What is in umbilicaria pustulata? A metagenomic approach to reconstruct the holo-genome of a lichen

Tzovaras, B.G.; Segers, F.H.I.D.; Bicker, A.; Grande, F. dal; Otte, J.; Anvar, S.Y.; ... ; Ebersberger, I.

What Is in *Umbilicaria pustulata*? A Metagenomic Approach to Reconstruct the Holo-Genome of a Lichen

Bastian Greshake Tzovaras[1,2,3], Francisca H.I.D. Segers[1,4], Anne Bicker[5], Francesco Dal Grande[4,6], Jürgen Otte[6], Seyed Yahya Anvar[7], Thomas Hankeln[5], Imke Schmitt[4,6,8], and Ingo Ebersberger [1,4,6,*]

[1]Applied Bioinformatics Group, Institute of Cell Biology and Neuroscience, Goethe University Frankfurt, Germany

[2]Lawrence Berkeley National Laboratory, Berkeley, California

[3]Center for Research & Interdisciplinarity, Université de Paris, France

[4]LOEWE Center for Translational Biodiversity Genomics, Frankfurt, Germany

[5]Institute for Organismic and Molecular Evolution, Molecular Genetics and Genome Analysis, Johannes Gutenberg University Mainz, Germany

[6]Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt, Germany

[7]Department of Human Genetics, Leiden University Medical Center, The Netherlands

[8]Molecular Evolutionary Biology Group, Institute of Ecology, Diversity, and Evolution, Goethe University Frankfurt, Germany

*Corresponding author: E-mail: ebersberger@bio.uni-frankfurt.de.

Accepted: March 9, 2020

Abstract

Lichens are valuable models in symbiosis research and promising sources of biosynthetic genes for biotechnological applications. Most lichenized fungi grow slowly, resist aposymbiotic cultivation, and are poor candidates for experimentation. Obtaining contiguous, high-quality genomes for such symbiotic communities is technically challenging. Here, we present the first assembly of a lichen holo-genome from metagenomic whole-genome shotgun data comprising both PacBio long reads and Illumina short reads. The nuclear genomes of the two primary components of the lichen symbiosis—the fungus *Umbilicaria pustulata* (33 Mb) and the green alga *Trebouxia* sp. (53 Mb)—were assembled at contiguities comparable to single-species assemblies. The analysis of the read coverage pattern revealed a relative abundance of fungal to algal nuclei of ~20:1. Gap-free, circular sequences for all organellar genomes were obtained. The bacterial community is dominated by *Acidobacteriaceae* and encompasses strains closely related to bacteria isolated from other lichens. Gene set analyses showed no evidence of horizontal gene transfer from algae or bacteria into the fungal genome. Our data suggest a lineage-specific loss of a putative gibberellin-20-oxidase in the fungus, a gene fusion in the fungal mitochondrion, and a relocation of an algal chloroplast gene to the algal nucleus. Major technical obstacles during reconstruction of the holo-genome were coverage differences among individual genomes surpassing three orders of magnitude. Moreover, we show that GC-rich inverted repeats paired with nonrandom sequencing error in PacBio data can result in missing gene predictions. This likely poses a general problem for genome assemblies based on long reads.

Key words: metagenome assembly, SPAdes, sequencing error, symbiosis, chlorophyta, gene loss, organellar ploidy levels, microbiome.

Introduction

The lichen symbiosis comprises a lichen-forming fungus (mycobiont) and a photosynthetic partner (photobiont), which is typically a green alga or a cyanobacterium. A bacterial microbiome and additional third-party fungi can also be part of the lichen consortium (Grube et al. 2015; Spribille et al. 2016). The bacterial microbiome in particular may contribute to auxin and vitamin production, nitrogen fixation, and stress protection (Erlacher et al. 2015; Grube et al. 2015; Sigurbjornsdottir et al. 2016). Lichenized fungi are well known for synthesizing diverse, bioactive natural products (reviewed by Muggia and Grube [2018]), which has recently stimulated research into biosynthetic pathways and gene clusters of these fungi (Armaleo et al. 2011; Abdel-Hameed et al. 2016; Bertrand and Sorensen 2018; Wang et al. 2018; Calchera et al. 2019). The estimated 17,500–20,000 species

Table 1

Genome Assembly Characteristics of a Selection of Lichenized Fungi and of Green Algae from the Class *Trebouxiophyceae*

	Species[a]	Size (Mb)	Scaffolds	N50 (Mb)	Genes	Missing BUSCO (%)[b]	FGMP: HCE (%)[c]	FGMP: Proteins (%)[c]
Fungus	*U. muehlenbergii*	34.6	7	7.0	8,822	1.3	90.3	94.9
	A. radiata	33.5	17	2.2	na	3.0	87.1	97.8
	U. pustulata^{M}	33.5	43	1.8	9,825	3.6	90.3	96.8
	G. flavorubescens	34.5	36	1.7	10,460*	1.5*	77.4	97.3
	X. parietina	31.9	39	1.7	11,065	1.4*	77.4	96.6
	C. metacorallifera	36.7	30	1.6	10,497*	3.0*	83.9	97.3
	C. macilenta	37.1	240	1.5	10,559*	2.7*	80.6	96.3
	P. furfuracea	37.8	46	1.2	8,842	1.8	93.5	97.1
	R. intermedia	26.2	198	0.3	na	3.3	87.1	97.3
	E. prunastri	40.3	277	0.3	10,992	1.3*	87.1	96.6
	C. rangiferina	35.7	1,069	0.3	na	2.5	80.6	98.0
	C. grayi	34.6	414	0.2	11,388	3.0*	87.1	96.8
	E. pusillum	36.8	908	0.2	9,238	3.9*	80.6	96.0
	L. hispanica	41.2	1,619	0.1	8,488	1.6	90.3	97.3
	R. peruviana	27.0	1,657	<0.1	9,338*	6.7*	80.6	95.4
	L. pulmonaria	56.1	1,911	<0.1	15,607	1.5*	83.9	97.0
	C. uncialis	32.9	2,124	<0.1	10,902*	5.3*	87.1	97.1
	C. linearis^{M}	19.5	2,703	<0.1	na	25.0	51.6	83.8
	A. sarmentosa^{M}	40.0	915	<0.1	na	21.9	58.1	83.3
Alga	*T. gelatinosa*^{L}	61.7	848	3.5	na	68.7	na	na
	C. subellipsoidea^{F}	48.8	29	2.0	9,851	2.4	na	na
	Chlorella sp. A99^{S}	40.9	82	1.7	8,298	18.4	na	na
	Trebouxia sp.^{L,M}	52.9	217	0.8	13,919	13.9	na	na
	A. glomerata^{L}	55.8	151	0.8	10,025	12.4	na	na
	A. protothecoides^{F}	22.9	374	0.3	7,016	12.2	na	na
	Trebouxia sp. TZW2008^{L}	69.3	677	0.2	na	14.8	na	na
	Helicosporidium sp.^{S}	12.4	5,666	<0.1	6,035	50.8	na	na

NOTE.—The species are sorted by descending scaffold N50. The lichen symbionts sequenced for this study are highlighted in gray. ^{F}Free-living algae, ^{L}lichen photobionts, ^{S}other symbiotic algae, and ^{M}assemblies resulting from metagenomic sequencing projects.

[a]Genome accession numbers are provided in supplementary table S5, Supplementary Material online.

[b]BUSCO analysis was performed on the assembly level. *Values taken from Calchera et al. (2019).

[c]FGMP assembly completeness was determined using 31 highly conserved noncoding elements (HCE) and 593 conserved fungal proteins.

of lichens (Kirk et al. 2008) are distributed across nearly all ecosystems (Ahmadjian 1993). Some lichens thrive as pioneering organisms in ecological niches that are otherwise adverse to eukaryotic life (Kranner et al. 2008; Hauck et al. 2009). The capability to inhabit such a diverse set of habitats is tightly connected with the lichen symbiosis itself. The nutritionally self-sustaining system harbors internal autotrophic photobionts, which provide carbohydrates to all other members of the association. Furthermore, some mycobiont species switch between different sets of environmentally adapted photobionts and can thus occupy broad ecological niches (Dal Grande et al. 2018).

There is an increasing interest in genomic resources on lichens, because lichens are valuable models in symbiosis research (Grube and Spribille 2012; Wang et al. 2014; Grube et al. 2015) and promising sources of biosynthetic genes for biotechnological applications (see above). Most lichenized fungi grow slowly, resist aposymbiotic cultivation, and are generally poor candidates for experimentation. Therefore, researchers increasingly use genomic data as sources of novel information on the lichen symbiosis (e.g., Armaleo et al. 2019). Genome sequences of about 19 lichenized fungi and of 2 algal photobionts have been published to date (table 1). Most genome sequences stem from lichens whose symbionts were grown in axenic culture. The few studies using metagenomic data to reconstruct the fungal genomes reported highly fragmented assemblies comprising >900 scaffolds (McDonald et al. 2013; Meiser et al. 2017; Allen et al. 2018; Liu et al. 2019). Some assemblies range in an expected total length (McDonald et al. 2013; Meiser et al. 2017) and achieve comparable BUSCO (Simao et al. 2015) scores to assemblies derived from single-species cultures (Meiser et al. 2017). However, the only two publicly available genome sequences of lichenized fungi that were assembled from metagenomics data, *Cetradonia linearis* and *Alectoria sarmentosa* (Allen et al. 2018; Liu et al. 2019), have >20% BUSCO genes missing (table 1). They are thus far from complete. Moreover, discontinuous assemblies are of limited use for functional genomics analyses, which rely on a comprehensive and accurate annotation of genes and even more so of gene clusters (Denton et al. 2014; Dunne and Kelly 2017). Attempts to assemble the entire holo-genome of a lichen have not been reported, thus

far. Also, a genome assembly strategy based on long-read sequencing technology, for example, PacBio, as well as hybrid approaches, has not yet been applied to lichens.

Obtaining the complete set of genome sequences from organisms forming obligate symbioses is challenging. Large-scale cultivation of the individual partners is often not feasible, or aposymbiotic cultivation of the symbionts is entirely impossible. This precludes efforts to obtain pure, single-species DNAs. The alternative approach, reconstructing high-quality genomes from multispecies, metagenomic samples, can be methodologically demanding (Greshake et al. 2016). For example, genomic representation can be skewed toward one partner in the association (e.g., the host species), resulting in uneven coverage of individual genomes (Greshake et al 2016). Further methodological challenges include the risk of creating chimeric contigs, that is, assemblies of reads from multiple genomes, or selecting the appropriate assembly software (Greshake et al. 2016; Meiser et al. 2017). Moreover, inaccurate postassembly taxonomic assignment (binning) can lead to chimeric draft genome sequences, which comprise contigs from multiple species (Sangwan et al. 2016). Thus, it is highly desirable to assess and develop methods for obtaining metagenome-assembled genomes of eukaryotes and eventually achieve similar assembly qualities and reporting standards as in prokaryotes (Bowers et al. 2017).

