1. INTRODUCTION

Solar radiation is an essential factor sustaining or limiting complex life on Earth. In the past decades, a drastic reduction in stratospheric ozone has occurred as a result of increased concentrations of chlorofluorocarbons (CFCs) (Aucamp 2007) and other halogen gases in the upper atmosphere (Russell et al. 1996). This depletion of ozone has resulted in an increase of biologically harmful UV radiation at the Earth’s surface, together with its detrimental effects on all life forms.

The amount of UV radiation reaching the ground comprises only a small proportion of global radiation: about 6–7% of UV-A (320–400 nm), less than 1.0% of UV-B (280–315 nm), and virtually no UV-C (100–280 nm) (Hu et al. 2008). Biological damage is wavelength-dependent: UV-A produces mainly reactive oxygen intermediates causing indirect damage to DNA, proteins, and lipids, while UV-B and UV-C (100–280 nm) cause both indirect and direct damage to DNA because of its strong absorption at wavelengths below 320 nm (Mitchell & Karentz 1993). In accordance, numerous studies have reported UV-B

Solar UV radiation can damage aquatic organisms and decrease the productivity of aquatic ecosystems. Beyond the targets for damage by UV radiation (DNA, lipids, protein) that are common for all biological systems, a major site of damage in primary producers is the photosynthetic machinery, including Photosystem II and the accessory pigments that funnel light energy to the reaction centers (Häder & Gao 2015). Subsequent damage will directly reduce primary production. The effects of UV radiation may also reduce the photosynthetic uptake of atmospheric carbon dioxide and affect species diversity, ecosystem stability, trophic interactions, and global biogeochemical cycles (Häder et al. 2011, Williamson et al. 2019).

Over evolutionary time, microorganisms exposed to high levels of UV and other DNA-damaging factors have developed specific and highly conserved DNA repair mechanisms. The most important are photoreactivation, nucleotide excision repair (NER), base excision repair (BER), mismatch repair (MMR), and homologous repair (HR). Other processes, such as damage tolerance (dimer bypass), SOS response, checkpoint activation, and programmed cell death efficiently act against DNA lesions, ensuring genomic integrity.

The finest and most efficient mechanism to repair DNA damage by UV-B is photoreactivation, executed by photoreactivating enzymes known as 'photolyases' which target the main products of UV-B, pyrimidine dimers (PDs). Such PD lesions bring polymerases to a standstill, eventually leading to cell death. Photolyases bind tightly to PDs in the dark and can be activated by light from different wavelengths, splitting the 2 PD bonds and resulting in the re-formation of the 2 separate pyrimidine bases (Weber 2005). Photolyases are classified according to the type of photoproduct that is repaired: cyclobutane pyrimidine dimer (CPD-Phr) photolyases or (6-4) pyrimidine–pyrimidone photolyases (6-4-Phr). These enzymes, together with the related cryptochromes (Cry), constitute the cryptochrome/photolyase family (CPF). However, cryptochromes have no photolyase activity, functioning as signaling molecules regulating diverse biological responses such as entrainment of circadian rhythms in plants and animals (Roenneberg & Merrow 2005, Harmer 2009). It has been argued that CPF evolved early in the history of life due to the need of primitive organisms to cope with extreme solar radiation in the time when no stratospheric ozone existed to protect the Earth (Kanai et al. 1997, Porter et al. 2019).

Investigations at the metagenome level about effects of sunlight on microbial communities have mostly involved photoreceptors such as microbial rhodopsins (Pushkarev & Béjà 2016, Pushkarev et al. 2018) and LOV-domains (Pathak et al. 2012). Likewise, Singh et al. (2009) reported abundances for several light-related genes in microbiomes from different environments. However, the abundance and diversity of CPF in microbiomes at a world-wide scale and light range have not been studied so far, even though these photoreceptors are a vital mechanism in the defense against UV-B. In this work, we studied aquatic microbiomes exposed to different intensities of UV-B under different photoperiods around the globe using metagenomic analysis of existing published DNA data sets. Our objective was to characterize the occurrence and diversity of CPF on microbiomes from diverse aquatic ecological niches. The results add a new dimension to our understanding of the short-term effects of climate change and atmospheric ozone depletion on environmental microbial communities. The focus of this work on microbial communities from extreme environments provides models for early-life research.

