous gene groups among green algae, representative transcriptome files were downloaded from Phytozome v.11 (http://www.phytozome.net/) for six available species: *Chlamydomonas reinhardtii*,

Coccomyxa subellipsoidea, *Micromonas pusilla*, *Micromonas sp.* RCC299, *Osterococcus lucimarinus* and *Volvox carteri*. The *de novo* assemblies of *C. acidophila* and *D. acidophila* (Puente-Sánchez et al., 2016) were translated to amino acids with TransDecoder (v. 3.0.0, The Broad Institute). Orthologous sequences from these eight species were grouped with the clustering software OrthoFinder (Emms and Kelly, 2015). The resulting alignments were filtered to contain only the longest isoform of *C. acidophila* and *D. acidophila* when several isoforms of the same gene (belonging to same component and subcomponent in the *de novo* assembly built with Trinity) were present in the same orthologous group. Orthologous groups related to heavy-metal tolerance were subject to further analyses while orthologous groups representing putative single-copy nuclear genes (an orthologous group with exactly one gene / species) present in all species were used to build a phylogeny.

2.7 Genes present in *C. acidophila* but not in *C. reinhardtii*

To find an explanation for the different responses to heavy metals in extremophiles and neutrophiles, two approaches to identify genes that are present in *C. acidophila* but not *C. reinhardtii* were employed. First, screening for genes related to heavy metal tolerance and detoxification was done based on keywords in the annotation of the *C. acidophila* transcriptome. Only transcripts that had other organisms than *C. reinhardtii* as first BLAST match were included. To verify that the identified candidate genes are not present in *C. reinhardtii*, a local BLASTn search against *C. reinhardtii* transcripts was performed. Reciprocal BLAST was performed to confirm the matches and confirmed isoforms were used in downstream analyses.

Secondly, to identify genes specific to acidophiles, orthologous groups containing both of the extremophiles (*C. acidophila* and *D. acidohila*) but not *C. reinhardtii* were extracted. As a precaution to exclude contaminant sequences, in the absence of reference genomes for the extremophiles, only the orthologous groups containing at least one additional green algal species were kept. In addition transcripts with organellar annotations (mitochondrial or chloroplast) were excluded. Phylogenetic analyses were made for the transcripts with an

annotation related to heavy metal tolerance and detoxification after confirming their absence in *C. reinhardtii* by a BLASTx against nr database with a cut-off E-value of ≤10−3.

2.8 Phylogenetic analyses

Sequences were aligned with Mafft (Katoh et al., 2002). For individual genes the alignments were manually edited in PhyDE® v1.0 (Müller et al., 2005) by excluding ends of the alignments which could not be confidently aligned due to length differences and ambiguities in homology assessment. The concatenated data matrix of 488 single orthologous groups was trimmed with Trimal (Capella-Gutierrez et al., 2009) using the option -gappyout. Bayesian analyses were performed with MrBayes v3.2.1 (Ronquist et al., 2012), applying the suggested search strategies for amino acids (Huelsenbeck et al., 2001; Ronquist and Huelsenbeck, 2003). For the individual genes four runs with four chains (1×10^6 iterations each) were run simultaneously while for the concatenated matrix of 488 single orthologous groups four runs with two chains (1×10^6 iterations each) were run. Chains were sampled every 1000 generations and the respective trees written to a tree file. Calculations of the consensus tree and of the posterior probability of clades were performed based upon the trees sampled after the chains converged. The concatenated matrix was also analyzed using RAxML (Stamatakis, 2006; Stamatakis et al., 2008) defining the used model automatically with the option -m PROTGAMMAAUTO. Consensus topologies and support values from the different methodological approaches were compiled and drawn using TreeGraph2 (Stöver and Müller, 2010).

2.9 Quantitative reverse transcription PCR (qRT-PCR)

For qRT-PCR protocols established by Díaz et al., (2007) were followed, applying the modifications detailed in Olsson et al., (2017). Actin (ACT1) and 18S were used as endogenous control genes. All qRT-PCR reactions were carried out in an iQTM5 multicolor Real-Time PCR detection System (Bio-Rad) apparatus with the following cycling conditions: (i) 5 min at 95°C to denature reverse transcriptase, (ii) 40 cycles of 95°C for 30 s, 55ºC for 30

s and 72ºC after 20 s. Both NTC (no template control) and RT minus control were negative. The real-time dissociation curve was used to check primer specificity and to confirm the presence of a unique PCR product. Standard curves were obtained using 10-fold serial cDNA dilutions and determining the Ct (cycle threshold) values. The standard line parameters (amplification efficiency, slope and correlation coefficient) are reported in Table 1. Analysis of relative gene expression was carried out according to the Standard-curve quantification method (Larinov et al., 2005) from, at least, four independent experiments (each performed in duplicates). Primers for qRT-PCR were designed using the program Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi) with default settings. All primers used in this study are listed in Table 2.

2.10 Codon usage bias and GC content analyses

Complete CDSs (coding DNA sequences) were extracted from the eight algal transcriptomes (*Chlamydomonas acidophila, Dunaliella acidophila, Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea*, *Micromonas pusilla*, *Micromonas sp.* RCC299, *Osterococcus lucimarinus* and *Volvox carteri*) by using the Transdecoder software included in the Trinity suite. GC content and codon and aminoacid usage for each CDS were calculated with GCUA (General Codon Usage Analysis; McInerney, 1998). For each gene, only the longest transcript was included in the analysis.

2.11 GO-terms enriched/depleted in particular aminoacids in acidophiles versus non-acidophiles

In order to detect Gene Ontology (GO) terms with a significant enrichment/depletion of particular aminoacids in acidophiles versus non-acidophiles, the following procedure was followed. Firstly, we selected the proteins with i) less than 2% of glutamate, ii) less than 2% of aspartate, iii) more than 4% of cysteine, iv) more than 15% of serine. These proteins will henceforth be referred to as “extreme” proteins. The particular aminoacids and the percentage cutoff values were selected after inspection of the aminoacid utilization profiles shown in the figure obtained in the previous section and shown in Additional File 6. For each of the four aminoacids, we then

counted the number of appearances of each GO-term in the “extreme” proteomes and in the “non-extreme” proteomes of the two acidophilic species and the six non-acidophilic species, respectively. This was information used to build the following contingency table for each GO-term, which was subjected to the Fisher’s Exact test in order to assess whether that particular GO-term was significantly enriched ($p<0.05$) in the extreme fraction of the proteome in acidophiles versus non-acidophiles, that is, was significantly enriched/depleted in that particular aminoacid in acidophiles versus non-acidophiles.

# GO-term appearances in the extreme proteome of acidophiles	# GO-term appearances in the non-extreme proteome of acidophiles
# GO-term appearances in the extreme proteome of non-acidophiles	# GO-term appearances in the non-extreme proteome of non-acidophiles

3 Results and Discussion

3.1 High-throughput sequencing, assembly and taxonomic annotation of *C. acidophila* transcripts

Six single-end Illumina Hi-Seq libraries were sequenced in order to monitor the transcriptomic response of *Chlamydomonas reinhardtii* and *Chlamydomonas acidophila* to cadmium stress right after cadmium exposure, three hours after exposure and six hours after exposure. A total of 131,128,472 raw reads were generated, of which 66,677,308 passed quality filtering, with the duplication level being consistent with that found in other studies (Gómez-Álvarez et al., 2009).

In order to obtain a high quality draft transcriptome for *C. acidophila*, the reads obtained in this study were pooled together and co-assembled with the reads obtained in Olsson *et al*., (2015). This yielded 151449 transcripts of unique isoforms corresponding to 47411 unique Trinity subcomponents (which can be interpreted as distinct genes), with a N50 of 3212 nucleotides, and average isoform coverage of 54.62X. The pre-processing

238 and assembly statistics are summarized in Tables 3A and 3B. The hybrid assembly significantly improved the
239 assembly results and genome fraction coverage over the existing assembly from the earlier study (GenBank
240 accession GBAH00000000) for which only 454 reads were used.
241
242 **3.2 Differential expression analysis of transcripts**
243 For both species, the gene expression after 3h and 6h of cadmium exposure was compared to the gene
244 expression right before cadmium exposure. H43 (Rubinelli et al., 2002) and Cds1 (Hanikenne et al., 2005) are
245 among the few genes that have been identified to be induced by cadmium in *C. reinhardtii*. In addition a novel
246 phytochelatin synthase CaPCS2 was recently showed to be strongly induced by Cd in *C. acidophila* (Olsson et al.,
247 2017). Transcripts homologous to these genes were not found to be differentially expressed in this study,
248 possibly due to the strict cutoff values applied. The time and concentration of the exposure might also greatly
249 affect the transcriptomic response of green algae to Cd. Hanikenne et al., (2005) observed a peak of expression
250 in the half-size ABC transporter gene *Cds1* at 4 hours after 200-400 µM Cd exposure and argued that the
251 transcript levels of this gene were too low to be detected under the experimental conditions (2 h exposure to
252 25 µM cadmium) used earlier by Rubinelli et al. (2002). On the other hand, Olsson et al., (2017) reported a very
253 strong induction of the gene CaPCS2 in as low concentration as 1 µM. Furthermore, different isoforms might
254 result in different expression values.
255
256 In this study we focused on genes showing differential gene expression when exposed to very strong Cd
257 exposure. To complement the gene expression profile of selected candidate genes qRT-PCR was performed
258 using different concentrations and time points.
259
260 **3.3 Differentially expressed genes in *C. reinhardtii***

The low number of transcripts detected to be differentially expressed in *C. reinhardtii* (Additional File 1) is likely due to the high amount of Cd used in the experiment, which was chosen to give a visible effect on the transcriptomic expression in *C. acidophila*.