Here, we report the reconstruction of the holo-genome for the lichen *Umbilicaria pustulata* entirely from metagenomic DNA. Details on the biology and distribution of *U. pustulata* have been published elsewhere (e.g., Hestmark 1992; Dal Grande et al. 2017). We inferred the genome sequences of the lichenized fungus *U. pustulata*, its green algal symbiont *Trebouxia* sp., and its bacterial microbiome. We combined Illumina short reads from different whole-genome shotgun library layouts with PacBio long reads and integrated results from complementary assembly strategies.

Specifically, we addressed the following questions: What is the quality of fungal and algal organellar and nuclear genomes based on hybrid short- and long-read assemblies obtained from a metagenomic lichen sample? What are the relative genome copy numbers and the relative taxon abundances of the microorganisms involved in the lichen symbiosis? What is the composition of the bacterial microbiome of a lichen individual? Is there evidence for horizontal gene transfer from algae or bacteria into the fungal genome? What are the methodological pitfalls associated with reconstructing the holo-genome of symbiotic communities from metagenomic reads, and with their integration into comparative genomics studies focusing on gene loss?

Materials and Methods

Sample Collection and DNA Extraction

Thalli of *U. pustulata* were collected near Olbia (Sardinia, Italy) and Orscholz (Saarland, Germany) between May 2013 and December 2014. DNA was extracted using the CTAB method (Cubero and Crespo 2002) and subsequently purified with the PowerClean DNA Clean-Up Kit (MO BIO, Carlsbad, CA).

Quantitative Polymerase Chain Reaction

Quantitative polymerase chain reaction (qPCR) targeted the fungal and algal single copy genes, *mcm7* (forward—gaatgcaaggcaaacaattc and reverse—ttgtactgttctatccgtcgg) and *g467* (COP-II coat subunit; forward—ccttcaagctgcctatctg and reverse—gcacctgaaggaaaagac), respectively. DNA concentrations extracted from four thalli were measured with the Qubit dsDNA High Sensitivity Kit (Life Technologies) according to the manufacturer's instructions. For qPCR measurements, we used the *GoTaq qPCR* Master Mix (Promega) at a total volume of 10 µl. PCR (95 °C for 2 min; 40 cycles of 95 °C for 15 s, 55 °C for 30 s, and 60 °C for 1 min) was carried out in an *ABI 7500 Fast Real Time PCR system cycler* (Applied Biosystems). Four lichen thalli were measured in three technical replicates. To determine the total copy numbers, we used a standard curve approach with serial 10-fold dilutions of plasmids engineered to contain single copy PCR templates (pGEM-T Easy Vector, Promega).

Whole-Genome Shotgun Sequencing

We generated a whole-genome paired-end library with the Illumina TruSeq DNA Sample Prep v2 (Illumina, San Diego, CA), selecting for a mean fragment length of 450 bp with the SPRIselect reagent kit (Beckman Coulter, Krefeld, Germany). A mate-pair library with an insert size of 5 kb was created with the Nextera Mate Pair Sample Prep Kit (Illumina). The paired-end and mate-pair libraries were sequenced on an Illumina MiSeq machine. Long-read sequencing was performed on the PacBio RS II system (Pacific Biosystems of California, Menlo Park, CA), using 16 SMRT cells in total.

Read Preprocessing

Low quality 3′-ends and adapter sequences were removed from the Illumina paired-end reads with Trimmomatic v0.32 (Bolger et al. 2014) (*ILLUMINACLIP: IlluminaAdapter.fasta: 2:30:10)*. Mate pairs were processed with nextclip v0.8 (Leggett et al. 2014) to remove adapters and to bin them according to read orientation. PacBio sequence reads were error corrected with two alternative strategies. For an intrinsic error correction, we used canu v1.20 (Koren et al. 2017). Because an intrinsic error correction requires a high long-read coverage, which might not be achieved for the less abundant genomes in the lichen holo-genome, we additionally corrected the PacBio reads using Illumina data as extrinsic information. We merged the Illumina paired-end reads with FLASH v1.2.8 (Magoc and Salzberg 2011), using standard parameters. The processed Illumina read- and mate-pair data were then assembled with MIRA v4.0, using the

genome, denovo, accurate flags (Chevreux et al. 1999). The resulting contigs were then used for correcting sequencing errors in the PacBio reads with ECTools (https://github.com/jgurtowski/ectools, last accessed February 27, 2020) requiring a minimum alignment length of 200 bp with a *WIGGLE_PCT* of 0.05 and a *CONTAINED_PCT_ID* of 0.8 for the read mappings. Only PacBio reads with lengths after correction of above 1,000 bp were retained.

De Novo Metagenome and Metatranscriptome Assembly

We employed a multilayered strategy to target different parts of the lichen holo-genome (see supplementary text, Supplementary Material online, for a detailed description of the assembly strategies and supplementary fig. S1, Supplementary Material online, for the workflow). In brief, we first generated an assembly of the *U. pustulata* metagenome with FALCON v0.2.1 (Chin et al. 2016) using the uncorrected PacBio reads. The resulting contigs were scaffolded with SSPACE-Long v.1.1 (Boetzer and Pirovano 2014). In parallel, we assembled the error-corrected PacBio reads with the Celera assembler wgs v8.3rc2 (Berlin et al. 2015). Finally, we made a hybrid assembly with SPAdes v3.5.0 (Bankevich et al. 2012) that made use of all Illumina reads, the ECTools error-corrected PacBio reads, and the uncorrected PacBio reads to support scaffolding. Subsequent to taxonomic assignment with MEGAN v.5.10 (Huson et al. 2016) (see below), we binned all algal and bacterial contigs, respectively. They were then merged into single assemblies using *minimus2* (Treangen et al. 2011) followed by a scaffolding step with SSPACE-Long with the help of the PacBio reads. For the genome of the fungus *U. pustulata*, the SPAdes contigs of at least 3 kb in length were used to further scaffold the FALCON assembly with SSPACE-Long. The final assemblies were polished with Pilon v1.15 (Walker et al. 2014) using the Illumina short reads.

For the reconstruction of the organellar genomes, we used a baiting strategy. We aligned the canu-corrected PacBio reads against the organellar genomes of the lecanoromycete fungus *Cladonia grayi* (JGI Clagr3 v2.0) and the green alga *Asterochloris glomerata* (JGI Astpho2 v2.0) (Armaleo et al. 2019) with BLAT v35 (Kent 2002), using no cutoffs. The baited reads were assembled with canu v1.20, and the resulting organellar genomes were circularized with the help of the canu-corrected PacBio reads and circlator v.1.2.0 (Hunt et al. 2015). Assembly polishing was performed as described above.

For the reconstruction of the metatranscriptome, we assembled the RNAseq data provided in (Dal Grande et al. 2017) with Trinity release 2013-11-10 (Haas et al. 2013), using the *–jaccard-clip –normalize_reads* parameters.

Reconstruction of 16S rRNA Gene Trees

16S rRNA genes were extracted from the bacterial fraction of the holo-genome assembly. These data were complemented with the 16S rRNA sequences from two new species recently found to be associated with lichens, *Lichenibacter ramalinae* gen. nov., sp. nov. (Pankratov et al. 2020) and *Lichenihabitans psoromatis* gen. nov., sp. nov. (Noh et al. 2019). Each gene served as a query for a BlastN search (Altschul et al. 1997) against the 16S rRNA database of NCBI. The best five hits were extracted for each query, except for the sole 16S rRNA gene representing a member of the *Chitonophagaceae*, where we considered the best ten hits. A nonredundant set of 16S rRNA sequences was generated, and we distinguished five taxonomic bins representing the *Rhizobiales*, *Acidobacteria*, *Chitinophagaceae*, *Actinobacteria*, and *Rhodospirillales*, respectively. Sequences in each bin were aligned with MUSCLE v.3.8.1551 (Edgar 2004) and maximum likelihood phylogenetic trees were computed with RAxML v.8.2.12 (Stamatakis 2014) using the GTRGAMMA model of sequence evolution. Branch support was assessed by performing 100 nonparametric bootstrap replicates. Phylogenetic trees were visualized and edited with FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/, last accessed March 29, 2020).

Taxonomic Assignment

All reads and contigs were used individually as query for a DIAMOND v.0.6.12.47 search (Buchfink et al. 2015). Contigs were searched against a custom database comprising 121 fungi, 20 plants, 8 animals, 1,471 bacteria, and 560 viruses (supplementary table S1, Supplementary Material online), and reads were searched against the NCBI nr database. All sequences were subsequently taxonomically classified with MEGAN v5.10 (Huson et al. 2016) requiring a minimum DIAMOND alignment score of 50. For MEGAN analyses including more than one read set, we normalized counts to the smallest read set in the analysis. Metagenomic compositions were visualized with *Krona* (Ondov et al. 2011).

Read Mapping and Coverage Distribution Analysis

Reads from the three WGS libraries were mapped to the assembled scaffolds with bowtie2 (Langmead and Salzberg 2012). RNAseq reads of *U. pustulata* (Dal Grande et al. 2017) were mapped with *HISAT2* (Kim et al. 2015), setting the maximal intron length to 3,000 bp and keeping standard parameter values otherwise. To visualize the variation of the WGS read coverages and of the GC content across the different genomes, we split all scaffolds into partitions of 20 kb in length, and subsequently clustered the individual partitions by their tetra-nucleotide frequencies. For each partition, we then plotted the mean read coverage for each WGS library and the mean GC content with Anvi'o (Eren et al. 2015).

Nuclear and Organellar Genome Annotation

Interspersed repeats were annotated with the RepeatModeler/RepeatMasker pipeline (Smit et al. 2015).