2. MATERIALS AND METHODS

2.1. Metagenome selection and UV-B data processing

All metagenomes were retrieved from the NCBI database. They were sequenced by a shotgun strategy through Illumina technology except the Lake Diamante metagenome, which was sequenced with 454 GS FLX Titanium chemistry. The accession numbers are as follows: Lake Diamante (DM) (ERR1824222), Socompa stromatolite (SS) (SRR3341855), Tibetan Plateau sediment (TB) (SRR3322106), Amazon River (AM) (SRR1790676), Lake Rauer (RA) (SRR6129205), Dewar Creek hot spring (HT) (SRR5580900), Greenland cryoconite (CR) (SRR5275901), Lake Montjoie (MT) (SRR5818193), Olkiluoto Island groundwater (OK) (SRR6976411), and human gut (GU) (SRR6517782).

An essential aspect of this work was to reanalyze published metagenomic data of globally distributed microbiomes taking into account their overall UV-B exposure (Fig. 1). Monthly mean UV-B data sets
Values were processed using QGIS (www.qgis.org) to calculate the UV-B exposure of the samples of the selected metagenomes according to the date and location of sampling available at the biosample-linked section of each SRA entry. To compensate for a lack of replicates in most samples, we performed the analysis by clustering the data sets into 3 groups according to UV-exposure regimes: high (UV\textsubscript{Low}), mid (UV\textsubscript{Mid}), and low-exposed (UV\textsubscript{High}). We considered the UV\textsubscript{Low} range as between intensities of 0 and 2000 J m\textsuperscript{-2} d\textsuperscript{-1}, the UV\textsubscript{Mid} as between intensities of 2000 and 6000 J m\textsuperscript{-2} d\textsuperscript{-1}, and UV\textsubscript{High} range as between intensities of 6000 and 12000 J m\textsuperscript{-2} d\textsuperscript{-1}. The metagenome of the GU was assumed as a UV-B-free environment, acting as a negative control. See Table 1 for the main characteristics of the environmental data sets. A brief description of the environmental conditions of each sampling site is summarized below, including citations of previous works describing the corresponding microbial communities in further detail.

DM samples were taken from red biofilms attached to the bottom of calcareous stones submerged in Lake Diamante on 15 February 2012. This lake is located inside the Galan Volcano crater, at 4589 m above sea level (a.s.l.) (26.008° S, 67.043° W). High pH (9–11), high arsenic concentrations (115–234 mg l\textsuperscript{-1}), high salinity (270 g l\textsuperscript{-1}, 217 mS cm\textsuperscript{-1}), high UV radiation (84 W m\textsuperscript{-2} of UVA-B at noon), large day–night temperature range (–20 to +20°C), and low O\textsubscript{2} pressure constitutes a set of unique extreme conditions that prevails in the lake (Rascovan et al. 2016).

SS samples were taken from stromatolites located along the southern shore of Socompa Lake (24.536° S, 68.202° W) at the base of the Socompa Volcano (3570 m a.s.l.). These structures were found around the border of Socompa Lake during the summer when they are partially exposed depending on the tide and hydrological regime; they are entirely submerged during winter and spring. The stromatolites are rounded, domal structures that present clear stratification in vertical sections. Samples were taken...
as 5 cm deep cylinders from the top of the stromatolite. The site was exposed to air at the moment of sampling in February 2011. Extreme environmental conditions at the lake include hypersalinity, high thermal amplitude with daily temperatures that range from −10 to 20°C in summer and −20 to 10°C in winter, UV solar irradiance that reaches 68 W m⁻², low O₂ pressure, low nutrient availability, and, primarily, high arsenic content (18.5 mg l⁻¹) (Kurth et al. 2017).

TB was collected on 1 August 2013 from sediments of the Qiangyong Glacier Lake (28.51° N, 90.12° E), located at the southern part of the Tibetan Plateau, in the Indian monsoon climate region. The temperature, pH, and O₂ were 55.5° W) at 33 m depth, by pumping water with a Shurflo submersible pump through a 297 μm stainless steel screen. The temperature, pH, and electric conductivity of surface water are, on average, 6.89°C, 8.32, and 136.4 μS cm⁻¹, respectively (Chen et al. 2016).

The AM water sample was collected on 19 May 2011 from the Amazon River at Obidos station (1.919° S, 55.525° W) at 33 m depth, by pumping water with a Shurflo submersible pump through a 297 μm stainless steel screen. The temperature, pH, and O₂ were 28.4°C, 6.6, and 3.2 m l⁻¹, respectively, at the site of sampling (Satinsky et al. 2015, Doherty et al. 2017).