The transcripts with highest increase in expression after Cd exposure between control and one of the cadmium treated samples include transcripts coding for an apoptosis-inducing factor, NSG6 protein, NifU-like protein 5, and a vacuolar protein sorting-associated protein, in addition to transcripts with unknown function. Induction of stress related genes, as well as genes operating in metal uptake and export as a response to cadmium has been observed previously in *C. reinhardtii* (Jamers *et al*., 2013) as well as in cyanobacteria (Houot *et al*., 2007). Other differentially expressed genes could not be directly linked to heavy metal detoxification but can nonetheless be related to stress responses. For example, NSG6 is involved in gametogenesis and induced under nitrogen starvation (Abe *et al*., 2004). Interestingly, gametogenesis in *C. reinhardtii* leads to an increased production of lipids with use as biofuels (Miller *et al.*, 2010). Here we show that, apart from nitrogen starvation, this process can also be induced by heavy metal stress, opening the way for novel engineering strategies in the search for high oil yields

3.4 Differentially expressed genes in *C. acidophila*

The top fifty up regulated higher transcripts are summarized in Additional File 2, and included several transposable elements. Significant upregulation was observed in a transcript annotated as retrotransposon copia (FPKM in 0h 0, 3h 240.83, 6h 840.66 in comp17295_c0_seq16), a transcript annotated as retrovirus-related Pol polyprotein from transposon TNT 1-94 (the annotation varies according to isoform, highest FPKM in comp17071_c1_seq27: 0h 1.84, 3h 156.11, 6h 676.71), retrovirus-related Pol polyprotein from transposon 297 (comp18064_c0_seq15) and Transposon Ty3-I Gag-Pol polyprotein (comp16440_c0_seq16). Retrotransposons are assumed to be a major driving force for genome evolution through genome organization and gene

285 regulation in plants (Flavell *et al*. 1992), some being transferred horizontally (Cheng *et al*. 2009 and references
286 therein). There are indications that retrovirus and retrotransposons are involved in gene regulation and
287 detoxification of heavy metals. Retrovirus-related Pol polyprotein from transposon TNT 1-94 has been shown to
288 alter its methylation status in *Populus alba* when grown on heavy metal contaminated soil (Cicatelli *et al*.,
289 2014). Castrillo *et al*., (2013) showed that heavy-metal stress induced transposon activity in plants. Exposure to
290 Cd strongly increased transposon expression in *C. acidophila*, which suggests for the first time that heavy-metal
291 stress induces transposon activity also in green algae.
292
293 There are several transcripts with the annotation arsenite resistance protein ArsB among the differentially
294 expressed genes (comp14907_c0 or comp15936_c0). The annotations are partly incongruent, comp14907_c0
295 getting annotated as arsenite resistance protein ArsB or ubiquitin-like modifier-activating enzyme ATG7, while
296 the annotations for comp15936_c0 are arsenite resistance protein ArsB or arsenate reductase. The automated
297 annotation is complicated by the fact that the nomenclature of the ACR3 family ArsB protein overlaps with ArsB
298 of *E.coli* belonging to the ArsB family (Wu *et al*., 1992). It was verified from the alignments including all isoforms
299 (data not shown) that all isoforms of one component belong together, and the different annotations are due to
300 lack of highly similar sequences in GenBank of some sequence parts. To avoid confusions in this manuscript the
301 isoforms of comp14907_c0 are referred to arsenical-resistance protein ACR3 and isoforms of comp15936_c0 as
302 ACR3 family arsenite transporter based on the annotation of the consensus sequences of these isoforms. Most
303 differentially expressed transcripts of both comp15936_c0 and comp14907_c0 are strongly induced by Cd, with
304 the exception of comp15936_c0_seq52. However, according to qRT-PCR analyses the ACR3 family arsenite
305 transporter comp_15936_c0 is down-regulated by Cd (Table 4). The incongruent results between the measures
306 based on the gene expression data and the qRT-PCR are likely due to the differences in the used Cd
307 concentrations.
308

An oil globule associated protein (comp13235_c0_seq1) was detected to be induced by copper by Olsson *et al*., (2015), and is now shown to be also induced by Cd (FPKM in 0h 0, 3h 10.058, 6 h 30.682). This again highlights the role of heavy metals as inductors of oil production in *Chlamydomonas* (see previous section). Furthermore, the ability of *C. acidophila* to tolerate extreme acidity and heavy metal concentrations might help it avoid the contamination issues that commonly hamper microalgal biodiesel production (Siaut et al., 2011; Wang et al., 2016). While a detailed study of the oil production potential of *C. acidophila* is beyond the scope of this manuscript, our findings warrant further investigation on its biotechnological applications.

3.5 Species phylogeny based on orthologous sequences

We identified 488 single orthologues present in all eight species (Additional File 3), which were used to build a species phylogeny. According to the phylogenomic analyses the genus *Chlamydomonas* is not monophyletic (Fig. 1). This is not so surprising since *Chlamydomonas* is known to be polyphyletic and in need for revision, first shown by Buchheim et al., (1990) and confirmed by several later studies (e.g. Leliaert et al., 2012; Nakada et al., 2016). However, earlier phylogenies have been based on few molecular markers and now the polyphyly of *Chlamydomonas* is shown for the first time on a phylogenomic level. *Micromonas pusilla* was resolved as best root in the species tree. All clades got full support both with MrBayes and RAxML.

3.6 Identification of genes unique to *C. acidophila*

Some genes can be important in heavy metal tolerance and metal homeostasis even if their expression is not altered in the presence of the metal. Most phytochelatin synthases, for example, are known to be constitutively expressed but post-translationally activated by heavy metals in plants (Cobbett and Goldsborough, 2002; Rea et al., 2004). To better understand the mechanisms that enable *C. acidophila* to live in its extremely acid environment we therefore identified genes involved in heavy metal tolerance and detoxification that are

present in *C. acidophila* but do not have an orthologue in *C. reinhardtii*, irrespective of their expression. Two approaches were employed.

First we identified thirteen candidate genes based on annotations of the transcripts and verified by reciprocal Blast searches as explained in material and methods (Table 5). Of these, in addition to the ACR3 family members discussed above, transcripts with following annotations were up-regulated: several isoforms of mitochondrial carrier protein MTM1 (comp10226_c0), which carries manganese for the mitochondrial superoxide dismutase, and of the MATE efflux family protein DETOXIFICATION 44 (comp12911_c0). To test for changes in gene expression caused by a low Cd exposure and metal specificity, qRT-PCR was performed on a selection of these candidate genes unique to *C. acidophila* (Table 4). Cells were collected for qRT-PCR at 1, 3 or 24 h after exposure to Cd (1 µM), Cu (1mM), Fe (10 mM), As (III) (1mM) or As(V) (5 mM). Due to degraded cDNA Cd and Fe are represented by only two time points each.

Cd was noted to somewhat affect the expression of 18S and therefore the relative mRNA expression levels of target genes were normalized against the levels of actin. Surprisingly, none of the tested transcripts were detected to be significantly induced by any of the added metals but significant down-regulation can be observed in most of them (Pair Wise Fixed Reallocation Randomisation test, $p<0.01$). These incongruences could be due to known methodological caveats in RNA-seq including different expression of the different isoforms, gene duplications or artifacts in assembly and annotation (Conesa et al. 2016). They demonstrate the importance of detailed functional studies of individual genes, although automated studies with massive input can offer useful information about general trends and serve as first step for further studies.

Secondly, we employed OrthoFinder to cluster genes and detect those unique to acidophiles (*C. acidophila* and *Dunaliella acidophila*). Eighteen genes present in both extremophilic species and at least one further algal

356 species were extracted from the resulting orthologous groups (Table 6). Three of them (phytochelatin synthase
357 CaPCS2, Arsenical-resistance protein ACR3 and multidrug efflux transporter AcrB) were detected both by the
358 first method based on key word search for metal tolerance from the annotations and the second method based
359 on filtering of orthologous groups.
360
361 Some of the key candidate genes highlighted in this study have been shown to enhance heavy metal tolerance
362 in *C. acidophila* or other organisms. The phytochelatin synthase CaPCS2 was shown to be strongly induced by
363 Cd in *C. acidophila* and cloning and expression of the gene in *Escherichia coli* clearly improved its cadmium
364 resistance (Olsson et al. 2017). Cobalamin has been shown to protect against oxidative stress in the acidophilic
365 iron-oxidizing bacterium *Leptospirillum* (Ferrer et al., 2016). Arsenical-resistance protein ACR3 is suggested to
366 be a key trait to its arsenic tolerance in the arsenic hyperaccumulator *Pteris vittata* (Indriolo et al., 2010) and it
367 might similarly enhance the tolerance to heavy metals in *C. acidophila*. Arsenite resistance efflux pump ArsB,
368 which pumps arsenite and antimonite, but not arsenate or cadmium, was first described in *E. coli* (Wu et al.,
369 1992. Our results suggest that these genes could be key traits for heavy-metal hypertolerance in *C. acidophila*. It
370 has been proposed in other extremophiles as well that just a few key genes would be responsible for their
371 hypertolerance to heavy-metals, for example Fer1 in the acidophilic archaeon *Ferroplasma acidarmanus* (Baker-
372 Austin et al., 2007).
373
374 **3.7 Phylogenetic distribution of candidate key genes involved in heavy-metal hyper-tolerance in *C. acidophila***
375 Most of the transcripts not present in *C. reinhardtii* with an annotation related to heavy metal tolerance are
376 most closely related to genes in other green algae or vascular plants (Fig. 2A and Additional File 4). However,
377 some transcripts get a first Blast hit in other algae, fungi, prokaryotes and amoebozoa. The phylogenetic
378 distribution patterns in these genes can be explained by ancient gene duplications, loss in some lineages or

horizontal gene transfer, and according to our results there are more than one explanation for the origin of these genes.

The mitochondrial carrier MTM1 (comp_10226, Fig. 2B), DETOXIFICATION 44 protein (comp_12911, Additional File 4) and arsenite-antimonite efflux family (comp_15332, Additional File 4) include both green algae and Chromalveolata among the most closely related genes. The phytochelatin synthase CaPCS2 (comp 11852, Additional File 4), which is located within a clade of prokaryotic genes, has been functionally characterized and shown to likely originate from horizontal gene transfer from bacteria (Olsson *et al*., 2017). The cobalamin biosynthesis protein CobW contains two gene copies in *C. acidophila* (comp15241_c0_seq3 and comp15241_c0_seq6), of which one is similar to *C. reinhardtii* but the other is more similar to bacterial homologues (Fig. 2C).

Similarly, the genes not present in *C. reinhardtii* extracted with OrthoFinder have variable phylogenetic distribution patterns (Fig. 3). Some of the genes are nested in a clade containing mainly bacteria (e.g. dioxygenase) and could be horizontally transferred. But for others, like transmembrane protein comp9629_c0_seq1 are closely related only to green algae and gene loss is a more likely explanation for their presence in *C. acidophila* but absence in *C. reinhardtii*.

3.8 Codon code and aminoacid usage analysis

The transcripts belonging to each of the analyzed species (*Chlamydomonas acidophila*, *Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea*, *Dunaliella acidophila*, *Micromonas pusilla*, *Micromonas sp. RC299*, *Ostreococcus lucimarinus* and *Volvox carteri*) clearly clustered together with regards to their Relative Synonymous Codon Usage (Fig. 4a), showing the presence of distinct codon usage biases, even within phylogenetically close species. For the most part, those differences did not result in different aminoacid usage

(Fig. 4b). The majority of the transcripts clustered together regardless of their source organism, except for a large set of transcripts from *C. reinhardtii* and the two *Micromonas* species, which clustered independently. Both *C. acidophila* and *D. acidophila* showed similar utilization profiles for several aminoacids, particularly an enrichment in serine and cysteine, and a depletion in glutamic and aspartic acids when compared to the non-acidophilic species (Fig. 5a, Additional File 5). This depletion in Glu and Asp in acidophiles was also observed by Goodarzi et al. (2008), but their study only included bacterial and archaeal genomes. To the best of our knowledge, this is the first study which proposes that the same can also be true in eukaryotes. We further calculated which GO-terms were significantly depleted in Glu and Asp, or enriched in Cys and Ser, in the acidophilic species when compared to the non-acidophilic species (Fig. 5b, Additional File 6). In the four cases, the significant GO-terms chiefly belonged to the binding, catalytic activity, and transporter activity base categories. These four modifications (lower Glu, lower Asp, higher Cys and higher Ser contents) are likely related to optimizations for acidic environments. For example, Glu and Asp are negatively charged in neutral conditions, but become neutral at lower pHs, which cancels their ability to stabilize proteins via salt bridges (Anderson et al., 1990). On the other hand, the higher content of cystein could contribute to metal detoxification and provide extra stability via disulfide bonds.

