A closely-related clade of globally distributed bloom-forming cyanobacteria within the Nostocales

Connor B. Driscoll[1], Kevin A. Meyer[2,3], Sigitas Šulčius[4], Nathan M. Brown[1], Gregory J. Dick[2], Huansheng Cao[5], Giedrius Gasiūnas[6], Albertas Timinskas[7], Yanbin Yin[8], Zachary C. Landry[1], Timothy G. Otten[1], Timothy W. Davis[9], Susan B. Watson[10], Theo W. Dreher[1,11*]

[1] Department of Microbiology, Oregon State University, 226 Nash Hall, Corvallis, OR, 97331, USA.
[2] Department of Earth & Environmental Sciences, University of Michigan, Ann Arbor, MI 48109-1005
[3] Cooperative Institute for Great Lakes Research (CIGLR), University of Michigan, Ann Arbor, MI 48109-1005
[4] Laboratory of Algology and Microbial Ecology, Akademijos Str. 2, LT-08412, Vilnius, Lithuania
[5] Biodesign Center for Fundamental and Applied Microbiomics, Arizona State University 427 E Tyler Mall, Tempe, AZ 85287, USA
[6] Department of Protein-DNA Interactions, Institute of Biotechnology, Vilnius University, Saulėtekio av. 7, LT-10257, Vilnius, Lithuania
[7] Department of Bioinformatics, Institute of Biotechnology, Vilnius University, Saulėtekio 7, LT-10257 Vilnius, Lithuania
[8] Department of Biological Sciences, Northern Illinois University, DeKalb, Illinois, USA.
[9] Department of Biological Sciences, Bowling Green State University, Bowling Green, OH 43402
[10] Environment and Climate Change Canada, Canada Centre for Inland Waters, Burlington ON L7S 1A1
[11] Center for Genome Research and Biocomputing, Oregon State University, Corvallis, OR 97331 USA.

* Correspondence: Theo Dreher, Department of Microbiology, Oregon State University, 226 Nash Hall, Corvallis, OR, 97331, USA; theo.dreher@oregonstate.edu ph: 541-737-1795 fax: 541-737-0496

Current addresses:
CBD: Department of Immunology, Center for Innate Immunity and Immune Disease, University of Washington, Seattle, WA 98109
TGO: Bend Genetics, LLC, 87 Scripps Drive, Ste. 108, Sacramento, CA 95825
GG: CasZyme, Saulėtekio al. 7c, LT-10257, Vilnius, Lithuania
ZCL: Institut für Umweltingenieurwissenschaften, ETH Zürich, HIL G37.2, Stefano-Franscini-Platz 5, 8093 Zürich, Switzerland

Running title: Genomics of a new clade of bloom-forming Nostocales

Keywords: genomics, metagenomics, filamentous cyanobacteria

Graphical abstract

Highlights

- Eight new Nostocales genomes derived from cultures or metagenomes from recent HABs
- New genus-level clade of HAB-forming *Anabaena/Dolichospermum/Aphanizomenon* genomes
- Analysis of differential gene content relevant to niche partitioning
- Widespread presence of ribosomally encoded peptides (esp. bacteriocins)

Abstract

In order to better understand the relationships among current Nostocales cyanobacterial blooms, eight genomes were sequenced from cultured isolates or from environmental metagenomes of recent planktonic Nostocales blooms. Phylogenomic analysis of publicly available sequences placed the new genomes among a group of 15 genomes from four continents in a distinct ADA clade (*Anabaena/Dolichospermum/Aphanizomenon*) within the Nostocales. This clade contains four species-level groups, two of which include members with both *Anabaena*-like and *Aphanizomenon flos-aquae*-like morphology. The genomes contain many repetitive genetic elements and a sizable pangenome, in which ABC-type transporters are highly represented. Alongside common core genes for photosynthesis, the differentiation of N_2-fixing heterocysts, and the uptake and incorporation of the major nutrients P, N

and S, we identified several gene pathways in the pangenome that may contribute to niche partitioning. Genes for problematic secondary metabolites—cyanotoxins and taste-and-odor compounds—were sporadically present, as were other polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) gene clusters. By contrast, genes predicted to encode the ribosomally generated bacteriocin peptides were found in all genomes.

Keywords: Genomics Metagenomics Nostocaceae

1. Introduction

Cyanobacteria are a diverse group of photoautotrophic bacteria with important roles in the biogeochemical cycles of aquatic and terrestrial habitats. They have played an important role in atmospheric oxygen accumulation on Earth through oxygenic photosynthesis, while assimilating carbon and, in some cases nitrogen, into food chains (Canfield, 2005; Karl et al., 1997). Their diversity encompasses growth in a range of environments, including saltwater, freshwater, soil, and deserts (Biller et al., 2015; Garcia-Pichel et al., 2001; Oliver and Ganf, 2000), as well as in symbioses with plants, animals and fungi (Raven, 2002). In recent years, potentially toxic blooms of cyanobacteria have increased in frequency and severity in many fresh and brackish water bodies, raising ecological and public health concerns (Davis and Gobler, 2016; Paerl et al., 2001). Such cyanobacterial harmful algal blooms (CyanoHABs) are frequently caused by members of the Order Nostocales, many of whose members are distinguished by their ability to produce differentiated cells enabling long-term dormancy (akinetes) and nitrogen fixation (heterocysts).

The Order Nostocales is comprised of a number of families, among which the *Nostocaceae* and *Aphanizomenonaceae* (Guiry and Guiry, 2016; Komarek et al., 2014) include most of the genera associated worldwide with nitrogen-fixing, filamentous CyanoHABs: *Anabaena, Aphanizomenon, Cylindrospermopsis/Raphidiopsis*, *Cylindrospermum*, *Dolichospermum, Nodularia, and Nostoc*. As is general of cyanobacteria, these members of the Nostocales are capable of synthesizing a rich diversity of

secondary metabolites from nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) gene clusters (Calteau et al., 2014; Wang et al., 2015), as well as ribosomally made peptides (Dittmann et al., 2015; Welker and Von Döhren, 2006). Cyanobacterial secondary metabolites of particular public health concern include the toxins anatoxin-a, cylindrospermopsin, microcystin, nodularin, and saxitoxin (Burford et al., 2016; Cirés and Ballot, 2016; Li et al., 2016; Pearson et al., 2016) as well as the offensive taste-and-odor compounds geosmin and 2-methylisoborneol that impair drinking and recreational water quality (Li et al., 2016; Watson et al., 2016). This arsenal is thought to benefit cyanobacteria at least in part through allelopathic interactions that inhibit grazers and competitors (Welker and Von Döhren, 2006), augmenting other mechanisms that allow CyanoHABs to occur, such as regulated buoyancy, colony formation, efficient nutrient acquisition and tolerance of extremes in irradiance and salinity.

The publicly available Nostocales genomes are mostly derived from cultures collected decades ago (Table S1) and only sparsely represent the many CyanoHABs that annually afflict inland waters. This study is intended to address this knowledge gap and add to our genomic knowledge of extant examples of bloom-forming Nostocales, centered on the *Anabaena/Dolichospermum* and *Aphanizomenon* genera. Although these genera are amongst the most common components of CyanoHABs (Li et al., 2016), *Anabaena* sp. 90 and *Anabaena* sp. WA102 are the only members whose genomes have been analyzed in detail (Brown et al., 2016; Wang et al., 2012). Comparative genomics can enhance our understanding of the Nostocales genetic repertoire and their evolutionary relationships, and may assist attempts to elucidate niche differentiating characteristics that might help to explain and predict the timing of bloom events. The paucity of reference genomes also limits the exploitation of molecular probes for monitoring or research needs, and the efficient interpretation of metagenomic and metatranscriptomic data that can describe the population structure and physiology of natural blooms (Harke et al., 2016; Otten et al., 2016). Finally, these cyanobacteria are currently grouped according to a taxonomic classification system that remains confused despite considerable revision in recent years, resulting in inconsistent nomenclature (Li et al., 2016).

Whole genome sequences of multiple members should provide the clearest guidance for taxonomic assignments. Recent taxonomic proposals have retained an emphasis on a polyphasic approach, in which phenotypic characteristics have significant weight alongside only limited use of phylogenetic criteria (Komárek, 2010; Komarek et al., 2014; Wacklin et al., 2009). This has resulted in the proposal to separate the genus *Anabaena* based on a phenotypic character (the presence or absence of gas vesicles), with benthic forms retaining their original name and planktonic forms assigned to the new genus *Dolichospermum* (Wacklin et al., 2009). There is, however, at present no phylogenetic rationale for such a distinction, since benthic and planktonic strains are phylogenetically intermixed (Rajaniemi et al., 2005), and some Nostocales may oscillate between these lifestyles (Halinen et al., 2008). A problem with both the preexisting and revised nomenclature is that *Aphanizomenon* and *Anabaena/Dolichospermum* are polyphyletic and intermixed (Gugger et al., 2002; Rajaniemi et al., 2005). Finally, some long-standing planktonic *Anabaena* isolates that have been well-studied but whose relationship to CyanoHABs is uncertain, are genetically close to the *Nostoc* genus (Shih et al., 2013)—indeed, *Anabaena* sp. PCC 7120 is now often referred to as *Nostoc* sp. PCC 7120—a genus that is itself polyphyletic (Shih et al., 2013).

Here, we report a comparative analysis of eight novel genomes and five additional genomes that have only been briefly reported (Cao et al., 2014; D'Agostino et al., 2014; Šulčius et al., 2015). These genomes cluster into a newly recognized *Anabaena/Dolichospermum/Aphanizomenon* (ADA) clade within the Nostocales whose members originate from CyanoHABs from three of the world's continents. We assessed the phylogenomic relationships within these genomes and assessed the distribution of gene content relevant to bloom formation and dominance.

2. Material and methods

2.1 Genome sequencing

The novel genome sequences included in our analyses originated from a number of lakes in the U.S.A., with each assembled from either environmental metagenomes or sequenced cultures (Table 1, Table S1). A uni-algal culture of *Aphanizomenon flos-aquae* LD13 was maintained in BG11 medium under white fluorescent illumination of approximately 20 $\mu Em^{-2}s^{-1}$ at 24 °C with a light/dark cycle of

16hr/8hr (Brown et al., 2016). The genomes of *Anabaena* sp. CRKS33, *Anabaena* sp. MDT14b, *Aphanizomenon flos-aquae* MDT14a, *Aphanizomenon* sp. WA102, and *Anabaena* sp. WA113 were obtained from environmentally sampled metagenomes in which the predominant morphotype and genotype could be correlated. After collection of cellular material on 1.2 µm glass fiber filters (VWR), DNA was extracted from filters using GeneRite DNA-EZ RW01 extraction kits. Samples were processed using a Nextera XT library preparation kit, with libraries sequenced using an Illumina HiSeq 2000 instrument with 101 bp, paired-end reads and 450 bp insert sizes (Otten et al., 2016). Sequencing reads were quality screened using Trimmomatic (Bolger et al., 2014), retaining those with Phred scores ≥30. Only sequences with mate pairs and a minimum length of 50 nt were retained. The genomes were assembled with IDBA-UD (Peng et al., 2012), and assembled contigs were binned using PhyloPythiaS+ (Gregor, 2014) and the mmgenome R package (Albertsen et al., 2013) as described (Otten et al., 2016).

Anabaena sp. AL09 and *Anabaena* sp. LE011-02 were maintained in unialgal culture in BG-11 medium under white fluorescent illumination of approximately 38 $\mu Em^{-2}s^{-1}$ at 20 °C with a light/dark cycle of 12hr/12hr. Cultures of 15 mL were spun down and the pelleted cellular material was frozen at -80 °C. Cell pellets were extracted using a Qiagen DNeasy® Blood and Tissue Kit, adding a lysate homogenization step (QiaShredder™ spin-column) prior to DNA purification. Shotgun DNA sequencing was performed using an Illumina HiSeq 2000 with 101 bp paired-end reads and 450 bp insert sizes at the University of Michigan DNA Sequencing Core. Sequence reads were quality controlled with FASTQC version 0.10.0 (http://www.bioinformatice.babraham.ac.uk/projects/fastqc/), dereplicated and trimmed. Genomes were assembled with IDBA-UD (Peng et al., 2012) and assembled contigs were binned using emergent self-organizing maps (ESOM) or tetranucleotide frequencies (Robust ZT transformation) with Databionics ESOM Tools (Dick et al., 2009) and the following parameters: contig length 4-10 kb, training with a K-Batch algorithm (k = 0.15%) for 40 training epochs, standard best match search method, local best match search radius of 8, a Gaussian weight initialization, Euclidean data space function, starting training radius of 204 with linear cooling to 1, and a starting learning rate of 0.5 with linear cooling to 0.1. Bin taxonomy was determined with a combination of BLASTN of contigs (Altschul et al.,

1990) against the Silva SSU Database version 119 (Quast et al., 2013) and phylogenetic analysis using the full marker set of the PhyloSift package (Darling et al., 2014).

Evaluating binned genomes

The new Nostocales genomes are all of draft quality, either binned from environmental metagenomes or from metagenomes derived from uni-algal cultures (Table 1, Supplemental Table 1). Contigs within binned genomes that were identified as contaminant NGS primer or control sequences by NCBI's WGS submission pipeline were removed, as were contaminant rRNA sequences identified by BLAST searches against the nt database (September 2015) that had been included in the original genome bin. Two methods were used to assess the completeness and degree of contamination for the novel genomes. We used CheckM (Parks et al., 2015) to assess the completeness and extent of comtamination for each genome. The mmgenome R package (Albertsen et al., 2013) was used to obtain universal gene counts and copy numbers for binned genomes (Suppl. Table 1).

2.2 Core and Pan-genome analysis

The core genomes of the 15 genomes in the ADA clade were analyzed using the GET_HOMOLOGUES software package (Contreras-Moreira and Vinuesa, 2013; Vinuesa and Contreras-Moreira, 2015). Homologous gene families were identified using the OrthoMCL clustering algorithm (OMCL) with sequence cluster reporting of t=0 and no Pfam-domain composition requirements (Fischer et al., 2011). Core genome size was calculated using the exponential decay models of Tettelin and Willenbrock and the pan-genome size was estimated with the exponential model of Tettelin (Tettelin et al., 2005; Willenbrock et al., 2007). A binomial mixture model (Snipen et al., 2009) classified genes based on distribution within all 15 analyzed genomes into core and pan genome categories (Kaas et al., 2012; Koonin and Wolf, 2008). Strain-specific genes of individual taxa were identified using the parse_pangenome_matrix.pl script in GET_HOMOLOGUES (Contreras-Moreira and Vinuesa, 2013).