RA was sampled on 12 January 2015 from Rauer Lake (Torckler Island; 68.555° S, 78.191° E). Water temperature and approximate depth of water at the sampling site were 9°C and 10–15 cm. The lake was shallow (<30 cm). There was a crust of salt crystals on the sediment (which makes the lake look frozen from the air). There also appeared to be stratification in the lake: a top clear layer of ~10 cm and a bottom layer of ~5 cm. These layers were not visible until they were disturbed, and a visible haze was produced when the layers were mixed (Tsichtshko et al. 2018).

CR came from a cryoconite sample collected on 29 August 2013 from the Greenland ice sheet at the TAS-U-A1 site (65.41° N, 38.51° W). The pH, altitude, and depth of sampling were 5.53, 580 m, and 20 cm, respectively. Cryoconite was collected from 2–18 holes at each site, depending on their availability, using sterile 50 ml syringes (Stibal et al. 2015a, b, Hauptmann et al. 2017).

The MT microbial sample was collected from Lake Montjoie, Canada (45.409° N, 72.099° W) in February 2014; the lake was covered in ice at the time of sampling. The sample was taken from the epilimnion strata of the lake (0–3 m); pH was 6.41 (Tran et al. 2018).

The HT sediment sample was collected from the Dewar Creek hot spring, Canada (49.954° N, 116.515° W), near the source of the hot spring on 28 September 2012. The pH and temperature at the site were 7.93 and 66.4°C (Eloe-Fadrosh et al. 2016). The OK sample was from groundwater collected from a deep subsurface site KR11_0.1 (61.241° N, 21.494° E) in Olkiluoto Island, on the south-west coast of Finland on 9 September 2016. Based on the environmental data provided by the authors of the original research (Bomberg et al. 2016), samples were collected from 60 drill holes from different fractures at different depths (296–798 m below sea level). The groundwater is stratified, with a salinity gradient extending from fresh to brackish water to a depth of 30 m and the highest salinity concentration of 125 g l⁻¹ total dissolved solids at 1000 m depth (Posiva 2013). Between 100 and 300 m depths, the groundwater originates from ancient (pre-Baltic) seawater and has high concentrations of SO₂⁻. The temperature rises linearly with depth, from ca. 5–6°C at 50 m to ca. 20°C at 1000 m (Ahokas et al. 2008). The pH of the groundwater is slightly alkaline throughout the depth profile. The bedrock of Olkiluoto consists mainly of mica gneiss and pegmatitic granite-type rocks (Kärki & Paulamäki 2006). The in situ temperature at 300 m depth in the Olkiluoto bedrock is stable at approximately 10°C and increases linearly to approximately 16°C at 800 m depth. GU corresponds to a gut metagenome sample collected from the feces of an active Behcet’s disease patient (Ye et al. 2018) and was used as a negative control with no exposure to UV.

2.2. Meta-analysis, quality control, and assembly of selected metagenomes

Adapters were removed from the Illumina raw reads using the ‘fastq-mcf’ tool of ‘ea-utils’ v.1.04.676 (Aronesty 2011). This step was not needed for the LD metagenome, as it was sequenced with 454 technologies. Quality filtering and trimming were performed with the same program, using parameters l > 50 and q > 20. The program ‘kneaddata’ v.0.6.1 (https://bitbucket.org/biobakery/kneaddata/wiki/Home) was used with the ‘—bypass-trim’ option to clean human contaminant sequences. Paired-end reads were merged with FLASH v.1.2.11 (Magoc & Salzberg 2011) to recover unpaired longer reads. The pair-end reads with a low percentage of merging were left in the paired state for assembly. Assembly of filtered reads was performed with MEGAHIT v.1.1.2 (Li et al. 2015). Genes were predicted in assembled contigs with Prodigal v.2.6.2 (Hyatt et al. 2010), which outputted translated protein sequences.

To perform quantitative metagenomics, pair-end reads that remained with a low percentage of merg-
ing were concatenated together by the forward and reverse using a script available at GitHub (https://github.com/LangilleLab/microbiome_helper/blob/master/concat_paired_end.pl).

2.3. Reference database building

Considering the main functional orthologous of the CPF group, a reference protein database was built. The 2 KEGG Orthology numbers corresponding to CPF members (K06876, K01669) were linked with the UniProt database, and then results were filtered by EC number, known molecular function or biological process, uniref 90 clusters, and sequence length >100.