While the differences in total amino acid usage between organisms in acidophilic environments are usually caused mostly by a limited number of amino acids (Goodarzi et al. 2008), the observed differences in codon usage might be due to stronger selection pressure for codon optimization in extreme environments. Natural selection acting through external environmental factors can shape the genomic pattern of synonymous codon usage in extremophilic prokaryotes (Lynn et al. 2002; Zeldovic 2007), and our study is the first to suggest this to be true also in eukaryotes.

The GC contents of a large amount of transcripts in all codon positions are different (Fig. 6) and can't be explained by, for example, GC-content differences in a few horizontally transferred genes. Differences in the first codon position affect the amino acid usage while differences in the third position are likely due to preferences for different synonymous codons. Our results confirm that the overall genome-wide GC content is the most significant parameter in explaining codon bias differences between organisms, suggested by Hershberg and Petrov (2008).

3.9 Conclusions

The results of this study, including the most complete published transcriptome of *C. acidophila* and a set of identified orthologous genes between eight green algae, increase the genomic information available on green algae and extremophilic eukaryotes, highlight the adaptations mechanisms used by algae to thrive in acidic environments, and provide a valuable resource for comparative studies on green algae from different habitats. Further work should focus on detailed analyses of individual genes and applied exploitation of the results, including engineering heavy metal tolerance in green algae for environmental and economic interests.

5 Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitivity (MINECO) [CGL2011-22540, AYA2011-24803]; the European Research Council (ERC) Advanced Grant [250350]. F. Puente-Sánchez was supported by the Spanish MINECO/FEDER [CTM2013-48292-C3-2-R]. We acknowledge the Data Intensive Academic Grid (DIAG) computing infrastructure (funded by National Science Foundation [0959894]) as well as CSC – Finnish IT Center for Science and the Finnish grid infrastructure (FGI) for the allocation of computational resources. Kimmo Mattila is acknowledged for help with setting up the OrthoFinder analysis pipeline. None of the co-authors declare a conflict of interest.

6 References

Abe, J., Kubo, T., Takagi, Y., Saito, T., Miura, K., Fukuzawa, H., and Matsuda, Y. (2004) The transcriptional program of synchronous gametogenesis in *Chlamydomonas reinhardtii*. Curr Genet 46: 304–315.

Aguilera, A., and Amils, R., 2005. Tolerance to cadmium in *Chlamydomonas sp*. (Chlorophyta) strains isolated from an extreme acidic environment, the Tinto River (SW, Spain). Aquatic Toxicol 75: 316–329.

Aguilera, A., Manrubia, S.C., Gómez, F., Rodríguez, N., and Amils, R. (2006) Eukaryotic community distribution and its relationship to water physicochemical parameters in an extreme acidic environment, Rio Tinto (SW, Spain). Appl Environ Microbiol 72: 5325–5330.

Aguilera, A., Zettler, E., Gomez, F., Amaral-Zettler, L., Rodríguez, N., and Amils, R. (2007). Distribution and seasonal variability in the benthic eukaryotic community of Río Tinto (SW, Spain), an acidic, high metal extreme environment. Syst Appl Microbiol 30: 531–546.

Anderson, D E., Becktel, W.J., and Dahlquist, F.W. (1990). pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. Biochem 29: 2403–2408.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acid Res 25: 3389–3402.

Amaral-Zettler, L.A., Zettler, E.R., Theroux, S.M., Palacios, C., Aguilera, A., and Amils, R. (2011) Microbial community structure across the tree of life in the extreme Río Tinto. ISME J 5: 42–50.

Baker-Austin, C., Dopson, M., Wexler, M., Sawers, G.R., Stemmler, A., Rosen, B.R., and Bond, P.L. (2007) Extreme arsenic resistance by the acidophilic archaeon '*Ferroplasma acidarmanus*' Fer1. Extremophiles 11: 425–434.

Brayner, R., Dahoumane, S.A., Nguyen, J.N., Yéprémian, C., Djediat, C., Couté, A., and Fiévet, F. (2011) Ecotoxicological studies of CdS nanoparticles on photosynthetic microorganisms. J Nanosci Nanotechnol 11: 1852–1858.

474 Buchheim, M.A., Turnel, M., Zimmer, E.A., and Chapman, R. (1990) Phylogeny of *Chlamydomonas* (Chlorophyta)
475 based on cladistic analysis of nuclear 18S rRNA sequence data. J Phycol 26: 689–699.
476 Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment
477 trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.
478 Castrillo, G., Sánchez-Bermejo, E., De Lorenzo, L., Crevillén, P., Fraile-Escaciano, A., T.C., M., *et al*. (2013) WRKY6
479 transcription factor restricts arsenate uptake and transposon activation in *Arabidopsis*. Plant Cell 25: 2944–
480 2957.
481 Cheng, X., Zhang, D., Cheng, Z., Keller, B., Ling, and H.-Q. (2009) A New Family of Ty1-copia-Like
482 retrotransposons originated in the tomato genome by a recent horizontal transfer event. Genetics 181: 1183–
483 1193.
484 Cicatelli, A., Todeschini, V., Lingua, G., Biondi, S., Torrigiani, P., and Castiglione, S. (2014) Epigenetic control of
485 heavy metal stress response in mycorrhizal versus non-mycorrhizal poplar plants. Environ Sci Pollut Res Int 21:
486 1723–37.
487 Cobbett, C.S., and Goldsbrough, P. (2002) Phytochelatins and metallothioneins: roles in heavy metal
488 detoxification and homeostasis. Annu Rev Plant Physiol Plant Mol Biol 53: 159–182.
489 Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Wojciech Szcześniak, M.,
490 Gaffney, D.J., Elo, L.L., Zhang, X., and Mortazavi, A. (2016) A survey of best practices for RNA-seq data analysis.
491 Genome Biol 17:13
492 De Wit, P., Pespeni, M.H., Ladner, J.T., Barshis, D.J., Seneca, F., Jaris, H., *et al*. (2012) The simple fool's guide to
493 population genomics via RNA-seq: an introduction to high-throughput sequencing data analysis. Mol Ecol
494 Resour 12: 1058–1067.
495 Díaz, S., Amaro, F., Rico, D., Campos, V., Benítez, L., Martín-González, A., *et al*. (2007) *Tetrahymena*
496 *metallothioneins* fall into two discrete subfamilies. PLoS ONE 2: e291

497 Emms, D.M., and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons
498 dramatically improves orthologous group inference accuracy. Genome Biol 16: 157.
499 Fernández-Remolar, D.C., Rodríguez, N., Gómez, F., and Amils, R. (2003) Geological record of an acidic
500 environment driven by the iron hydrochemistry: the Tinto River system. J Geophys Res 108: 5080–5095.
501 Ferrer, A., Rivera, J., Zapata, C., Norambuena, J., Sandoval, Á., Chávez, R., Orellana, O., and Levicán, G. (2016)
502 Cobalamin protection against oxidative stress in the acidophilic iron-oxidizing bacterium *Leptospirillum* group II
503 CF-1. Front Microbiol 7: 748.
504 Flavell, A.J., Dunbar, E., Anderson, R., Pearce, S.R., Hartley, R., and Kumar, A. (1992) Ty1-copia group
505 retrotransposons are ubiquitous and heterogeneous in higher plants. Nucl Acids Res 20:3639–44.
506 García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., *et al*. (2012) Qualimap:
507 evaluating next generation sequencing alignment data. Bioinformatics 28: 2678–2679.
508 Ghamsari, L., Balaji, S., Shen, Y., Yang, X., Balcha, D., Fan, C., *et al*. (2011) Genome-wide functional annotation
509 and structural verification of metabolic ORFeome of *Chlamydomonas reinhardtii*. BMC Genomics, 12: S4.
510 Gómez-Álvarez V, Teal T.K., and Schmidt, T.M. (2009) Systematic artifacts in metagenomes from complex
511 microbial communities. ISME J 3:1314–1317.
512 González-Toril, E., Llobet-Brossa, E., Casamayor, E.O., Amann, R., and Amils, R. (2003) Microbial ecology of an
513 extreme acidic environment, the Tinto River. Appl Environ Microbiol 69: 4853–4865.
514 Goodarzi, H., Torabi, N., Najafabadi, H.S., and Archetti, M. (2008) Amino acid and codon usage profiles: adaptive
515 changes in the frequency of amino acids and codons. Gene 407: 30–41.
516 Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., *et al*. (2011) Full-length
517 transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644–52.
518 Hanikenne, M., Motte, P., Wu, M.C.S., Wang, T., Loppes, R., and Matagne, R.F. (2005) A mitochondrial half-size
519 ABC transporter is involved in cadmium tolerance in *Chlamydomonas reinhardtii*. Plant Cell Environ 28: 863–
520 873.

521 Huelsenbeck, J.P., Ronquist, F., Nielsen, R., and Bollback, J.P. (2001) Bayesian inference of phylogeny and its
522 impact on evolutionary biology. Science 294: 2310–2314.
523 Hershberg, R., and Petrov, D.A. (2008) Selection on codon bias. Annu Rev Genet 42: 287–99.
524 Houot, L., Floutier, M., Marteyn, B., Michaut, M., Picciocchi, A., Legrain, P., *et a*l. (2007) Cadmium triggers an
525 integrated reprogramming of the metabolism of *Synechocystis* PCC6803, under the control of the Slr1738
526 regulator. BMC Genomics 8: 350.
527 Hutchins, C.M., Simon, D.F., Zerges, W., Wilkinson, K.J. (2010) Transcriptomic signatures in *Chlamydomonas*
528 *reinhardtii* as Cd biomarkers in metal mixtures. Aquat Toxicol 100: 120–127.
529 Indriolo, E., Na, G., Ellis, D., Salt, D.E., and Banks, J.O. (2010) A Vacuolar arsenite transporter necessary for
530 arsenic tolerance in the arsenic hyperaccumulating fern *Pteris vittata* is missing in flowering plants. The Plant
531 Cell 6: 2045–2057.
532 Jamers, A., Blust, R., Coen, W.D., Griffin, J.L., and Jones O.A.H. (2013) An omics based assessment of cadmium
533 toxicity in the green alga *Chlamydomonas reinhardtii*. Aquat Toxicol 126: 355–364.
534 Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence
535 alignment based on fast Fourier transform. Nucleic Acid Res 30: 3059–3066.
536 Keller, M.D., Selvin, R.C., Claus, W., and Guillard, R.R.L. (1987) Media for the culture of oceanic
537 ultraphytoplankton. J Phycol 23: 633–638.
538 Lamai, C., Kruatrachue, M., Pokethitiyook, P., Upatham, E.S., and Soonthornsarathool, V. (2005) Toxicity and
539 accumulation of lead and cadmium in the filamentous green alga *Cladophora fracta* Kützing: a laboratory study.
540 Science Asia 31: 121–127.
541 Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short
542 DNA sequences to the human genome. Genome Biol 10: R25.
543 Larinov, A., Krause, A., Miller, W. (2005) Standard curve based method for relative real time PCR data
544 processing. BMC Bioinformatics 6: 62

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., and De Clerck, O. (2012) Phylogeny and molecular evolution of the green algae. Crit Rev Plant Sci 31: 1–46.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al*., (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078–2079.