The fungal nuclear genome was annotated with funannotate (https://funannotate.readthedocs.io, last accessed February 27, 2020). As training data, we used the proteomes of *Xanthoria parietina* JGI v1.1 and *C. grayi* JGI v2.0 (Armaleo et al. 2019), together with *U. pustulata* transcripts. The transcripts were obtained in the following way. RNAseq data from *U. pustulata* (Dal Grande et al. 2017) were de novo assembled with Trinity (Haas et al. 2013). In addition, we performed a second, reference-based assembly of the RNAseq data using Trinity's reference-guide mode together with the fungal genome assembly. Both assemblies, together with the raw read sets, were used to identify transcripts with PASA (Haas et al. 2008).

The nuclear genome of *Trebouxia* sp. was annotated with Maker v2.31.8 (Holt and Yandell 2011), utilizing GeneMark (Besemer and Borodovsky 2005), AUGUSTUS v3.1 (Stanke et al. 2006), and SNAP v2006-07-28 (Korf 2004). CEGMA (Parra et al. 2007), RNAseq data (Dal Grande et al. 2017), and the proteome of *A. glomerata* (JGI Astpho2 v2.0) were used for model training. The organelle genomes were annotated using MFannot via the web service provided at http://megasun.bch.umontreal.ca/RNAweasel/ (last accessed February 27, 2020). BLAST2GO (Gotz et al. 2008) and BlastKOala (Kanehisa et al. 2016) were used to assign Gene Ontology terms and KEGG identifiers to the predicted genes. The graphic representation of the organellar genomes was generated with OGDraw (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html, last accessed February 27, 2020).

Manual Curation of Gene Loss

To assess whether the absence of evolutionary old genes from the *U. pustulata* draft genome sequence is likely a methodological artifact or indeed indicates a gene loss, we performed a gene neighborhood analysis (see Supplementary Material online for more detailed methods). In brief, we determined the ortholog to the missing LCA_{Lec} gene in the close relative, *Umbilicaria hispanica* (Dal Grande et al. 2018), and identified its flanking genes. Next, we searched for the orthologs of these flanking *U. hispanica* genes in *U. pustulata*. We decided on a methodological artifact, if any of these orthologs reside at the terminus of either a contig or a scaffold. Otherwise, we extracted the genomic regions flanking the *U. pustulata* orthologs and used it as a query of a BlastX search (Altschul et al. 1997) against NCBI nr-prot. In addition, we used the *U. hispanica* protein as query for a TBlastN search in the *U. pustulata* genome assembly. Only when both searches provided no evidence of the missing gene, we inferred gene loss.

Data Accessibility

The raw Illumina and PacBio sequence reads have been deposited in the NCBI Sequence Read Archive (SRR8446862–SRR8446881). The assemblies have been deposited at GenBank under the accession numbers VXIT00000000 (*U. pustulata* A1-1) and VXIU00000000 (*Trebouxia* sp. A1-2), respectively. The orthologous groups representing the LCA_{Lec} gene set together with the gene annotation of *U. hispanica* are available via https://applbio.biologie.uni-frankfurt.de/download/lichen/ (last accessed February 27, 2020).

Fig. 1.—The lichen *U. pustulata*.

Results and Discussion

Reconstructing the Holo-Genome Sequence of *U. pustulata*

Umbilicaria pustulata is a rock-dwelling lichen (fig. 1), for which all attempts to cultivate the mycobiont in isolation have failed so far. This leaves a metagenomic approach as currently the only option to reconstruct the genome sequences of the lichen symbionts. qPCR revealed an average ratio of fungal to algal genomes in the lichen thallus of 16.2, with individual replicates varying from a minimum of 13 to a maximum of 24 (supplementary table S2, Supplementary Material online). The heterogeneity between the replicates most likely reflects natural variation in the thickness of the algal layer, and thus algal cell number, within and between lichen thalli (Kummerova et al. 2006). Such skewed data challenge individual assemblers to an extent that no single tool is capable to faithfully reconstruct all genomes (Bradnam et al. 2013; Greshake et al. 2016). We therefore devised a sequencing and assembly scheme to reconstruct the lichen holo-genome at high contiguity (for details on the workflow, see supplementary fig. S1 and text, Supplementary Material online). In brief, we used both Illumina short reads and PacBio long-read data and integrated three assemblers: FALCON (Chin et al. 2016) for assembling uncorrected full-length PacBio data, the Celera assembler (Berlin et al. 2015) for assembling the extrinsically error-corrected—and thus often fragmented—PacBio reads, and SPAdes (Bankevich et al. 2012) for a hybrid assembly of both Illumina and PacBio reads

Table 2
Metrics of the Metagenome Assembly

Assembly Method	Taxonomic Classification	Number of Scaffolds	Total Length (Mb)	N50 (kb)
FALCON	All	2,343	62	323
	Fungal	120	32	551
	Algal	709	9	17
	Bacterial	790	15	56
SPAdes	All	21,900	123	225
	Fungal	5,736	35	159
	Algal	257	47	461
	Bacterial	1,193	26	91
Celera	All	22,216	216	11
	Fungal	12,230	113	10
	Algal	3,557	52	17
	Bacterial	2,804	17	8
Merged (Minimus)	Fungal	43	33	1,808
	Algal	217	53	848
	Bacterial	483	35	251

Table 3
Mean Read Coverages for the Fungal and Algal Components of the *U. pustulata* Holo-Genome

		***U. pustulata* (Mycobiont)**		***Trebouxia* sp.**		
Sequencing Technology	**Library**	**Nuclear**	**mtGenome**	**Nuclear**	**mtGenome**	**cpGenome**
IlluminaMiSeq	Mate pair	40.7	573.3	2.5	25.1	48.6
	Paired end	123.4	2,472.4	12.8	239.5	214.2
PacBio RS II	16 SMRT cells	195.5	4,685.5	20.1	754.8	776.7

(supplementary fig. S1, Supplementary Material online). No individual method sufficed to reconstruct all genomes. A taxonomic assignment of the contigs revealed, however, that the tools complement each other in assembling different parts of the holo-genome at different contiguities (table 2). Interestingly, SPAdes performed substantially better on the low coverage algal reads than on the more abundant fungal data, both with respect to N50 and number of scaffolds. The difference in N50 reproduced findings from a previous study where NG(A)50 values produced by SPAdes from a simulated lichen holo-genome were consistently about an order of magnitude smaller for the fungal than for the algal parts of the assembly (Greshake et al. 2016). Because reads from both species were simulated with the same software, ART (Huang et al. 2012), this performance difference must be due to an intrinsic characteristic of the fungal genome, most likely its considerably high content of interspersed repeats (25%; see below). The average read coverage of 360× for the fungal genome (table 3) might represent an additional confounding factor. Anecdotal evidence exists that a too high read coverage impairs the performance of SPAdes. To follow up this point, we used ART (Huang et al. 2012) to simulate MiSeq whole-genome shotgun read sets with average read coverages ranging between 10× and 450× using the *U. pustulata* scaffolds as template. The corresponding read sets were then individually assembled with SPAdes, and we determined assembly size, number of scaffolds, and the scaffold N50 (supplementary table 3, Supplementary Material online). This revealed that coverages around 50× allow excellent genome reconstructions, which only very modestly improve upon increase of the read coverage. More importantly, increasing the coverage beyond 100× results in a constant increase of the number of scaffolds without increasing either assembly size or scaffold N50.

A joint scaffolding of all fungal contigs resulted in a *U. pustulata* mycobiont genome sequence of 33 Mb comprising 43 scaffolds with a scaffold N50 of 1.8 Mb. Merging and scaffolding of the algal contigs generated 217 scaffolds with an N50 of 0.8 Mb and a total assembly length of 53 Mb. The assembly lengths for both the fungal and the algal genomes fall well in the diversity of other lichenized fungi and members of the *Trebouxiophyceae*, respectively (table 1). Merging and scaffolding the bacterial fraction of the three assemblies resulted in 483 contigs amounting up to 35 Mb. Two bacterial scaffolds with lengths of 3.6 and 3.4 Mb represent major parts of two genomes from the genus *Acidobacterium*. We refer to them as *Acidobacterium BS 16* and *Acidobacterium BS 35*, respectively.

No scaffold in the final assembly represented the full-length genomes of the fungal and algal mitochondria, or of

the algal chloroplast. We therefore used the organellar genome sequences of *C. grayi* and of *A. glomerata* as baits to identify PacBio reads originating from the organellar genomes. The baited reads were assembled individually for each genome, resulting in a circular, gap-free sequence for each of the three organelles (supplementary figs. S2–S4, Supplementary Material online). The fungal mitochondrial genome (mt genome) comprises 95.4 kb. It ranks third in length among 23 mt genomes from lecanoromycete lichens (Pogoda et al. 2018; Armaleo et al. 2019), superseded only by *Leptogium hirsutum* (120 kb) and *Parmotrema stuppeum* (109 kb). The algal mitochondrion and chloroplast have lengths of 99.9 and 272.0 kb, respectively. They are larger than the organellar genomes in other *Trebouxiophyceae*, both symbiotic and free living (Fan et al. 2017), with the exception of *A. glomerata*, which has an even larger mitochondrial genome of 110 kb in length (Armaleo et al. 2019).

Taxon Abundance in the Lichen Holo-Genome

The metagenomic reconstruction of the lichen holo-genome allows, for the first time, to infer average genome copy numbers in a lichen thallus from the read coverage distribution (table 3, fig. 2, and supplementary table S4, Supplementary Material online). The coverage for the fungal nuclear genome assembly, and thus the genomic copy number, is on average about 20 times higher than that of the algal nuclear genome assembly. Similar to the results from the qPCR analysis, the individual estimates vary from a minimum of 9.6 to a maximum of 29.7, which is expected when the thickness of the algal layer varies within and between lichen thalli (Kummerova et al. 2006). Because both symbionts are haploid, this translates into an average abundance of 20 (SD: 7.2) fungal nuclei per algal nucleus. In the mycobiont, there are 15.4 (SD: 4.5) copies of the mitochondrial genome per nuclear genome. This value is substantially lower than the around 60 mtGenome copies per nucleus reported for *Aspergillus fumigatus* (*Eurotiomycetes*) (Neubauer et al. 2015). It is tempting to speculate that the small number of mitochondrial genomes in the mycobiont is connected to its slow growth. Yet, too little is known about temporal fluctuations and interindividual differences in mtGenome content in either species to draw conclusions from this difference. In each *Trebouxia* sp. cell, there are 20 (SD: 7.9) copies of the mitochondrial genome. *Trebouxia* sp. possesses only a single chloroplast. Thus, similar to many other green microalgae (Gallaher et al. 2018), the *Trebouxia* sp. chloroplast genome is polyploid and contains, on average, 20 (SD: 7.5) copies. To our knowledge, this is the first report of ploidy level for the chloroplast in a lichenized green alga. The two *Acidobacterium* spp. are each represented with about one cell per algal cell.