2.4. Estimation of CPF gene abundance and microbial diversity

To search and quantify the abundance of genes in the different metagenomes, we aligned the pre-processed unpaired reads against the CPF reference database previously created. Protein alignment was performed with PALADIN v.1.3.1 (Westbrook et al. 2017), which outputs a table with alignment counts for each gene. Alignments were filtered by maximum quality (= 60). The outputted read counts were normalized to reads per kb per genome, and the number of genomes was computed through Microbe-Census v.1.1.0 (Nayfach & Pollard 2015). The following formula describes the normalization:

\[
\text{Read Counts ÷ Number of Genomes ÷ Gene Length}
\]

Diversity analysis of each metagenome was performed with MetaPhlAn v.2.7.7 (Truong et al. 2015) with the ‘–bt2_ps’ parameter set to ‘sensitive’. The MetaPhlAn program profiles the composition of microbial communities from metagenomic shotgun sequencing data at the species level. It relies on ~1 M unique clade-specific marker genes identified from ~17 000 reference genomes (~13 500 bacterial and archaeal, ~3500 viral, and ~110 eukaryotic). The Shannon diversity index was obtained with the R package ‘vegan’ v.2.5-2 (Oksanen et al. 2018).

2.5. Analysis of the annotated genes

The protein sequences were aligned, using Diamond v.0.9.22 software (Buchfink et al. 2015), against the CPF genes of the reference database built earlier. Alignment parameters were 50% identity, 70% query coverage, and e < 10^{-5}. The retrieved files coming from each metagenome were modified at the sequence headers to hold the name of its respective metagenome. Next, all files were merged and filtered by sequence length of >400 amino acid residues.

The filtered protein sequences were used for phylogenetic analysis. The phylogenetic tree was built with ‘FastME’ v.2.0 (Lefort et al. 2015) using the Jones-Taylor-Thornton rate matrix (Jones et al. 1992) with 1000 bootstrap replicates. A consensus of the 1000 resulting trees was selected for further processing and visualization using ‘iTOL’ (Letunic & Bork 2016). Additionally, a taxonomical identification of the sequences was performed through the BLAST web server (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

3. RESULTS AND DISCUSSION

3.1. UV-B intensity profiles on world-wide microbiomes

UV-B geo-referenced data helps us to group microorganisms according to their UV-exposure regimes as UV_{High}, UV_{Mid}, and UV_{Low}. The UV_{High} samples were DM, SS, and TB, with UV-B intensities of 9677, 9536, and 8885 J m^{-2} d^{-1}, respectively. In contrast, UV_{Low} microorganisms OK, MT, and CR were exposed to low UV-B intensities of 77, 719, and 1759 J m^{-2} d^{-1}, respectively. The UV_{Mid} microbiomes AM, RA, and HT were exposed to intensities of 5630, 2289, and 2281 J m^{-2} d^{-1}, respectively. It is important to note that these intensity values were registered taking into account the sampling dates; for instance, the samples of the UV_{High} group were registered during summer, when insolation is maximum at the high-altitude environments of the Argentinean Puna and Tibetan Plateau (February and August) and UV-B incidence becomes the highest on the planet. On the other hand, the UV_{Low} microbiomes were sampled during August, November, and February which corresponds to the summer, autumn, and winter seasons in the Northern Hemisphere, with lower levels of irradiation compared to the Southern Hemisphere. In turn, the UV_{Mid} samples included the AM microbiome from a tropical environment and RA and HT from high-latitude extreme cold and hot environments, respectively.

The calculated day length from the UV data sets for each geo-referenced metagenome (https://www.suncalc.org) revealed that 2 microbiomes experienced peculiar photoperiods at the time of sampling. While the OK microbiome was experiencing a short
photoperiod (only 7.8 h), RA was sampled when Antarctica was experiencing 24 h of daylight.

We assumed that the environments had been exposed to similar conditions of solar irradiance for a considerable period before sampling. As a consequence, the diversity of species and genes on genomes is a reflection of the ecological pressure of radiation in the environment (among other factors), since the geographic conditions (latitude, altitude, and orography) did not change considerably in those sampled regions for decades. Although not available, transcriptomic data from these sites would be an excellent tool to evaluate the reflection of environmental pressure on the dynamics of the microbial communities.

