Title

Resequencing and annotation of the Nostoc punctiforme ATTC 29133 genome: facilitating biofuel and high-value chemical production

Permalink

https://escholarship.org/uc/item/1fq9x99n

Journal

AMB Express, 7(1)

ISSN

2191-0855

Authors

Moraes, Luis E

Blow, Matthew J

Hawley, Erik R

et al.

Publication Date

2017-12-01

DOI

10.1186/s13568-017-0338-9

Peer reviewed

AMB Express

ORIGINAL ARTICLE

Open Access

Resequencing and annotation of the *Nostoc punctiforme* ATTC 29133 genome: facilitating biofuel and high-value chemical production

Luis E. Moraes[1], Matthew J. Blow[2], Erik R. Hawley[3], Hailan Piao[4], Rita Kuo[2], Jennifer Chiniquy[2], Nicole Shapiro[2], Tanja Woyke[2], James G. Fadel[1] and Matthias Hess[1,2*]

Abstract

Cyanobacteria have the potential to produce bulk and fine chemicals and members belonging to *Nostoc* sp. have received particular attention due to their relatively fast growth rate and the relative ease with which they can be harvested. *Nostoc punctiforme* is an aerobic, motile, Gram-negative, filamentous cyanobacterium that has been studied intensively to enhance our understanding of microbial carbon and nitrogen fixation. The genome of the type strain *N. punctiforme* ATCC 29133 was sequenced in 2001 and the scientific community has used these genome data extensively since then. Advances in bioinformatics tools for sequence annotation and the importance of this organism prompted us to resequence and reanalyze its genome and to make both, the initial and improved annotation, available to the scientific community. The new draft genome has a total size of 9.1 Mbp and consists of 65 contiguous pieces of DNA with a GC content of 41.38% and 7664 protein-coding genes. Furthermore, the resequenced genome is slightly (5152 bp) larger and contains 987 more genes with functional prediction when compared to the previously published version. We deposited the annotation of both genomes in the Department of Energy's IMG database to facilitate easy genome exploration by the scientific community without the need of in-depth bioinformatics skills. We expect that an facilitated access and ability to search the *N. punctiforme* ATCC 29133 for genes of interest will significantly facilitate metabolic engineering and genome prospecting efforts and ultimately the synthesis of biofuels and natural products from this keystone organism and closely related cyanobacteria.

Keywords: *Nostoc punctiforme*, Cyanobacteria, Carbon cycle, Nitrogen cycle, Natural product synthesis, Single molecule real-time sequencing

Introduction

Cyanobacteria are capable of converting carbon dioxide into oxygen via photosynthesis and have been proposed as significant players in the evolution of current atmospheric oxygen levels during the Precambrian era (Meeks et al. 2001; Schirrmeister et al. 2013) and as origin of chloroplasts via endosymbiosis with eukaryotic cells (Agapakis et al. 2011; Chu et al. 2004). The economic significance of cyanobacteria centers on their potential as synthesis platform for biofuels and other natural products (Machado and Atsumi 2012; Rosgaard et al. 2012); and gene clusters coding for novel enzymes that catalyze unique chemical reactions have been identified from several cyanobacterial genomes (Kleigrewe et al. 2016; Zarzycki et al. 2013). Access to cyanobacterial genomes with up-to-date annotation will provide an improved framework for studying and understanding the biology of this phylogenetic group that is essential for the global carbon and nitrogen cycle and to develop genetic tools that will facilitate the engineering of cyanobacteria for

*Correspondence: mhess@ucdavis.edu
[1] Department of Animal Science, University of California, Davis, 2251 Meyer Hall, Davis, CA 95616, USA
Full list of author information is available at the end of the article