2.3 Genome annotations

All genomes were annotated with the NCBI Prokaryotic Genome Annotation Pipeline (PGAP). This pipeline includes rRNA and tRNA annotations by BLAST and tRNAscan, respectively. Gene clusters

from the pan-genome analysis were annotated with KEGG BlastKOALA using the "genus_prokaryotes" database (March 23, 2016). Differences in gene content were assessed by the distribution of KO annotations, while specific gene categories (e.g., sulfur metabolism and photosynthesis, carotenoid-, vitamin-, and glutathione-synthesis pathways) were also analyzed. All protein-coding sequences were assigned to COG categories using RAPSearch 2.16 (Zhao et al., 2012) with the COG database and a 1E-30 E-value cutoff. Genes involved in nitrogen metabolism and heterocyst differentiation were identified by BLASTN relationships to characterized genes in *Nostoc* sp. PCC 7120 together with manual inspection guided by synteny and gene alignments observed using Geneious software together with whole-genome alignments (ADA genomes and *Nostoc* sp. PCC 7120) generated by progressiveMauve (Darling et al., 2010). Annotations of other selected genes were similarly manually curated.

Secondary metabolite genes were identified with antiSMASH 3.04 (Weber et al., 2015) without the inclusive option for cluster identification for all genomes. Toxin synthesis gene clusters were also identified by BLASTN using a custom database containing secondary metabolite synthesis gene clusters previously identified (Dittmann et al., 2015). An E-value of 1E-30 cutoff was used to filter non-significant hits; further manual curation was guided by synteny and gene alignments observed using Geneious software. Extracellular polymeric synthesis (EPS) genes were identified by using genes previously characterized (Pereira et al., 2009; Pereira et al., 2015) in BLASTP searches with an E-value cutoff of 1E-30.

Insertion sequences (IS) were identified using HMMSEARCH with the TnPred IS Hidden Markov Model database (Riadi et al., 2012) and a 1E-30 E-value cutoff. This database contains 47 HMMs for 19 IS families. The components of restriction-modification (R-M) systems within the genomes were identified by performing protein sequence searches with TBLASTN (e-value of ≤1E-100) against known R-M system protein sequences obtained from the REBASE database (Roberts et al., 2015) (accessed on May 8, 2016). VirSorter 1.0.3 (Roux et al., 2015) and PHAST (Zhou et al., 2011) were used to identify regions of putative viral or prophage origin.

CRISPR arrays were identified using CRISPR-finder (http://crispr.i2bc.paris-saclay.fr) (Grissa et al., 2007) with manual proofreading; a minimum of three nearly identical repeats was required. The identification of *cas* genes was performed using BLAST. Spacer and gene sequence analysis was performed within a group. The type of CRISPR-Cas systems was attributed manually according to gene cluster architecture and Cas protein sequences (Makarova et al., 2015). The repeats and Cas protein sequences were aligned using ClustalOmega (http://www.ebi.ac.uk/Tools/msa/clustalo/). The phylogenetic trees were created using ClustalW2 – Phylogeny (http://www.ebi.ac.uk/Tools/phylogeny/clustalw2_phylogeny/). A BLASTN search was performed against the publicly available cyanophage genomes using spacer sequences as a query.

2.4 Phylogenomic tree construction and genome-wide composition analysis

A phylogenomic tree of the 29 Nostocales genomes in Table S1a was generated using a re-implementation of the Hal phylogenomics pipeline (Robbertse et al., 2006; Brown et al., 2016; Landry et al., 2017), resulting in a phylogenomic tree built from all single-copy orthologues shared between all genomes (279 genes). Pairwise genome comparisons were made to calculate genome-wide nucleotide identities (gANI; based on pair-wise shared genes) and the fraction of common genes within each genome (alignment fraction, AF) as described (Varghese et al., 2015).

3. Results and Discussion

3.1 Evaluating binned genomes

One of the goals of our study was to obtain genome sequences relevant to current CyanoHAB events, focusing on blooms in the northwestern USA (Oregon and Washington) and the Great Lakes (Lakes Erie and Ontario), which have experienced massive CyanoHABs in recent years (Bullerjahn et al., 2016). Advances in DNA sequencing and genome assembly (Escobar-Zepeda et al., 2015) have facilitated the extraction of genome sequences from environmental shotgun metagenomes. This can avoid some disadvantages associated with determining genome sequences after the establishment of laboratory cultures, such as bottlenecking, differential selection of genotypes depending on the growth medium used (Gorski, 2012), and culture-derived gene inactivation or loss (Koskiniemi et al., 2012; Mlouka et al.,

2004; Wang et al., 2012). An important disadvantage of an exclusively metagenomic approach is the lack of a reference culture that can be used for experimentation to exploit a newly derived genome sequence. We used both approaches, determining five genome sequences from environmental metagenomes and three from established cultures, each produced using the Illumina platform (Table 1; Suppl. Table 1). Since the cultures were uni-algal rather than axenic, genome assembly in all cases involved binning procedures to discriminate the target genome from other sequences present (see Materials and Methods).

To assess the completeness and degree of contamination for the novel genomes and for Nostocaceae and Aphanizomenonaceae genomes available as of September 2016 (Table 1; Suppl. Table 1) we used CheckM, which uses a taxonomically refined set of marker genes (Parks et al., 2015). The 10 finished genomes were estimated by CheckM to be >98.9% complete with <0.36% contamination by other sequences. The three novel cultured genomes were of similar quality (>98.1% completeness with <0.37% contamination). The five genomes extracted from environmental metagenomes were likewise of high quality (>97.2% complete, with all but one >99% complete), although the estimated contamination levels were slightly higher (0.44-4.2%; Table 1). Three additional cultured genomes that have been only briefly reported and are interpreted for the first time in this study (AFA NIES-81, Dol 131C, Dol 310F; see Table 1 for definition of abbreviated names) were of high quality (>99.5% complete with <0.56% contamination), while a fourth (AFA KM1D3) was less complete (87.5% complete, 7.2% contamination). The high contamination estimate for this last genome may have been affected by the unexplained presence of 450 kbp of duplicated sequence (7.8% of genome) (Supp. Table 1C).

The eight novel genomes had 101-106 of the 107 universal marker genes used by mmgenome to assess genome completeness; seven of those genes are not present in all cyanobacterial genomes (Albertsen et al., 2013), suggesting that absence may not mean genomes are in fact incomplete. We conclude that the novel genomes reported in this study are of high quality in terms of completeness and contamination. Nevertheless, it is prudent to remember that all genome sequences that are incomplete may include small errors of gene content or arrangement (e.g., Brown et al., 2016) and underestimated repetitive sequences (including rRNAs), which are the cause of fragmented assemblies.

3.2 Phylogenomic analysis places the novel genomes from extant CyanoHABs in a distinct ADA clade

We assessed the evolutionary relationships among the available Nostocales genomes (Suppl. Table 1A) by generating a phylogenomic tree based on alignments of single-copy shared orthologues (279 genes) (Robbertse et al., 2011) (Fig. 1). The newly sequenced genomes are part of a well-separated clade that contains additional CyanoHAB-associated isolates, four of which are producers of cyanotoxins of major concern (microcystin, anatoxin-a or saxitoxin). We refer to this as the ADA clade in recognition of its component *Anabaena, Dolichospermum* and *Aphanizomenon* genomes. The clade forms a distinct branch within the Nostocales phylogenomic tree (Fig. 1) (Shih et al., 2013) and is cosmopolitan, with genomes originating from North America, Europe, Asia and Australia.

Four groups, each with three or four members, are represented within the ADA clade; we refer to these as Groups ADA-1 through ADA-4 (Fig. 1). Genome pairs within Groups ADA-2, -3 and -4 have gANI values (average genome-wide nucleotide identity in shared genes) >96.5% and AF values (alignment fraction, representing the extent of shared genes) of >0.65 (Fig. 2; Suppl. Table 2). Based on the proposed values for delineation between bacterial species of 95-96.5% gANI and 0.6 AF (Kim et al., 2014; Varghese et al., 2015), Groups ADA-2 to -4 could be considered distinct species. Group ADA-1 could represent another species, although Ana CRKS33 is slightly more distantly related to the Dol 131C and 310F genomes (96.0% gANI and 0.75 and 0.73 AF, respectively; Suppl. Table 2).

Based on single gene or multi-locus sequence typing relationships, it has previously been observed that *Anabaena* and *Aphanizomenon* isolates are intermixed in the phylogenetic tree and neither genus is monophyletic (Gugger et al., 2002; Rajaniemi et al., 2005). Our whole genome comparisons confirm these findings (Fig. 1). The only fully sequenced *Aphanizomenon* genomes (all *A. flos-aquae,* AFA) fall within the ADA clade, although more divergent *Aphanizomenon* isolates are expected on the basis of 16S rDNA analysis (Gugger et al., 2002; Rajaniemi et al., 2005). Phylogroup ADA-3 includes two cultured *Anabaena* isolates and two cultured *Aphanizomenon flos-aquae* isolates. Despite the morphological distinctions (Suppl. Fig. 1) that have guided their classification, these isolates are close genetic relatives. AFA is readily distinguished from the various *Anabaena* morphotypes by its parallel-sided filaments and

large fascicles that are composed of parallel stacks of filaments often visible to the naked eye. Cyanobacteria with this characteristic morphology (AFA LD13 and AFA MDT14a) (Suppl. Fig. 1) are found in two branches of the ADA clade (Groups ADA-3 and -4), while each has close relative(s) with typical *Anabaena* morphology within their own phylogroup.

3.3 General properties of ADA clade genomes

The genome sizes of ADA clade members range between 4.4 Mbp and 5.9 Mbp, with ADA-3 genomes about 20% larger than most other ADA genomes (Table 1). The ADA genomes are smaller — and have fewer predicted genes—than the 6 to 9 Mbp genomes of several other Nostocales, but considerably larger than the 3.2 to 3.9 Mbp genomes of the HAB-forming *Cylindrospermopsis* and *Raphidiopsis* (Suppl. Table 1A). The ADA genomes have a relatively high proportion of pseudogenes; about twice the number of pseudogenes corrected for genome size are present compared to the non-ADA Nostocales (other than the obligate endosymbiont *Nostoc azollae*, which has a high number of recently inactivated genes; Ran et al., 2010). Considerably more pseudogenes are present in the AFA KM1D3 genome than the other ADA genomes. The G+C contents of ADA genomes are 37-39%, lower than the 40-42% of many of the other Nostocales genomes (Table 1, Suppl. Table 1A).

The completed genomes of two ADA clade members, Ana 90 (ADA-2) and Ana WA102 (ADA-3) both indicate the presence of five rRNA operons (5S-16S-23S), and 44 and 43 tRNA genes, respectively (Table 1, Suppl. Table 1A). Other completed Nostocales genomes contain 1-5 ribosomal RNA operons and 38-49 tRNA genes (Suppl. Table 1A). Our draft genomes contained incomplete and missing rRNA genes (Suppl. Tables 1, 3), a consequence of their repetitive nature interrupting contig assembly. tRNA genes associated with rRNA operons are therefore also at risk of missing from draft genomes ($tRNA^{Ile}$ genes are absent from several draft genomes; Suppl. Table 4). On the other hand, tRNA genes might also be occasionally present on contaminating contigs and not recognizable as contaminants. For these reasons, the rRNA gene number and tRNA gene number and identifications in our draft genomes are provisional. Three members of the ADA-4 phylogroup appear to have 6 or 7 rRNA operons. The largest number of ribosomal operons previously reported in a Nostocales genome is five (Suppl. Table 1A).

Core and pan-genome analysis using orthologous gene clustering was conducted for the 15 members of the ADA clade. The number of core genes approached the asymptote, with about 1500 genes constituting the common gene pool (Fig. 3A). Recent estimates for the core genome sizes of the bloom-forming cyanobacteria *Cylindrospermopsis/Raphidiopsis*, *Microcystis aeruginosa* and *Planktothrix* are about 2000, 2500 and 3000, respectively (Humbert et al., 2013; Meyer et al., 2017; Pancrace et al., 2017; Abreu et al., 2018). Across all of the cyanobacteria, the core genome size has been estimated at 500-560 (Simm et al., 2015). The pan-genome curve, which reached about 9,000 genes for the 15 genomes, remained linear in the 8-15 genome range (Fig. 3B), with about 216 additional genes for each newly sequenced genome. This is a common observation, each new *Escherichia coli* genome in a set of 64 strains adding about 190 new genes (Lukjancenko et al., 2010). The known pan-genomes for *M. aeruginosa* and *Planktothrix* are considerably larger, at about 12,000 and 14,000 genes, respectively (Humbert et al., 2013; Meyer et al., 2017; Pancrace et al., 2017). For *Planktothrix* this may be a consequence of some very large genomes (up to 6.7 Mbp), while *M. aeruginosa* appears to possess an inherently more open pan-genome. *Cylindropspermopsis/Raphidiopsis* is considered to be undergoing genome streamlining, and has a pangenome estimate at only about 4700 genes (Abreu et al., 2018).

Since most of the ADA clade genomes are in draft form and may be missing some key genes (e.g., gene absences at contig breaks, especially in the Ana MDT14b and AFA KM1D3 genomes; see Suppl. Tables 6, 7), we viewed core genes as those present in all but one of the ADA genomes ("soft-core" as defined by Kaas et al., 2012). Of the 2,158 gene clusters identified by OrthoMCL, 751 (34.8%) were assigned to KEGG functional groups (only 6.7% of variable genes found in fewer than 13 of these 15 genomes were assigned). Genes associated with protein synthesis and oxidative phosphorylation were preferentially part of the soft-core genome, while ABC transporter and cysteine and methionine metabolism genes were more abundant in the variable genome (Fig. 3C).

3.4 *Genome architecture and mobilome genes*

Previous comparison of two complete ADA genomes (Ana WA102 and Ana 90; Brown et al., 2016) demonstrated a lack of genome-wide synteny. These genomes, like those of the CyanoHAB genus

Microcystis (Humbert et al., 2013), carry high loads of mobile genetic elements. The extent of synteny can be approximated statistically by computing locally collinear block (LCB) lengths with progressiveMauve (Darling et al., 2010), where LCBs can include some extent of gene insertion/deletion. Average LCB lengths for pairwise comparisons within each ADA group are between 4.1 kbp and 9.6 kbp (Table 2), generally not long enough to accommodate more than one operon. LCB estimates are not necessarily limited by the fragmentation of draft genomes, whose N50 values (50% of assembly is in contigs longer than the given value) are 8-72 kbp (Suppl. Table 1B); further, the closely related Ana WA102 and Ana AL93 share LCBs averaging 21.9 kbp although the latter is a draft genome (Supp. Table 5A).