The intensity of solar irradiation over each microbiome also depends on the on-site spatial disposition, being at a maximum for samples taken from soil or from shallow waters of lakes, springs, oceans, and streams but much lower in sediments, groundwater, or deep water. Some of the microbiome samples were not directly exposed to UV, but instead were sheltered in a protective environment. Nevertheless, during the biofilm or stromatolite formation process, microbes must have been directly exposed to the harsh UV radiation without a protective shelter, and managed to survive due to their strong capacity to deal with high UV irradiation. Even many constitutive species of these communities could present migration behavior, a life cycle, or an evolutionary history in which they could have been exposed to radiation.

It is well known that UV-B negatively affects microbial diversity (Balaré et al. 2011); nevertheless, this effect has not yet been assessed in a systematic way across different irradiation regimes at a global scale. Our results for species richness and the Shannon index in diverse metagenomes suggest an interesting trend (Fig. 2); TB, SS, and DM from the UV_{High} group had the lowest diversity, while the GU microbiome, which received no irradiation, had the maximum diversity. Moreover, microbiomes exposed to intensities of less than 4000 J m^{-2} d^{-1} revealed a high range of species richness (number of species: 11–103), whereas the SS microbiome with intensities above 8000 J m^{-2} d^{-1} only showed 21 species.

Species richness (Fig. 2A) was variable for microbiomes whose UV-B exposure was below a certain threshold (6000 J m^{-2} d^{-1}). In microbiomes receiving higher UV doses, the number of species decreased, indicating selective pressure. Those species that do not possess efficient molecular mechanisms to defend themselves from UV-B become unable to adapt to the new assemblage of species. A similar situation was observed when applying the Shannon index

![Graph A](image1)

![Graph B](image2)

Fig. 2. Impact of UV-B on the microbial diversity in the whole metagenomic data set. Two parameters were measured: (A) species richness and (B) the Shannon index. Metagenomes were colored according the group to which they belong: red: UV_{High}; green: UV_{Mid}; blue: UV_{Low}. DM: Lake Diamante; SS: Socompa stromatolite; TB: Tibetan Plateau sediment; AM: Amazon River; RA: Lake Rauer; HT: Dewar Creek hot spring; CR: Greenland cryoconite; MT: Lake Montjoie; OK: Olkiluoto Island groundwater; GU: human gut
(Fig. 2B) (which incorporates equitability of the species abundance in addition to its ability to detect rare species): microbiomes of the UVHigh group and HT (UVMid) presented the lowest Shannon index values. These results suggest that those microorganisms with full capacity to defend themselves against UV-B radiation reach dominance, while less fit species vanish. Nevertheless, this correlation is not conclusive, as many other chemical and physical factors (salinity, nutrient availability, heavy metals) may cause selective pressure on microbial diversity.

3.2. The meta-analysis, quality control, and assembly of selected metagenomes

Quality filtering and merging yielded a range between 0.64 and 13.66 Gb for further analysis (Table 1). Only the TB and AM data sets reported a high percentage of merging with FLASH (Magoč & Salzberg 2011); however, the low-merged data set from OK was also used for further downstream steps, as it was a reasonable size (between the range mentioned above). MEGAHIT (Li et al. 2015) was used for assembly since that program can deal with large and complex data sets in a time- and cost-efficient manner. Protein prediction with Prodigal (Hyatt et al. 2010) over contigs with >999 bp outputted a rate of 1000–1500 proteins Mb⁻¹, which is congruent with the high coding density expected for microbial species.

3.3. Quantitative analysis of CPF

In order to evaluate the abundance of CPF genes in each microbiome, a database was built using Uniprot sequences linked to KEGG orthology and then aligned to the query metagenomes. Our results show that the distribution of CPF genes in metagenomes followed a predicted ecological pattern; CPF genes were abundant in microbiomes with