natural products synthesis. *Nostoc punctiforme* is a filamentous cyanobacterium found in the majority of illuminated environments and it plays a key role in the global nitrogen cycle through its facultative diazotrophic traits. Its life cycle is complex with vegetative cells differentiating into heterocysts, hormogonia and akinetes. The genome of *N. punctiforme* has been described previously as a large bacterial genome approaching 10 million (M) base pairs (bp) and the most recent genome sequence of the type strain *N. punctiforme* ATCC 29133 was released into the National Center for Biotechnology Information's (NCBI's) public nonredundant (nr) database more than 15 years ago (Meeks et al. 2001). Since then, this article has been cited more than 170 times and the initial annotation of this genome has been extensively utilized to identify genes involved in the carbon and nitrogen fixation process mediated by *N. punctiforme* and related species. This highlights the importance of this microorganism as a model system and the value of an up-to-date annotation of its genome. The *N. punctiforme* ATCC 29133 genome that was available through NCBI's nr database (GenBank ID: CP001037.1) prior to this study contained 9,059,191 bp with approximately three quarters (77.44%) of its genome sequence coding for genes. Of the 6690 protein-coding genes, 2601 (38.3%) had no functional annotation, which subsequently renders a comprehensive understanding of their physiological role and the metabolic capability of the complete genome rather challenging. The rapid improvements of sequencing technologies and sequence analysis tools in the past decade has extended our ability of generating and mining genomic information to several levels beyond what was possible more than a decade ago (Gomez-Escribano et al. 2016; Shendure and Lieberman Aiden 2012) when the first version of the *N. punctiforme* ATCC 29133 genome was released. For instance, sequencing of bacterial genomes using Pacific Biosciences's (PacBio's) single molecule sequencing approach that simultaneously allows the detection of deoxyribonucleic acid (DNA) methylation patterns in bacterial genomes is now routinely performed in many laboratories (Flusberg et al. 2010; Gomez-Escribano et al. 2016; Koren et al. 2013). Moreover, the ample development of bioinformatics tools with various strategies for genome assembly and annotation has dramatically increased our ability to understand the biology of prokaryotes and eukaryotes at the genomic level and to mine their genomes for natural products (Weber et al. 2015; Medema et al. 2015; Yandell and Ence 2012). In this context, the objective of this study is to provide an update of the *N. punctiforme* genome through resequencing and reannotating of its genome with state-of-the-art technologies. The data presented here were generated and analyzed as part of Community Science Program (Project ID 1393) at the Department of Energy's Joint Genome Institute (DOE's JGI) and the sequenced and annotated genome are publically available through the Joint Genome Institute's Integrated Microbial Genomes and Microbiomes (IMG/M; https://img.jgi.doe.gov/cgi-bin/mer/main.cgi) system (Markowitz et al. 2012) samples warehouse under the IMG Submission Identifier (ID) 62757. The data can be explored and analyzed using the IMG/M system or downloaded for further analyses using stand-alone tools.

Nostoc punctiforme ATCC 29133 is an aerobic, Gram-negative, filamentous and motile cyanobacterium with photosynthetic and nitrogen-fixing capabilities. *Nostoc punctiforme* is found in illuminated terrestrial environments and has a complex life cycle with vegetative cells differentiating into heterocysts, hormogonia and akinets. Vegetative cells are often larger than hormogonium cells with a diameter between 5 and 6 mm. The hormogonium cells institute infection for the symbiosis between cyanoba cterium and plants and are 1.5–2 mm in diameter (Meeks et al. 2001; Meeks and Elhai 2002). The heterocysts are 6–10 mm in diameter whereas akinetes are 10–20 mm. The functioning of oxygen sensitive nitrogenases is ensured by the heterocyst's glycolipid layer, which provides a barrier for gases. Moreover, the differentiation of vegetative cells into the spore-like akinetes is determined by environmental factors for example through light limitation (Adams and Duggan 1999). Besides its photoautotrophic mode, *N. punctiforme* growth has been reported in the absence of light but with availability of sucrose, glucose or fructose (Meeks et al. 2001). General features and information about *N. punctiforme* ATCC 29133 are summarized in Table 1.

We selected *N. punctiforme* ATCC 29133 for genome resequencing and reannotation due to its importance in the global carbon and nitrogen cycle, its potential biotechnological applications and the value and impact of the first publicly available version of its genome on the scientific community. It is very likely that an updated annotation of the *N. punctiforme* ATTC 29133 genome will benefit future genome explorations that target genes associated with carbon and nitrogen fixation and the production of biofuels and value-added chemicals.

Materials and methods

Growth conditions and genomic DNA preparation

Nostoc punctiforme ATCC 29133 was obtained from the American Type Culture Collection (ATCC) and grown at 26 °C in a 75 cm^2 corning vent cap tissue flask using 30 mL of ATCC's 616 medium under diurnal conditions with 16 h of light exposure followed by 8 h of darkness. After 12 days, 10 mL of cell culture were concentrated by centrifugation at 10,000 rpm for 10 min. Cells were resuspended

Table 1 Classification and general features of *Nostoc punctiforme* ATCC 29133

Property	Term	Evidence code[a]
Classification	Domain *Bacteria*	TAS (Woese et al. 1990)
	Phylum *Cyanobacteria*	TAS (Castenholz 2015)
	Class *Cyanophyceae*	
	Order *Nostocales* (Subsection IV)	TAS (Rippka et al. 2015)
	Family *Nostocaceae* (Subsection IV.I)	TAS (Whitman 2015)
	Genus *Nostoc*	TAS (Herdman et al. 2015)
	Species *punctiforme*	TAS (Herdman et al. 2015)
	Strain ATCC 29133/PCC 73102	
Gram stain	Negative	TAS (Hoiczyk and Hansel 2000)
Cell shape	Filamentous	TAS (Herdman et al. 2015)
Motility	Motile	TAS (Lehner et al. 2011)
Growth temperature	26 °C	IDA
pH	7.1	IDA
Habitat	Fresh water, Soil	TAS (Herdman et al. 2015)
Oxygen requirement	Aerobic	TAS (Herdman et al. 2015)
Biotic relationship	Symbiotic	TAS (Herdman et al. 2015)
Pathogenicity	Non-pathogen	NAS
Geographic location	USA/Washington	
Sample collection	October 10th 2014	
Latitude	46.3119	
Longitude	−119.263	