Li, B., and Dewey, C.N. (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.

Lynn, D.J., Singer, G.A.C., and Hickey, D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. Nucl Acid Res 30: 4272–4277.

Manichaikul, A., Ghamsari, L., Hom, E.F., Lin, C., Murray, R.R., Chang, R.L., *et al*. (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. Nat Methods 6: 589–592.

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770.

McInerney, J.O. (1998) GCUA: general codon usage analysis. Bioinformatics 14: 372–373.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., *et al*. (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318: 245–250.

Miller, R., Wu, G., Deshpande, R. R., Vieler, A., Gärtner, K., Li, X., *et al*. (2010). Changes in transcript abundance in *Chlamydomonas reinhardtii* following nitrogen deprivation predict diversion of metabolism. Plant Physiol 154: 1737–1752.

Müller, K., Quandt, D., Müller, J., and Neinhuis, C. (2005) PhyDE®: Phylogenetic Data Editor, version 0.995. www.phyde.de

Nakada, T., Tomita, M., Wu, J.-T., and Nozaki, H. (2016) Taxonomic revision of *Chlamydomonas* subg. *Amphichloris* (Volvocales, Chlorophyceae), with resurrection of the genus *Dangeardinia* and descriptions of *Ixipapillifera* gen. nov. and *Rhysamphichloris* gen. nov. J Phycol 52: 283–304.

568 Okamoto, O.K., Asano, C.S., Aidar, E., and Colepicolo, P. (1996) Effects of cadmium on growth and superoxide
569 dismutase activity of the marine microalga Tetraselmis gracilis (Prasinophyceae). J Phycol 32: 74–79.
570 Olsson, S., Puente-Sánchez, F., Gómez-Rodriguez, and M., Aguilera, A. (2015) Transcriptional response to copper
571 excess and identification of genes involved in heavy metal tolerance in the extremophilic microalga
572 *Chlamydomonas acidophila*. Extremophiles 19: 657–672.
573 Olsson, S., Penacho, V., Puente-Sánchez, F., Díaz, S., and Aguilera, A. (2017) Horizontal gene transfer of
574 phytochelatin synthases from bacteria to extremophilic green algae. Microbial Ecol 73: 50–60.
575 Puente-Sánchez, F., Olsson, S., and Aguilera, A. (2016). Comparative transcriptomic analysis of the response of
576 *Dunaliella acidophila* (Chlorophyta) to short-term cadmium and chronic natural metal-rich water exposures.
577 Microbial Ecol 72: 595–607.
578 Prasad, M.N.V., and Strzalka, K. (1999) Impact of heavy metals on photosynthesis. In: Heavy metal stress in
579 plants: from molecules to ecosystems. Prasad MNV, and Hagemeyer J, (eds). Berlin, Germany: Springer, pp.
580 117–128.
581 Rea, P.A., Vatamaniuk, O.K., and Rigden DJ. (2004) Weeds, Worms, and More. Papain's Long-Lost Cousin,
582 Phytochelatin Synthase. Plant Physiol 136: 2463–2474.
583 Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010) EdgeR: a Bioconductor package for differential
584 expression analysis of digital gene expression data. Bioinformatics 26: 139–140.
585 Ronquist, F.,and Huelsenbeck, J.P. (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models.
586 Bioinformatics 19: 1572–1574.
587 Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D., Darling, A., Höhna, S., *et al*. (2012) MrBayes 3.2: efficient
588 Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61: 539–542.
589 Rubinelli, P., Siripornadulsil, S., Gao-Rubinelli, F., and Sayre, R.T. (2002) Cadmium- and iron-stress-inducible gene
590 expression in the green alga *Chlamydomonas reinhardtii*: evidence for H43 protein function in iron assimilation.
591 Planta 215: 1–13.

592 Schmieder, R., and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets.
593 Bioinformatics 27: 863–864.
594 Siaut, M., Cuiné, S., Cagnon, C., Fessler, B., Nguyen, M., Carrier, P., *et al*. (2011) Oil accumulation in the model
595 green alga *Chlamydomonas reinhardtii*: characterization, variability between common laboratory strains and
596 relationship with starch reserves. BMC Biotechnol 11: 7.
597 Stamatakis, A. (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa
598 and mixed models. Bioinformatics 22: 2688–2690.
599 Stamatakis, A., Hoover, P., and Rougemont, J. (2008) A rapid bootstrap algorithm for the RAxML Web servers.
600 Syst Biol 57: 758–771.
601 Stöver BC, and Müller KF. (2010) TreeGraph 2: Combining and visualizing evidence from different phylogenetic
602 analyses. BMC Bioinformatics 11: 7.
603 Wang, L., Yang, F., Chen, H., Fan, Z., Zhou, Y., Lu, J., and Zheng, Y. (2016) Antimicrobial cocktails to control
604 bacterial and fungal contamination in *Chlamydomonas reinhardtii* cultures. Biotechniques 60: 145–149.
605 Wang, S., Zhang, D., and Pan, X. (2013) Effects of cadmium on the activities of photosystems of *Chlorella*
606 *pyrenoidosa* and the protective role of cyclic electron flow. Chemosphere 93: 230–237.
607 Wit Wu, J., Tisa, L.S., and Rosen, B.P. (1992) Membrane topology of the ArsB protein, the membrane subunit of
608 an anion-translocating ATPase. J Biol Chem 267: 12570–12576.
609 Zeldovich, K.B., Berezovsky, I.N., and Shakhnovich, E.I. (2007) Protein and DNA Sequence Determinants of
610 Thermophilic Adaptation. PLoS Comput Biol 3: e5.
611 Zhang, W., Tana, N., and Li, S.F. (2014) NMR-based metabolomics and LC-MS/MS quantification reveal metal-
612 specific tolerance and redox homeostasis in Chlorella vulgaris. Mol Biosyst 10: 149–160.

FIGURE CAPTIONS

Figure 1. Phylogenetic relationships based on 488 nuclear single orthologous genes clustered with OrthoFinder and present in all eight species (*Chlamydomonas acidophila, Dunaliella acidophila, Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea*, *Micromonas pusilla*, *Micromonas sp.* RCC299, *Osterococcus lucimarinus* and *Volvox carteri*). The trees represent the majority consensus of trees sampled after stationarity in the Bayesian analysis. Posterior probability values from the Bayesian inference are indicated above, the corresponding bootstrap values of the maximum likelihood analysis below the branches.

Figure 2. Simplified phylogenetic analyses of transcripts coding for genes with an annotation related to heavy metal tolerance and present in *C. acidophila* but not in *C. reinhardtii*. The phylograms represent the majority consensus of trees sampled after stationarity in the Bayesian analysis. PP values equal or greater than 0.50 are shown above branches. The scale bar indicates relative distance between different sequences based on mutation rate. A) peroxisome isogenesis protein comp_10128 B) mitochondrial carrier comp 10226 C) cobalamin biosynthesis protein CobW comp_15241.

Figure 3. Simplified phylogenetic analyses of transcripts coding for genes that are involved in heavy metal tolerance and are present in *C. acidophila* but not in *C. reinhardtii* extracted from the results from the search for orthologous genes with OrthoFinder. The phylograms represent the majority consensus of trees sampled after stationarity in the Bayesian analysis. PP values equal or greater than 0.50 are shown above branches. The scale bar indicates relative distance between different sequences based on mutation rate. A) 2-hydroxyacyl-CoA lyase comp18202_c0_seq7 B) Dioxygenase comp13804_c0_seq4 C) Transmembrane protein 230 comp9629_c0_seq1 D) Cocaine esterase comp3348_c0_seq1 E) SDR-family protein with acetoacetyl-CoA reductase activity comp_14433_c0_seq1.

Figure 4. a) Correspondence analysis showing the distribution of transcripts (points) according to their Relative Synonymous Codon Usage distribution and b) Correspondence analysis showing the distribution of transcripts (points) according to the aminoacid usage bias of their predicted ORFs. The percentage of inertia explained by each axis is indicated in the axis caption. Transcripts are coloured by their source genome.

Figure 5. a) Distribution of Glu, Asp, Cys and Ser contents in the predicted ORFs from the eight species analyzed in this study. The ORFs included in the green area (low Glu, low Asp, high Cys and high Ser) were subjected to a GO-term enrichment analysis between the two acidophilic species and the rest. b) Summary of the Molecular Function GO-terms found to be significantly enriched ($p<0.05$) in acidophiles versus non-acidophiles, when focusing in the low Glu, low Asp, high Cys, and high Ser fractions of the proteomes. Full results are provided in Additional File 6.

Figure 6. Scatterplots showing the GC content on each transcript (points) in the three different codon positions. Transcripts are coloured by their source genome.

TABLES WITH CAPTIONS

Table 1. Quantitative real-time RT-PCR standard-curve parameters for selected transcripts present in C. *acidophila* but not in *C. reinhardtii* and the expression control (housekeeping) genes 18S rNA and actin. S= slope, R^2= correlation coefficient, E= amplification efficiency.

Gene	**S**	**R2**	**E**
ACR3	-2.858	0.99	2.24
Arsenite transporter	-2.885	0.97	2.22
AcrB	-3.04	0.99	2.13
Glutathione-regulated potassium-efflux family	-2.823	1	2.26
MATE efflux protein	-3.079	0.99	2.11
Arsenite-antimonite efflux family	-2.96	0.96	2.18

Table 2. Primers used for quantitative real-time RT-PCR used in this study. For each region, forward (F) and reverse (R) primers are indicated, as well as product size.

Gene	**Primer name**	**5' Sequence 3'**	**F/R**	**Product size (bp)**
Multidrug efflux transporter AcrB comp_16471	comp16471_AcrB-F	GTAGGCATTCCCTTGCTGTC	F	89
	comp16471_AcrB-R	CCAAGGACCAAAACAAGCAT	R	
ACR3 comp_14907	comp14907_ACR3-F	ACTTTTGGCTTCTGGGAGGT	F	106
	comp14907_ACR3-R	TTTCACCATAAGCCCAGACC	R	
Arsenite transporter comp_15936	comp15936_ArsB-F	AATGTTACGGCAAAGCGAAC	F	100
	comp15936_ArsB-R	CAGTCACTGGCGAGCTCATA	R	
MATE efflux protein comp_12911	comp12911_MATE-F	ACTTTGGGTTCATGGCTTTG	F	98
	comp12911_MATE-R	CACTCCTGCCAGTCCTAACC	R	
Arsenite-antimonite efflux family comp_15332	comp15332_MATE-F	CTAACACTCCTGTGGCAGCA	F	125
	comp15332_MATE-R	CAGCCTGTTAAGCCCTTTTG	R	
Glutathione-regulated potassium-efflux family comp_16013	comp16013_K-efflux-F	CGCTAGAAATTCCCAACCAG	F	87
	comp16013_K-efflux-R	GCATTTCTTTGCACCTCCAT	R	

Table 3. Sequence statistics on A) Illumina sequencing and for comparison, statistics on 454 reads from Olsson et al., (2015) are also shown. B) transcriptome assembly with Trinity using a hybrid assembly strategy combining 454 reads with Illumina reads.