<table>
<thead>
<tr>
<th>Metagenome</th>
<th>Description</th>
<th>Depth of sampling (cm)</th>
<th>Reference</th>
<th>Size after merging and quality filtering (Gb)</th>
<th>Size of raw metagenomic data (Gb)</th>
<th>Size of assembled contigs (&gt;999 bp) (Mb)</th>
<th>No. of annotated proteins</th>
<th>Assembly No. of Mb⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake Diamante</td>
<td>Red biofilm from Lake Diamante, located inside the Galan Volcano crater at 4589 m a.s.l.</td>
<td>−</td>
<td>Rascovan et al. (2016)</td>
<td>0.64</td>
<td>30 047 993</td>
<td>19 537</td>
<td>47 731</td>
<td>1388,492</td>
</tr>
<tr>
<td>Socompa</td>
<td>Sample from a high-altitude stromatolite in the surroundings of Lake Socompa</td>
<td>0−5</td>
<td>Kurth et al. (2017)</td>
<td>11.4</td>
<td>7.57</td>
<td>77 652</td>
<td>219 571</td>
<td>1151,254</td>
</tr>
<tr>
<td>Tibetan Plateau</td>
<td>Sediment sample collected from Qiangyong</td>
<td>−</td>
<td>Chen et al. (2017)</td>
<td>11.4</td>
<td>42 453</td>
<td>80 284 053</td>
<td>1218,554</td>
<td></td>
</tr>
<tr>
<td>Amazon River</td>
<td>Water sample from the Amazon River at a site facing Óbidos city, Brazil</td>
<td>0−3300</td>
<td>Brandon et al. (2015)</td>
<td>3.4</td>
<td>5.7</td>
<td>41 430</td>
<td>1218,554</td>
<td></td>
</tr>
<tr>
<td>Lake Rauer</td>
<td>Saline lake microbial sample from Rauer</td>
<td>0−30</td>
<td>Tschitschko et al. (2018)</td>
<td>3.4</td>
<td>5.7</td>
<td>41 430</td>
<td>1218,554</td>
<td></td>
</tr>
<tr>
<td>Dewar Creek hot spring sediment</td>
<td>Groundwater microbial communities from the Greenland ice sheet</td>
<td>&gt;29 600</td>
<td>Hauptmann et al. (2017)</td>
<td>3.4</td>
<td>5.7</td>
<td>41 430</td>
<td>1218,554</td>
<td></td>
</tr>
<tr>
<td>Lake Montjoie</td>
<td>Freshwater microbial communities from the epilimnion strata of Lake Montjoie</td>
<td>0−300</td>
<td>Tran et al. (2018)</td>
<td>14.2</td>
<td>13.66</td>
<td>240 051</td>
<td>1270,47</td>
<td></td>
</tr>
<tr>
<td>Olkiluoto Island</td>
<td>Groundwater microbial communities from the Indian Islands, Antarctica</td>
<td>&gt;29 600</td>
<td>Ye et al. (2018)</td>
<td>10</td>
<td>8.85</td>
<td>74 876</td>
<td>1453,049</td>
<td></td>
</tr>
<tr>
<td>Human gut</td>
<td>Metagenome of fecal sample collected from an active Behet’s disease patient</td>
<td>−</td>
<td>Ye et al. (2018)</td>
<td>10</td>
<td>8.85</td>
<td>74 876</td>
<td>1453,049</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. General traits of the metagenomes used in the present study
UV$_{\text{High}}$ and UV$_{\text{Mid}}$, while they were insignificant or non-existent in the UV$_{\text{Low}}$ microbiomes OK and GU, respectively. In the DM sample, CPF genes appeared profusely. Their abundance was 249 e$^{-02}$ RPKG, which was higher than in the rest of the microbiomes.

The relationship between the abundance of CPF and the intensity/time of insolation is provided in Fig. 3A, and shows an upward trend of CPF abundance as UV-B irradiation increases. However, RA (a UV$_{\text{Mid}}$ microbiome) completely deviates from the trend because of its high abundance. We had previously mentioned that RA had the most extended photoperiod of all the samples, so we set out to verify if there was a connection between the abundance of CPF and photoperiod. Fig. 3B shows that there is an upward trend of CPF as the photoperiod increases, with RA quite in line with the trend. Thus, both factors — intensity and photoperiod — could be influencing the abundance of CPF in the microbial communities.

Our study revealed a rising trend of CPF abundance in microbial communities in environments with higher radiation levels or longer day durations. That result is likely due to an increase in DNA damage caused by UV-B, which mainly of CPD and [6-4] photoproducts. In such situations, populations could increase the expression of CPF by increasing the gene copy number, or species lacking CPF in their genome would be replaced by species that contain these genes. In either case, the overall increase in CPF copy abundance indicates that the community improves its defense capabilities against UV-B by using the highly efficient mechanism of photoreactivation or modulating enzymatic mechanisms triggered by light-sensing cryptochromes.