[a] Evidence codes—IDA: Inferred from direct assay; TAS: traceable author statement (i.e., a direct report exists in the literature); NAS: non-traceable author statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). Evidence codes are from the Gene Ontology project (Ashburner et al. 2000)

in 1 mL fresh media and used for DNA extraction. DNA was extracted using MP Biomedicals' FastDNA™ SPIN Kit for Soil DNA extraction kit according to the manufacturer's protocol. DNA was suspended in 100 µL DNase/pyrogen-free H_2O and DNA concentration was determined using a Qubit 2.0 fluorometer (Life Technologies, Grand Island, NY) according to the manufacturer's protocol.

Genome sequencing and assembly

The draft genome of *N. punctiforme* ATCC 29133 was generated at DOE's JGI using PacBio's single molecule real-time sequencing technology (Eid et al. 2009). A > 10 kbp PacBio SMRTbell™ library was constructed and sequenced on the PacBio RS platform, which generated 446,884 filtered subreads totaling 915.4 Mbp. A description of the library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov. The raw reads were assembled using hierarchical genome-assembly process (HGAP version: 2.3.0, protocol version = 2.3.0 method = RS HGAP Assembly.3, smrtpipe.py v1.87.139483) (Chin et al. 2013). The final draft assembly contained 65 contigs in 65 scaffolds, totaling 9.064 Mbp in size. The obtained read coverage was 96.8-fold. Sequencing and assembly statistics are summarized in Tables 2 and 3.

Table 2 Sequencing and assembly information

MIGS ID[a]	Property	Term
MIGS 31	Finishing quality	High quality draft
MIGS-28	Libraries used	>10 kbp PacBio SMRTbell
MIGS 29	Sequencing platform	PacBio SMRT RS
MIGS 31.2	Fold coverage	96.8-fold
MIGS 30	Assembler	HGAP 2.3.0

[a] Field et al. 2008

Genome annotation

Genes were identified using Prodigal (Hyatt et al. 2010) and the predicted coding DNA sequences (CDSs) were translated and used to search the NCBI nr database, as well as the UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAScanSE tool (Lowe and Eddy 1997) was used to find transfer ribonucleic acid (tRNA) genes, whereas ribosomal RNA (rRNA) genes were found by searches against models of the rRNA genes built from SILVA (Pruesse et al. 2007). Other non-coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL (Nawrocki et al. 2009). Additional gene prediction analysis and manual functional

Table 3 Genome statistics for *Nostoc punctiforme* ATCC 29133

Attribute	Meeks et al. (2001)		This study	
	Value	% of total	Value	% of total
Genome size (bp)	9059191	100	9064343	100
DNA coding (bp)	7015747	77.44	7393120	81.56
DNA G+C (bp)	3746385	41.35	3751137	41.38
DNA scaffolds	6	100	65	100
Total genes	6791	100	7775	100
Protein coding genes	6690	98.51	7664	98.57
RNA genes	101	1.49	111	1.43
Genes with function prediction	4089	60.21	5076	65.29
Genes without assigned function	2601	38.3	2588	33.29
Genes assigned to COGs	3432	50.54	3598	46.28
Genes with Pfam domains	5010	73.77	5381	69.21
Genes with signal peptides	274	4.03	297	3.82
Genes with transmembrane helices	1525	22.46	1635	21.03
Genes in biosynthetic clusters	602	8.86	1080	13.89
CRISPR repeats	8		10	

annotation was performed within the JGI's Integrated Microbial Genome platform (Markowitz et al. 2010).

Results

Genome properties

The genome of *N. punctiforme* ATCC 29133 generated in this study is 9,064,343 bp in length with 7,393,120 (81.56%) coding base pairs and a GC content of 41.38%. Assembly of the sequences produced 65 DNA scaffolds of 65 contigs and a total of 7775 genes for which 7664 (98.57%) were identified as protein coding genes and 111 (1.43%) were identified as RNA genes. In addition, 5076 genes (65.29%) had a predicted function, which increases the number of *N. punctiforme* ATCC 29133 genes with functional annotation by ~24%. Of the identified genes, 3598 (46.28%) were assigned to Clusters of Orthologous Groups (COGs) categories. Genome statistics and distribution of genes into COGs functional categories are presented in Tables 3 and 4 respectively. In short, the COGs functional category with the largest number of assigned genes was identified as signal transduction mechanisms with 313 genes (7.71%), followed by the cell wall/membrane biogenesis functions with 277 genes (6.83%). Furthermore, 564 genes (13.9%) had a "general function" prediction, 270 genes with unknown function (6.65%) and 4177 genes (53.72%) were not assigned to any COG category.