Illumina sequencing					
Library	**Condition**	**Raw reads**	**Input bases (Gb)**	**Trimmed reads**	**Discarded sequences (including duplicates)**
J1	Reinhardtii-0h	21243002	1.61	451530	9504620
J2	Reinhadtii-3h	21819884	1.66	459756	10645037
J3	Reinhadtii-6h	19474228	1.48	438115	9709139
J4	Acidophila-0h	23656624	1.80	477750	11356978
J5	Acidophila-3h	22928585	1.74	503621	11287054
J6	Acidophila-6h	22006149	1.67	466332	11948336
454 data from Olsson *et al.,* (2015)					
Input 454 reads		**Simulated Illumina reads**		**Normalized simulated pairs**	
1021062		458306001		7717263	

B)

Hybrid assembly	
Input pairs	7717263
Input SE reads	33998990
Bases in assembly (Mb)	293
Trinity genes	47411
Trinity isoforms	151449
Isoform median length	1398
Isoform mean length	1936.66
Range of isoform lengths	201-19360
Isoform N50	3212
Isoform mean coverage	54.62X
Isoform std coverage	107.15X
Isoforms after filtering	129188
Isoforms after filtering with nr BLAST matches	87676

Table 4. Results of gene expression analysis by qRT-PCR of selected genes present in *C. acidophila* but not in *C. reinhardtii*. The cells were collected at 1, 3 or 24 hours after exposure to 1 µM Cd solution ($CdCl_2$ X 2 ½ H_2O), 1mM Cu ($CuSO_4$ X $5H_2O$), 10 mM Fe ($FeSO_4$ X 7H2O), 1mM As (III) ($AsNaO_2$) or 5 mM As(V) (Na_2HAsO_4). The relative

mRNA expression levels of target genes were normalized against the levels of actin and 18S. The fold induction and SD for each target gene is shown. ACR3 comp14907 = arsenical-resistance protein ACR3, ACR3 comp15936 = ACR3 family arsenite transporter, MATE comp12911 = MATE efflux protein, Arsenite-antimonite comp15332 = Arsenite-antimonite efflux family, AcrB comp16471 = Multidrug efflux transporter AcrB, MATE comp12911 = MATE efflux protein. Nd = Not defined.

	ACR3 comp14907	**ACR3 comp15936**	**MATE comp12911**	**Arsenite-antimonite comp15332**	**AcrB comp16471**	**MATE comp12911**
Cd 1h	-12,98 ± 1,97	-513 ± 126,1	-675,6 ± 79,1	-2327 ± 574	-761 ± 112	-8,1 ± 2,5
Cd 24h	-6,36 ± 0,8	-14,3 ± 3,6	-20,62 ± 3,0	-2862 ± 599	-609 ± 113	-12,5± 3,3
As(III) 1h	0,52 ± 0,06	-1,19 ± 0,09	-1,04 ± 0,3	-10,2 ± 1,4	-13,5 ± 2	-0,7 ± 0,1
As(III) 3h	-293,79 ± 36,59	-1824 ± 138	-731,7 ± 135,4	-15158 ± 1135	-3142 ± 378	-4,2 ± 1,2
As(III) 24h	-11,7 ± 1,45	-3445 ± 261	-314,9 ± 46,1	-3531 ± 864	-476 ± 70	-18 ± 5,4
Cu 1h	4,15 ± 0,52	-3687 ± 1012	3,38 ± 1,1	-3,8 ± 1,2	-6639 ± 786	-55 ± 11,7
Cu 3h	4,65 ± 0,5	-6358 ± 2424	5,6 ± 0,4	-11,48 ± 3,3	-690 ± 81	-1,5 ± 0,4
Cu 24h	1,04 ± 0,12	-170,2 ± 35,7	-26,02 ± 3,9	-554,9 ± 89	-207 ± 30	-47 ± 12
Fe 3h	Nd	Nd	1,31 ± 0,2	-819,8 ± 94,7	Nd	Nd
Fe 24h	0,36 ± 0,05	1,18 ± 0,2	0,8 ± 0,1	-2,2 ± 0,3	Nd	Nd

Table 5. Transcripts coding for genes that are involved in heavy metal tolerance present in *C. acidophila* but not in *C. reinhardtii* based on transcript annotations.

Contig name	**Putative function**	**BLAST top match organism**	**BLAST match accession**	**E-value**
comp10128_c0_seq1	Peroxisome isogenesis	*Coccomyxa subellipsoidea*	XP_005647114	3.01E-39
comp10226_c0_seq5	Mitochondrial carrier	*Coccomyxa subellipsoidea*	XP_005652123	3.95E-27
comp11852_c0_seq1	Phytochelatin synthase	*Calothrix sp.*	YP_007140091	6.44E-30
comp12911_c0_seq40	Protein DETOXIFICATION 44	*Chlorella variabilis*	EFN56963	6.62E-22
comp13602_c0_seq11	NRAMP family protein	*Volvox carteri*	XP_002947173	5.89E-153
comp14042_c0_seq2	ABC-ATPase	*Coccomyxa subellipsoidea*	XP_005643834	1.62E-92
comp14907_c0_seq53	Arsenical-resistance protein ACR3	*Coccomyxa subellipsoidea*	XP_005649016	1.99E-29
comp15241_c0_seq3	Cobalamin biosynthesis CobW	*Chlamydomonas reinhardtii*	XP_001699037	7.26E-60
comp15241_c0_seq6	Cobalamin biosynthesis CobW	*Burkholderia vietnamiensis*	YP_001117931	1.28E-80
comp15332_c0_seq3	Arsenite-antimonite efflux family	*Guillardia theta*	EKX52062	3.20E-66
comp15936_c0_seq52	ACR3 family arsenite transporter ArsB	*Coccomyxa subellipsoidea*	XP_005649501	6.10E-54
comp16013_c0_seq1	Glutathione-regulated potassium-efflux system	*Volvox carteri*	XP_002953483	8.63E-37
comp16471_c0_seq3	Multidrug efflux transporter AcrB	*Zea mays*	AFW59203	4.15E-58
comp17557_c1_seq9	Multidrug resistance-associated protein	*Coccomyxa subellipsoidea*	XP_005651467	1.59E-144

Table 6. Transcripts coding for genes that are involved in heavy metal tolerance present in *C. acidophila* but not in *C. reinhardtii* filtered from orthologous groups defined with OrthoFinder. Orthologous groups with annotations related to heavy metal tolerance and detoxification are marked with *. In the case of groups including several transcripts, the Blast hit organism and accession refers to the first one.

Orthologous group	**Contig name**	**Putative function**	**BLAST top match organism**	**BLAST match accession**	**E-value**
OG0001276	comp13064_c0_seq2, comp15004_c0_seq1	Tripeptidyl-peptidase 1	*Polysphondylium pallidum*	EFA84081	1.38e-17
OG0001752	comp12567_c0_seq1	Amino acid permease 2	*Capsella rubella*	EOA20485	3.33e-60
OG0001782	comp15790_c0_seq1	Alpha-1,3-glucosyltransferase	*Coccomyxa subellipsoidea*	XP_005651392	9.52e-92
OG0003420	comp16380_c0_seq1	Metal-nicotianamine transporter	*Amborella trichopoda*	ERN09450	5.88e-32
OG0003495	comp18202_c0_seq7	2-hydroxyacyl-CoA lyase	*Galdieria sulphuraria*	XP_005708092	0.0
OG0004374	comp18062_c0_seq1	Abhydrolase domain-containing protein	*Dictyostelium purpureum*	XP_002957250	2.33e-05
OG0004475*	comp11852_c0_seq1	Phytochelatin synthase	*Calothrix sp.*	YP_007140091	6.44E-30
OG0005070	comp16077_c0_seq9	G-box-binding factor 1	*Brassica napus*	CAA58774	1.33e-10
OG0005487*	comp14907_c0_seq15	Arsenical-resistance protein ACR3	*Coccomyxa subellipsoidea*	XP_005649016	1.45e-85
OG0005928*	comp13804_c0_seq4	dioxygenase	*Volvox carteri*	XP_002957190	2.32e-53
OG0006489	comp13735_c0_seq3	Snurportin-1	*Physcomitrella patens*	XP_001763666	5.34e-49
OG0006590*	comp16471_c0_seq1	multidrug efflux transporter AcrB	*Arabidopsis thaliana*	OAP00250	5.35e-63
OG0007003	comp14473_c0_seq1	Ankyrin-1	*Aegilops tauschii*	EMT31987	3.47e-56
OG0007890*	comp9629_c0_seq1	Transmembrane protein 230	*Physcomitrella patens*	XP_001772694	2.89e-17
OG0008459*	comp3348_c0_seq1	Cocaine esterase	*Achromobacter xylosoxidans*	WP_006387564	1.35e-80
OG0009876*	comp14433_c0_seq1	SDR-family protein with acetoacetyl-CoA reductase activity	*Sphingobium japonicum*	YP_003545425	9.92e-41
OG0010052	comp17871_c0_seq9	Hisitidine kinase	*Synechocystis sp.*	WP_009631601	1.298e-39

SUPPLEMENTARY MATERIAL

Additional File 1. *C. reinhardtii* transcripts with a log2 fold change higher than 6 and FPKM of more than 20 in at least one sample. Normalized expression of the transcripts in each condition (0h, 3h and 6h after exposure to cadmium), their putative annotation and sequence length are shown. Transcripts are ordered by FPKM values at 6h.