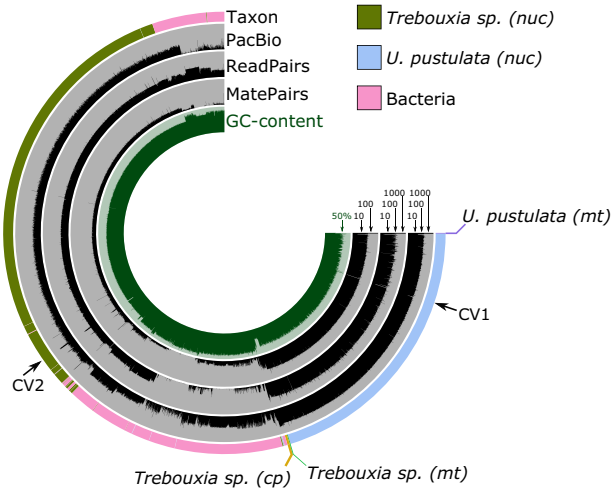

Fig. 2.—Read coverages and GC content distribution across the genomes in the lichen holo-genome and the three whole-genome shotgun libraries. The genome assemblies were split into nonoverlapping bins of 20 kb in length and were subsequently clustered according to their tetra-nucleotide frequency. Bins representing the same taxon share the same color. The bar height indicates mean read coverage (black) or mean GC content (green) for each bin. Read coverages are represented on a log scale. The arrows indicate 10×, 100×, and 1,000× read coverage, respectively. The mitochondrial genome of the fungus (*U. pustulata* (mt)) is, with a mean read coverage (PacBio) of 3,713, the most abundant component of the holo-genome. The read coverages across the nuclear genome reconstructions of the alga and the fungus are considerably even with only few notable coverage variations (CV). CV1 represents a GC-rich (>70%) repetitive region at the terminus of scaffold 8 paired with an assembly gap in this scaffold. The local increase in read coverage of the algal genome assembly combined with a drop in GC content (CV2) represents a nuclear copy of the algal mitochondrial genome (NUMT).

Characterization of the Bacterial Community

In a first, high-resolution approach to characterize individual members of the bacterial community, we identified 21 bacterial scaffolds harboring a 16S rRNA gene. Phylogenetic analyses integrating the 21 16S rRNAs with the most similar sequences represented in the NCBI 16S rRNA database (supplementary table S6, Supplementary Material online) grouped the sequences into five major clades, representing *Rhizobiales*, *Rhodospirillales*, *Actinobacteria*, *Chitinophagaceae*, and *Acidobacteria*, respectively (fig. 3). Notably, the *Rhizobiales* tree reveals that the *U. pustulata* microbiome harbors a close relative of *Lichenibacter ramalinae*, which has been previously identified as an endophytic bacterium in the thalli of subarctic lichens (Pankratov et al. 2020). Moreover, we found eight 16S rRNA genes that stem from Acidobacteria closely related to *Edaphobacter lichenicola*, *Granulicella aggregans*, *Acidipila rosea*, and *Acidobacterium capsulatum*. All taxa have been described to inhabit thalli of tundran lichens (Pankratov and Dedysh 2010; Pankratov 2012; Belova et al. 2018). The remaining 16S rRNA genes represent members of the *Rhodospirillales*

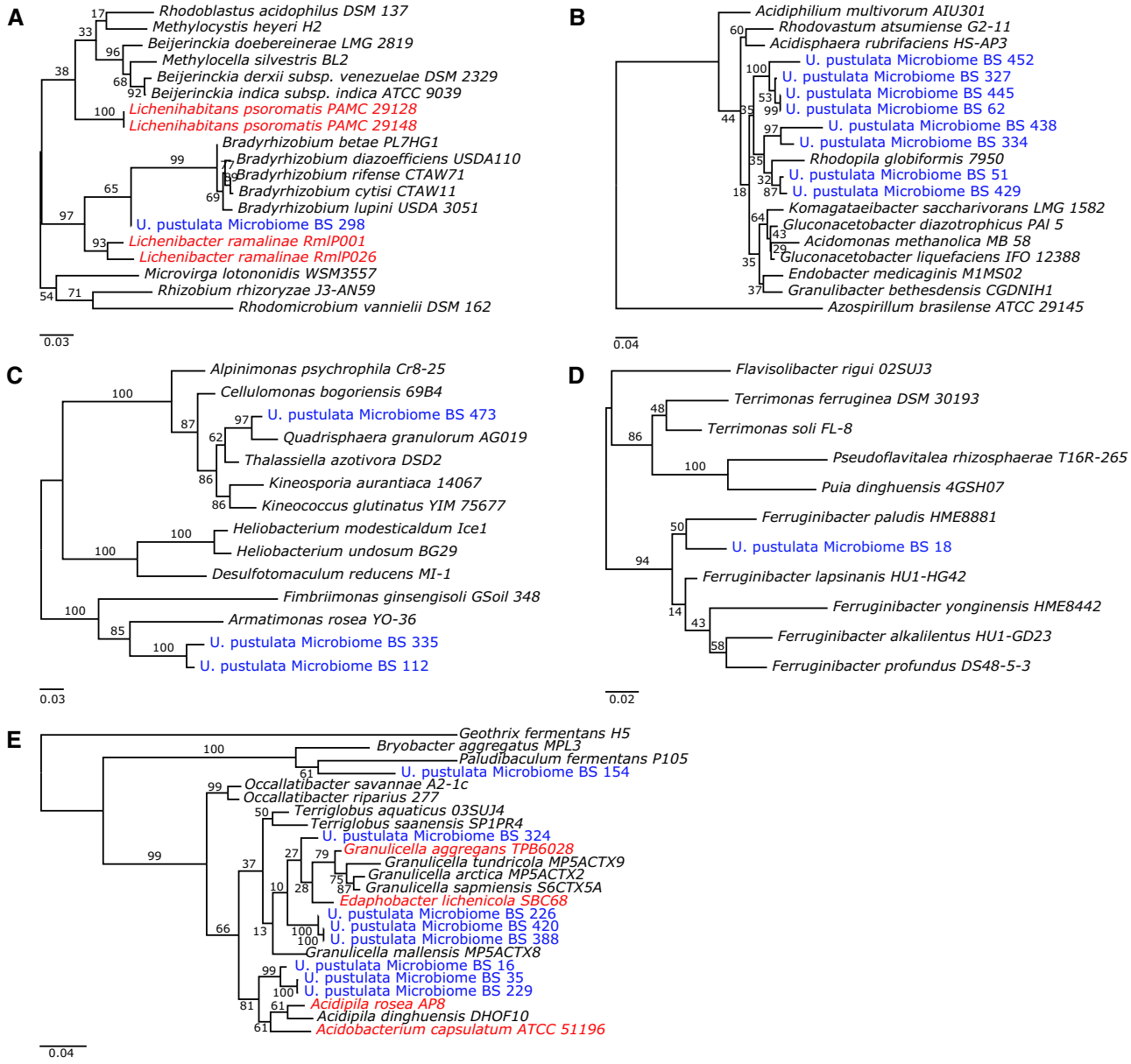

Fig. 3.—16S rRNA phylogenies for bacterial taxa represented in the *U. pustulata* microbiome. (*A*) *Rhizobiales*, (*B*) *Rhodospirillales*, (*C*) *Actinobacteria*, (*D*) *Chitinophagaceae*, and (*E*) *Acidobacteria*. 16S rRNA genes for the taxa in blue were extracted from the bacterial fraction of the *U. pustulata* holo-genome reconstruction. The trees reveal that the *U. pustulata* microbiome harbors close relatives to bacterial taxa that have been previously associated with microbiomes of tundran and subarctic lichens (red). Branch labels denote percent bootstrap support. NCBI accession numbers of the sequences are provided in the supplementary table S6, Supplementary Material online.

(*Alphaproteobacteria*; eight sequences), the *Actinobacteria* (three sequences), and the *Chitinophagaceae* (one sequence). To our knowledge, neither of these taxa has so far been associated with lichen microbiomes.

To obtain a more comprehensive overview of the bacterial community that is associated with *U. pustulata*, we performed a taxonomic assignment at the read level (fig. 4 and supplementary fig. S5, Supplementary Material online). *Acidobacteriaceae*, *Actinobacteria*, and *Alphaproteobacteria* are the three most abundant bacterial phyla. This is in line with the findings from the 16S rRNA analysis, and it is similar to what has been observed for Antarctic lichens (Park et al. 2016). In general, the taxonomic composition resembles closely typical rock-inhabiting bacterial communities (Choe et al. 2018). Yet, other studies suggested that *Alphaproteobacteria* and not *Acidobacteria* dominate lichen microbiomes (e.g., Grube et al. 2009; Bates et al. 2011; Aschenbrenner et al. 2014), with abundances of up to 32%