Ample mobile genetic elements capable of supporting genome rearrangements exist in these genomes. There are from 19 to 129 intact or partial genes annotated as transposases, and from 3 to 16 additional HNH homing endonuclease genes per genome, accounting for 0.35% to as high as 1.9% of the genome (Table 2; Suppl. Table 5B). Repetitive sequence elements, which can support rearrangements via homologous recombination, are highly abundant in the completed genomes Ana WA102 and Ana 90, numbering 1483 and 739 and accounting for 8.3 and 5.1% of these genomes (Table 2, Suppl. Table 5B). Almost 90% of these elements are <500 bp in length and they are highly distributed around the genomes. These elements are underrepresented in many draft genomes; they are almost entirely missing from short-read libraries (Illumina, 101 nt, as in all of the newly sequenced genomes), but abundant in medium-read libraries (~400 nt Roche 454 or Ion Torrent PGM) as used for AFA NIES-81 and AFA KM1D3 sequencing (Table 2). It is noted that the AFA KM1D3 genome has 450 kbp of coding region duplicated (Suppl. Table 1C); this duplication is of unknown origin or significance, consisting of duplicated elements >1 kbp lacking highly repetitive elements.

Phage mediated genome rearrangements may have had limited impact in the recent evolution of these genomes, as no intact prophages were found, and prophage remnants were detected in limited number and not in all genomes (Table 2). The greater prominence of prophage remnants in the two completed genomes (Ana WA102 and Ana 90) does, however, suggest that prophage sequences may have been lost

during draft genome clustering. A 9.1 kb part of one prophage element in the Ana 90 genome exists in two near-identical copies (Suppl. Table 5C) (Wang et al., 2012). A 3.3 kb, 7-gene fragment from the Ana WA102 genome shares regions of homology (>89% nucleotide identity) with the Ana 90 repeats as well as with two contigs from Ana AL93 and one from AFA WA102 (Suppl. Table 5C). BLAST analysis retrieved matches in none of the other ADA genomes. Thus, closely related phages have at one point integrated into three genomes isolated from two lakes from Washington State, USA, and a genome from the Baltic Sea.

3.5 ***Nutrient acquisition systems and assimilation of N, P and S***

Since phosphorus and nitrogen are the key nutrients that drive CyanoHAB population expansions (Conley et al., 2009; Paerl and Otten, 2013), we documented the genes for acquisition and utilization of P and N, as well as S, and compared these gene sets to those present in *Nostoc/Anabaena* sp. PCC 7120 (*Nostoc* PCC 7120), the best characterized Nostocales in terms of gene function (Malatinszky et al., 2017; Muro-Pastor and Hess, 2012). The 15 ADA genomes share two homologous gene clusters annotated as phosphate-specific ATP transporters, with 2 to 4 free-standing phosphate transporter genes (Suppl. Table 6A). A polyphosphate kinase gene for high energy phosphate storage (Rao et al., 2009) is present, as are the utilization genes exopolyphosphatase and two *ppnK* genes for NAD phosphorylation. Alkaline phosphatase and pyrophosphatase genes allow abstraction of phosphate from various molecules, and a phosphonate-specific ABC transporter operon is present to allow import of organic P molecules. Genes that regulate the phosphorus utilization regulon—*phoH* and *phoUSR/sphUSR*—are likewise present in all genomes. The P utilization genes in ADA genomes share homologs of those possessed by *Nostoc* PCC 7120 and represent similar physiological capacities. *Nostoc* PCC differs from the ADA genomes in possessing a large set of *phn* genes that encode a C-P bond lyase system with associated phosphonate ABC transporters (absent in the ADA genomes). Of the other genes encoding enzymes that can liberate P from phosphonates (Villarreal-Chiu et al., 2012), only *palA* (phosphonopyruvate hydrolase) was identified; it was present in all except the ADA-2 genomes (Fig. 4). There are only minor differences among other P utilization genes among the ADA genomes (Suppl. Table 6A).

Many elements of the N utilization gene network are also conserved across the ADA genomes, with clearly orthologous relationships to *Nostoc* PCC 7120, in several cases emphasized by conserved operon design (Suppl. Table 6B, C). There are also several cases in which important differences in gene content exist. Strikingly, these differences are not necessarily congruent with the clustering of ADA genomes into discrete species. A clear case is the set of genes for nitrite and nitrate uptake. In all cases, they are positioned between conserved *nirA* and *narB* nitrite and nitrate reductase genes, but two types of nitrite/nitrate transporter genes exist (Fig. 5, Suppl. Table 6B). The four ADA-3 genomes, as well as Ana CRKS33 (ADA-1) and AFA MDT14a and AFA LD13 (ADA-4), possess the *nrtABCD* genes encoding an ABC-type transporter complex with presumed high affinity for both nitrite and nitrate (Ohashi et al., 2011), as found in *Nostoc* PCC 7120. The other genomes—Dol 131C and Dol 310F (ADA-1), all four ADA-2 genomes and the ADA-4 genomes Ana WA113 and AFA WA102—possess the *nrtP* gene, encoding a nitrite/nitrate MFS family permease. Until recently, it was thought that *nrtP*-dependent uptake was a characteristic of marine cyanobacteria (Bird and Wyman, 2003; Ohashi et al., 2011; Wang et al., 2000), but *nrtP* is present in *Nostoc punctiforme* (Aichi et al., 2006) and has scattered representation among the Nostocales (Fig. 5) and other freshwater cyanobacteria (not shown).

All the ADA genomes have multiple genes for amino acid transport via ABC transporters, for which four transporters with differing amino acid specificities have been identified in *Nostoc* PCC 7120 (Pernil et al., 2015); homologs to two of these genes are present in all ADA genomes, while some genomes lack one or both of the other two (see Suppl. Table 6C for details). Amino acid import offers an alternative to endogenous synthesis, for which genes are present in all genomes. There are less dramatic differences in the genes for utilization of environmental ammonium and urea (Suppl. Table 6B, C). Either two or three *amt* ammonium transporter genes are present, and urease *ureABCDEFG* genes are present in all ADA genomes. Urea-specific ABC transporter genes *urtABCDE* (Valladares et al., 2002) are present in all genomes except Ana 90 but are not universally present in the Nostocales (Fig. 5).

Most of the regulatory genes in *Nostoc* PCC 7120 that influence the expression of the N-gene regulon and the differentiation and specialized gene expression of the N-fixing heterocysts (Ehira and Ohmori,

2006; Flores and Herrero, 2010; Muro-Pastor and Hess, 2012; Ramírez et al., 2005; Wang and Xu, 2005; Xu et al., 2008; Zhang et al., 2007) are conserved in the ADA genomes (Suppl. Table 6B, C). There is also strong conservation of the genes for heterocyst-specific glycolipid and envelope polysaccharide synthesis (Fan et al., 2005; Huang et al., 2005; Nicolaisen et al., 2009) and of the set of genes necessary for nitrogen fixation (three nitrogenases and the universally contiguous gene cluster *nifB-fdxN-nifSUHDKENXW-hesAB-fdxH-feoA*). Remarkable differences were, however, observed in the excision elements that interrupt genes and are removed to facilitate gene expression in heterocysts (Kumar et al., 2010). While the *fdxN*, *hupL* and *nifD* genes are interrupted in *Nostoc* PCC 7120, the *hupL*, *nifH* and *nifD* genes are interrupted in most of the ADA genomes, but in a variety of combinations (Fig. 5). Depending on the genome, *nifH* is either intact or split near nucleotide 150, 430 or both, and *nifD* is either intact or split near nucleotide 1355 and in one case also near nucleotide 895 (Ana CRKS33). Candidate *xis* genes for recombinases catalyzing each rearrangement are present near the target genes and are absent or inactivated in cases when no recombination is necessary. The distribution of split gene design is not congruent with phylogenomic relationships (Fig. 5).

Genes for sulfate uptake (*cysPTW*) and assimilation via adenylylphosphosulfate (APS, *sat*) to phosphoadenylyl-phosphosulfate (PAPS, *cysC*), followed by reduction to sulfite (*cysH*) and hydrogen sulfide (*sir*) are present in all ADA genomes and in *Nostoc* PCC 7120 (Fig. 4, Suppl. Table 6D), but for one exception: the AFA KM1D3 genome lacks the *sat* gene at a site that is rearranged relative to sister genomes. Perhaps this critical gene has been translocated to another site and is missing from the genome assembly. Some ADA genomes possess a set of genes for S-assimilation from organic forms of sulfur: sulfonate uptake (*ssuABC*), sulfonate reduction to sulfite (*ssuD* and FMN reductase), and taurine dioxygenase (*tauD*; Fig. 4; Suppl. Table 6D). The *ssuABCD/tauD* genes appear to be part of a larger genetic unit or genomic island (a 45 kbp fragment in Ana WA102) containing a number of other genes related to S metabolism (4Fe-4S ferredoxin, *metXY*, glutathione S-transferase, SAM methylase, cysteine synthase; Suppl. Fig. 2; Suppl. Table 6D). The capability to utilize organic S is sporadic in the ADA-1, -2 and -3 groups, and not present in ADA-4 genomes. Single genes are disrupted or missing in three cases. It

is not known whether the missing genes are dispensable or whether gene erosion has occurred. Sulfonate detergent pollution in wastewater may in some cases serve as an alternative S source, although sulfate levels have also risen through anthropogenic activities (Thompson and Hutton, 1985).

3.6 *Gene differences affecting general metabolism and physiology*

All ADA genomes contain the complete gene sets in support of photosynthesis (not shown) and for synthesis of phycocyanin (*cpcABCDEFG*), the light-harvesting pigment that is ubiquitous in cyanobacteria and absorbs primarily orange/red light at 620 nm (Suppl. Table 7B; see note on *cpc* gene absences in Ana MDT14b). The additional pigments phycoerythrin (λ_{max} ~560 nm) and phycoerythrocyanin (λ_{max} ~570 nm) allow cyanobacteria to adjust the wavelengths of absorbed incident light (Bryant, 1982). Genes for phycoerythrin synthesis (*cpeABCRSTUYZ*) were not identified in any of the ADA genomes (Fig. 4) and only occur in the symbiotic Nostocales (*Richelia* and *Nostoc punctiforme* PCC 73102) (Meeks et al., 2001). Genes encoding the green-light harvesting pigment phycoerythrocyanin (*pecABCEF*) are present in several Nostocales but in only two genomes from the ADA clade (Ana LE011 and Ana AL93)(Brown et al., 2016)(Fig. 4). Close relatives (both genetically and geographically) of these ADA isolates—Ana AL09 and Ana WA102—lack the *pec* genes. These genes are induced in *Nostoc* PCC 7120 at low light levels (Swanson et al., 1992) and thus would allow photosynthesis to continue in deeper water, when shaded by chlorophyll-containing cells or scums, or earlier in the season when light levels are lower. The genomes containing *pec* genes were derived from deeper waters (Suppl. Table 1B).

Uptake systems predicted to be specific for cobalamin (vitamin B12) are differentially represented across the ADA genomes (Fig. 4; Suppl. Table 7A). Cobalamin, together with sugars, ferric-siderophore complexes and some other substrates, are imported via TonB-dependent transporters (Noinaj et al., 2010). *Nostoc* PCC 7120 has genes for TonB-dependent cobalamin uptake across the outer membrane (alr4028/4029) and genes encoding an associated ABC transporter for inner membrane transport are also present (Mirus et al., 2009). Homologs of these genes are present in 9 of the ADA genomes (Fig. 4; Suppl. Table 7A), with the pathway missing in some members of each ADA group; however, all ADA

genomes have homologs of a second *Nostoc* TonB-dependent cobalamin transporter (all3310), allowing at least outer membrane passage. The all3310 gene is constitutively expressed in *Nostoc* PCC 7120, whereas alr4028/4029 is induced by iron limitation (Mirus et al., 2009). It thus appears that some of the ADA members may have limited ability to scavenge extracellular cobalamin, such as in cases of iron-deficiency that might limit growth rates in dense blooms when resource competition is high. It is interesting to note that the ADA genomes appear to have a low reliance on TonB-dependent importers, particularly in comparison to *Nostoc* PCC 7120, which has four *tonB* genes (Stevanovic et al., 2012) (only one in the ADA genomes, two of which have C-terminally divergent variants that may not be active) and 22 TonB-dependent transporter genes, most of which are probably devoted to iron complex (incl. siderophore) uptake (Dong and Xu, 2009; Mirus et al., 2009; Stevanovic et al., 2012). TonB-dependent transport appears only to be used for cobalamin uptake in the ADA isolates, suggesting that they do not acquire iron via siderophores or citrate complexes. Iron may be acquired via iron-specific ABC transporters (Ana WA102 gene AA650_RS01060 and homologs). In addition to the cobalamin importer genes, all ADA genomes do possess cobalamin biosynthetic genes, but the likely product is pseudocobalamin, as may be general for cyanobacteria (Helliwell et al., 2016).

A number of metabolic genes are differentially represented in the ADA genomes (Fig. 4; Suppl. Table 7A). These are: *metYX*, which incorporate methanethiol (a product of anoxic freshwater sediments; Lomans et al., 1997) into methionine (Kiene et al., 1999); carboxymethylbutenolidase, which has been detected in the extracellular proteome of cyanobacteria (Stuart et al., 2016); *ggt*, gamma-glutamyltranspeptidase, involved in glutathione turnover, which can be triggered by N and S starvation (Cameron and Pakrasi, 2010), or amino acid glutamylation to possibly reduce amino acid loss by leakage from cells (Baran et al., 2013); genes for molybdopterin-containing xanthine dehydrogenase (involved in purine recycling) or *yagTSR* (which oxidizes aromatic aldehydes; Neumann et al., 2009), which are present only in the four ADA-3 genomes and no other Nostocales; tyramine oxidase (present only in the four ADA-2 genomes); prolycopene cis-trans isomerase *crtH*, which in *Synechocystis* allows beta-

carotene synthesis in darkness while non-enzymatic photoisomerization acts in the light (Masamoto et al., 2001); the alternative terminal respiratory oxidase *cydAB* (Jones and Haselkorn, 2002).

There are also differences in the representation of sensory genes. Homologs of *pixJ*, a red/green photosensory cyanobacteriochrome (Fukushima et al., 2011), and adjacent chemotaxis-like *cheYYW* genes in *Nostoc* PCC 7120 are present only in ADA-3 genomes (Fig. 4, Suppl. Table 7A). These genes are related to phototaxis genes sll0038-sll0041 in *Synechocystis* sp. PCC 6803 (Schuergers et al., 2016; Yoshihara and Ikeuchi, 2004) and NpF2161-2164 in *Nostoc punctiforme* (Campbell et al., 2015). The ADA-3 isolates may be the only ADA clade members capable of phototaxis, although some type of motility seems to be a general property, as all the ADA genomes have annotated motility genes (not shown). Another photoprotein, phytochrome A *aphA*, exists in the ADA-2 and a few other genomes (though not ADA-4) together with a two-component regulator, although three of the genomes have an incomplete set of genes (Fig. 4, Suppl. Table 7A).