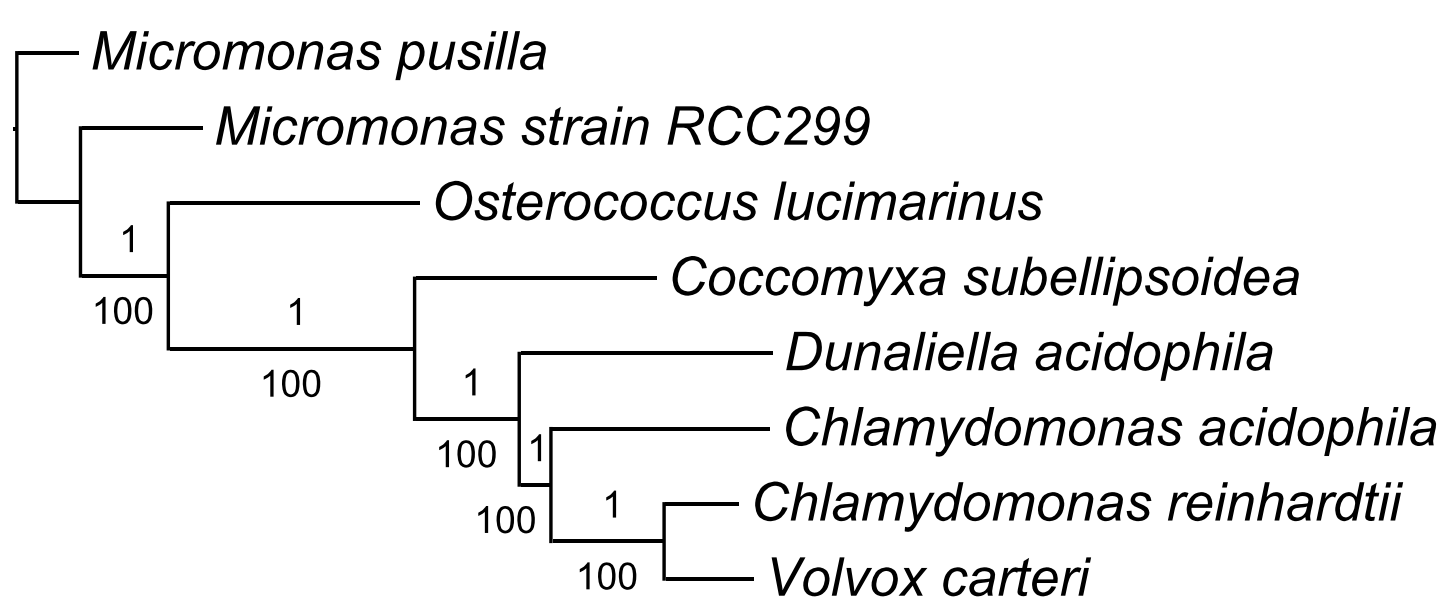
Additional File 2. *C. acidophila* RT46 transcripts with a log2 fold change higher than 6 and FPKM of more than 20 in at least one sample. Normalized expression of the transcripts in each condition (0h, 3h and 6h after exposure to cadmium) and their putative annotation are shown. Transcripts are ordered by FPKM values at 6h.

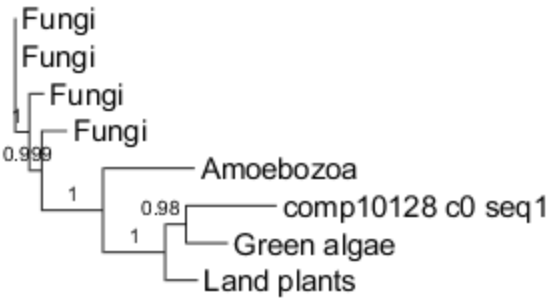
Additional File 3. Data matrix used for the phylogenomic analysis, containing 488 single orthologous groups of proteins clustered with OrthoFinder present in all species *(Chlamydomonas acidophila, Chlamydomonas reinhardtii, Coccomyxa subellipsoidea, Dunaliella acidophila, Micromonas pusilla, Micromonas sp. RC299, Osterosoccus lucimarinus and Volvox carteri)*.

Additional File 4. Simplified phylogenetic analyses of transcripts coding for genes with an annotation related to heavy metal tolerance and present in *C. acidophila* but not in *C. reinhardtii*. The trees represent the majority consensus of trees sampled after stationarity in the Bayesian analysis. PP values equal or greater than 0.50 are shown above branches. A) ACR3 comp_14907 and comp_15936 B) DETOXIFICATION 44 protein comp_12911 C) NRAMP family comp_13602 D) ABC-ATPase comp_14042 E) Arsenite-antimonite efflux family comp_15332 F) Glutathione-regulated potassium-efflux comp_16013 G) Multidrug efflux transporter AcrB comp_16471 H) Multidrug resistance-associated protein comp_17557.

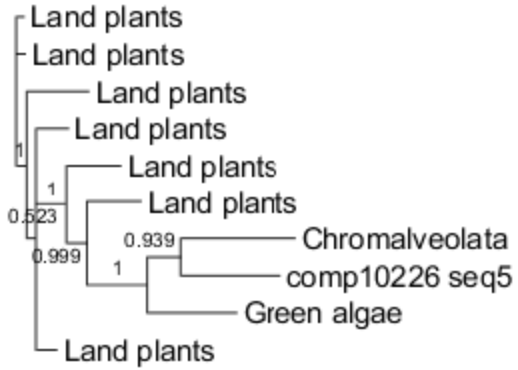
Additional File 5. Aminoacid utilization profiles for the different species. There is one cumulative frequency plot for each aminoacid.

Additional File 6. GO terms enriched in acidophiles versus non-acidophiles in the low-Glu, low-Asp, high-Cys, and high-Ser fractions of the proteome.

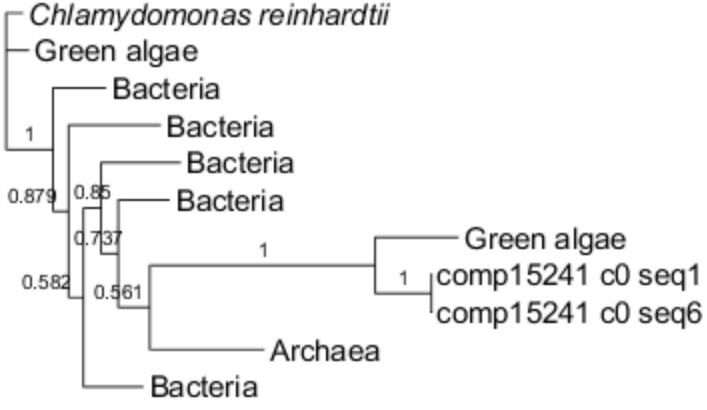

Micromonas pusilla
Micromonas strain RCC299
Osterococcus lucimarinus
Coccomyxa subellipsoidea
Dunaliella acidophila
Chlamydomonas acidophila
Chlamydomonas reinhardtii
Volvox carteri
1
100
1
100
1
100
1
100
1
100


Fungi
Fungi
Fungi
Fungi
1
0.999
Amoebozoa
1
0.98
comp10128 c0 seq1
1
Green algae
Land plants

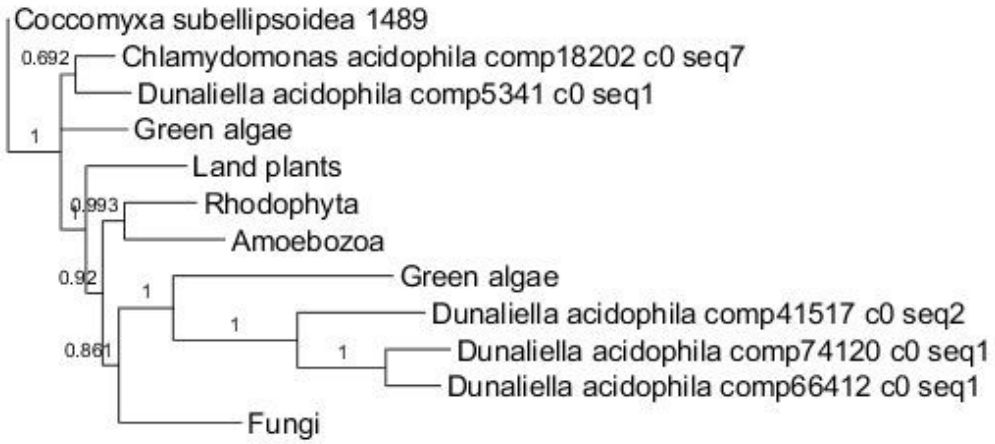
A


Land plants
Land plants
Land plants
Land plants
1
Land plants
1
Land plants
0.523
0.939
Chromalveolata
0.999
1
comp10226 seq5
Green algae
Land plants

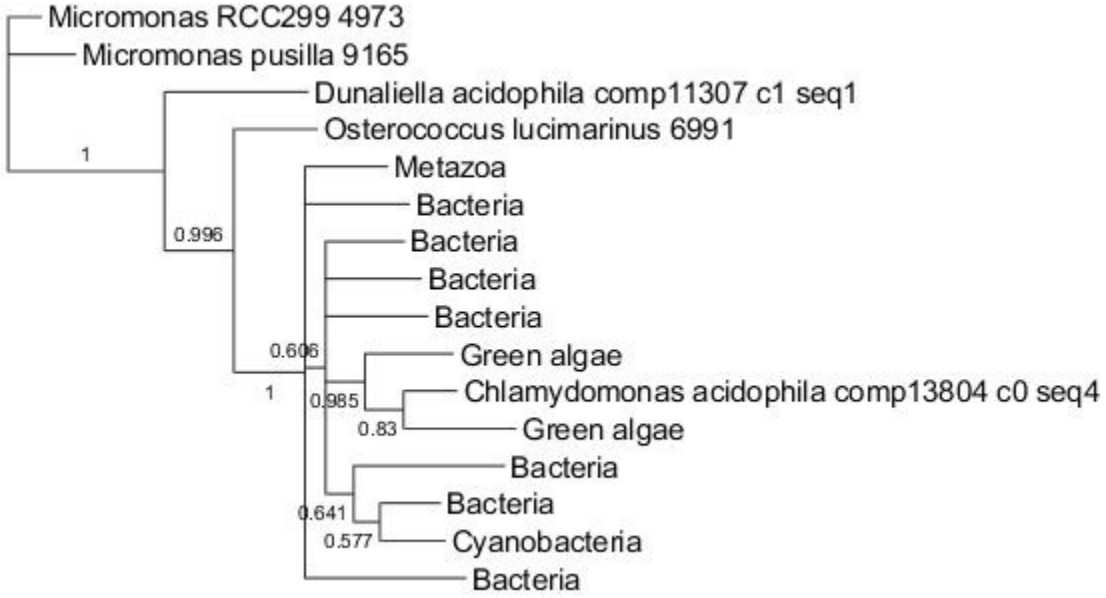
B


Chlamydomonas reinhardtii
Green algae
Bacteria
Bacteria
1
Bacteria
0.879
0.85
Bacteria
0.737
Green algae
1
0.582
0.561
1
comp15241 c0 seq1
comp15241 c0 seq6
Archaea
Bacteria

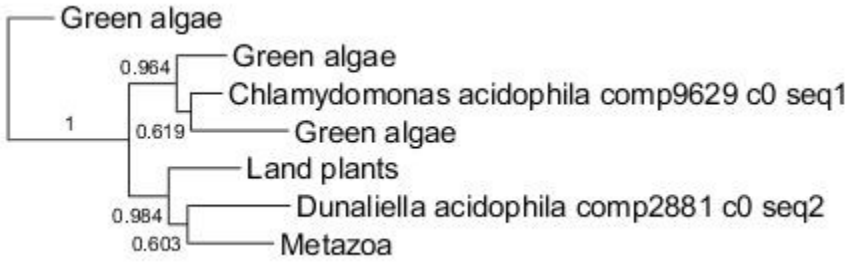
C



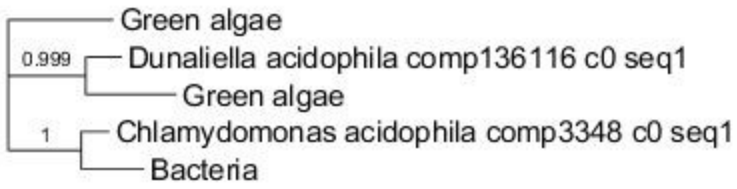
A) OG0003495: 2-hydroxyacyl-CoA lyase comp18202_c0_seq7



