genomeA
Journals.ASM.org

Extreme Sensory Complexity Encoded in the 10-Megabase Draft Genome Sequence of the Chromatically Acclimating Cyanobacterium *Tolypothrix* sp. PCC 7601

Shaila Yerrapragada,[a]* Animesh Shukla,[b]* Kymberlie Hallsworth-Pepin,[c] Kwangmin Choi,[d]* Aye Wollam,[c] Sandra Clifton,[c] Xiang Qin,[a] Donna Muzny,[a] Sriram Raghuraman,[b,d]* Haleh Ashki,[b,d]* Akif Uzman,[e] Sarah K. Highlander,[a]* Bartlomiej G. Fryszczyn,[e]* George E. Fox,[f] Madhan R. Tirumalai,[f] Yamei Liu,[f]* Sun Kim,[d]* David M. Kehoe,[b] George M. Weinstock[a]*

Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA[a]; Department of Biology, Indiana University, Bloomington, Indiana, USA[b]; The Genome Institute, Washington University in St. Louis, St. Louis, Missouri, USA[c]; School of Informatics, Indiana University, Bloomington, Indiana, USA[d]; College of Sciences and Technology, University of Houston-Downtown, Houston, Texas, USA[e]; Department of Biology and Biochemistry, University of Houston, Houston, Texas, USA[f]

* Present address: Shaila Yerrapragada, College of Sciences and Technology, University of Houston-Downtown, Houston, Texas, USA; Animesh Shukla, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland, USA; Kwangmin Choi, Experimental Hematology and Cancer Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA; Sriram Raghuraman, Two Sigma Solutions, LLC, Houston, Texas, USA; Haleh Ashki, Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA; Sarah K. Highlander, The J. Craig Venter Institute, La Jolla, California, USA; Bartlomiej G. Fryszczyn, EMD Millipore, Billerica, Massachusetts, USA; Yamei Liu, National Research Centre of Veterinary Biological Engineering and Technology, Jiangsu Academy of Agricultural Sciences, Nanjing, Jiangsu, China; Sun Kim, Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea; George M. Weinstock, The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA.

***Tolypothrix* sp. PCC 7601 is a freshwater filamentous cyanobacterium with complex responses to environmental conditions. Here, we present its 9.96-Mbp draft genome sequence, containing 10,065 putative protein-coding sequences, including 305 predicted two-component system proteins and 27 putative phytochrome-class photoreceptors, the most such proteins in any sequenced genome.**

Received 11 March 2015 **Accepted** 30 March 2015 **Published** 7 May 2015

Citation Yerrapragada S, Shukla A, Hallsworth-Pepin K, Choi K, Wollam A, Clifton S, Qin X, Muzny D, Raghuraman S, Ashki H, Uzman A, Highlander SK, Fryszczyn BG, Fox GE, Tirumalai MR, Liu Y, Kim S, Kehoe DM, Weinstock GM. 2015. Extreme sensory complexity encoded in the 10-megabase draft genome sequence of the chromatically acclimating cyanobacterium *Tolypothrix* sp. PCC 7601. Genome Announc 3(3):e00355-15. doi:10.1128/genomeA.00355-15.

Address correspondence to George M. Weinstock, george.weinstock@jax.org, or David M. Kehoe, dkehoe@indiana.edu.

Cyanobacteria have a tremendous capacity to acclimate to changing environments (1). The filamentous cyanobacterium *Tolypothrix* sp. PCC 7601 is studied for its phenotypic plasticity in changing environments (2). Isolated from a Connecticut lake and named *Fremyella diplosiphon* (UTEX 481), it was added to the Pasteur Culture Collection as *Calothrix* sp. PCC 7601 and renamed *Tolypothrix* sp. PCC 7601. It is a model organism for studying the mechanism and regulation of chromatic acclimation, the reversible modification of photosynthetic light-harvesting antennae in response to changes in ambient light color, shifting the cell phenotype between red and blue-green (2–4).

Tolypothrix sp. PCC 7601 responds to many additional abiotic conditions. It acclimates to low sulfate conditions by producing antennae proteins depleted in sulfur-containing amino acids (5, 6). It has multiple developmental pathways, changes its cellular morphology and average filament length in different ambient light colors, and historically could reduce atmospheric nitrogen (7). Thus, it possesses an extensive repertoire of environmental responses. Its genome contains large numbers of genes encoding regulatory components, particularly two-component system proteins.

Shortened filament mutant SF33 (also called Fd33) was generated from *F. diplosiphon* (UTEX 481) (8) and used for genome sequencing. The sequence was generated using a full 3-kb paired-end Titanium 454 sequencing run, representing 46.6-fold genome coverage. A Newbler draft assembly was generated using Newbler version vMapAsmResearch-02/17/2010. The draft genome is 9,963,861 bp in length and contains 157 contigs (>139 bp in length) with a mean contig size of 63,464 bp and a maximum length of 955,511 bp. The draft was not further joined due to the presence of approximately 150 repetitive regions (probable endogenous transposable elements). The mean G+C genome content is 40.6%. Annotation and gene prediction used the TIGR Gene Indices gene annotation process (9). Coding sequences were predicted using GeneMark (10) and Glimmer3 (11). Intergenic regions not spanned by GeneMark and Glimmer3 were blasted against NCBI's nonredundant bacterial (NR) database. Loci were then defined by clustering predictions with the same reading frame. The best prediction at each locus was selected by evaluating all predictions against nonredundant bacterial, NR, and Pfam evidence (12) and resolving overlaps between adjacent coding genes. tRNA genes were determined using tRNAscan-SE (13) and noncoding RNA genes by RNAmmer (14) and Rfam (15). The final gene set was processed through KEGG (16), psortB (17), and Interproscan (18) to determine possible function. Gene product names were determined by BLAST Extend Repraze (http://sourceforge.net/projects/ber/).