### 3.4. Phylogenetic analysis of CPF genes

The CPF group of proteins comprises mostly genes of photolyases with different specializations and cryptochromes, whose functions are mostly unknown. The family has been divided into different subgroups considering phylogeny, chromophore type, specialization, host organism, and structure. Alignment, editing, and subsequent phylogenetic analysis of 214 CPF sequences found in the metagenomes configured a robust tree with numerous clades (Fig. 4). The majority of sequences grouped within well-known subfamilies: the CPD photolyases classes I, II, and III, DASH cryptochromes (cry-DASH), the FeS-BCPs group, and the thermostable CPD photolyases, a new group with an unstable position in phylogeny (Ueda et al. 2005, Portero et al. 2019). This work also reports 3 novel clades: unidentified I, II, and III (UI, UII, and UIII), which have 98, 53, and 85% of bootstrap support, respectively. The 3 new clades group 24.76% of the whole sequences in the analysis (Fig. 5), thus offering a large pool of candidates for...
Fig. 4. Phylogenetic tree built with the cryptochrome/photolyase family sequences from the whole metagenomic data set. Bootstrap values for the main groups are shown. Each leaf is colored and labeled by sample **procedence**: DM: Lake Diamante (purple); SS: Socompa stromatolite (dark green); TB: Tibetan Plateau sediment (brown); AM: Amazon River (blue); RA: Lake Rauer (grey); HT: Dewar Creek hot spring (orange); CR: Greenland cryoconite (cyan); MT: Lake Montjoie (red); OK: Olkiluoto Island groundwater (light green). CPD I, II, and III: class I, II, and III cyclobutane pyrimidine dimer photolyases; CPD THR: thermostable CPD photolyase; DASH: DASH cryptochrome; FeS-BCPs: FeS bacterial cryptochromes and photolyases; UI, UII and UIII: unidentified group I, II and III, respectively.
new functions, specializations, or molecular specifici-
ties. Interestingly, clades UI and UII are mainly com-
posed of sequences of a single microbiome; se-
quencies from the RA microbiome constitute 85.7% of
UI, while 87.5% of UII corresponds to MT sequences.
The UI subfamily is the only clade in DM but domi-
nant in RA; both microbiomes displayed the highest
content of CPF genes in this study. Furthermore,
clades UI and UII are taxonomically homogeneous,
with sequences belonging to Halobacteria and Acti-
nobacteria, respectively, while clade UIII is formed
by sequences of different taxonomic origins, includ-
ing the phyla Proteobacteria, Bacteroidetes, and Ver-
rucomicrobia.

The widespread presence of cryptochrome-hold-
ing subfamilies, cry-DASH and Fes-BCPs, among
the microbiomes is evident (Fig. 6). The abundance
of cry-DASH followed a pattern consistent with
higher irradiation conditions; the highest number of
cry-DASH sequences were found in the SS and TB
communities (UV\text{High}), AM (UV\text{Mid}), and RA (the
most prolonged photoperiod). It has been reported
in previous work that these cryptochromes are in
fact photolyases with an affinity for single-stranded
DNA and, in some cases, RNA (Selby & Sancar
2006). Cry-DASH may play a complementary role
of photoreactivation in these microbiomes. The
case for the FeS-BCPs group is intriguing, being
present in almost all the communities with similar
relative abundance. It has been suggested that this
class complements other photolyases by performing
the function of a prokaryotic [6-4] photolyase
(Zhang et al. 2013, Graf et al. 2015), in this way
avoiding or decreasing the use of the inefficient
NER system. In previous work on 3 isolated strains
from the high-altitude Andean Lakes (where DM
and SS originate) whose genomes were sequenced,
we reported the presence of both CPD photolyases
and members of the FeS-BCP subfamily in all
genomes (Portero et al. 2019), even the presence of
a third CPF member belonging to cryptochrome
DASH in one of the strains. Since the formation of
the [6-4] photoproduct depends on the dose of UV
radiation and increases with increasing dose reach-
ing up to 40% of the photoproduct fraction (Sancar
2003), we suggested the possibility that acquisition
of prokaryotic [6-4] photolyases (included in the
new group FeS-BCP) by bacteria from the high-
altitude Andean Lakes is a compensatory adapta-
tion to the high dose of UV radiation present in
their original environment. Our present work pro-
vides further insight into the FeS-BCP subfamily by
showing its ubiquitous and homogenous presence
in world-wide microbiomes.
3.5. Relationship between CPF and taxa dominance

The assignment of CPF genes to a particular taxonomic group was also considered in this work. CPF sequences were classified by genus and paired with the same information previously obtained for each microbiome through MetaPhlAn. The CPF sequences grouped with their homolog with a BLAST identity lower than 70% clustered as an unclassified category. Genus diversity from the whole microbiome (top bar) and the set of cryptochrome/photolyase family (CPF) sequences recovered from the same metagenome (lower bar). The diversity of the microbial community was estimated with MetaPhlAn, while the diversity of the set of CPF sequences were computed with the BLAST online platform. See Fig. 2 for sample location abbreviations.