Buoyancy control afforded by gas vesicles provides cyanobacteria an important competitive advantage in still water over other phytoplankton (Walsby, 1994). All ADA genomes (and *Nostoc* PCC 7120) contain single copies of the gas vesicle genes *gvpCNJKF*/*LGVW* (Suppl. Table 7C)(Mlouka et al., 2004; Pfeifer, 2012), although single genes are disrupted and potentially inactivated in two cultured genomes. A partial *gvpG* deletion that arose during culturing and inactivated buoyancy was described in Ana 90 previously (Wang et al., 2012); Ana AL93 has a partial deletion in the *gvpF/L* gene that likely also abrogates buoyancy (this culture has unfortunately been lost). The number of copies of the *gvpA* gene varies widely between genomes: one in Ana CRKS33 and Dol 310F (ADA-1), 3, 4 and 7 in the ADA-3 genomes AFA KM1D3, AFA NIES-81 and Ana WA102, and 7 in Ana 90 (ADA-2); the number of copies in the other genomes is uncertain because of contig fragmentation at these repeated sequences; there were no *gvpA* genes in two of the draft genome assemblies (Suppl. Table 7C). While GvpA subunits construct the basic gas vesicle, GvpC attaches to the outer surface to provide stabilization (Pfeifer, 2012). Smaller GvpC proteins (16-20 kDa, c.f. 28 kDa) are thought to provide increased stabilization, allowing buoyancy control over a greater depth range (Beard et al., 2000). The ADA *gvpC* genes encode 22-26 kDa proteins,

except for AFA KM1/D3, where a deletion of 66 nucleotides between two internal 44-nt repeats results in a 15 kDa protein. This should provide highly stable gas vesicles, although their utility at the isolation site in the shallow margins of the Baltic Sea is uncertain.

3.7 *Cyanotoxin and secondary metabolite synthesis genes*

Secondary metabolites are important in diverse roles as toxins, allelopathic molecules and taste-and-odor compounds (Leão et al., 2009; Pearson et al., 2016; Watson et al., 2016). The following sections report an in-depth survey of nonribosomal peptide synthetase (NRPS), polyketide synthase (PKS) and other genes producing secondary metabolites (Suppl. Table 8).

3.7.1 Cyanotoxin and hassalidin NRPS or NRPS/PKS products

As among all cyanobacteria, toxin production is sporadically represented among the ADA genomes (Figs. 1, 4). Dol 131C is a saxitoxin producer, Ana 90 is a microcystin producer, and Ana WA102 and Ana AL93 are anatoxin-a producers; these biosynthetic gene clusters have been described previously (Brown et al., 2016; Mihali et al., 2009; Wang et al., 2012); see also Suppl. Table 8A, B). None of the ADA genomes had incomplete or partial toxin gene clusters. Another NRPS-synthesized compound with sporadic presence in the ADA clade is the anti-fungal hassalidin, produced by Ana 90 (Wang et al., 2012); although produced by a variety of Nostocales (Vestola et al., 2014), hassalidin biosynthetic genes were not detected in ADA genomes other than Ana 90.

3.7.2 Other NRPS products

Three other classes of bioactive compounds that are produced by NRPS gene cassettes have been described for the Nostocales: aeruginosin, anabaenopeptin and anabaenopeptilide, together with closely allied products. Aeruginosins are linear tetrapeptides with characteristic 2-carboxy-6-hydroxyoctahydroindole (Choi) moieties that have protease inhibitor activities (Ersmark et al., 2008); they may serve as zooplankton anti-feeding deterrents. These compounds are also known to be produced by *Microcystis* and *Planktothrix* (Ishida et al., 2009). Homologous gene clusters were identified in all ADA-1 and ADA-4 genomes, but in none from ADA-2 or ADA-3 (Fig. 4; Suppl. Table 8C), and these clusters are related to one in *Nodularia spumigena* (Voß et al., 2013). Aeruginosins can be glycosylated

by the action of the *aerI* gene to form aeruginosides, but no such gene was found associated with the *aer* clusters in the ADA genomes. Gene absences—*aerD* or *aerDEF* involved in Choi synthesis (Ishida et al., 2009)—in three of the genomes suggest that the synthesized products could be distinct aeruginosin-like compounds. In three of the ADA-4 genomes, gene disruptions further suggest that the gene cluster is inactive. The complete and intact AFA LD13 cluster exists on a single contig, whereas the *aer* genes are distributed over 2 or 3 contigs in the other ADA-4 genomes (Suppl. Table 8C), perhaps as a consequence of inserted repetitive sequence elements responsible for gene degradation.

Anabaenopeptins are a diverse group of cyclic hexapeptides that also have protease inhibitor activity and are common products of Nostocales and other cyanobacteria (Rouhiainen et al., 2010). The *apt* gene cluster described for Ana 90 has an unusual design featuring two NRPS starter module genes, allowing the synthesis of peptides differing in one position (Rouhiainen et al., 2010). Two other ADA-2 genomes (Ana AL09 and Ana LE011) and one ADA-3 genome (AFA NIES-81) have anabaenopeptin gene clusters (Fig. 4) that appear to be fully functional and that have only a single starter module (Suppl. Table 8D, as is true of *apt* clusters in the *Nostoc punctiforme* and *Nodularia spumigena* genomes (Rouhiainen et al., 2010).

Anabaenopeptilides are yet another class of protease-inhibiting cyclic peptides, containing the distinct amino acid 3-amino-6-hydroxy-2-piperidone (Ahp) and a cyclizing ester bond involving the hydroxy group of a terminal threonine (Rouhiainen et al., 2000; Tooming-Klunderud et al., 2007). Ana 90 has been shown to produce anabaenopeptolide from the *apd* gene cluster (Rouhiainen et al., 2000)(Suppl. Table 8D). Homologous clusters are present in two other ADA-2 genomes, Ana AL09 and Ana LE011, but there are gene-inactivating (frame-shifting) internal deletions or insertions in at least one gene in each of these clusters (Fig. 4, Suppl. Table 8E). For each of the aeruginosin, anabaenopeptin and anabaenopeptilide gene clusters, there is a general conservation of adjacent flanking genes that do not include transposon genes (Suppl. Tables 8C, D, E).

Most of the ADA genomes contain additional NRPS genes that could represent the capacity to synthesize products that are yet to be identified or perhaps are isolated remnant genes of degraded NRPS

gene clusters (Suppl. Table 8F). There can be substantial sequence similarity between the reiterated domains of NRPS genes (e.g., adenylation, condensation, peptide carrier domains), facilitating recombinogenic rearrangements (Tooming-Klunderud et al., 2007), which can either create novel active NRPSs or lead to gene fragmentation or inactivation. The 3.4 kb adenylation-condensation domain insertion that inactivated the Ana LE011 *apdA* NRPS gene (Suppl. Table 8E) appears to be such an example.

3.7.3 PKS products

Another important class of biosynthetic genes in cyanobacteria are the polyketide synthases (PKS). The *hglEFDCAB* PKS gene cluster supporting the synthesis of heterocyst cell wall glycolipids is conserved across all of the ADA genomes (Suppl. Table 8G). Another cluster containing both PKS and NRPS genes, predicted by the antiSMASH program (Weber et al., 2015) to produce glycolipid-like compounds, is also conserved across all ADA genomes (Suppl. Table 8H).

3.7.4 Ribosomal peptides

Bacteriocins are ribosomally produced peptides that are released from a precursor by the action of C39 peptidases. They often have anti-microbial activities and their genes are usually associated with Hly secretion protein genes (Wang et al., 2011). Searches primarily for C39 and Hly genes (Wang et al., 2011) identified five candidate bacteriocin gene clusters that are widely conserved across the ADA genomes, an additional one found in most of the ADA-2 and ADA-3 genomes, and some further genes that might be involved in bacteriocin synthesis (Fig. 4; Suppl. Table S8I). The five conserved clusters are among the seven described for Ana 90 (Wang et al., 2012). Transposase genes are commonly associated with these clusters, suggesting the ability to move within or between genomes.

The cyanobactins also constitute a group of ribosomally produced peptides; they are typically cyclized after their release from precursor peptides (Sivonen et al., 2010). Cyanobactin gene clusters include genes for the two proteases that produce the N and C termini, and a precursor peptide gene. These cyclic peptides are present in a diversity of cyanobacteria (Leikoski et al., 2013), with varied and uncertain biological activities (Sivonen et al., 2010). In multiple *Anabaena* isolates, anacyclamide

cyanobactins produced from *acyCBAEFG* gene clusters were found to be common and diverse in length and sequence (Leikoski et al., 2010). BLAST searches for similarity to the *acyA* and *acyG* protease genes of the Ana 90 anacyclamide cluster (Wang et al., 2012) identified cyanobactin gene clusters in most of the ADA-1, ADA-2 and ADA-3 genomes, but clusters were absent in Ana CRKS33 (ADA-1), AFA NIES-81 (ADA-3) and all ADA-4 genomes (Fig. 4; Suppl. Table 8J). A variety of mature peptide sequences is predicted (Suppl. Table 8J). The gene context surrounding all but two of the gene clusters is fully or partially conserved and flanking transposon genes are not evident. Among the other Nostocales in this study, only the genome of *Nodularia spumigena* has been reported to contain cyanobactin genes (Leikoski et al., 2013; Voß et al., 2013), although these appear to be non-functional. This is likely also the case for Dol 131C and Ana LE011, due to *acyA* gene disruptions (Suppl. Table 8J).

3.7.5 Taste-and-odor compounds

Genomes were screened for the presence of genes related to the production of geosmin (Giglio et al., 2008) and 2-methylisoborneol (2-MIB)(Giglio et al., 2010). Geosmin synthase genes were identified in all three ADA-1 genomes and in AFA NIES-81 (ADA-3) (Fig. 4; Suppl. Table 8J). No genes for 2-MIB synthesis were found, consistent with the fact that this compound has not been reported from Nostocales (Watson et al., 2016).

3.8 Protection against invading genetic elements

3.8.1 Restriction-modification systems

The distribution of identified restriction-modification (R-M) systems and their predicted DNA targets are summarized in Suppl. Table 9. The analysis revealed generally higher numbers of predicted R-M systems (Types I – III) and of DNA sequence specificities in the ADA genomes than in the other Nostocales (Suppl. Table 9A). R-M systems are particularly abundant in ADA-1 genomes. High numbers of R-M systems have been reported previously in filamentous cyanobacteria and in *Microcystis*, another bloom-forming cyanobacterium (Meyer et al., 2017; Wang et al., 2012; Zhao et al., 2006). They seem to be more abundant in bacteria with more mobile genetic elements and higher rates of genetic exchange, and more abundant in larger genomes, which are assumed to be large because of net DNA gain by

horizontal gene transfer (Oliveira et al., 2016). The frequent association of R-M systems with mobile genetic elements drives acquisition by horizontal gene transfer events (Kobayashi, 2001; Kobayashi et al., 1999) as well as losses of systems that no longer confer advantageous protection (Matveyev et al., 2001). The abundance of R-M systems in the ADA genomes is consistent with the density of mobile genetic elements (discussed above), and suggests that these genomes are especially active in DNA exchange, perhaps even more so than the other Nostocales with larger genomes (7-9 Mbp; Suppl. Table 1A) that would be expected to harbor more R-M systems.

3.8.2 CRISPR-Cas systems

Widely varying numbers of CRISPR arrays and spacers exist in Nostocales genomes (Suppl. Table 10A). Among ADA genomes, two to four CRISPR arrays were found per genome, except in ADA-3 genomes, where numbers range from 6 to 13 (Suppl. Table 10). AFA KM1D3 and ADA NIES-81 harbored more than 150 spacers each, while spacer numbers in most of the other ADA genomes were between about 30 and 90. ADA-2 genomes had fewer arrays and spacers (8-33 spacers across 2-3 arrays; the absence of CRISPRs from the Ana AL09 genome is assumed to be an anomaly related to draft genome assembly and clustering). Arrays and spacer numbers are generally higher among the non-ADA Nostocales (excepting the obligate symbionts *Nostoc azollae* and *Richelia*) (Suppl. Table 10A).

Among the ADA genomes, there is considerable heterogeneity in identified CRISPR-Cas arrays with regard to direct repeat length, spacers, *cas* gene sequences, and organization. The CRISPR arrays were classified into 21 groups according to direct repeat similarities (CRISPR1 to CRISPR21, Suppl. Table 10A); not all arrays are associated with *cas* genes. All genomes appear to have fully functional *cas* gene clusters with modules for spacer insertion (*cas1/cas2*) and target interference (various *cas, csc, cmr* genes): Type I-D and/or Type III-B (Makarova et al., 2015) (Fig. 4; see Suppl. Table 10B for full details). Each CRISPR cluster type is associated with specific consensus repeat sequences regardless of ADA group membership, while the genomic context (identity of flanking genes) is mostly specific to each ADA group (and differing between clusters). The Type III-B and one of the Type I-D clusters have chimeric designs, with both runs of homologous genes and subsets of genes shared by only some clusters (Suppl.

Table 10B); the arrangement of clusters suggests that considerable genetic cross-talk could exist between the four ADA groups.

Most of the CRISPR spacers are unique, suggesting that each strain has been exposed to diverse types of invading DNA. On the other hand, all ADA-4 genomes share the same terminal spacer in one of their CRISPR4 arrays (CRISPR4-b2, b3, b5, b7) and share a different terminal spacer in their CRISPR6 arrays (Suppl. Table 10C). Similarly, three of the ADA-4 genomes (AFA MDT14a, AFA LD13 and Ana WA113) have identical terminal spacers in their CRISPR6 arrays (AFA MDT14a and AFA LD13 CRISPR6 additionally share the same sub-terminal spacer). The ADA-4 genomes are derived from the adjacent states of Oregon and Washington (USA) and seem to have been challenged by the same phage or plasmids. The shared spacers are probably at the distal, older ends of the arrays and are assumed to be shared as a result of inheritance from common ancestors or acquisition by lateral exchange.

A BLAST search using all observed CRISPR spacers as a query against more than 200 publicly available cyanophage genome sequences (mostly from www.ebi.ac.uk, 2016-05-14) found hits only in AFA KM1D3 and AFA NIES-81. Both genomes contain spacers that match sequences from the recently isolated cyanophage vB-AphaS-CL131, which has been shown to infect AFA KM1D3 (Šulčius et al. 2015).

4 Conclusions

The availability of multiple new Nostocales genome sequences derived from recent CyanoHAB events has allowed new understanding of the phylogenetic relationships among these increasingly troublesome cyanobacteria. Fifteen genomes form a well-separated clade that we have designated the ADA Clade and which we view as representing a genus (Fig. 1). The fifteen genomes cluster into four groups whose genomes are related closely enough (ANI >96%) to propose the existence of four species (Varghese et al., 2015). Following current nomenclature and taxonomic guidance (Wacklin et al., 2009), the ADA clade embraces three genus names—*Anabaena, Dolichospermum* and *Aphanizomenon*—and those designations are intermixed in the phylogenomic tree (Fig. 1).