A total of 10,065 coding sequences were predicted, including 305 two-component-system proteins and 27 phytochrome-class photoreceptors, the largest number of each of these groups of sensory proteins reported for any bacterial genome to date. These results suggest the presence of complex sensory and regulatory systems that are required for the extensive environmental responsiveness and phenotypic plasticity of this cyanobacterium.

These sequence data will be useful for elucidating the regulatory systems of prokaryotes with large, complex genomes.

Nucleotide sequence accession numbers. This whole-genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession number AGCR00000000. The version described in this paper is the first version, AGCR01000000.

ACKNOWLEDGMENTS

This work was supported by grant no. EF-0626927 to G.W. from the National Science Foundation.

We thank The Genome Institute, Washington University in St. Louis, specifically Bob Fulton, for data production and Chad Tomlinson, Sai Lek, and Wesley Warren for genome assembly. We also thank Brendon Bingwa and Catherine (Horton) Schaffer for their early efforts on this project.

REFERENCES

1. **Grossman AR, Schaefer M, Chiang G, Collier J.** 1994. The responses of cyanobacteria to environmental conditions: light and nutrients, p 641–675. *In* Bryant D (ed), The molecular biology of *Cyanobacteria*. Kluwer Academic Publishers, Dordrecht, Netherlands.
2. **Kehoe DM, Gutu A.** 2006. Responding to color: the regulation of complementary chromatic adaptation. Annu Rev Plant Biol **57:**127–150. http://dx.doi.org/10.1146/annurev.arplant.57.032905.105215.
3. **Tandeau de Marsac N.** 1983. Phycobilisomes and complementary chromatic adaptation in cyanobacteria. Bull Inst Pasteur **81:**201–254.
4. **Grossman AR.** 2003. A molecular understanding of complementary chromatic adaptation. Photosynth Res **76:**207–215. http://dx.doi.org/10.1023/A:1024907330878.
5. **Mazel D, Marlière P.** 1989. Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature **341:**245–248. http://dx.doi.org/10.1038/341245a0.
6. **Gutu A, Alvey RM, Bashour S, Zingg D, Kehoe DM.** 2011. Sulfate-driven elemental sparing is regulated at the transcriptional and posttranscriptional levels in a filamentous cyanobacterium. J Bacteriol **193:**1449–1460. http://dx.doi.org/10.1128/JB.00885-10.
7. **Tandeau de Marsac N, Houmard J.** 1993. Adaptation of cyanobacteria to environmental stimuli: new steps towards molecular mechanisms. FEMS Microbiol Rev **104:**119–189. http://dx.doi.org/10.1111/j.1574-6968.1993.tb05866.x.
8. **Cobley JG, Zerweck E, Reyes R, Mody A, Seludo-Unson JR, Jaeger H, Weerasuriya S, Navankasattusas S.** 1993. Construction of shuttle plasmids which can be efficiently mobilized from *Escherichia coli* into the chromatically adapting cyanobacterium, *Fremyella diplosiphon*. Plasmid **30:**90–105. http://dx.doi.org/10.1006/plas.1993.1037.
9. **Quackenbush J, Liang F, Holt I, Pertea G, Upton J.** 2000. The TIGR Gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res **28:**141–145. http://dx.doi.org/10.1093/nar/28.1.141.
10. **Borodovsky M, Mills R, Besemer J, Lomsadze A.** 2003. Prokaryotic gene prediction using GeneMark and GeneMark.hmm. Curr Protoc Bioinformatics. Unit 4.5. http://dx.doi.org/10.1002/0471250953.bi0405s01.
11. **Delcher AL, Harmon D, Kasif S, White O, Salzberg SL.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res **27:**4636–4641. http://dx.doi.org/10.1093/nar/27.23.4636.
12. **Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A.** 2008. The Pfam protein families database. Nucleic Acids Res **36:**D281–D288. http://dx.doi.org/10.1093/nar/gkm960.
13. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:**955–964. http://dx.doi.org/10.1093/nar/25.5.0955.
14. **Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res **35:**3100–3108. http://dx.doi.org/10.1093/nar/gkm160.
15. **Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A.** 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res **33:**D121–D124. http://dx.doi.org/10.1093/nar/gki081.
16. **Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M.** 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res **32:**D277–D280. http://dx.doi.org/10.1093/nar/gkh063.
17. **Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FS.** 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics **26:**1608–1615. http://dx.doi.org/10.1093/bioinformatics/btq249.
18. **Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R.** 2005. InterProScan: protein domains identifier. Nucleic Acids Res **33:**W116–W120. http://dx.doi.org/10.1093/nar/gki442.