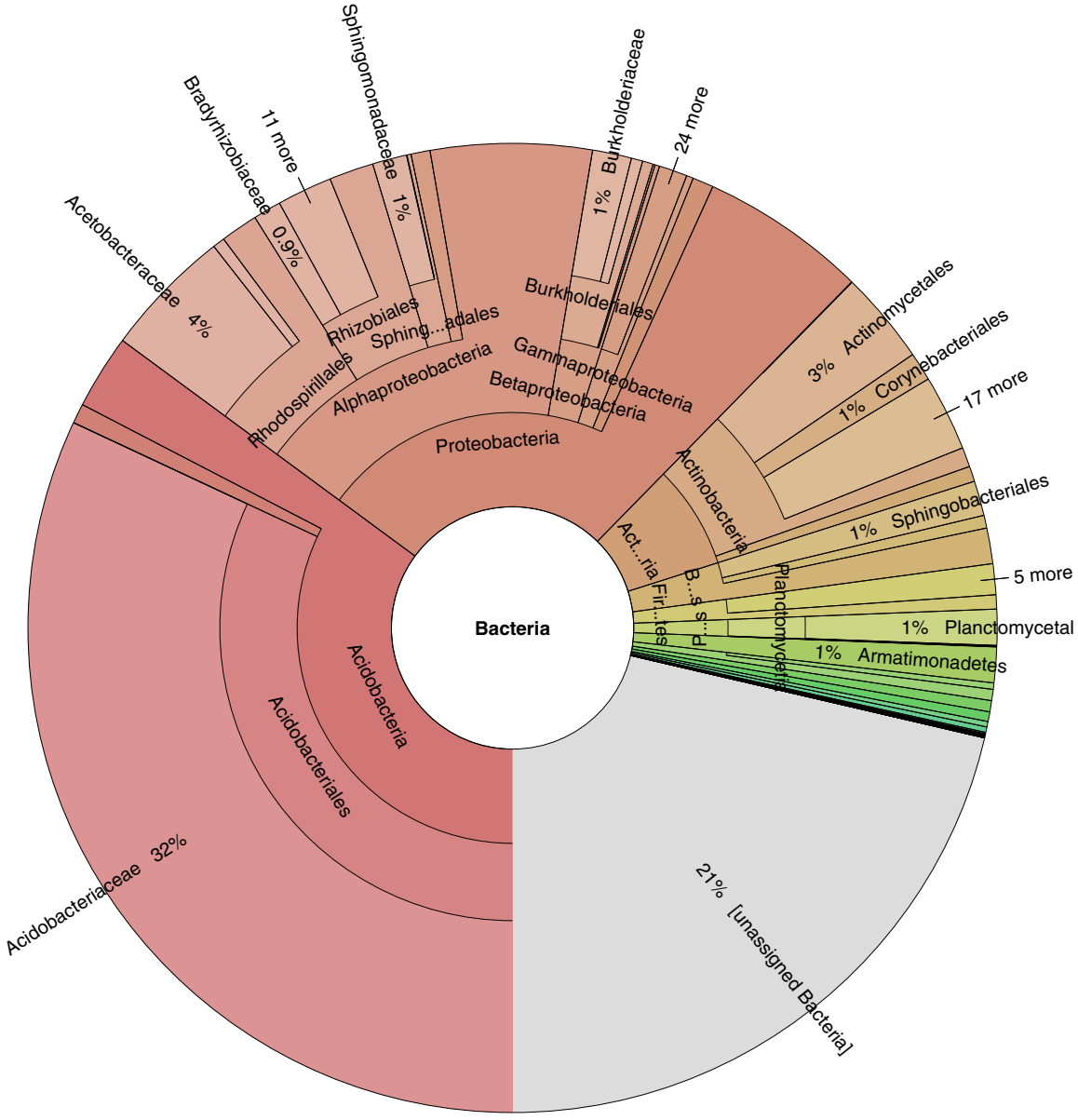

Fig. 4.—Composition of the bacterial fraction represented in the *U. pustulata* metagenomic reads. Reads from the two Illumina whole-genome shotgun libraries and the PacBio reads were pooled and taxonomically assigned with MEGAN (Huson et al. 2016). *Acidobacteria*, uniting 35% of the read counts, *Proteobacteria* (27%), and *Actinobacteria* (8%) are the three most abundant phyla. Notably, a single family, the *Acidobacteriaceae* (32%), dominates the microbiome. Its most abundant genera are *Granulicella*, *Terriglobus*, and *Acidobacterium* to which the two largest bacterial contigs belong to. Among the *Proteobacteria*, *Rhodospirillales* (6%), and therein the *Acetobactereaceae* (4%) take the largest share, followed by the *Rhizobiales* (4%). Within the *Actinobacteria*, *Actinomycetales* are the dominant family (3%). See supplementary figure 5, Supplementary Material online, for a species-level resolution of the microbiome.

for the *Rhizobiales* in the lichen *Lobaria pulmonaria* (Erlacher et al. 2015). This indicates that microbiome compositions can vary considerably between lichen species. However, differences in the methodology for assessing taxon frequencies can also result in substantially deviating results (Nayfach and Pollard 2016). The microbiome analyses by Erlacher et al. (2015) were performed at the level of assembled contigs. Although this eases the taxonomic assignment, due to the use of longer sequences (Vollmers et al. 2017), it is bound to result in distorted abundance estimates. The high read coverage for abundant taxa in a microbiome generally results in more contiguous assemblies comprising only few contigs. In a typical MEGAN analysis, taxon abundance is assessed by the number of sequences that are assigned to that taxon. As a consequence, common taxa with contiguous genome assemblies will receive low counts, and their abundance will be

underestimated. Rare taxa, in turn, whose lower read coverage results in more fragmented genome reconstructions with many short contigs will receive high counts. Their abundance will be overestimated (supplementary fig. S6, Supplementary Material online). We demonstrate the effect of the chosen methodology on the reconstruction of the *U. pustulata* microbiome. Applying the method of Erlacher et al. (2015) increased the estimated abundance of the *Rhizobiales* to 11% and decreased that of the *Acidobacteriaceae* to 18% (supplementary fig. S7*A*, Supplementary Material online). The dominance of the *Acidobacteriaceae* was restored when pursuing a hybrid approach, in which the taxonomic assignment was done at the contig level and the abundance estimates were based on the reads mapping to the contigs (supplementary fig. S7*B*, Supplementary Material online). We conclude that the methodological impact on the taxon abundance estimates is substantial and needs to be taken into account when comparing microbiome community composition in different studies.

Annotation of the Nuclear Genomes

The nuclear genome of *U. pustulata* (mycobiont) has an average GC content of 51.7%, and interspersed repeats account for 25.5% of the sequence. We identified 9,825 protein-coding genes (table 2), with on average 3.3 exons, and a mean transcript length of 1,406 bp. A BUSCO analysis (Simao et al. 2015) revealed that 94.4% of the 1,315 genes in the "Ascomycota" data set are represented over their full length in the genome sequence. Similarly, FGMP (Cisse and Stajich 2019) found 90% of the 31 highly conserved fungal noncoding elements and 96.8% of the 593 conserved fungal proteins that are represented in the FGMP search set. Both tools indicate a level of assembly completeness that is in the same range of what has been, thus far, achieved only for fungal genomes reconstructed from axenic cultures (table 1 and supplementary table S5, Supplementary Material online). Contrasting to the situation in many other lichens (cf., Spribille et al. 2016), we found no evidence for the presence of a second fungus in the lichen thalli (supplementary text, Supplementary Material online).

The genome of *Trebouxia* sp. has an average GC content of 50.0%, and interspersed repeats account for only 4.9% of the sequence. We predicted 13,919 genes with on average 6.7 exons per gene and a mean transcript length of 1,221 bp. With 13.9%, the fraction of genes from the "Chlorophyta" BUSCO (2,168 genes) that were not found in the genome sequence is considerably high. However, similar results were obtained when analyzing other representatives of the *Trebouxiophyceae* with both free living and symbiotic lifestyles (table 1). A notable exception, with only 2.4% missing BUSCOs, is *Coccomyxa subellipsoidea*. This is, however, not surprising because this species was used for the initial compilation of the "Chlorophyta" BUSCO set. We have shown previously that even highly fragmented genome assemblies can recover most of the BUSCO genes (Greshake et al. 2016). Thus, our results indicate that the plasticity of the algal gene set might be higher than hitherto acknowledged.

No Evidence for Horizontal Gene Transfer in *U. pustulata*

The lichen symbiosis, an evolutionarily old, obligate, and stable association of individuals from different species, should provide an optimal basis for the mutual exchange of genetic material. We therefore screened the fungal genome assembly for indications of horizontal acquisitions of either algal or bacterial genes. Ten fungal genes were classified as of algal and further 12 as of bacterial origin. All genes are located amidst fungal genes in the genome assembly. However, a subsequent case-by-case curation of these 22 genes revealed that the taxonomic assignments by MEGAN are, in all instances, borderline cases (supplementary table S7, Supplementary Material online). The sequence similarity of the corresponding genes to an algal or bacterial protein, which served as basis for the classification, was low, and only slightly higher than the similarity to the closest fungal gene. Only a slight shift in the parameterization of MEGAN's taxonomic classification algorithm left these genes essentially taxonomically unassigned. Thus, the true evolutionary origin remains unknown for all 22 genes. Individual examples of genetic exchange between lichenized fungi and their algal partners have been reported before (e.g., Wang et al. 2014; Beck et al. 2015). Here, we find no convincing evidence for the horizontal acquisition of either algal or bacterial genes by *U. pustulata*.

Lineage-Specific Absence of Evolutionarily Old Genes in *U. pustulata*

We subsequently increased the resolution of the gene set analysis to search for 9,081 genes that were present in the last common ancestor of the *Lecanoromycetes* (LCA_{Lec}; see supplementary text, Supplementary Material online). For 142 LCA_{Lec} genes, we were missing an ortholog only in the *U. pustulata* gene set, suggesting, on the first sight, an exclusive loss on the *U. pustulata* lineage. On closer scrutiny, however, all but 33 of these genes had been either missed during genome annotation or reside in assembly gaps because an ortholog could be detected in the transcript data. A corresponding analysis in genes exclusively missing in *C. grayi* and *U. muehlenbergii* obtained similar results (supplementary text, table S8, and fig. S8, Supplementary Material online). Taking the absence of genes in annotated gene sets at face value can, therefore, lead to wrong evolutionary inferences (Deutekom et al. 2019). However, for 33 LCA_{Lec} genes, we could find, to this point, no indication of an experimental artifact, and they appear genuinely absent from the *U. pustulata* genome assembled by us (supplementary table S9, Supplementary Material online). Four of these genes are represented by an ortholog in the closely related *U. hispanica*

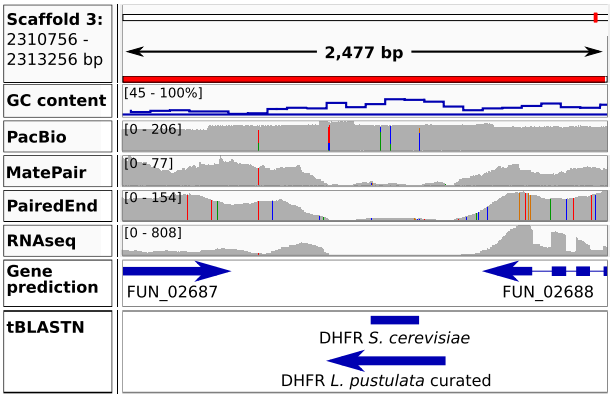

Fig. 5.—Read coverage distribution in the DHFR locus. Coverage pattern at the DHFR locus (scaffold 3: 2,310,756–2,313,256). Although the read coverage is consistently high for PacBio (~200×), there is a marked decrease for the two Illumina whole-genome shotgun libraries toward the center of this region. This decrease coincides with a marked increase of the GC content up to 79%. A TBlastN search using the dihydrofolate synthase of *Saccharomyces cerevisiae* (UniProt-ID: P07807) obtains a partial hit in the central part of region. Eight frameshift mutations in the coding sequences of DHFR were manually corrected (supplementary figs. S9–S19, Supplementary Material online) resulting in a curated putative protein of 210 aa in length.