Our studies firmly support earlier conclusions (Gugger et al., 2002; Rajaniemi et al., 2005) that *Anabaena* and *Aphanizomenon flos-aquae* are tightly related, despite their distinct morphology (Suppl. Fig. 1). We view the introduction of a new genus name—*Dolichospermum* (Wacklin et al., 2009)—as having been premature in the absence of the extensive genome sequencing that should be used to guide a definitive taxonomy based predominantly on phylogenomic relationships based on the relatedness of multiple core genes. Our study is a step towards mapping the relationships among members of this branch of the Nostocales, and ultimately it might be appropriate for the ADA clade to adopt the *Dolichospermum* genus nomenclature. Issues that should be resolved with further genome sequences include (a) clarification of relationships between *Aphanizomenon flos-aquae* and other *Aphanizomenon* isolates and use of the *Aphanizomenon* name, (b) clarification of relationships between benthic and planktonic "*Anabaena*" isolates, and (c) clarification of relationships between the ADA clade and the *Chrysosporum* and *Sphaerospermopsis* genera, to which transferal of some *Anabaenas* has been advocated (Li et al., 2016; Zapomělová et al., 2009; Zapomělová et al., 2012).

The genomes of ADA members share the core Nostocales characteristics, best studied in *Nostoc* PCC 7120, of possessing genes that support photosynthesis and the ability to fix nitrogen in differentiated heterocysts (Suppl. Tables 6, 7). All possess uptake systems for the P and N nutrients that drive bloom growth—phosphate, phosphonate, nitrite/nitrate, ammonium and amino acids (Fig. 4; Suppl. Table 6)—and that are among a multitude of transporter genes, particularly of the ABC type, in these genomes (Fig. 3C). Differences exist, however, in the number of ammonium and amino acid uptake systems and in uptake and/or utilization genes for organic forms of P and S: phosphonates and sulfonates (Fig. 4). Like other cyanobacteria, these genomes are also rich in genes for the production of varied secondary metabolites, some of which are found in all species, such as genes for glycolipids needed for heterocyst cell wall maturation or for bacteriocins that may regulate interactions with other microbes (Aharonovich and Sher, 2016). Others are found in only some genomes, often with ADA species-specific representation, while still others are only sporadically present across the four ADA species (Fig. 4; Suppl. Table 8). Understanding trait distinctions among these cyanobacteria that are major contributors to extant

CyanoHABs will be important in determining each organism's preferred niche and in unravelling the influences that lead to blooms and successional changes across a season.

Acknowledgments

The research at Oregon State University was supported by U.S. Geological Survey grant 2012OR127G, the Oregon State University Agricultural Experiment Station, the Mabel E. Pernot Trust and the NL Tartar Research Fellowship. We thank Alexandra Weisberg for assistance with Mauve alignments. Research at the University of Michigan was supported by a grant from the Erb Family Foundation made through the University of Michigan Water Center (Grant N017871) and from the National Science Foundation (NSF OCE 1736629). The research at Nature Research Centre and Vilnius University was funded by a grant (No. S-LJB-17-1) from the Research Council of Lithuania to SŠ and GG.

Contributions of authors

CBD, TGO, SS, TWDreher devised the study concept. CBD, NMB, TGO produced draft genomes from environmental metagenomes. KAM, GJD, TWDavis, SBW produced draft genomes from Great Lakes cultures. All authors contributed to data analyses and manuscript revision. CBD and TWDreher wrote the initial manuscript drafts.

Conflict of interest

GG is inventor on patent applications related to CRISPR, co-founder and employee of CasZyme. No other potential conflicts of interest exist.

References

Abreu, V.A.C., Popin, R.V., Alvarenga, D.O., Schaker, P.D.C., Hoff-Risseti, C., Varani, A.M., and Flore, M.F. (2018) Genomic and genotypic characterization of *Cylindrospermopsis raciborskii*: Toward an intraspecific phylogenetic evaluation by comparative genomics. Frontiers Microbiol. 9: 306.

Aharonovich, D., and Sher, D. (2016) Transcriptional response of Prochlorococcus to co-culture with a marine Alteromonas: differences between strains and the involvement of putative infochemicals. The ISME J 10: 2892-2906.

Aichi, M., Yoshihara, S., Yamashita, M., Maeda, S.-I., Nagai, K., and Omata, T. (2006) Characterization of the nitrate-nitrite transporter of the major facilitator superfamily (the nrtP gene product) from the cyanobacterium Nostoc punctiforme strain ATCC 29133. Bioscience, Biotechnology, and Biochemistry 70: 2682-2689.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology 31: 533-538.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. J Molec Biol 215: 403-410.

Baran, R., Ivanova, N.N., Jose, N., Garcia-Pichel, F., Kyrpides, N.C., Gugger, M., and Northen, T.R. (2013) Functional genomics of novel secondary metabolites from diverse cyanobacteria using untargeted metabolomics. Marine Drugs 11: 3617-3631.

Beard, S., Davis, P., Iglesias-Rodrıguez, D., Skulberg, O., and Walsby, A. (2000) Gas vesicle genes in Planktothrix spp. from Nordic lakes: strains with weak gas vesicles possess a longer variant of gvpC. Microbiology 146: 2009-2018.

Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015) Prochlorococcus: the structure and function of collective diversity. Nature Reviews Microbiology 13: 13-27.

Bird, C., and Wyman, M. (2003) Nitrate/nitrite assimilation system of the marine picoplanktonic cyanobacterium Synechococcus sp. strain WH 8103: effect of nitrogen source and availability on gene expression. Appl Environ Microb 69: 7009-7018.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics: btu170.

Brown, N.M., Mueller, R.S., Shepardson, J.W., Landry, Z.C., Morre, J.T., Maier, C.S., Hardy, F.J., and Dreher, T.W. (2016) Structural and functional analysis of the finished genome of the recently isolated toxic *Anabaena* sp. WA102. BMC Genomics 17: 457.

Bryant, D.A. (1982) Phycoerythrocyanin and phycoerythrin: properties and occurrence in cyanobacteria. Microbiology 128: 835-844.

Burford, M.A., Beardall, J., Willis, A., Orr, P.T., Magalhaes, V.F., Rangel, L.M., Azevedo, S.M., and Neilan, B.A. (2016) Understanding the winning strategies used by the bloom-forming cyanobacterium Cylindrospermopsis raciborskii. Harmful Algae 54: 44-53.

Calteau, A., Fewer, D.P., Latifi, A., Coursin, T., Laurent, T., Jokela, J., Kerfeld, C.A., Sivonen, K., Piel, J., and Gugger, M. (2014) Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. BMC Genomics 15: 977.

Cameron, J.C., and Pakrasi, H.B. (2010) Essential role of glutathione in acclimation to environmental and redox perturbations in the cyanobacterium Synechocystis sp. PCC 6803. Plant Physiology 154: 1672-1685.

Campbell, E.L., Hagen, K.D., Chen, R., Risser, D.D., Ferreira, D.P., and Meeks, J.C. (2015) Genetic analysis reveals the identity of the photoreceptor for phototaxis in hormogonium filaments of Nostoc punctiforme. J Bacteriol 197: 782-791.

Canfield, D.E. (2005) The early history of atmospheric oxygen: homage to Robert M. Garrels. Annu. Rev. Earth Planet. Sci. 33: 1-36.

Cao, H., Shimura, Y., Masanobu, K., and Yin, Y. (2014) Draft genome sequence of the toxic bloom-forming cyanobacterium Aphanizomenon flos-aquae NIES-81. Genome Announcements 2: e00044-00014.

Cirés, S., and Ballot, A. (2016) A review of the phylogeny, ecology and toxin production of bloom-forming Aphanizomenon spp. and related species within the Nostocales (cyanobacteria). Harmful Algae 54: 21-43.

Conley, D.J., Paerl, H.W., Howarth, R.W., Boesch, D.F., Seitzinger, S.P., Havens, K.E., Lancelot, C., and Likens, G.E. (2009) Controlling eutrophication: nitrogen and phosphorus. Science 323: 1014-1015.

Contreras-Moreira, B., and Vinuesa, P. (2013) GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microb 79: 7696-7701.

D'Agostino, P.M., Song, X., Neilan, B.A., and Moffitt, M.C. (2014) Comparative proteomics reveals that a saxitoxin-producing and a nontoxic strain of Anabaena circinalis are two different ecotypes. J Proteome Res 13: 1474-1484.

Darling, A.E., Jospin, G., Lowe, E., Matsen IV, F.A., Bik, H.M., and Eisen, J.A. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2: e243.

Darling, A.E., Mau, B., and Perna, N.T. (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5: e11147.

Davis, T.W., and Gobler, C.J. (2016) Preface for Special Issue on “Global expansion of harmful cyanobacterial blooms: Diversity, ecology, causes, and controls”. Harmful Algae 54: 1-3.

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., and Banfield, J.F. (2009) Community-wide analysis of microbial genome sequence signatures. Genome Biol 10: 1.

Dittmann, E., Gugger, M., Sivonen, K., and Fewer, D.P. (2015) Natural product biosynthetic diversity and comparative genomics of the cyanobacteria. Trends Microbiol 23: 642-652.

Dong, Y., and Xu, X. (2009) Outer membrane proteins induced by iron deficiency in Anabaena sp. PCC 7120. Progress in Natural Science 19: 1477-1483.

Ehira, S., and Ohmori, M. (2006) NrrA, a nitrogen-responsive response regulator facilitates heterocyst development in the cyanobacterium Anabaena sp. strain PCC 7120. Mol Microbiol 59: 1692-1703.

Ersmark, K., Del Valle, J.R., and Hanessian, S. (2008) Chemistry and biology of the aeruginosin family of serine protease inhibitors. Angewandte Chemie International Edition 47: 1202-1223.

Fan, Q., Huang, G., Lechno-Yossef, S., Wolk, C.P., Kaneko, T., and Tabata, S. (2005) Clustered genes required for synthesis and deposition of envelope glycolipids in Anabaena sp. strain PCC 7120. Molecular Microbiol 58: 227-243.

Fischer, S., Brunk, B.P., Chen, F., Gao, X., Harb, O.S., Iodice, J.B., Shanmugam, D., Roos, D.S., and Stoeckert, C.J. (2011) Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new Ortholog groups. Current Protocols in Bioinformatics: 6.12. 11-16.12. 19.

Flores, E., and Herrero, A. (2010) Compartmentalized function through cell differentiation in filamentous cyanobacteria. Nature Reviews Microbiol 8: 39-50.

Fukushima, Y., Iwaki, M., Narikawa, R., Ikeuchi, M., Tomita, Y., and Itoh, S. (2011) Photoconversion mechanism of a green/red photosensory cyanobacteriochrome AnPixJ: time-resolved optical spectroscopy and FTIR analysis of the AnPixJ-GAF2 domain. Biochemistry 50: 6328-6339.

Garcia-Pichel, F., López-Cortés, A., and Nübel, U. (2001) Phylogenetic and Morphological Diversity of Cyanobacteria in Soil Desert Crusts from the Colorado Plateau. Appl Environ Microb 67: 1902-1910.

Giglio, S., Chou, W., Ikeda, H., Cane, D., and Monis, P. (2010) Biosynthesis of 2-methylisoborneol in cyanobacteria. Environ Sci Technol 45: 992-998.

Giglio, S., Jiang, J., Saint, C.P., Cane, D.E., and Monis, P.T. (2008) Isolation and characterization of the gene associated with geosmin production in cyanobacteria. Environ Sci Technol 42: 8027-8032.

Gorski, L. (2012) Selective enrichment media bias the types of Salmonella enterica strains isolated from mixed strain cultures and complex enrichment broths. PLoS One 7: e34722.

Gregor, I.D., J.; Schirmer, M.; Quince, C.; McHardy, A. C. (2014) PhyloPythiaS+: A self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. arXiv: 1406.7123.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35: W52-W57.

Gugger, M., Lyra, C., Henriksen, P., Coute, A., Humbert, J.F., and Sivonen, K. (2002) Phylogenetic comparison of the cyanobacterial genera Anabaena and Aphanizomenon. Int J Syst Evol Microbiol 52: 1867-1880.

Guiry, M.D., and Guiry, G.M. (2016) AlgaeBase, National University of Ireland, Galway.

Halinen, K., Fewer, D.P., Fewer, L.M., Lyra, C., Eronen, E., and Sivonen, K. (2008) Genetic diversity in strains of the genus Anabaena isolated from planktonic and benthic habitats of the Gulf of Finland (Baltic Sea). FEMS Microbiol Ecol 64: 199-208.

Harke, M.J., Davis, T.W., Watson, S.B., and Gobler, C.J. (2016) Nutrient-controlled niche differentiation of western Lake Erie cyanobacterial populations revealed via metatranscriptomic surveys. Environ Sci Technol 50: 604-615.

Helliwell, K.E., Lawrence, A.D., Holzer, A., Kudahl, U.J., Sasso, S., Kräutier, B., Scanlan, D.J., Warren, M.J, and Smith, A.G. (2016) Cyanobacteria and eukaryotic algae use difference chemical variants of vitamin B_{12}. Current Biol 26: 999-1008.

Huang, G., Fan, Q., Lechno-Yossef, S., Wojciuch, E., Wolk, C.P., Kaneko, T., and Tabata, S. (2005) Clustered genes required for the synthesis of heterocyst envelope polysaccharide in Anabaena sp. strain PCC 7120. J Bacteriol 187: 1114-1123.

Humbert, J.F., Barbe, V., Latifi, A., Gugger, M., Calteau, A., Coursin, T., Lajus, A., Castelli, V., Oztas, S., Samson, G., Longin, C., Medigue, C., and de Marsac, N.T. (2013) A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium Microcystis aeruginosa. PloS One 8: e70747.

Ishida, K., Welker, M., Christiansen, G., Cadel-Six, S., Bouchier, C., Dittmann, E., Hertweck, C., and de Marsac, N.T. (2009) Plasticity and evolution of aeruginosin biosynthesis in cyanobacteria. Appl Environ Microb 75: 2017-2026.

Jones, K.M., and Haselkorn, R. (2002) Newly identified cytochrome c oxidase operon in the nitrogen-fixing cyanobacterium Anabaena sp. strain PCC 7120 specifically induced in heterocysts. J Bacteriol 184: 2491-2499.

Kaas, R.S., Friis, C., Ussery, D.W., and Aarestrup, F.M. (2012) Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. BMC Genomics 13: 577.

Karl, D., Letelier, R., Tupas, L., Dore, J., Christian, J., and Hebel, D. (1997) The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. Nature 388: 533-538.

Kiene, R.P., Linn, L.J., González, J., Moran, M.A., and Bruton, J.A. (1999) Dimethylsulfoniopropionate and methanethiol are important precursors of methionine and protein-sulfur in marine bacterioplankton. Appl Environ Microb 65: 4549-4558.

Kim, M., Oh, H.-S., Park, S.-C., and Chun, J. (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International journal of systematic and evolutionary microbiology 64: 346-351.

Kobayashi, I. (2001) Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. Nucleic Acids Res 29: 3742-3756.