To evaluate how the assembled genome sequence from this project compares to the *N. punctiforme* genome released by the JGI in 2014 (GenBank ID: CP001037.1; not published), we aligned both assembled genomes to each other. Alignment was performed using MUMmer 3.0 (Kurtz et al. 2004) with the previously released genome sequence as the reference. The genomic sequence similarity plot (Fig. 1) suggests high agreement between the genome sequences. The JGI IMG portal genome comparison tools indicate an average nucleotide identity of 99.99%, a fraction of orthologous genomic regions of 0.98 and 0.92 and 6606 bidirectional best hits. In contrast, the genome presented here consists of 9064343 bp compared to the previously reported 9059191 bp, which might be due to the larger amount of sequence generated, but it is also possible that this is caused by the improved assembly algorithms that was developed since the first genome assembly was performed and that were employed for our data analysis. Likewise, we have identified 7775 genes with 7664 genes identified as protein-coding genes and 5076 genes with function prediction. In the initial assembly performed by the JGI in collaboration with Meeks and colleagues only 6791 genes were identified, with 6690 of these genes identified as protein-coding genes and 4089 of these genes with a functional prediction. The initial assembly has also been uploaded and can be accessed and utilized for future analyses and studies through the JGI's IMG system using the IMG Genome ID 642555144 (https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=TaxonDetail&page=tax onDetail&taxon_oid=642555144).

Discussion

The advent of affordable second and third-generation DNA sequencing resulted in a significant reduction of costs per sequence supporting the democratization of genome sequencing (Caporaso et al. 2012). Sequencing microbial genomes is now a standard procedure performed in many laboratories and third-generation sequencing technologies, such as PacBio's SMRT Sequencing platform, facilitate to generate almost complete high-quality genomes from archaea and bacteria in only a few hours (Eid et al. 2009; Rice et al. 2016). This development resulted in a vast amount of genomic information that is now available through public database such as NCBI's GenBank (https://www.ncbi.nlm.nih.gov/genome/browse/) or JGI's IMG (https://img.jgi.doe.gov/). These two databases alone contained a total of 89971 and 47009 prokaryotic genome submissions, including 5590 and 26707 non-redundant complete prokaryotic genomes, when accessed on December 27th 2016 respectively. With this almost infinite and continuously growing amount of sequence information, identification and extraction of relevant and accurate data becomes a major

Table 4 Number of genes associated with general COG functional categories

Code	Meeks et al. (2001)		This study		Description
	Count	%[a]	Count	%[a]	
E	246	6.23	245	6.04	Amino acid transport and metabolism
G	185	4.69	192	4.73	Carbohydrate transport and metabolism
D	43	1.09	39	0.96	Cell cycle control, cell division, chromosome partitioning
N	63	1.6	67	1.65	Cell motility
M	276	6.99	277	6.83	Cell wall/membrane/envelope biogenesis
B	2	0.05	2	0.05	Chromatin structure and dynamics
H	240	6.08	240	5.91	Coenzyme transport and metabolism
V	144	3.65	153	3.77	Defense mechanisms
C	207	5.24	210	5.17	Energy production and conversion
W	20	0.51	23	0.57	Extracellular structures
S	242	6.13	270	6.65	Function unknown
R	553	14.01	564	13.9	General function prediction only
P	233	5.9	237	5.84	Inorganic ion transport and metabolism
U	37	0.94	38	0.94	Intracellular trafficking, secretion, and vesicular transport
I	143	3.62	144	3.55	Lipid transport and metabolism
–	43	1.09	68	1.68	Mobilome: prophages, transposons
F	76	1.93	74	1.82	Nucleotide transport and metabolism
O	198	5.02	204	5.03	Posttranslational modification, protein turnover, chaperones
A	1	0.03	1	0.02	RNA processing and modification
L	138	3.5	140	3.45	Replication, recombination and repair
Q	180	4.56	194	4.78	Secondary metabolites biosynthesis, transport and catabolism
T	318	8.05	313	7.71	Signal transduction mechanisms
K	155	3.93	159	3.92	Transcription
J	205	5.19	204	5.03	Translation, ribosomal structure and biogenesis
–	3359	49.46	4177	53.72	Not in COG

[a] Based on the total number of protein coding genes

challenge. Microbiologists are nowadays limited mostly by the relatively low throughput of molecular experiments that remain necessary to test the functionality and biological role of proteins newly identified using in silico approaches. This renders accurate functional prediction as key to an enhanced understanding of biological processes (Jiang et al. 2016). The importance of an accurate function prediction is even more pronounced in organisms that function as model systems and whose genomes serve as reference for the analyses of phylogenetically related genomes (Cormier et al. 2016); which explains the growing interest in improving the genome annotation of model system microorganisms such as *Bacillus pumilus* (Gioia et al. 2007; Stepanov et al. 2016) and *Pichia pastoris* (*Komagataella phaffi*) (Valli et al. 2016; De Schutter et al. 2009), whose genomes were published during the early phase of the genomics era.