Fig. 6. Comparative barplot showing, for each metagenome, the genus diversity both for the whole microbial community (top bar) and the set of cryptochrome/photolyase family (CPF) sequences recovered from the same metagenome (lower bar). The diversity of the microbial community was estimated with MetaPhlAn, while the diversity of the set of CPF sequences were computed with the BLAST online platform. See Fig. 2 for sample location abbreviations.
poorly referenced diversity of the community itself (top bar). Despite this, MT had the most substantial diversity of genera (12) with CPF hits along with SS and RA (11 each).

The most abundant taxa in the DM, SS, RA, and HT communities possess CPF genes. In AM, only *Ralstonia* and *Vibrio* displayed CPF; both genera together represented 35% of the community. In the case of CR, only *Hymenobacter* and *Agromyces* possessed these genes, with both groups adding up to 7%. *Pedobacter* was the unique taxon in TB with CPF genes, and it represented less than 4% of the community. Finally, *Escherichia*, with an abundance of just 0.44%, was the unique CPF contributor in GU. Neither in MT nor OK found matches between the **MetaPhlAn** and **BLAST classifications**.

These results show that CPF, and thus photoreactivation, is usually present in the most abundant organisms of the UV$_{\text{high}}$ microbiomes (Fig. 6). Also, SS and RA show a high number of taxa in which photoreactivation is present. That suggests that photoreactivation may act as a successful system for assuring the survival and predominance of taxa in UV-stressed environments. This fact contrasts with those less isolated samples, where natural selection does not promote the diversification and dominance of taxa carrying CPF genes.

### 4. CONCLUSIONS

In this work, UV-B was considered for the first time as an ecological variable in a sequenced-based metagenomic study of microbial communities on a worldwide scale. Our results consistently showed concordance of high UV exposure of a given microbiome with its low microbial diversity and high CPF abundance.

We also evaluated CPF diversity in the world-wide microbiomes and reported 3 novel CPF clades not identified in previous analyses. CPF was more likely present in the most abundant organisms of the UV$_{\text{high}}$ microbiomes, suggesting a significant evolutionary force for survival and dominance in highly irradiated environments. Also, cryptochrome-like genes were much more abundant in the most exposed microbiomes, indicating a complementary role to standard photolyases.

Finally, metagenomics proved to be an excellent tool to reveal a clear correspondence between microbiomes, UV-exposure, and UV-B resistance mechanisms, measured in the number of gene copies. Additional methods, such as metatranscriptomics and metaproteomics, should be implemented in order to unveil the molecular dynamics of the CPF upon different light conditions.

**Acknowledgements.** V.H.A. and M.E.F. are staff researchers from the National Research Council (CONICET) in Argentina. D.A. is a recipient of a doctoral fellowship from CONICET. The authors have produced this manuscript despite the delays and shortages of funding execution from National Agencies in Argentina, mainly FONCyT (PICT 2013 2991) and CONICET (PIP 2015 0519, PIO-UNCA y PICT V 3825-2016), and the drastic devaluation of Argentinean currency during the period 2016–2019. This manuscript has been released as a pre-print at bioRxiv (Alonso-Reyes et al. preprint doi:10.1101/701565).

**LITERATURE CITED**


Doherty M, Yager PL, Moran MA, Coles VJ and others (2017) Bacterial biogeography across the Amazon River–ocean continuum. Front Microbiol 8:882


Alonso-Reyes et al.: Metagenomics of photolyase/cryptochromes in aquatic microbiomes

Graf D, Wesslowski J, Ma H, Scheerer P and others (2015) Key amino acids in the bacterial (6-4) photolyase PhrB from Agrobacterium fabrum. PLOS ONE 10:e0140953


Submitted: May 6, 2020; Accepted: August 17, 2020
Proofs received from author(s): ******, 2020