Kobayashi, I., Nobusato, A., Kobayashi-Takahashi, N., and Uchiyama, I. (1999) Shaping the genome–restriction–modification systems as mobile genetic elements. Current Opinion in Genetics & Development 9: 649-656.

Komárek, J. (2010) Recent changes (2008) in cyanobacteria taxonomy based on a combination of molecular background with phenotype and ecological consequences (genus and species concept). Hydrobiologia 639: 245-259.

Komarek, J., Kastovsky, J., Mares, J., and Johansen, J.R. (2014) Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. Preslia 86: 295-335.

Koonin, E.V., and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36: 6688-6719.

Koskiniemi, S., Sun, S., Berg, O.G., and Andersson, D.I. (2012) Selection-driven gene loss in bacteria. PLoS Genet 8: e1002787.

Kumar, K., Mella-Herrera, R.A., and Golden, J.W. (2010) Cyanobacterial heterocysts. Cold Spring Harbor Perspectives in Biology 2: a000315.

Landry, Z., Swan, B.K., Herndl, G.J. Stepanauskas, R., and Giovannoni, S.J. (2017) SAR202 Genomes from the dark ocean predict pathways for the oxidation of recalcitrant dissolved organic matter. mBio 8: e00413-17.

Leão, P.N., Vasconcelos, M.T.S., and Vasconcelos, V.M. (2009) Allelopathy in freshwater cyanobacteria. Critical Rev Microbiol 35: 271-282.

Leikoski, N., Fewer, D.P., Jokela, J., Wahlsten, M., Rouhiainen, L., and Sivonen, K. (2010) Highly diverse cyanobactins in strains of the genus Anabaena. Appl Environ Microb 76: 701-709.

Leikoski, N., Liu, L., Jokela, J., Wahlsten, M., Gugger, M., Calteau, A., Permi, P., Kerfeld, C.A., Sivonen, K., and Fewer, D.P. (2013) Genome mining expands the chemical diversity of the cyanobactin family to include highly modified linear peptides. Chemistry & Biology 20: 1033-1043.

Li, X., Dreher, T.W., and Li, R. (2016) An overview of diversity, occurrence, genetics and toxin production of bloom-forming *Dolichospermum* (*Anabaena*) species. Harmful Algae 54: 54-68.

Lomans, B.P., Smolders, A., Intven, L.M., Pol, A., Op, D., and Van Der Drift, C. (1997) Formation of dimethyl sulfide and methanethiol in anoxic freshwater sediments. Appl Environ Microb 63: 4741-4747.

Lukjancenko, O., Wassenaar, T.M., and Ussery, D.W. (2010) Comparison of 61 sequenced Escherichia coli genomes. Microbial Ecology 60: 708-720.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., and Haft, D.H. (2015) An updated evolutionary classification of CRISPR-Cas systems. Nature Reviews Microbiol. 13:722-36

Malatinszky, D., Steuer, R., and Jones, P.R. (2017) A comprehensively curated genome-scale two-cell model for the heterocystous cyanobacterium *Anabaena* sp. PCC 7120. Plant Physiol 173: 509-523.

Masamoto, K., Wada, H., Kaneko, T., and Takaichi, S. (2001) Identification of a gene required for cis-to-trans carotene isomerization in carotenogenesis of the cyanobacterium Synechocystis sp. PCC 6803. Plant Cell Physiol 42: 1398-1402.

Matveyev, A.V., Young, K.T., Meng, A., and Elhai, J. (2001) DNA methyltransferases of the cyanobacterium Anabaena PCC 7120. Nucleic Acids Res 29: 1491-1506.

Meeks, J.C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., Predki, P., and Atlas, R. (2001) An overview of the genome of Nostoc punctiforme, a multicellular, symbiotic cyanobacterium. Photosynthesis Res 70: 85-106.

Meyer, K.A., Davis, T.W., Watson, S.B., Denef, V.J., Berry, M.A., and Dick, G.J. (2017) Genome sequences of lower Great Lakes Microcystis sp. reveal strain-specific genes that are present and expressed in western Lake Erie blooms. PloS One 12: e0183859.

Mihali, T.K., Kellmann, R., and Neilan, B.A. (2009) Characterisation of the paralytic shellfish toxin biosynthesis gene clusters in Anabaena circinalis AWQC131C and Aphanizomenon sp. NH-5. BMC Biochem 10: 8.

Mirus, O., Strauss, S., Nicolaisen, K., von Haeseler, A., and Schleiff, E. (2009) TonB-dependent transporters and their occurrence in cyanobacteria. BMC Biology 7: 68.

Mlouka, A., Comte, K., Castets, A.-M., Bouchier, C., and de Marsac, N.T. (2004) The gas vesicle gene cluster from Microcystis aeruginosa and DNA rearrangements that lead to loss of cell buoyancy. J Bacteriol 186: 2355-2365.

Muro-Pastor, A.M., and Hess, W.R. (2012) Heterocyst differentiation: from single mutants to global approaches. Trends Microbiol 20: 548-557.

Neumann, M., Mittelstädt, G., Seduk, F., Iobbi-Nivol, C., and Leimkühler, S. (2009) MocA is a specific cytidylyltransferase involved in molybdopterin cytosine dinucleotide biosynthesis in Escherichia coli. J Biological Chem 284: 21891-21898.

Nicolaisen, K., Hahn, A., and Schleiff, E. (2009) The cell wall in heterocyst formation by Anabaena sp. PCC 7120. J Basic Microbiol 49: 5-24.

Noinaj, N., Guillier, M., Barnard, T.J., and Buchanan, S.K. (2010) TonB-dependent transporters: regulation, structure, and function. Annual Review of Microbiol 64: 43-60.

Ohashi, Y., Shi, W., Takatani, N., Aichi, M., Maeda, S.-i., Watanabe, S., Yoshikawa, H., and Omata, T. (2011) Regulation of nitrate assimilation in cyanobacteria. J Experimental Botany 62: 1411-1424.

Oliveira, P.H., Touchon, M., and Rocha, E.P. (2016) Regulation of genetic flux between bacteria by restriction–modification systems. Proceedings of the National Academy of Sciences (USA) 113: 5658-5663.

Oliver, R.L., and Ganf, G.G. (2000) Freshwater blooms, The ecology of cyanobacteria. Springer, pp. 149-194.

Otten, T.G., Graham, J.L., Harris, T.D., and Dreher, T.W. (2016) Elucidation of taste-and-odor producing bacteria and toxigenic cyanobacteria by shotgun metagenomics in a Midwestern drinking water supply reservoir. Appl Environ Microbiol 82: 5410-5420.

Paerl, H.W., Fulton, R.S., Moisander, P.H., and Dyble, J. (2001) Harmful freshwater algal blooms, with an emphasis on cyanobacteria. The Scientific World Journal 1: 76-113.

Paerl, H.W., and Otten, T.G. (2013) Harmful cyanobacterial blooms: causes, consequences, and controls. Microbial Ecology 65: 995-1010.

Pancrace, C., Barny, M.-A., Ueoka, R., Calteau, A., Scalvenzi, T., Pédron, J., Barbe, V., Piel, J., Humbert, J.-F., and Gugger, M. (2017) Insights into the Planktothrix genus: Genomic and metabolic comparison of benthic and planktic strains. Scientific Reports 7.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25: 1043-1055.

Pearson, L.A., Dittmann, E., Mazmouz, R., Ongley, S.E., D'Agostino, P.M., and Neilan, B.A. (2016) The genetics, biosynthesis and regulation of toxic specialized metabolites of cyanobacteria. Harmful Algae 54: 98-111.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28: 1420-1428.

Pereira, S., Zille, A., Micheletti, E., Moradas-Ferreira, P., De Philippis, R., and Tamagnini, P. (2009) Complexity of cyanobacterial exopolysaccharides: composition, structures, inducing factors and putative genes involved in their biosynthesis and assembly. FEMS Microbiology Rev 33: 917-941.

Pereira, S.B., Mota, R., Vieira, C.P., Vieira, J., and Tamagnini, P. (2015) Phylum-wide analysis of genes/proteins related to the last steps of assembly and export of extracellular polymeric substances (EPS) in cyanobacteria. Scientific reports 5.

Pernil, R., Picossi, S., Herrero, A., Flores, E., and Mariscal, V. (2015) Amino acid transporters and release of hydrophobic amino acids in the heterocyst-forming cyanobacterium Anabaena sp. strain PCC 7120. Life 5: 1282-1300.

Pfeifer, F. (2012) Distribution, formation and regulation of gas vesicles. Nature Rev Microbiol 10: 705.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41: D590-D596.

Rajaniemi, P., Hrouzek, P., Kastovska, K., Willame, R., Rantala, A., Hoffmann, L., Komarek, J., and Sivonen, K. (2005) Phylogenetic and morphological evaluation of the genera Anabaena, Aphanizomenon, Trichormus and Nostoc (Nostocales, Cyanobacteria). Intl J System Evol Microbiol 55: 11-26.

Ramírez, M.E., Hebbar, P.B., Zhou, R., Wolk, C.P., and Curtis, S.E. (2005) Anabaena sp. strain PCC 7120 gene devH is required for synthesis of the heterocyst glycolipid layer. J Bacteriol 187: 2326-2331.

Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J.A., Ininbergs, K., Zheng, W.-W., Lapidus, A., Lowry, S., Haselkorn, R., and Bergman, B. (2010) Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS One 5: e11486.

Rao, N.N., Gómez-García, M.R., and Kornberg, A. (2009) Inorganic polyphosphate: essential for growth and survival. Annual Rev Biochem 78: 605-647.

Raven, J.A. (2002) Evolution of cyanobacterial symbioses, Cyanobacteria in symbiosis. Springer, pp. 329-346.

Riadi, G., Medina-Moenne, C., and Holmes, D.S. (2012) TnpPred: A web service for the robust prediction of prokaryotic transposases. Comparative and Functional Genomics 2012:678761.

Robbertse, B., Reeves, J.B., Schoch, C.L., and Spatafora, J.W. (2006) A phylogenomic analysis of the Ascomycota. Fungal Genet Biol 43: 715-725.

Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2015) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 43: D298-D299.

Rouhiainen, L., Jokela, J., Fewer, D.P., Urmann, M., and Sivonen, K. (2010) Two alternative starter modules for the non-ribosomal biosynthesis of specific anabaenopeptin variants in Anabaena (Cyanobacteria). Chemistry & Biology 17: 265-273.

Rouhiainen, L., Paulin, L., Suomalainen, S., HyytiaÈinen, H., Buikema, W., Haselkorn, R., and Sivonen, K. (2000) Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in Anabaena strain 90. Molecular Microbiol 37: 156-167.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. PeerJ 3: e985.

Schuergers, N., Lenn, T., Kampmann, R., Meissner, M.V., Esteves, T., Temerinac-Ott, M., Korvink, J.G., Lowe, A.R., Mullineaux, C.W., and Wilde, A. (2016) Cyanobacteria use micro-optics to sense light direction. Elife 5: e12620.

Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., Calteau, A., Cai, F., Tandeau de Marsac, N., Rippka, R., Herdman, M., Sivonen, K., Coursin, T., Laurent, T., Goodwin, L., Nolan, M., Davenport, K.W., Han, C.S., Rubin, E.M., Eisen, J.A., Woyke, T., Gugger, M., and Kerfeld, C.A. (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci U S A 110: 1053-1058.

Simm, S., Keller, M., Selymesi, M., and Schleiff, E. (2015) The composition of the global and feature specific cyanobacterial core-genomes. Frontiers in microbiology 6: 219.

Sivonen, K., Leikoski, N., Fewer, D.P., and Jokela, J. (2010) Cyanobactins—ribosomal cyclic peptides produced by cyanobacteria. Appl Microbiol Biot 86: 1213-1225.

Snipen, L., Almøy, T., and Ussery, D.W. (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10: 385.

Stevanovic, M., Hahn, A., Nicolaisen, K., Mirus, O., and Schleiff, E. (2012) The components of the putative iron transport system in the cyanobacterium Anabaena sp. PCC 7120. Environmental Microbiol 14: 1655-1670.

Stuart, R.K., Mayali, X., Lee, J.Z., Everroad, R.C., Hwang, M., Bebout, B.M., Weber, P.K., Pett-Ridge, J., and Thelen, M.P. (2016) Cyanobacterial reuse of extracellular organic carbon in microbial mats. The ISME J 10: 1240.

Šulčius, S., Alzbutas, G., Kvederavičiūtė, K., Koreivienė, J., Zakrys, L., Lubys, A., and Paškauskas, R. (2015) Draft genome sequence of the cyanobacterium Aphanizomenon flos-aquae strain 2012/KM1/D3, isolated from the Curonian Lagoon (Baltic Sea). Genome Announcements 3: e01392-01314.

Swanson, R.V., de Lorimier, R., and Glazer, A.N. (1992) Genes encoding the phycobilisome rod substructure are clustered on the Anabaena chromosome: characterization of the phycoerythrocyanin operon. J Bacteriol 174: 2640-2647.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., and Durkin, A.S. (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102: 13950-13955.

Thompson, M.E., and Hutton, M.B. (1985) Sulfate in lakes of eastern Canada: calculated yields compared with measured wet and dry deposition. Water, Air, & Soil Pollution 24: 77-83.

Tooming-Klunderud, A., Rohrlack, T., Shalchian-Tabrizi, K., Kristensen, T., and Jakobsen, K.S. (2007) Structural analysis of a non-ribosomal halogenated cyclic peptide and its putative operon from Microcystis: implications for evolution of cyanopeptolins. Microbiology 153: 1382-1393.

Valladares, A., Montesinos, M.L., Herrero, A., and Flores, E. (2002) An ABC-type, high-affinity urea permease identified in cyanobacteria. Molecular Microbiol 43: 703-715.

Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C., and Pati, A. (2015) Microbial species delineation using whole genome sequences. Nucleic Acids Res 43: 6761-6771.

Vestola, J., Shishido, T.K., Jokela, J., Fewer, D.P., Aitio, O., Permi, P., Wahlsten, M., Wang, H., Rouhiainen, L., and Sivonen, K. (2014) Hassallidins, antifungal glycolipopeptides, are widespread among cyanobacteria and are the end-product of a nonribosomal pathway. Proc Natl Acad Sci USA 111: E1909-E1917.

Villarreal-Chiu, J.F., Quinn, J.P., and McGrath, J.W. (2012) The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. Frontiers Microbiol 3.

Vinuesa, P., and Contreras-Moreira, B. (2015) Robust Identification of Orthologues and Paralogues for Microbial Pan-Genomics Using GET_HOMOLOGUES: A Case Study of pIncA/C Plasmids. Bacterial Pangenomics: Methods and Protocols: 203-232.