(Dal Grande et al. 2018), dating their putative loss to after the split of the two *Umbilicaria* species. In three cases, a subsequent manual curation found no evidence against the gene loss assumption. The three genes encode an oxidoreductase with a significant sequence similarity to gibberellin-20-oxidases, a putative methyl-transferase, and a protein with unknown function. The functional consequences of these alleged losses remain to be determined. Moreover, it is not yet clear whether the absence of these genes is fixed within *U. pustulata*, or whether it represents a copy number variation between different populations of this species (Zhao and Gibbons 2018). For the fourth gene encoding a dihydrofolate reductase (DHFR), however, our curation revealed an error source in the gene identification, which is typically neglected. *DHFR* encodes a protein, which is involved in the basal nucleotide metabolism. This gene is almost ubiquitously present throughout fungi and animals. Its absence in *U. pustulata* therefore would imply far-reaching changes in metabolism (Huang et al. 1992). Our manual curation could exclude assembly errors and genomic rearrangements as likely explanations for the absence of DHFR (fig. 5). A TBlastN search with the *Saccharomyces cerevisiae* DHFR (UniProt-ID: P07807) as query obtained a partial hit in this region, which indicated that the open reading frame (ORF) of DHFR is disrupted by several frameshift mutations. Because this region is covered by about 200 PacBio reads, sequencing errors appeared unlikely suggesting a recent pseudogenization of *DHFR* in the lineage leading to *U. pustulata*. However, we noted a very low Illumina read coverage at the DHFR locus (fig. 5). This coverage drop coincides with an extraordinary high GC content of up to 79% paired with the presence of extended stretches of self-complementarity (fig. 6). In combination, this can lead to the formation of stable stem loops that can interfere with both DNA amplification and sequencing (Benjamini and Speed 2012; Ross et al. 2013; Schirmer et al. 2016). We suspected that the low Illumina read coverage rendered assembly polishing with Pilon less effective. Indeed, a visual inspection exploiting the few Illumina reads that map to the DHFR locus identified six of eight frameshift mutations as recurrent sequencing errors in the underlying PacBio reads (supplementary figs. S9–S14, Supplementary Material online). The remaining two frameshifts toward the 3′-end of the ORF, which are not covered by any Illumina reads, coincide with runs of Gs. Thus, they are very likely to be also sequencing errors (supplementary figs. S15 and S16, Supplementary Material online). Correcting all frameshifts resulted in an uninterrupted ORF (supplementary fig. S17, Supplementary Material online) encoding a full-length DHFR.

To assess the extent to which GC-rich inverted repeats may interfere in general with the correct identification of genes, we annotated inverted repeats (IR) throughout the genome draft sequence of *U. pustulata* with the Inverted Repeat Finder (Warburton et al. 2004). This revealed 1,464 IR, with a median length of 819.5 bp. The GC content of these repeats follows a bimodal distribution peaking at 51% and 75%. Although the number of inverted repeats falls within the values obtained for other genomes of lichenized fungi, IRs with a GC content of over 70% are largely unique to *U. pustulata* (fig. 7). Whether this is due to the fact that only *U. pustulata* was sequenced with a long-read technology that is less sensitive to GC-rich inverted repeats, or whether the other genomes are devoid of such repeats remains to be determined. Overlaying the IR regions with the Illumina and the PacBio read coverage information reveals 467 IR with a mean GC content of 67.8% for which the Illumina read coverage drops to <10×, whereas the PacBio coverage remains uniformly high. Any gene residing in such a region has a considerable chance to be either incorrectly predicted or overlooked due to remaining sequencing errors in the genome draft sequence.

Organellar Genome Annotation

Annotation of the *L. pustulata* mitochondrial genome resulted in 15 protein-coding genes, a small subunit rRNA gene, 33 additional ORFs, and 31 tRNA genes encoding 24 distinct tRNAs (supplementary fig. 2, Supplementary Material online). All 15 fungal core protein-coding genes (Pogoda et al. 2018) are represented, among them atp9, which was found to be frequently missing in the mt genomes of lichenized fungi (supplementary table S10, Supplementary Material online). Although this suggests, on the first sight, a considerably standard layout of the mt genome, a closer look at the annotated genes revealed a number of interesting findings. Most notably, *cox2*, the gene encoding the cytochrome c oxidase

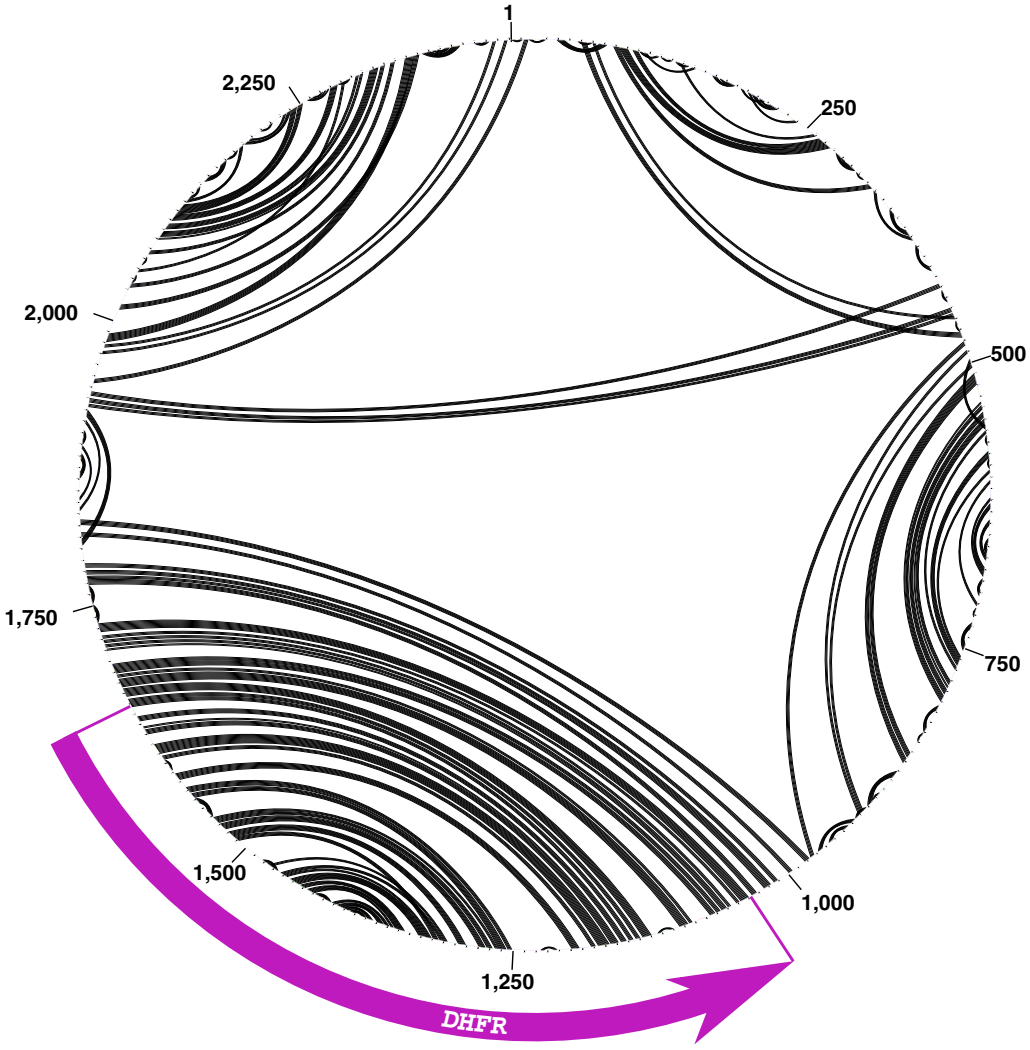

FIG. 6.—Inverted repeats in the DHFR locus. We assessed the potential of the DHFR locus to form secondary structures that may interfere with the Illumina sequencing technology. The plot shows self-complementarity predicted by ProbKnot (Bellaousov and Mathews 2010) as black arcs. The pattern reveals that the DHFR gene in *U. pustulata* is embedded in an inverted repeat spanning ~800 bp.

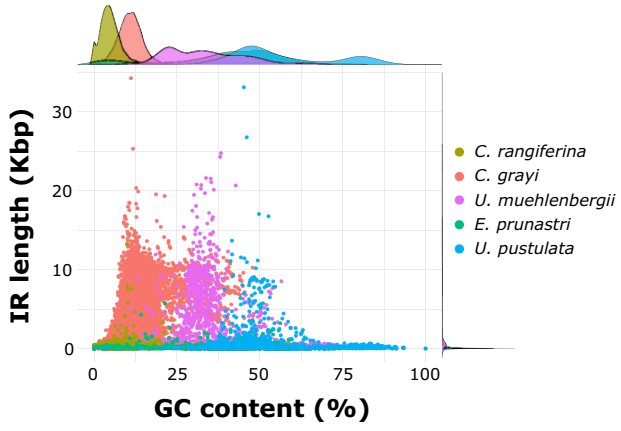

FIG. 7.—The distribution of inverted repeats in the draft genome sequences of five lichenized fungi. Inverted repeats with a GC content above 70% are observed only in *U. pustulata*.

subunit II is fused head-to-tail to *cob*, which encodes cytochrome b, into one transcription unit (supplementary fig. 18, Supplementary Material online). The corresponding Trinity transcript contains an uninterrupted reading frame, suggesting that it is translated into a single fusion protein. To the best of our knowledge, such a fusion as never been reported before, although at least the lecanoromycete *Usnea ceratina* contains a similar fusion (NCBI Gene ID: 34569213). Future studies will have to reveal when during evolution this gene fusion emerged, and at what stage during gene expression—and via what mechanism—the two proteins are separated. Moreover, we noted that nad6, the gene encoding the NADH dehydrogenase subunit 6, is disrupted by the integration of a 2.4-kb long segment, most likely a mobile Group II intron (Lambowitz and Belfort 1993) (supplementary fig. S19, Supplementary Material online). Eventually, three protein-

coding genes do not possess a recognizable stop codon (supplementary table S10, Supplementary Material online). One example is the gene encoding the NADH dehydrogenase subunit 3 (*nad3*). The predicted ORF is covered by three distinct transcripts, indicating that it is not a single transcription unit (supplementary fig. S20, Supplementary Material online). A search against the MitoFun database (http://mitofun.biol.uoa.gr, last accessed February 27, 2020) reveals that the coding sequences encoding nad3 spans approximately the first 396 bp of this ORF. In this region, no canonical stop codon is detected, and the agreement between the about 100 individual RNAseq reads and the genomic sequence suggests that no stop codon is generated posttranscriptionally via RNA editing. BlastP and BlastN searches (Altschul et al. 1997) against the NCBI databases *nr-prot* and *nr*, respectively, revealed no significant hits for the parts of the ORF downstream of *nad3*. The absence of recognizable stop codons in the gene encoding nad3 can be found in the mt genome annotations of other *Lecanoromycetes*, for example, in *Usnea mutabilis* (NCBI GeneID: 38289161) and *Parmotrema ultralucens* (NCBI GeneID: 38466336). It remains unclear how lichenized fungi achieve an accurate termination of the translation for such genes. Of the remaining 36 ORFs annotated in the *U. pustulata* mt genome, 9 encode homing endonucleases that have been proposed to act as selfish genetic elements driving changes in both mt genome size and gene order (Aguileta et al. 2014; Kanzi et al. 2016).