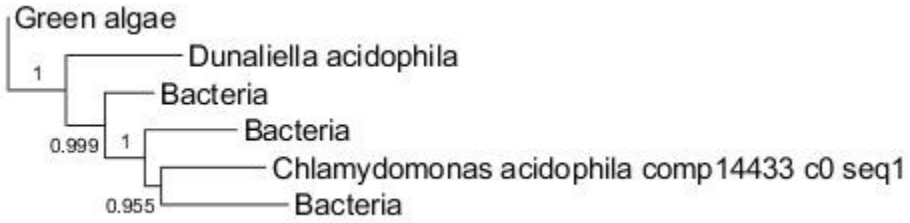
B) OG0005928: Dioxygenase comp13804_c0_seq4



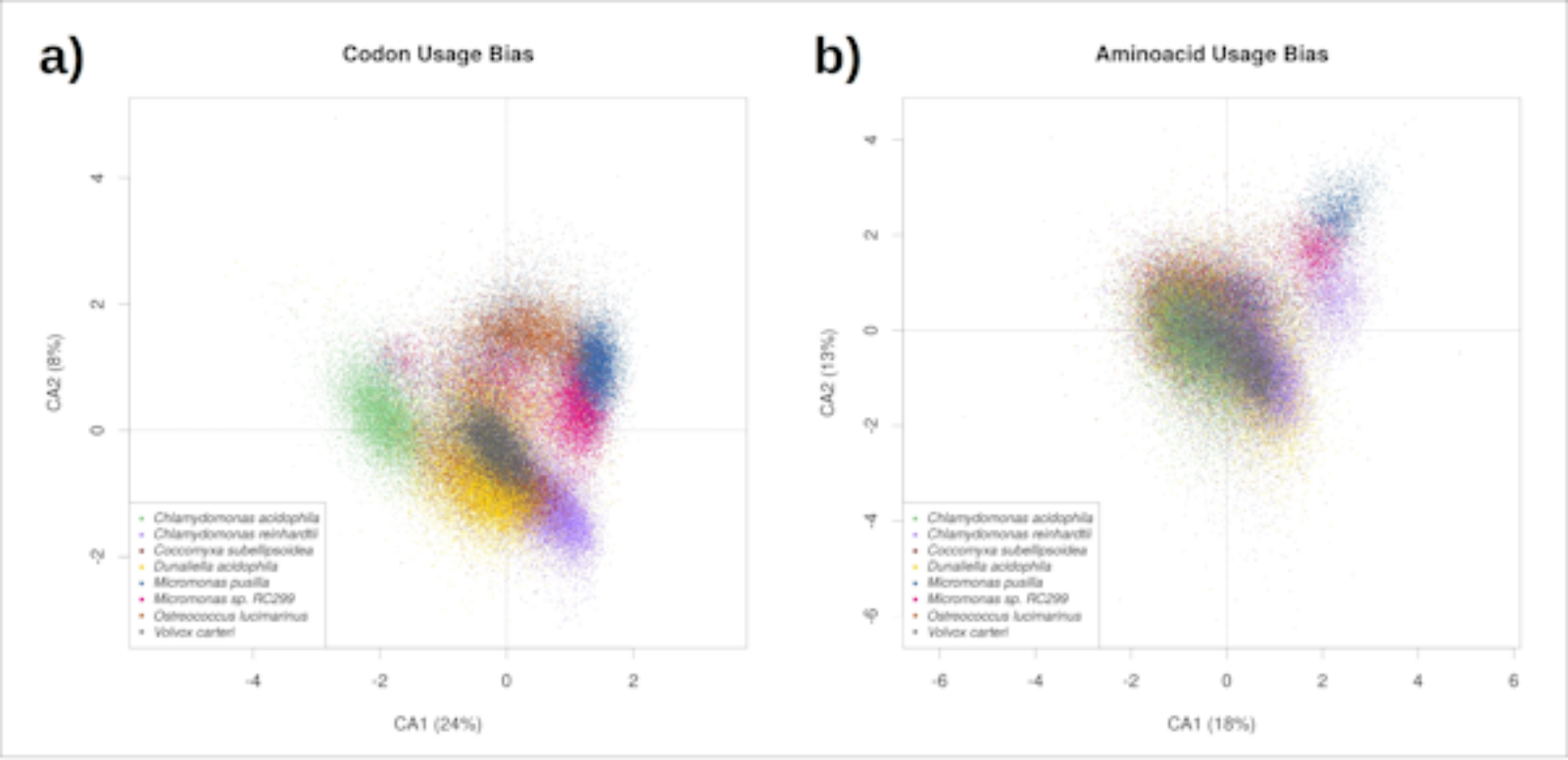
C) OG0007890: Transmembrane protein 230 comp9629_c0_seq1

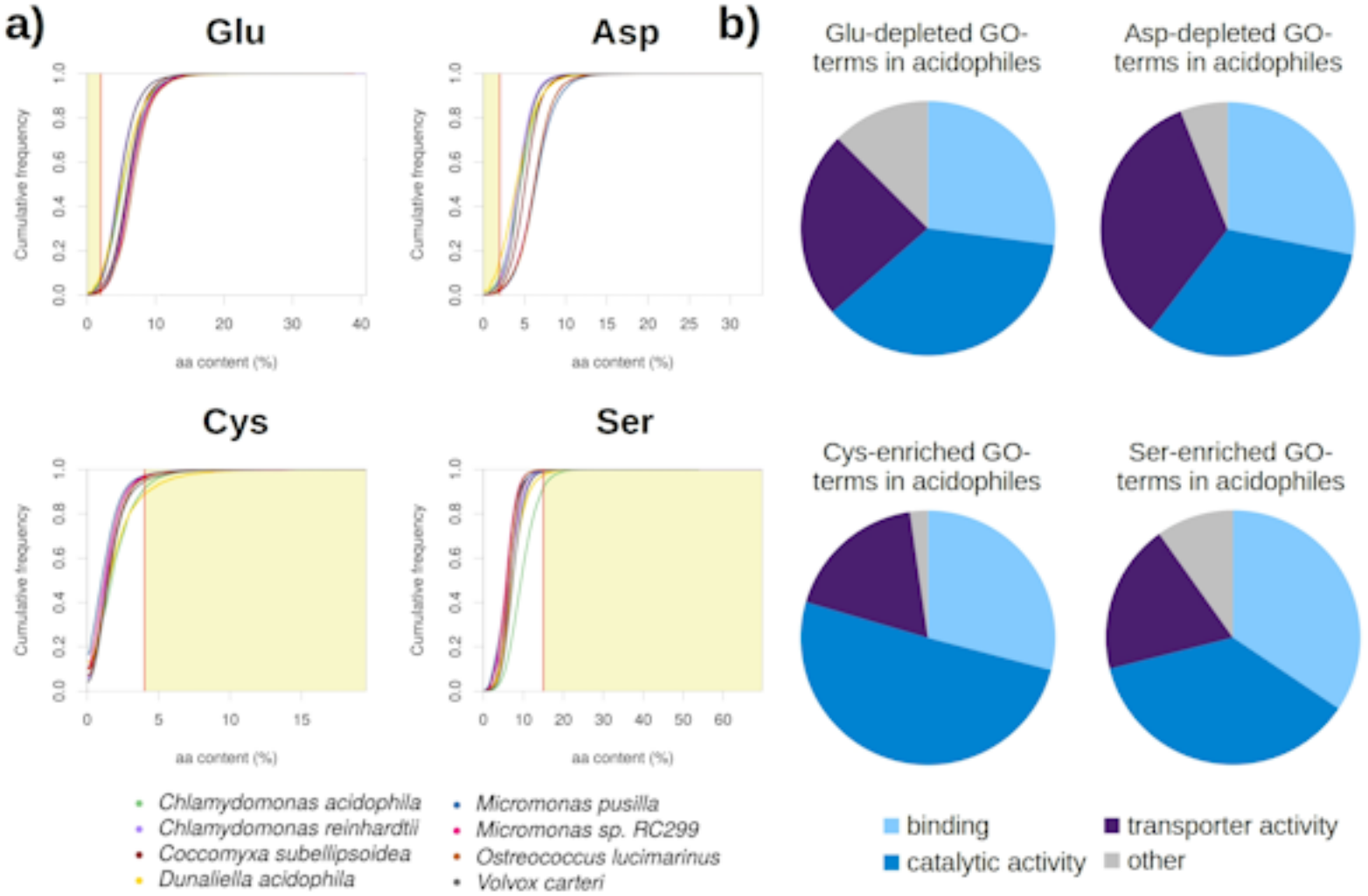


D) OG0008459: Cocaine esterase comp3348_c0_seq1

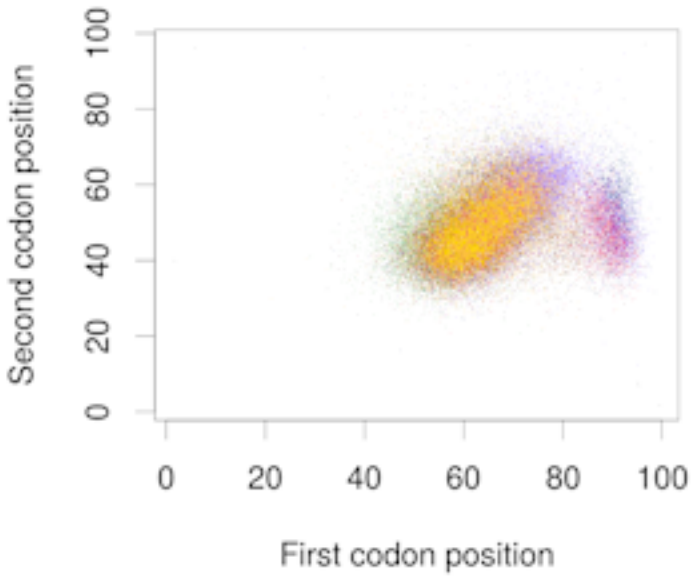


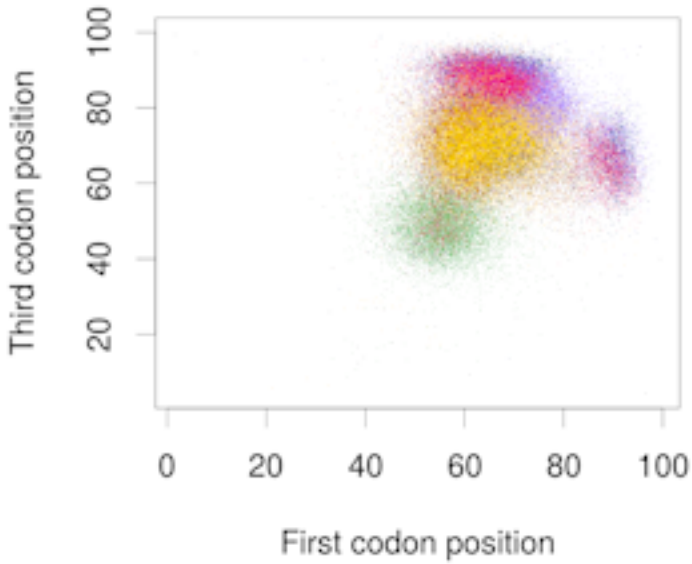
E) OG0009876: SDR-family protein with acetoacetyl-CoA reductase activity comp_14433_c0_seq1