Nostoc punctiforme is an important model system for examining genomic and phenotypic properties of cyanobacteria and their biological carbon and nitrogen fixation capabilities (Sandh et al. 2014). An additional interest in the enhanced understanding of the *N. punctiforme* genome has been sparked by the increased utilization of cyanobacteria as synthesis platform of biofuels and other natural products (Machado and Atsumi 2012; Rosgaard et al. 2012). For these reasons we selected the type strain *N. punctiforme* ATCC 29133 for resequencing and reannotation with the assumption that an improved genome sequence and annotation would be of significant value to the scientific community. The newly sequenced, assembled and annotated *N. punctiforme* ATCC 29133 genome presented here, consists of 984 more genes and 987 more genes with predicted function compared to the previously available *N. punctiforme* ATCC 29133 genome. The improved gene calling and ability to predict gene function is most likely the result of combination of improved sequence quality and gene annotation algorithm that are now available. It can be anticipated that increasing the repertoire of genes with assigned function within the new *N. punctiforme* ATCC 29133 genome by ~24% might allow to fillin some of the knowledge-gaps that exist in our current understanding of the molecular processes

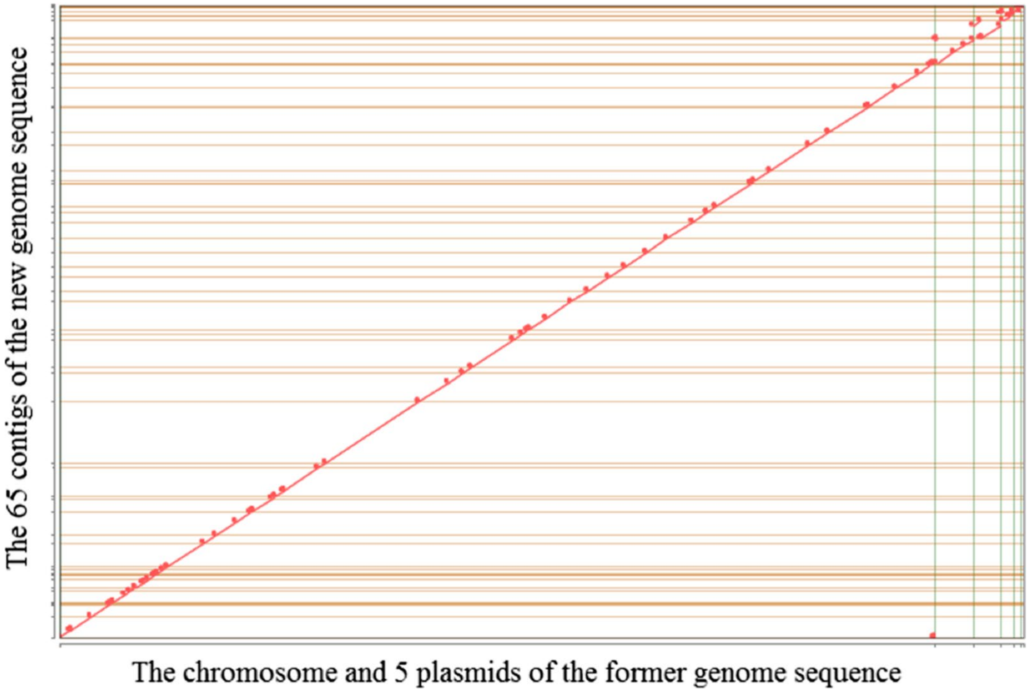

Fig. 1 Alignment plot of *Nostoc punctiforme* ATCC 29133 genomes

during microbial carbon and nitrogen fixation. The new *N. punctiforme* genome also contained ~79% more biosynthetic clusters when compared to the previous version, opening new opportunities to explore the capability of *N. punctiforme* and its phylogenetically related neighbors for secondary metabolite synthesis.

Access to a reannotated *N. punctiforme* ATCC 29133 genome through a versatile genome browser that allows to explore DNA sequence data without any or only limited bioinformatics skills will enhance the ability of the scientific community to mine this genome for new insights into genes and gene pathways associated with carbon and nitrogen fixation and secondary metabolite synthesis. It might furthermore facilitate the identification of the genetic properties that might be responsible for some of the metabolic functions and phenotypes associated with *N. punctiforme* and close relatives. In the case presented here, we opted to load and provide the genome sequence and annotation through IMG/M, a system that has been maintained by the Department of Energy for over a decade and that facilitates comparative analysis and visualization of multilayered omics data (Chen et al. 2016). Subsequent analysis can be performed within the system using a variety of the visualization and analysis tools that have been implemented since the first release of this genome data analysis system in 2005 (Markowitz et al. 2008). Of particular interest for users might be the gene neighborhood function within the Biosynthetic Cluster option, through which the global structure as well as the nucleotide sequence of individual genes and gene clusters can be retrieved for subsequent wet-lab experiments. As new omics data, such as proteomic and transcriptomic data, become available, they can also be integrated easily into the IMG/M structure to complement the reannotated *N. punctiforme* ATCC 29133 genome data generated during this study. To increase the value of the genome data presented here, sequence data and annotation can also be downloaded through IMG/M and the JGI's website and analyzed with other genome analysis software, including genome annotation pipelines and analysis systems such as SEED, RAST (Overbeek et al. 2014; Aziz et al. 2008) and the Department of Energy Systems Biology Knowledgebase (KBase; http://kbase.us).