The annotation of the *Trebouxia* sp. mitochondrial genome revealed 32 protein-coding genes, 20 additional ORFs, and 26 tRNAs, which agrees with previous findings in the *Trebouxiophyceae* (Fan et al. 2017). Similar to other plant and algal species (Ko and Kim 2016), we found a nuclear copy of the mtGenome (NUMT), which was identified via a local increase of the read coverage in the Anvio'o plot shown in figure 2. In the chloroplast genome, we could annotate 78 protein-coding genes, 3 ribosomal RNAs, 52 additional ORFs, and 31 tRNA. The set of annotated genes comprises all green algal core genes, and additionally 15 out of 16 common algal chloroplast genes showing sporadic lineage-specific gene loss (Turmel et al. 2015). Interestingly, the missing ribosomal protein, rps4, is encoded on scaffold 44 of the algal nuclear genome assembly. Here, it is flanked by two genes, whose counterparts in other green algae are located in the nucleus (supplementary fig. S21, Supplementary Material online), and the read coverage pattern provides no hint for any assembly error. This indicates a relocation of rps4 from the chloroplast to the nucleus in *Trebouxia* sp. Recently, it was hypothesized that a fission of the tRNA-Ile lysidine synthase encoding gene, *tilS* (Suzuki and Miyauchi 2010), observed in mutualistic or parasitic species of the *Trebouxiophyceae* might be connected to symbiosis (Armaleo et al. 2019). The corresponding gene *ycf62* in the chloroplast genome of *Trebouxia* sp. encodes a 725 aa long polypeptide (supplementary fig. S4, Supplementary Material online). It harbors the full Pfam domain ATP_bind_3 (PF1171.20) representing the TilS/TtcA_N domain (IPR011063) (supplementary fig. S22, Supplementary Material online), similar to the situation in most chlorophyte and streptophyte *tilS* proteins. The two further domains of bacterial tRNA-Ile lysidine synthases described by Suzuki and Miyauchi (2010), tilS (PF09179.11) and tilS-C (PF11734.8) (supplementary fig. S23, Supplementary Material online), are absent from all eukaryotic *tilS* proteins described thus far. In essence, we found no evidence for a fission of this gene in *Trebouxia* sp.

Conclusion

Here, we have shown that the reconstruction of the holo-genome for an obligate symbiotic community purely from metagenomic sequence reads at contiguities comparable to assemblies for single-species samples is feasible. The greatly varying coverage ratios for the individual genomes, spanning three orders of magnitude, emerged as the most challenging task. Key to success was the combination of short Illumina and long PacBio reads with a comprehensive assembly scheme. In particular, we had to 1) target different components of the holo-genome with different assembly methodologies, 2) include taxonomic assignments on the contig level, 3) perform a merging of contigs from different assembly approaches that were assigned to the same taxonomic group, and 4) perform a final scaffolding step. Numerous benchmark studies have indicated that there is no general gold standard for a genome assembly procedure (Dominguez Del Angel et al. 2018). Thus, our workflow should be considered a template that can be adapted to the needs of the precise symbiotic community under study. The initial analysis of the *U. pustulata* holo-genome already revealed a number of genetic changes both in the nuclear and in the organellar genomes whose functional relevance for this obligate lichen symbiosis will be interesting to determine. However, we encountered also a number of pitfalls that, if remain unnoticed, lead to wrong conclusions. One of the main advantages of metagenomic approaches is that holo-genome reconstruction, relative genomic copy number assessment, taxonomic classification and relative taxon abundance estimation will be performed on the same data. It is tempting to use the assembled contigs for the taxonomic assignments, because longer sequences will allow a classification with greater confidence. If the aim is, however, to assess the abundance of individual taxa in microbial community, the analysis has to take the read data into account. Either by performing the taxonomic assignment at the read level—bearing the risk that a fraction of reads will remain unclassified—or by taking the read coverage of the taxonomically assigned contigs into account, which will miss rare taxa covered by only few reads. From an evolutionary perspective, the availability of genome sequences for an obligate symbiotic community is the relevant starting point for determining the genetic changes underlying the dependency

of the symbionts. A comprehensive gene annotation is essential for such analyses, which have a strong focus on detecting loss of individual genes. BUSCO and FGMP analyses provide an initial indication for the completeness of gene annotations. However, a number of genes in both BUSCO and FGMP sets are compared with the gene set of a species, typically small, and they are often not designed for the phylogenetic clade in focus, that is, *Lecanoromycetes* and *Trebouxiophyceae* in this study. On the example of the *Trebouxiophyceae*, we showed that the latter aspect makes it difficult to differentiate between the absence of BUSCO genes due to an incomplete gene set reconstruction, or due to a higher than expected number of BUSCO gene losses. The use of tailored core gene sets for the clade of interest, paired with targeted ortholog searches both in the annotated gene set and in the assembled transcriptome data, is an alternative that substantially increases resolution. Genes that then remain undetected are good candidates for a lineage-specific loss with all its consequences for the symbionts' metabolism. Still, this does not exclude an artifact. It was only the suspicious deviation in coverage between the PacBio reads and the Illumina reads, which eventually revealed that the gene encoding the DHFR was not lost in *U. pustulata*. Ultima ratio remains, therefore, expert candidate curation considering all evidences that can hint toward an artifact mimicking gene loss.

Supplementary Material

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

Acknowledgments

The present study is a result of the Centre for Translational Biodiversity Genomics (LOEWE-TBG). The authors thank Anjuli Calchera (Frankfurt) for technical assistance, and Pavel Škaloud (Dpt. Botany, Charles University, Prague) for useful discussion on *Trebouxia* organellar genomics and ontogeny. Moreover, they acknowledge Daniele Armaleo and Basil Britto for providing access to the organelle genomes of *C. grayi* and *A. glomerata*, and Olafur S. Andresson for sharing the *L. pulmonaria* data. Moreover, we acknowledge two anonymous reviewers for their constructive comments. This study was funded through the programme "LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts.

Literature Cited

Abdel-Hameed M, Bertrand RL, Piercey-Normore MD, Sorensen JL, 2016. Putative identification of the usnic acid biosynthetic gene cluster by de novo whole-genome sequencing of a lichen-forming fungus. Fungal Biol. 120(3):306–316.

Aguileta G, et al. 2014. High variability of mitochondrial gene order among fungi. Genome Biol Evol. 6(2):451–465.

Ahmadjian V. 1993. The lichen symbiosis. New York: John Wiley and Sons, Inc.

Allen JL, McKenzie SK, Sleith RS, Alter SE, 2018. First genome-wide analysis of the endangered, endemic lichen *Cetradonia linearis* reveals isolation by distance and strong population structure. Am J Bot. 105(9):1556–1567.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Armaleo D, Sun X, Culberson C, 2011. Insights from the first putative biosynthetic gene cluster for a lichen depside and depsidone. Mycologia 103(4):741–754.

Armaleo D, et al. 2019. The lichen symbiosis re-viewed through the genomes of *Cladonia grayi* and its algal partner *Asterochloris glomerata*. BMC Genomics. 20(1):605.

Aschenbrenner IA, Cardinale M, Berg G, Grube M, 2014. Microbial cargo: do bacteria on symbiotic propagules reinforce the microbiome of lichens? Environ Microbiol. 16(12):3743–3752.

Bankevich A, et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bates ST, Cropsey GW, Caporaso JG, Knight R, Fierer N, 2011. Bacterial communities associated with the lichen symbiosis. Appl Environ Microbiol. 77(4):1309–1314.

Beck A, Divakar PK, Zhang N, Molina MC, Struwe L, 2015. Evidence of ancient horizontal gene transfer between fungi and the terrestrial alga *Trebouxia*. Org Divers Evol. 15:235–248.

Bellaousov S, Mathews DH, 2010. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA 16(10):1870–1880.

Belova SE, Suzina NE, Rijpstra WIC, Sinninghe Damsté JS, Dedysh SN, 2018. *Edaphobacter lichenicola* sp. nov., a member of the family *Acidobacteriaceae* from lichen-dominated forested tundra. Int J Syst Evol Microbiol. 68(4):1265–1270.

Benjamini Y, Speed TP, 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40(10):e72.

Berlin K, et al. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 33(6):623–630.

Bertrand RL, Sorensen JL, 2018. A comprehensive catalogue of polyketide synthase gene clusters in lichenizing fungi. J Ind Microbiol Biotechnol. 45(12):1067–1081.

Besemer J, Borodovsky M, 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 33(Web Server):W451–W454.

Boetzer M, Pirovano W, 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15(1):211.

Bolger AM, Lohse M, Usadel B, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Bowers RM, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol. 35(8):725–731.

Bradnam KR, et al. 2013. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2(1):10.

Buchfink B, Xie C, Huson DH, 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Calchera A, Dal Grande F, Bode HB, Schmitt I, 2019. Biosynthetic gene content of the 'perfume lichens' *Evernia prunastri* and *Pseudevernia furfuracea*. Molecules 24(1):203.

Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. Proc German Conf Bioinformatics 99:45–56.
Chin CS, et al. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 13(12):1050–1054.
Choe YH, et al. 2018. Comparing rock-inhabiting microbial communities in different rock types from a high arctic polar desert. FEMS Microbiol Ecol. 94:fiy070.
Cisse OH, Stajich JE, 2019. FGMP: assessing fungal genome completeness. BMC Bioinformatics 20:184.
Cubero OF, Crespo A. 2002. Isolation of Nucleic Acids From Lichens. In: Kranner IC, Beckett RP, Varma AK, editors. Protocols in Lichenology. Springer Lab Manuals. Berlin, Heidelberg: Springer.
Dal Grande F, et al. 2017. Adaptive differentiation coincides with local bioclimatic conditions along an elevational cline in populations of a lichen-forming fungus. BMC Evol Biol. 17:93.
Dal Grande F, et al. 2018. The draft genome of the lichen-forming fungus *Lasallia hispanica* (Frey) Sancho & A. Crespo. Lichenologist 50:329–340.
Denton JF, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 10(12):e1003998.
Deutekom ES, Vosseberg J, van Dam TJP, Snel B, 2019. Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences. PLoS Comput Biol. 15(8):e1007301.
Dominguez Del Angel V, et al. 2018. Ten steps to get started in Genome Assembly and Annotation. F1000Research 7:148.
Dunne MP, Kelly S, 2017. OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations. BMC Genomics. 18(1):390.
Edgar RC, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.
Eren AM, et al. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.
Erlacher A, et al. 2015. Rhizobiales as functional and endosymbiontic members in the lichen symbiosis of *Lobaria pulmonaria* L. Front Microbiol. 6:53.
Fan W, Guo W, Van Etten JL, Mower JP, 2017. Multiple origins of endosymbionts in *Chlorellaceae* with no reductive effects on the plastid or mitochondrial genomes. Sci Rep. 7(1):10101.
Gallaher SD, et al. 2018. High-throughput sequencing of the chloroplast and mitochondrion of *Chlamydomonas reinhardtii* to generate improved de novo assemblies, analyze expression patterns and transcript speciation, and evaluate diversity among laboratory strains and wild isolates. Plant J. 93(3):545–565.
Gotz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420–3435.
Greshake B, et al. 2016. Potential and pitfalls of eukaryotic metagenome skimming: a test case for lichens. Mol Ecol Resour. 16(2):511–523.
Grube M, Cardinale M, de Castro JV, Müller H, Berg G, 2009. Species-specific structural and functional diversity of bacterial communities in lichen symbioses. ISME J. 3(9):1105–1115.
Grube M, Spribille T, 2012. Exploring symbiont management in lichens. Mol Ecol. 21(13):3098–3099.
Grube M, et al. 2015. Exploring functional contexts of symbiotic sustain within lichen-associated bacteria by comparative omics. ISME J. 9(2):412–424.
Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 9(1):R7.
Haas BJ, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(8):1494–1512.
Hauck M, Willenbruch K, Leuschner C, 2009. Lichen substances prevent lichens from nutrient deficiency. J Chem Ecol. 35(1):71–73.
Hestmark G, 1992. Sex, size, competition and escape-strategies of reproduction and dispersal in *Lasallia pustulata* (Umbilicariaceae, Ascomycetes). Oecologia 92(3):305–312.
Holt C, Yandell M, 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12(1):491.
Huang T, Barclay BJ, Kalman TI, von Borstel RC, Hastings PJ, 1992. The phenotype of a dihydrofolate reductase mutant of *Saccharomyces cerevisiae*. Gene 121(1):167–171.
Huang W, Li L, Myers JR, Marth GT, 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28(4):593–594.
Hunt M, et al. 2015. Circlator: automated circularization of genome assemblies using long sequencing reads. Genome Biol. 16(1):294.
Huson DH, et al. 2016. MEGAN community edition—interactive exploration and analysis of large-scale microbiome sequencing data. PLoS Comput Biol. 12(6):e1004957.
Kanehisa M, Sato Y, Morishima K, 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol. 428(4):726–731.
Kanzi AM, Wingfield BD, Steenkamp ET, Naidoo S, van der Merwe NA, 2016. Intron derived size polymorphism in the mitochondrial genomes of closely related *Chrysoporthe* species. PLoS One 11(6):e0156104.
Kent WJ, 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12(4):656–664.
Kim D, Langmead B, Salzberg SL, 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 12(4):357–360.
Kirk P, Minter DW, Stalpers JA. 2008. Dictionary of the fungi. Wallingford (United Kingdom): CAB International.
Ko YJ, Kim S, 2016. Analysis of nuclear mitochondrial DNA segments of nine plant species: size, distribution, and insertion loci. Genomics Inform. 14(3):90–95.
Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.
Korf I, 2004. Gene finding in novel genomes. BMC Bioinformatics 5(1):59.
Kranner I, Beckett R, Hochman A, Nash TH, 2008. Desiccation-tolerance in lichens: a review. Bryologist 111(4):576–593.
Kummerova M, et al. 2006. Inhibitory effect of fluoranthene on photosynthetic processes in lichens detected by chlorophyll fluorescence. Ecotoxicology 15:121–131.
Lambowitz AM, Belfort M, 1993. Introns as mobile genetic elements. Annu Rev Biochem. 62(1):587–622.
Langmead B, Salzberg SL, 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359.
Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M, 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics 30(4):566–568.
Liu F, et al. 2019. Draft genome sequences of five *Calonectria* species from *Eucalyptus* plantations in China, *Celoporthe dispersa*, *Sporothrix phasma* and *Alectoria sarmentosa*. IMA Fungus 10(1):22.
Magoc T, Salzberg SL, 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27(21):2957–2963.
McDonald TR, Mueller O, Dietrich FS, Lutzoni F, 2013. High-throughput genome sequencing of lichenizing fungi to assess gene loss in the ammonium transporter/ammonia permease gene family. BMC Genomics. 14(1):225.
Meiser A, Otte J, Schmitt I, Grande FD, 2017. Sequencing genomes from mixed DNA samples—evaluating the metagenome skimming approach in lichenized fungi. Sci Rep. 7(1):14881.
Muggia L, Grube M, 2018. Fungal diversity in lichens: from extremotolerance to interactions with algae. Life (Basel) 8:pii: E15.

Nayfach S, Pollard KS, 2016. Toward accurate and quantitative comparative metagenomics. Cell 166(5):1103–1116.
Neubauer M, et al. 2015. Mitochondrial dynamics in the pathogenic mold *Aspergillus fumigatus*: therapeutic and evolutionary implications. Mol Microbiol. 98(5):930–945.
Noh HJ, et al. 2019. *Lichenihabitans psoromatis* gen. nov., sp. nov., a member of a novel lineage (*Lichenihabitantaceae* fam. nov.) within the order of *Rhizobiales* isolated from Antarctic lichen. Int J Syst Evol Microbiol. 69(12):3837–3842.
Ondov BD, Bergman NH, Phillippy AM, 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12(1):385.
Pankratov TA, 2012. Acidobacteria in microbial communities of the bog and tundra lichens. Microbiology 81(1):51–58.
Pankratov TA, Dedysh SN, 2010. *Granulicella paludicola* gen. nov., sp. nov., *Granulicella pectinivorans* sp. nov., *Granulicella aggregans* sp. nov. and *Granulicella rosea* sp. nov., acidophilic, polymer-degrading acidobacteria from *Sphagnum* peat bogs. Int J Syst Evol Microbiol. 60(12):2951–2959.
Pankratov TA, et al. 2020. *Lichenibacterium ramalinae* gen. nov, sp. nov., *Lichenibacterium minor* sp. nov., the first endophytic, *beta*-carotene producing bacterial representatives from lichen thalli and the proposal of the new family *Lichenibacteriaceae* within the order *Rhizobiales*. Antonie Van Leeuwenhoek. 113(4):477–489.
Park C, Kim K, Kim O, Jeong G, Hong S, 2016. Bacterial communities in Antarctic lichens. Antarct Sci. 28(6):455–461.
Parra G, Bradnam K, Korf I, 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061–1067.
Pogoda CS, Keepers KG, Lendemer JC, Kane NC, Tripp EA, 2018. Reductions in complexity of mitochondrial genomes in lichen-forming fungi shed light on genome architecture of obligate symbioses. Mol Ecol. 27(5):1155–1169.
Ross MG, et al. 2013. Characterizing and measuring bias in sequence data. Genome Biol. 14(5):R51.
Sangwan N, Xia F, Gilbert JA, 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4(1):8.
Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C, 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17(1):125.
Sigurbjornsdottir MA, Andresson OS, Vilhelmsson O, 2016. Nutrient scavenging activity and antagonistic factors of non-photobiont lichen-associated bacteria: a review. World J Microbiol Biotechnol. 32:68.
Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM, 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212.
Smit AFA, et al. 2015. RepeatMasker Open-4.0 [Internet]. Seattle: Institute for Systems Biology. Available from: http://www.repeatmasker.org.
Spribille T, et al. 2016. Basidiomycete yeasts in the cortex of ascomycete macrolichens. Science 353(6298):488–492.
Stamatakis A, 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Stanke M, Schoffmann O, Morgenstern B, Waack S, 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7(1):62.
Suzuki T, Miyauchi K, 2010. Discovery and characterization of tRNAIle lysidine synthetase (TilS). FEBS Lett. 584(2):272–277.
Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M, 2011. Next generation sequence assembly with AMOS. Curr Protoc Bioinformatics Chapter 11:Unit 11.8.
Turmel M, Otis C, Lemieux C, 2015. Dynamic evolution of the chloroplast genome in the green algal classes Pedinophyceae and Trebouxiophyceae. Genome Biol Evol. 7(7):2062–2082.
Vollmers J, Wiegand S, Kaster AK, 2017. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective—not only size matters! PLoS One 12(1):e0169662.
Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.
Wang Y, et al. 2018. Identification of a putative polyketide synthase gene involved in usnic acid biosynthesis in the lichen *Nephromopsis pallescens*. PLoS One 13(7):e0199110.
Wang YY, Liu B, Zhang XY, Zhou QM, 2014. Genome characteristics reveal the impact of lichenization on lichen-forming fungus *Endocarpon pusillum* Hedwig (Verrucariales, Ascomycota). BMC Genomics. 15:1–18.
Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G, 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. Genome Res. 14(10a):1861–1869.
Zhao S, Gibbons JG, 2018. A population genomic characterization of copy number variation in the opportunistic fungal pathogen *Aspergillus fumigatus*. PLoS One 13(8):e0201611.

Associate editor: Richard Cordaux