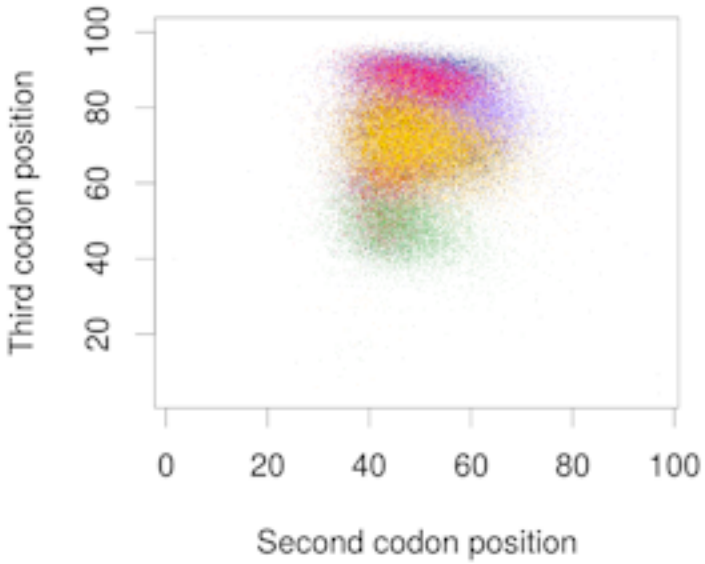

a)
Codon Usage Bias
CA2 (8%)
CA1 (24%)
Chlamydomonas acidophila
Chlamydomonas reinhardtii
Coccomyxa subellipsoidea
Dunaliella acidophila
Micromonas pusilla
Micromonas sp. RCC299
Ostreococcus lucimarinus
Volvox carteri
b)
Aminoacid Usage Bias
CA2 (13%)
CA1 (18%)
Chlamydomonas acidophila
Chlamydomonas reinhardtii
Coccomyxa subellipsoidea
Dunaliella acidophila
Micromonas pusilla
Micromonas sp. RCC299
Ostreococcus lucimarinus
Volvox carteri

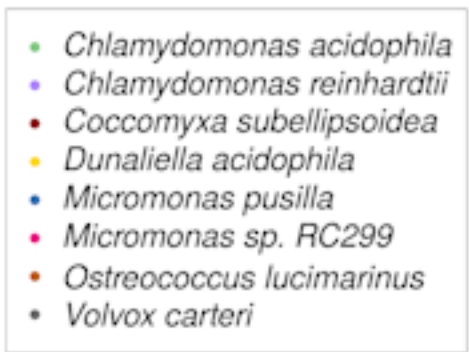

a)
Glu
Asp
Cys
Ser
Cumulative frequency
aa content (%)
b)
Glu-depleted GO-terms in acidophiles
Asp-depleted GO-terms in acidophiles
Cys-enriched GO-terms in acidophiles
Ser-enriched GO-terms in acidophiles
Chlamydomonas acidophila
Chlamydomonas reinhardtii
Coccomyxa subellipsoidea
Dunaliella acidophila
Micromonas pusilla
Micromonas sp. RC299
Ostreococcus lucimarinus
Volvox carteri
binding
catalytic activity
transporter activity
other

Average GC content in the three codon positions

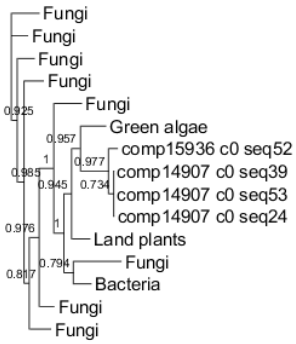




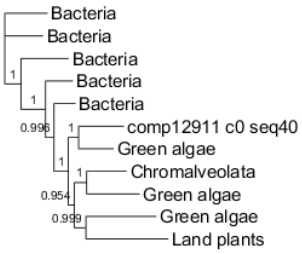




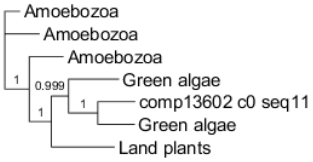
Additional File 4. Simplified phylogenetic analyses of additional transcripts coding for genes with an annotation related to heavy metal tolerance and present in *C. acidophila* but not in *C. reinhardtii*. The phylograms represent the majority consensus of trees sampled after stationarity in the Bayesian analysis. PP values equal or greater than 0.50 are shown above branches. The scale bar indicates relative distance between different sequences based on mutation rate. A) ACR3 comp_14907 and comp_15936 B) MATE efflux protein comp_12911 C) NRAMP family comp_13602 D) ABC-ATPase comp_14042 E) arsenite-antimonite efflux family comp_15332 F) glutathione-regulated potassium-efflux comp_16013 G) multidrug efflux transporter AcrB comp_16471 H) multidrug resistance-associated protein comp_17557.



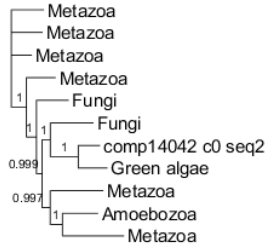
A) ACR3 comp_14907 and comp_15936



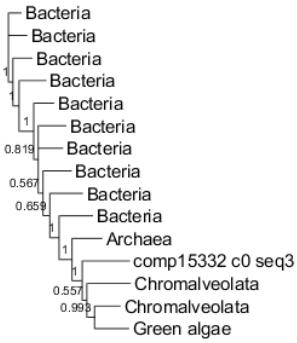
B) MATE efflux protein comp_12911



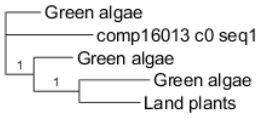
C) NRAMP family comp_13602



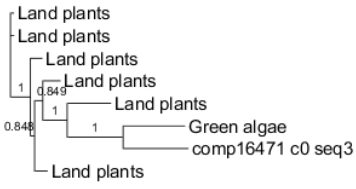
D) ABC-ATPase comp_14042



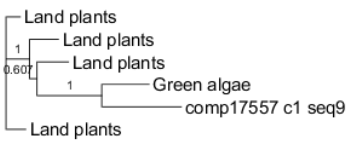
E) arsenite-antimonite efflux family comp_15332



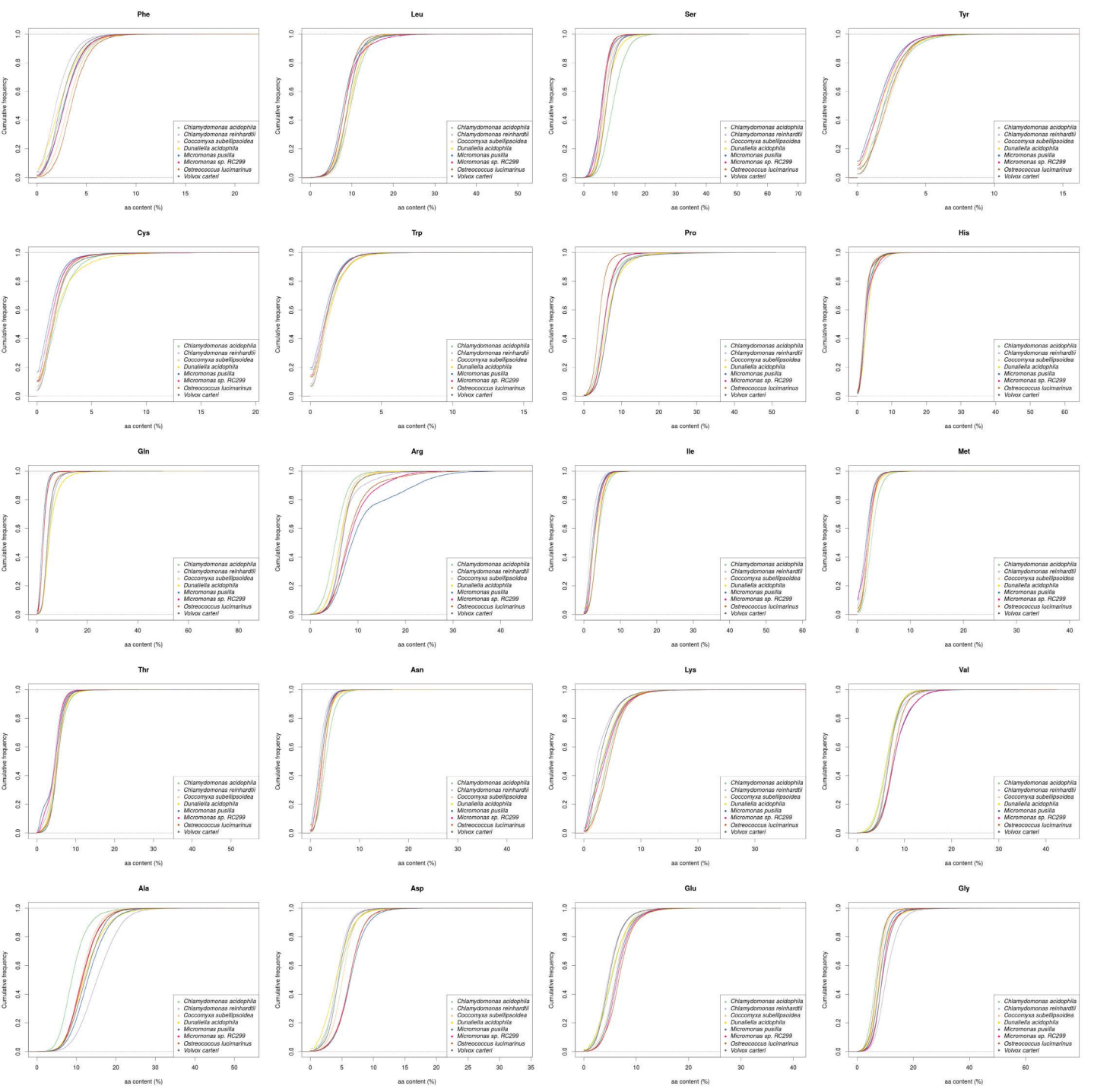
F) glutathione-regulated potassium-efflux comp_16013



G) multidrug efflux transporter AcrB comp_16471



H) multidrug resistance-associated protein comp_17557



Additional File 5. Aminoacid utilization profiles for the different species. There is one cumulative frequency plot for each aminoacid.