To maximize the benefit of genomes generated with federal funding and to provide access to up-to-date sequence information without the need of in-depth bioinformatics skills, we advocate to resequence and annotate microbial type strains that were sequenced in the early stage of the genomics era, with up-to-date sequencing platforms and improved annotation software and to make these genomes and their annotations available to the scientific community through user-friendly genome browsers, such as the JGI's IMG/M, at the time of publication.

Abbreviations

M: million; bp: base pairs; nr: nonredundant; NCBI: National Center for Biotechnology Information; PacBio: Pacific Biosciences; DNA: deoxyribonucleic acid; DOE: Department of Energy; JGI: Joint Genome Institute; IMG/M: Integrated Microbial Genomes and Microbiomes; ID: identifier; ATCC: American Type Culture Collection; CDSs: coding deoxyribonucleic acid sequences; tRNA: transfer ribonucleic acid; rRNA: ribosomal ribonucleic acid; COGs: Clusters of Orthologous Groups.

Authors' contributions

Conceived and designed the experiments: MH; Performed the experiments: MH, ERH, HP; Generated and analyzed the data: LEM, MH; Wrote the paper: LEM, MH, ERH, JGF; Performed PacBio library prep: RK; Performed PacBio sequencing and sequence analysis: JC and MJB; Project management: NS and TW. All authors read and approved the final manuscript.

Author details

[1] Department of Animal Science, University of California, Davis, 2251 Meyer Hall, Davis, CA 95616, USA. [2] Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA. [3] ZeaChem, Boardman, OR 97818, USA. [4] Washington State University, Richland, WA 99354, USA.

Acknowledgements

We would like to thank Dave Culley for the fruitful discussions and guidance as well as the staff at the JGI who are essential for data generation through the JGI's Community Science Program.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The *N. punctiforme* ATCC 21933 nucleotide sequence data and the associated genome annotation are publicly available and can be downloaded or explored directly without any bioinformatics skills through the Department of Energy Joint Genome Institute's Integrated Microbial Genomes and Microbiome Sample (https://img.jgi.doe.gov/cgi-bin/m/main.cgi) under the Taxon ID 642555144 and 2617270889. The nucleotide sequence of the initial assembly can be downloaded at from NCBI's GenBank under the Accession Number CP001037.1.

Funding

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Worked performed at UC Davis and WSU was supported by the College of Agricultural and Environmental Sciences and the College of Arts and Sciences respectively.

Received: 11 November 2016 Accepted: 3 February 2017
Published online: 16 February 2017

References

Adams DG, Duggan PS (1999) Heterocyst and akinete differentiation in cyanobacteria. New Phytol 144(1):3–33. doi:10.1046/j.1469-8137.1999.00505.x

Agapakis CM, Niederholtmeyer H, Noche RR, Lieberman TD, Megason SG, Way JC, Silver PA (2011) Towards a synthetic chloroplast. PLoS ONE 6(4):e18877. doi:10.1371/journal.pone.0018877

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29. doi:10.1038/75556

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genom 9:75. doi:10.1186/1471-2164-9-75

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6(8):1621–1624. doi:10.1038/ismej.2012.8

Castenholz RW (2015) General characteristics of the cyanobacteria. Bergey's Man Syst Archaea Bacteria. doi:10.1002/9781118960608.cbm00019

Chen IA, Markowitz VM, Chu K, Palaniappan K, Szeto E, Pillay M, Ratner A, Huang J, Andersen E, Huntemann M, Varghese N, Hadjithomas M, Tennessen K, Nielsen T, Ivanova NN, Kyrpides NC (2016) IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. doi:10.1093/nar/gkw929

Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10(6):563–569. doi:10.1038/nmeth.2474

Chu KH, Qi J, Yu ZG, Anh V (2004) Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Mol Biol Evol 21(1):200–206. doi:10.1093/molbev/msh002

Cormier A, Avia K, Sterck L, Derrien T, Wucher V, Andres G, Monsoor M, Godfroy O, Lipinska A, Perrineau MM, Van De Peer Y, Hitte C, Corre E, Coelho SM, Cock JM (2016) Re-annotation, improved large-scale assembly and establishment of a catalogue of noncoding loci for the genome of the model brown alga *Ectocarpus*. New Phytol. doi:10.1111/nph.14321

De Schutter K, Lin YC, Tiels P, Van Hecke A, Glinka S, Weber-Lehmann J, Rouze P, Van de Peer Y, Callewaert N (2009) Genome sequence of the recombinant protein production host *Pichia pastoris*. Nat Biotechnol 27(6):561–566. doi:10.1038/nbt.1544

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138. doi:10.1126/science.1162986