Voß, B., Bolhuis, H., Fewer, D.P., Kopf, M., Möke, F., Haas, F., El-Shehawy, R., Hayes, P., Bergman, B., and Sivonen, K. (2013) Insights into the physiology and ecology of the brackish-water-adapted cyanobacterium Nodularia spumigena CCY9414 based on a genome-transcriptome analysis. PLoS One 8: e60224.

Wacklin, P., Hoffmann, L., and Komárek, J. (2009) Nomenclatural validation of the genetically revised cyanobacterial genus Dolichospermum (Ralfs ex Bornet et Flahault) comb. nova. Fottea 9: 59-64.

Walsby, A. (1994) Gas vesicles. Microbiological reviews 58: 94-144.

Wang, H., Fewer, D.P., and Sivonen, K. (2011) Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. PloS One 6: e22384.

Wang, H., Sivonen, K., and Fewer, D.P. (2015) Genomic insights into the distribution, genetic diversity and evolution of polyketide synthases and nonribosomal peptide synthetases. Current Opinion in Genetics & Development 35: 79-85.

Wang, H., Sivonen, K., Rouhiainen, L., Feweer, D.P., Lyra, C., Rantala-Ylinen, A., Vestola, J., Jokela, J., Rantasarkka, K., Li, Z., and Liu, B. (2012) Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium *Anabaena* sp. strain 90. BMC Genomics 13: 613.

Wang, Q., Li, H., and Post, A.F. (2000) Nitrate assimilation genes of the marine diazotrophic, filamentous cyanobacterium Trichodesmium sp. strain WH9601. J Bacteriol 182: 1764-1767.

Wang, Y., and Xu, X. (2005) Regulation by hetC of genes required for heterocyst differentiation and cell division in Anabaena sp. strain PCC 7120. J Bacteriol 187: 8489-8493.

Watson, S.B., Monis, P., Baker, P., and Giglio, S. (2016) Biochemistry and genetics of taste-and odor-producing cyanobacteria. Harmful Algae 54: 112-127.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., Lee, S.Y., Fischbach, M.A., Muller, R., Wohlleben, W., Breitling, R., Takano, E., and Medema, M.H. (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43: W237-243.

Welker, M., and Von Döhren, H. (2006) Cyanobacterial peptides—nature's own combinatorial biosynthesis. FEMS Microbiol Rev 30: 530-563.

Willenbrock, H., Hallin, P.F., Wassenaar, T.M., and Ussery, D.W. (2007) Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray. Genome Biol 8: 1.

Xu, X., Elhai, J., and Wolk, C.P. (2008) Transcriptional and developmental responses by Anabaena to deprivation of fixed nitrogen, in: Herrero, A., Flores, E. (Eds.), The cyanobacteria: molecular biology, genomics and evolution. Caister Academic Press, Norfolk, UK, pp. 383-422.

Yoshihara, S., and Ikeuchi, M. (2004) Phototactic motility in the unicellular cyanobacterium Synechocystis sp. PCC 6803. Photochemical & Photobiological Sciences 3: 512-518.

Zapomělová, E., Jezberová, J., Hrouzek, P., Hisem, D., Řeháková, K., and Komárková, J. (2009) Polyphasic characterization of three strains of Anabaena reniformis and Aphanizomenon aphanizomenoides (Cyanobacteria) and their reclassification to Sphaerospermum gen. nov.(incl. Anabaena kisseleviana). J Phycology 45: 1363-1373.

Zapomělová, E., Skácelová, O., Pumann, P., Kopp, R., and Janeček, E. (2012) Biogeographically interesting planktonic Nostocales (Cyanobacteria) in the Czech Republic and their polyphasic evaluation resulting in taxonomic revisions of Anabaena bergii Ostenfeld 1908 (Chrysosporum gen. nov.) and A. tenericaulis Nygaard 1949 (Dolichospermum tenericaule comb. nova). Hydrobiologia 698: 353-365.

Zhang, W., Du, Y., Khudyakov, I., Fan, Q., Gao, H., Ning, D., Wolk, C.P., and Xu, X. (2007) A gene cluster that regulates both heterocyst differentiation and pattern formation in Anabaena sp. strain PCC 7120. Molecular Microbiol 66: 1429-1443.

Zhao, F., Zhang, X., Liang, C., Wu, J., Bao, Q., and Qin, S. (2006) Genome-wide analysis of restriction-modification system in unicellular and filamentous cyanobacteria. Physiological Genomics 24: 181-190.

Zhao, Y., Tang, H., and Ye, Y. (2012) RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics 28: 125-126.

Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011) PHAST: A fast phage search tool. Nucleic Acids Res 39: W347-W352.

Tables

Table 1. General features of the Nostocales genomes of the newly recognized ADA clade, including eight genomes newly reported in this work. Shading indicates membership of the phylogroups ADA-1 through ADA-4 delineated in Fig. 1.

Table 2. Genome fragmentation in ADA clade genomes.

Figures

Figure 1. Phylogenomic tree of the Nostocales emphasizing the distinct ADA clade comprised of bloom-forming members from four continents. The tree was built using a concatenated alignment of all single-copy orthologues that are found in all genomes (279 genes). Genome names are colored based on distinct phylogroups (potential species) delineated by genomic ANI cutoff of 96% and the aligned genome fraction (AF) cutoff of 0.6 (Varghese et al., 2015). Genomes new to this study are highlighted with an asterisk. The presence of genes for key secondary metabolites is indicated.

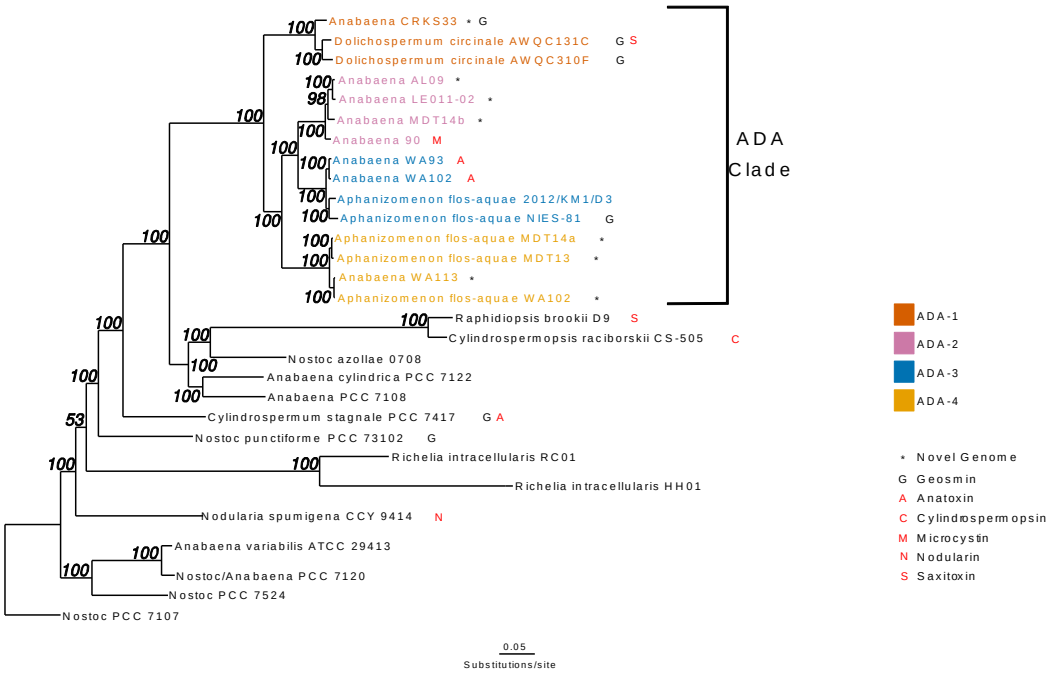
Figure 2. Genome-wide average nucleotide identities (gANI) for the Nostocales genomes shown in Fig. 1. Cladograms plot the relationships between gANI properties of the genomes.

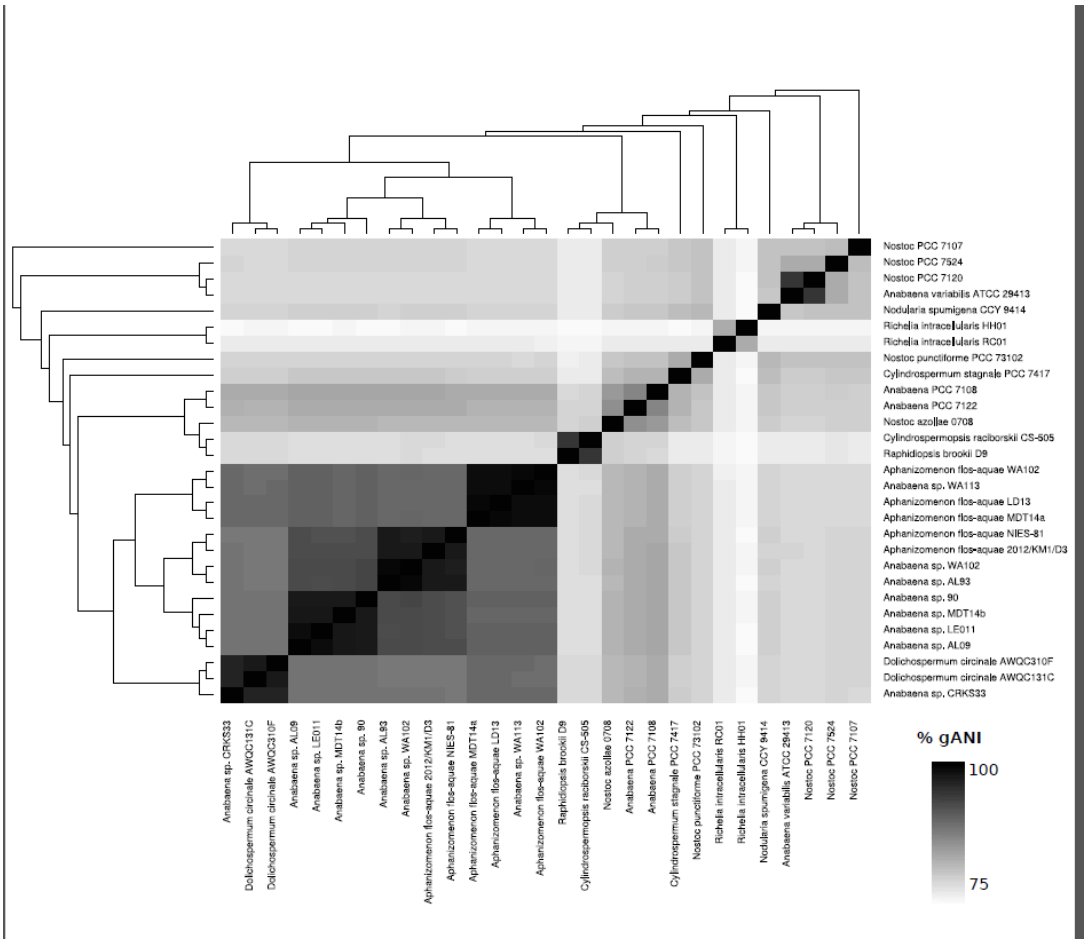
Figure 3. Core and pan genome analysis for the 15 ADA clade genomes. **A.** Core genome curves generated by the Tettelin (red line) and Willenbrock (blue line) exponential decay models, estimating 1559 (standard error = 261) and 1478 (standard error = 225) core genes, respectively. **B.** Pan-genome analysis with Tettelin estimation of 8956 genes with a residual standard error of 314. Dots represent single iterations of the core and pan genome calculations. **C.** Representation of KEGG gene groups in core and pan genomes. Soft-core genes are genes found in all but one genome.

Figure 4. Commonalities and differences across the ADA clade in genes affecting nutrient acquisition, metabolic and physiological traits that could influence niche partitioning. Each row lists the presence or absence of orthologous genes/pathways in the ADA genomes and in Nostoc PCC 7120, in which the roles of many genes have been functionally tested. Instances in which genes are disrupted or pathways are incomplete are indicated, as are cases in which gene presence is uncertain because of contig fragmentation in draft genomes.

Figure 5. Differential presence of select genes across the Nostocales, including the ADA clade. *nifD*, *nifH*, *fdxN* and *hupL* are variously interrupted by excision elements that are removed during heterocyst differentiation by *xis* recombinases that act at the specific sites indicated. Uninterrupted genes occur in some cases, as indicated by the absence of specific *xis* genes; in addition, multiple copies of intact (non-identical) *nifH* genes exist in the following genomes: *Anabaena variabilis* ATCC 29413 (4), *Cylindrospermum stagnale* (3), *Nostoc* PCC 7120 (2), *Nostoc azollae* (3), *Nostoc punctiforme* (3). For more details, see Suppl. Tables 6B, 6C, 10.

ɟ. 1


Anabaena CRKS33 * G
Dolichospermum circinale AWQC131C G S
Dolichospermum circinale AWQC310F G
Anabaena AL09 *
Anabaena LE011-02 *
Anabaena MDT14b *
Anabaena 90 M
Anabaena WA93 A
Anabaena WA102 A
Aphanizomenon flos-aquae 2012/KM1/D3
Aphanizomenon flos-aquae NIES-81 G
Aphanizomenon flos-aquae MDT14a *
Aphanizomenon flos-aquae MDT13 *
Anabaena WA113 *
Aphanizomenon flos-aquae WA102 *
ADA Clade
Raphidiopsis brookii D9 S
Cylindrospermopsis raciborskii CS-505 C
Nostoc azollae 0708
Anabaena cylindrica PCC 7122
Anabaena PCC 7108
Cylindrospermum stagnale PCC 7417 G A
Nostoc punctiforme PCC 73102 G
Richelia intracellularis RC01
Richelia intracellularis HH01
Nodularia spumigena CCY 9414 N
Anabaena variabilis ATCC 29413
Nostoc/Anabaena PCC 7120
Nostoc PCC 7524
Nostoc PCC 7107
0.05
Substitutions/site
ADA-1
ADA-2
ADA-3
ADA-4
* Novel Genome
G Geosmin
A Anatoxin
C Cylindrospermopsin
M Microcystin
N Nodularin
S Saxitoxin


Nostoc PCC 7107
Nostoc PCC 7524
Nostoc PCC 7120
Anabaena variabilis ATCC 29413
Nodularia spumigena CCY 9414
Richelia intracellularis HH01
Richelia intracellularis RC01
Nostoc punctiforme PCC 73102
Cylindrospermum stagnale PCC 7417
Anabaena PCC 7108
Anabaena PCC 7122
Nostoc azollae 0708
Cylindrospermopsis raciborskii CS-505
Raphidiopsis brookii D9
Aphanizomenon flos-aquae WA102
Anabaena sp. WA113
Aphanizomenon flos-aquae LD13
Aphanizomenon flos-aquae MDT14a
Aphanizomenon flos-aquae NIES-81
Aphanizomenon flos-aquae 2012/KM1/D3
Anabaena sp. WA102
Anabaena sp. AL93
Anabaena sp. 90
Anabaena sp. MDT14b
Anabaena sp. LE011
Anabaena sp. AL09
Dolichospermum circinale AWQC310F
Dolichospermum circinale AWQC131C
Anabaena sp. CRKS33
% gANI
100
75