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, dePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glockner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, Gil IS, Wilson G, Wipat A (2008) The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol 26(5):541–547. doi:10.1038/Nbt1360

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7(6):461–465. doi:10.1038/nmeth.1459

Gioia J, Yerrapragada S, Qin X, Jiang H, Igboeli OC, Muzny D, Dugan-Rocha S, Ding Y, Hawes A, Liu W, Perez L, Kovar C, Dinh H, Lee S, Nazareth L, Blyth P, Holder M, Buhay C, Tirumalai MR, Liu Y, Dasgupta I, Bokhetache L, Fujita M, Karouia F, Moorthy PE, Siefert J, Uzman A, Buzumbo P, Verma A, Zwiya H, McWilliams BD, Olowu A, Clinkenbeard KD, Newcombe D, Golebiewski L, Petrosino JF, Nicholson WL, Fox GE, Venkateswaran K, Highlander SK, Weinstock GM (2007) Paradoxical DNA repair and peroxide resistance gene conservation in *Bacillus pumilus* SAFR-032. PLoS ONE 2(9):e928. doi:10.1371/journal.pone.0000928

Gomez-Escribano JP, Alt S, Bibb MJ (2016) Next generation sequencing of Actinobacteria for the discovery of novel natural products. Mar Drugs. doi:10.3390/md14040078

Herdman M, Castenholz RW, Rippka R (2015) Cyanobacteria/Subsection IV/Subsection IV.I/Form-Nostoc. Bergey's Manual of Systematics of Archaea and Bacteria. doi:10.1002/9781118960608.gbm00459

Hoiczyk E, Hansel A (2000) Cyanobacterial cell walls: news from an unusual prokaryotic envelope. J Bacteriol 182(5):1191–1199. doi:10.1128/Jb.182.5.1191-1199.2000

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 11:119. doi:10.1186/1471-2105-11-119

Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, da Koo CE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Toronen P, Koskinen P, Holm L, Chen CT, Hsu WL, Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Bkc D, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent LC, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasidharan R, Yang H, Paccanaro A, Gillis J, Sedeno-Cortes AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcet-Houben M, Supek F, Gong Q, Ning W, Zhou Y, Tian W, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo M, Piovesan D, Tosatto SC, Del Pozo A, Fernandez JM, Maietta P, Valencia A, Tress ML, Benso A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, van Dijk AD, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida ESDC, Vencio RZ, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJ, Wass MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC, Brenner SE, Linial M, Orengo CA, Rost B, Greene CS, Mooney SD, Friedberg I, Radivojac P (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 17(1):184. doi:10.1186/s13059-016-1037-6

Kleigrewe K, Gerwick L, Sherman DH, Gerwick WH (2016) Unique marine derived cyanobacterial biosynthetic genes for chemical diversity. Nat Prod Rep 33(2):348–364. doi:10.1039/c5np00097a

Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, Radune D, Bergman NH, Phillippy AM (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14(9):R101. doi:10.1186/gb-2013-14-9-r101

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5(2):R12. doi:10.1186/gb-2004-5-2-r12

Lehner J, Zhang Y, Berendt S, Rasse TM, Forchhammer K, Maldener I (2011) The morphogene AmiC2 is pivotal for multicellular development in the cyanobacterium *Nostoc punctiforme*. Mol Microbiol 79(6):1655–1669. doi:10.1111/j.1365-2958.2011.07554.x

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5):955–964

Machado IM, Atsumi S (2012) Cyanobacterial biofuel production. J Biotechnol 162(1):50–56. doi:10.1016/j.jbiotec.2012.03.005

Markowitz VM, Szeto E, Palaniappan K, Grechkin Y, Chu K, Chen IM, Dubchak I, Anderson I, Lykidis A, Mavromatis K, Ivanova NN, Kyrpides NC (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. Nucleic Acids Res Database issue(1):D528–D533. doi:10.1093/nar/gkm846

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Anderson I, Lykidis A, Mavromatis K, Ivanova NN, Kyrpides NC (2010) The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res 38(Database issue):D382–D390. doi:10.1093/nar/gkp887

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40(Database issue):D115–D122. doi:10.1093/nar/gkr1044

Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Dusterhus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich EJ, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kotter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nutzmann HW, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, Yan X, Yim G, Yu F, Xie Y, Aigle B, Apel AK, Balibar CJ, Balskus EP, Barona-Gomez F, Bechthold A, Bode HB, Borriss R, Brady SF, Brakhage AA, Caffrey P, Cheng YQ, Clardy J, Cox RJ, De Mot R, Donadio S, Donia MS, van der Donk WA, Dorrestein PC, Doyle S, Driessen AJ, Ehling-Schulz M, Entian KD, Fischbach MA, Gerwick L, Gerwick WH, Gross H, Gust B, Hertweck C, Hofte M, Jensen SE, Ju J, Katz L, Kaysser L, Klassen JL, Keller NP, Kormanec J, Kuipers OP, Kuzuyama T, Kyrpides NC, Kwon HJ, Lautru S, Lavigne R, Lee CY, Linquan B, Liu X, Liu W, Luzhetskyy A, Mahmud T, Mast Y, Mendez C, Metsa-Ketela M, Micklefield J, Mitchell DA, Moore BS, Moreira LM, Muller R, Neilan BA, Nett M, Nielsen J, O'Gara F, Oikawa H, Osbourn A, Osburne MS, Ostash B, Payne SM, Pernodet JL, Petricek M, Piel J, Ploux O, Raaijmakers JM, Salas JA, Schmitt EK, Scott B, Seipke RF, Shen B, Sherman DH, Sivonen K, Smanski MJ, Sosio M, Stegmann E, Sussmuth RD, Tahlan K, Thomas CM, Tang Y, Truman AW, Viaud M, Walton JD, Walsh CT, Weber T, van Wezel GP, Wilkinson B, Willey JM, Wohlleben W, Wright GD, Ziemert N, Zhang C, Zotchev SB, Breitling R, Takano E, Glockner FO (2015) Minimum information about a biosynthetic gene cluster. Nat Chem Biol 11(9):625–631. doi:10.1038/nchembio.1890

Meeks JC, Elhai J (2002) Regulation of cellular differentiation in filamentous cyanobacteria in free-living and plant-associated symbiotic growth states. Microbiol Mol Biol Rev 66(1):94–121

Meeks JC, Elhai J, Thiel T, Potts M, Larimer F, Lamerdin J, Predki P, Atlas R (2001) An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. Photosynth Res 70(1):85–106. doi:10.1023/A:1013840025518

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25(10):1335–1337. doi:10.1093/bioinformatics/btp157

Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R (2014) The SEED and the rapid annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res 42(Database issue):D206–D214. doi:10.1093/nar/gkt1226

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35(21):7188–7196. doi:10.1093/nar/gkm864

Rice MC, Norton JM, Valois F, Bollmann A, Bottomley PJ, Klotz MG, Laanbroek HJ, Suwa Y, Stein LY, Sayavedra-Soto L, Woyke T, Shapiro N, Goodwin LA, Huntemann M, Clum A, Pillay M, Kyrpides N, Varghese N, Mikhailova N, Markowitz V, Palaniappan K, Ivanova N, Stamatis D, Reddy TB, Ngan CY, Daum C (2016) Complete genome of *Nitrosospira briensis* C-128, an ammonia-oxidizing bacterium from agricultural soil. Stand Genomic Sci 11:46. doi:10.1186/s40793-016-0168-4

Rippka R, Castenholz RW, Herdman M (2015) Cyanobacteria/subsection IV. Bergey's Man Syst Archaea Bacteria. doi:10.1002/9781118960608.gbm00450

Rosgaard L, de Porcellinis AJ, Jacobsen JH, Frigaard NU, Sakuragi Y (2012) Bioengineering of carbon fixation, biofuels, and biochemicals in cyanobacteria and plants. J Biotechnol 162(1):134–147. doi:10.1016/j.jbiotec.2012.05.006

Sandh G, Ramstrom M, Stensjo K (2014) Analysis of the early heterocyst Cys-proteome in the multicellular cyanobacterium *Nostoc punctiforme* reveals novel insights into the division of labor within diazotrophic filaments. BMC Genom 15:1064. doi:10.1186/1471-2164-15-1064

Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC (2013) Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. Proc Natl Acad Sci USA 110(5):1791–1796. doi:10.1073/pnas.1209927110

Shendure J, Lieberman Aiden E (2012) The expanding scope of DNA sequencing. Nat Biotechnol 30(11):1084–1094. doi:10.1038/nbt.2421

Stepanov VG, Tirumalai MR, Montazari S, Checinska A, Venkateswaran K, Fox GE (2016) *Bacillus pumilus* SAFR-032 genome revisited: sequence update and re-annotation. PLoS ONE 11(6):e0157331. doi:10.1371/journal.pone.0157331

Valli M, Tatto NE, Peymann A, Gruber C, Landes N, Ekker H, Thallinger GG, Mattanovich D, Gasser B, Graf AB (2016) Curation of the genome annotation of *Pichia pastoris* (*Komagataella phaffii*) CBS7435 from gene level to protein function. FEMS Yeast Res. doi:10.1093/femsyr/fow051

Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Muller R, Wohlleben W, Breitling R, Takano E, Medema MH (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 43(W1):W237–W243. doi:10.1093/nar/gkv437

Whitman WB (2015) Cyanobacteria/subsection IV/subsection IV.I In: Bergey's Man Syst Archaea Bacteria. doi:10.1002/9781118960608.gbm00451

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA 87(12):4576–4579

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13(5):329–342. doi:10.1038/nrg3174

Zarzycki J, Axen SD, Kinney JN, Kerfeld CA (2013) Cyanobacterial-based approaches to improving photosynthesis in plants. J Exp Bot 64(3):787–798. doi:10.1093/jxb/ers294