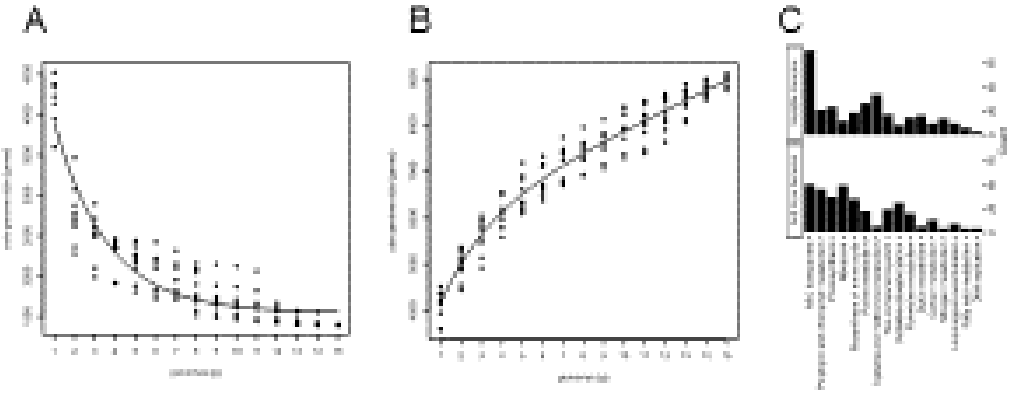

A
B
C

Fig. 4


Disrupted gene present or some genes missing from operon
Some genes incomplete or missing due to contig break
Ana CRK33
Dol 131C
Dol 310F
Ana AL09
Ana LE011
Ana MDT 14b
Ana 90
Ana AL93
Ana WA102
AFA KM1/D3
AFA NIES-81
AFA MDT14a
AFA LD13
Ana WA113
AFA WA102
Nostoc PCC 7120
Phosphate ABC uptake system
Phosphonate ABC uptake system
Phosphonate C-P lyase genes (phnCDEGHIMJKLM)
Phosphonopyruvate hydrolase (palA)
Nitrite & nitrate reduction (nirA, narB)
ABC nitrite & nitrate uptake system (nrtABCD)
MFS nitrite & nitrate uptake system (nrtP)
Ammonium transport (amt)
Urea uptake (urtABCDE)
Urease (ureABCDEFG)
natB amino acid ABC transporter
natI amino acid ABC transporter
bgtB amino acid ABC transporter
natF amino acid ABC transporter
Sulfate uptake (cysPTW)
Sulfate adenylyl transferase (sat)
APS kinase (cysC)
Reduction to sulfite & sulfide from APS/PAPS (cysH, sir)
Alkane sulfonate uptake to produce sulfite (ssuBACDE, tauD, FMN red)
ABC-type cobalamin transporters
BtuB-like TonB-dependent cobalamin transporter (single gene)
tonB
Phycocyanin (cpc)
Phycoerythrin (cpe)
Phycoerythrocyanin (pec)
metYX; methionine synthesis via methanethiol
Carboxymethylenebutenolidase
Gamma-glutamyltranspeptidase (ggt)
Mo-dependent oxidation (yagTSR-like, mocA)
Darkness beta carotene synthesis (crtH)
Cytochrome d ubiquinol oxidase (cydAB)
Phototaxis (pixJ, cheYW)
Phytochrome A (aphA) and 2-component regulator
Anatoxin-a
Cylindrospermopsin
Microcystin
Saxitoxin
Hassalidin
Aeruginosin
Anabaenopeptin
Anabaenopeptilide
Bacteriocins
Cyanobactin (anacyclamide)
Geosmin
Type I-C CRISPR-Cas (CRISPR6)
Type I-D CRISPR-Cas (CRISPR5)
Type I-D CRISPR-Cas (CRISPR1)
Type III-B CRISPR-Cas (CRISPR4)

ig. 5

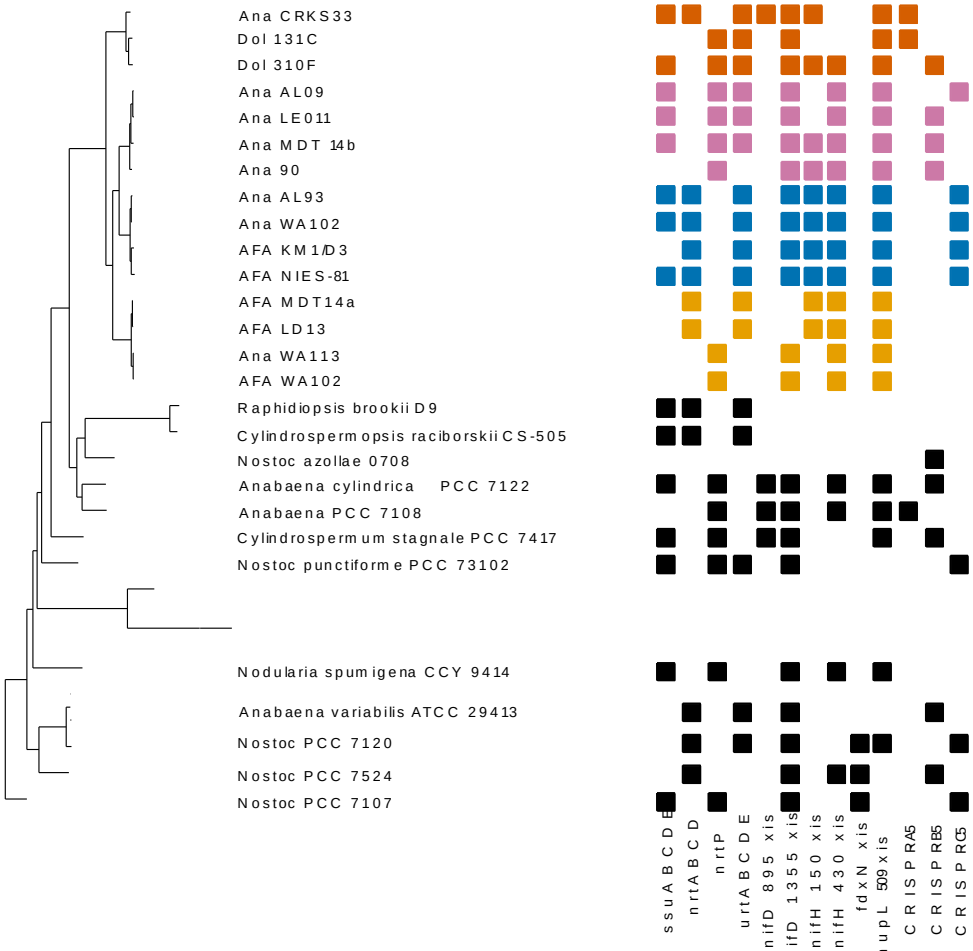

Ana CRKS33
Dol 131C
Dol 310F
Ana AL09
Ana LE011
Ana MDT 14b
Ana 90
Ana AL93
Ana WA102
AFA KM1/D3
AFA NIES-81
AFA MDT14a
AFA LD13
Ana WA113
AFA WA102
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS-505
Nostoc azollae 0708
Anabaena cylindrica PCC 7122
Anabaena PCC 7108
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102
Nodularia spumigena CCY 9414
Anabaena variabilis ATCC 29413
Nostoc PCC 7120
Nostoc PCC 7524
Nostoc PCC 7107
ssuA B C D E
nrtA B C D
nrtP
urtA B C D E
nifD 895 xis
nifD 1355 xis
nifH 150 xis
nifH 430 xis
fdxN xis
hupL 509 xis
CRISPRA5
CRISPRB5
CRISPRC5

Table 1 General features of the Nostocales genomes of the newly recognized ADA clade, including eight genomes newly reported in this work. Shading indicates membership of the phylogroups ADA-1 through -4 delineated in Fig. 1.

Genome	Abbrev. name	Isolation Site & Date	Genome Size (Mbp)	[1]Source/ Status/ Contig No.	NCBI Accession Number	[2]Estimated % Genome Completeness/ Contamination	GC%	No. Protein Coding Sequences	[3]No. rRNA operons (5S-16S-23S)	Reference
Anabaena sp. CRKS33	Ana CRKS33	Cheney Reservoir, KS, USA; 30-Aug-2013	4.95	EM / D 1109	LJOT00000000	99.44 1.78	37.6	4638	2	This work
Dolichospermum circinale AWQC131C	Dol 131C	Lake Cargelligo, NSW, Australia	4.45	C / D 121	NZ_KE384588	99.56 0.00	37.01	3750	1	D'Agostino et al., 2014
Dolichospermum circinale AWQC310F	Dol 310F	Farm Dam, Milawa, VIC, Australia; 1995	4.41	C / D 82	NZ_KE384663	99.56 0.00	37.33	3676	2	D'Agostino et al., 2014
Anabaena sp. AL09	Ana Al09	Lake Ontario, Canada; 1-Aug-2005	4.66	C / D 109	LJOQ00000000	98.11 0.00	38.1	3988	-	This work
Anabaena sp. LE011-02	Ana LE011	Lake Erie, USA; 12-Jul-2011	4.74	C / D 122	LJOP00000000	99.22 0.11	38.06	4072	-	This work
Anabaena sp. MDT14b	Ana MDT14b	Upper Klamath Lake, OR, USA; 4-Jun-2014	4.96	EM / D 1227	LJOV00000000	97.17 4.22	38.9	4546	4	This work
Anabaena sp. 90	Ana 90	Lake Vesijärvi, Finland; 30-Jul-1986	5.31	C / F 5(2 + 3)	NC_019427	- -	38.1	4444	5	Wang et al., 2012
Anabaena sp. AL93	Ana AL93	American Lake, WA, USA; 1993	5.66	C / D 217	LJOU00000000	99.67 0.52	38.4	4693	5	Brown et al., 2016
Anabaena sp. WA102	Ana WA102	Anderson Lake, WA, USA; 20-May-2013	5.78	C / F 2(1 + 1)	NZ_CP011456	- -	38.4	4880	5	Brown et al., 2016
Aphanizomenon flos-aquae 2012/KM1/D3	AFA KM1D3	Curonian Lagoon, Lithuania, 2012	5.74	C / D 325	NZ_JSDP01000254	87.52 7.22	38.22	4601	5	Sulcius et al., 2015
Aphanizomenon flos-aquae NIES-81	AFA NIES-81	Lake Kasumigaura, Japan; 1978	5.85	C / D 103	NZ_KI928192	99.67 0.56	37.37	4744	5	Cao et al., 2014
Aphanizomenon flos-aquae MDT14a	AFA MDT14a	Upper Klamath Lake, OR, USA; 4-Jun-2014	4.63	EM / D 193	LJOX00000000	99.00 1.00	37.11	3936	6	This work
Aphanizomenon flos-aquae LD13	AFA LD13	Upper Klamath Lake, OR, USA; 6-Aug-2013	4.44	C / D 307	LJOY00000000	99.67 0.37	37.05	3787	3	This work
Anabaena sp. WA113	Ana WA113	Cranberry Lake, WA, USA; 11-Aug-2014	4.70	EM / D 279	LJOS00000000	99.89 0.44	37.22	4002	7	This work
Aphanizomenon flos-aquae WA102	AFA WA102	Anderson Lake, WA, USA; 20-May-2013	5.48[4]	EM / D 1160	LJOW00000000	99.89 3.60	39.12	4685[4]	7	This work

[1] EM, environmental metagenome-assembled genome; C, culture; D, draft genome; F, finished genome. No plasmids were identified for the draft genomes. Plasmid(s) were associated with both finished genomes; number of chromosomes and plasmids indicated in parentheses.
[2] % Genome completeness and proportion of contaminating sequences estimated by CheckM (Parks et al., 2015).
[3]Multiple copy ribosomal RNA genes are susceptible to absence or incomplete representation in shotgun metagenomes; thus, rRNA operon numbers are provisional for draft genomes (See Supplementary Table 3 for more details). tRNAs are likewise commonly incompletely represented in draft genomes: see Supplementary Table 4 for tRNA genes.
[4]Reflects removal of 464,126 bp of likely viral sequences in 20 contigs of original draft genome cluster, including 611 coding sequences and 13 tRNA genes.

Table 2 Genome fragmentation in ADA clade genomes

	ADA-1			ADA-2				ADA-3				ADA-4			
	Ana CRKS33	**Dol 131C**	**Dol 310F**	**Ana AL09**	**Ana LE011**	**Ana MDT14b**	**Ana 90**	**Ana AL93**	**Ana WA102**	**AFA KM1D3**	**AFA NIES-81**	**AFA MDT14a**	**AFA LD13**	**Ana WA113**	**AFA WA102**
Av. within-group LCB length (kb)	4.89	4.89	4.88	4.51	4.54	4.61	4.14	9.32	9.11	7.47	9.6	8.42	8.33	8.47	8.39
% genome in LCBs >10 kb	45.0	49.2	48.8	26.6	27.5	25.8	22.8	60.7	58.7	43.7	63.2	60.1	60.9	58.4	49.2
Total Tn genes (kb)	28.7	13.9	12.7	25.9	21.1	47.7	69.3	30.9	104.1	92.3	43.2	21.0	17.3	31.3	59.0
Total HNH genes (kb)	6.3	1.7	2.7	3.3	3.8	6.1	7.8	6.2	8.0	7.7	6.7	2.1	3.4	3.2	4.4
% genome in Tn + HNH genes	0.71	0.35	0.35	0.63	0.53	1.08	1.45	0.66	1.94	1.74	0.85	0.50	0.47	0.73	1.16
% genome as repeats >100 bp[1]	(0.14)	(0.53)	(0.77)	—	(0.05)	(0.04)	5.11	(0.09)	8.34	(7.84)	(5.55)	(0.11)	(0.11)	(0.08)	(0.07)
No. regions of possible phage origin[2]	0 (4)	0	1 (3)	(0) 2	0	1 (4)	3 (5)	1 (2)	3 (4)	0 (1)	0 (2)	2 (5)	0 (1)	0	0 (8)
Prophage-like regions (kb, #)[3]	—	—	8.1	—	—	—	60.3 (4x)	—	26.5 (3x)	—	—	87	—	—	—

LCB, locally collinear block (Darling et al., 2010); Tn, transposon; HNH, HNH homing endonuclease
Tn and HNH-HE genes and repeat regions were identified by annotations
[1]Numbers in parentheses refer to values for draft genomes, for which Illumina short-read sequencing can severely underestimate the presence of repeat sequences
[2]Number of prophage fragments identified by PHAST; no intact prophages were found. The numbers in parentheses refer to the number of phage-like fragments identified by VirSorter, some of which may be contaminating contigs representing phage fragments.
[3]Prophage fragments identified by PHAST
See Suppl Table 5 for more details