

MIT Open Access Articles

ProPortal: a resource for integrated systems biology of Prochlorococcus and its phage

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Kelly, L. et al. "ProPortal: a Resource for Integrated Systems Biology of Prochlorococcus and Its Phage." *Nucleic Acids Research* 40.D1 (2011): D632–D640. Web.

As Published: <http://dx.doi.org/10.1093/nar/gkr1022>

Publisher: Oxford University Press

Persistent URL: <http://hdl.handle.net/1721.1/72444>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution Non-Commercial



ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage

Libusha Kelly¹, Katherine H. Huang¹, Huiming Ding^{1,2} and Sallie W. Chisholm^{1,2,*}

¹Department of Civil and Environmental Engineering and ²Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received August 15, 2011; Revised October 14, 2011; Accepted October 21, 2011

ABSTRACT

ProPortal (<http://proportal.mit.edu/>) is a database containing genomic, metagenomic, transcriptomic and field data for the marine cyanobacterium *Prochlorococcus*. Our goal is to provide a source of cross-referenced data across multiple scales of biological organization—from the genome to the ecosystem—embracing the full diversity of ecotypic variation within this microbial taxon, its sister group, *Synechococcus* and phage that infect them. The site currently contains the genomes of 13 *Prochlorococcus* strains, 11 *Synechococcus* strains and 28 cyanophage strains that infect one or both groups. Cyanobacterial and cyanophage genes are clustered into orthologous groups that can be accessed by keyword search or through a genome browser. Users can also identify orthologous gene clusters shared by cyanobacterial and cyanophage genomes. Gene expression data for *Prochlorococcus* ecotypes MED4 and MIT9313 allow users to identify genes that are up or downregulated in response to environmental stressors. In addition, the transcriptome in synchronized cells grown on a 24-h light–dark cycle reveals the choreography of gene expression in cells in a ‘natural’ state. Metagenomic sequences from the Global Ocean Survey from *Prochlorococcus*, *Synechococcus* and phage genomes are archived so users can examine the differences between populations from diverse habitats. Finally, an example of cyanobacterial population data from the field is included.

INTRODUCTION

The cyanobacterium *Prochlorococcus* is an ideal model system for understanding the ecology and evolution of microorganisms. The genomic and metabolic simplicity of these organisms—on average 2000 genes, with a very streamlined regulatory system (1–5)—coupled with their abundance and intra-species diversity, make them an ideal system for understanding microbial metabolism and the origins and consequences of biological diversity. Furthermore, as the smallest and most abundant photosynthetic cell in the oceans—an estimated 10^{27} cells globally—*Prochlorococcus* metabolic processes have far-reaching biogeochemical consequences (6,7). As a result, *Prochlorococcus* has attracted attention from research groups spanning many disciplines, which motivated us to create a comprehensive database to facilitate research on this organism.

Prochlorococcus is composed of several clades, referred to as ‘ecotypes’, defined by the intergenic transcribed spacer region between their 16S and 23S rRNA sequences (8–11). These ecotypes are both phylogenetically and physiologically distinct with respect to their light and temperature optima and tolerance ranges (3,12–16). The temporal and spatial (depth) distribution of six *Prochlorococcus* ecotype clades show robust patterns in abundance over a 5-year time-series both in the Pacific and Atlantic Oceans (17) that are consistent with what has been learned from physiological studies of isolated strains (13–15). There is also tremendous variability within ecotypes, particularly with regard to nutrient acquisition (18–22), susceptibility to predation or phage infection (23,24), and modes of interaction with other members of the community (25).

Comparative genomics of ecotypes within *Prochlorococcus*, (1–3) and its sister group *Synechococcus*

*To whom correspondence should be addressed. Tel: +1 617 253 1771; Fax: +1 617 324 0336; Email: chisholm@mit.edu
Present address:
Katherine H. Huang, The Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

(26–29) have provided a rich resource for analysis of genome evolution (3,30), and for recruitment of genome fragments from environmental metagenomic studies such as the Global Ocean Survey (GOS) (31,32). The combination of cultured genomes and field genomics-based datasets has illuminated how selective pressures such as light, temperature and limiting nutrients have shaped the diversity of these populations (17–20,26,29,31–36). *Prochlorococcus* always co-occurs with *Synechococcus* (but not vice versa) thus the prevalence of cyanophage that cross-infect them is not surprising. Because cyanophage carry many homologs of host genes in their genomes (33,37–39), an integrated database for host and phage allows identification of shared phage/host genes and facilitates interrogation into the evolution and exchange of these genes.

Prochlorococcus transcriptome analyses reflect the tight synchrony of the cells when grown on light–dark cycles and the partitioning of metabolism into daytime and nighttime activities (4). The expression levels of 80% of the genes in *Prochlorococcus* MED4 oscillate with a 24-h periodicity, and clustering of gene expression into synchronized groups lends insight into the coordination of metabolic activities in the cell in its ‘natural’ state. In addition, there is data on the transcriptional response of *Prochlorococcus* to environmental perturbations such as phosphate (18), nitrogen (40), iron starvation (22), light shifts (41) and phage infection (42).

The genomic and metabolic capabilities of these cells are intimately linked to their population distributions and dynamics in the oceans (17,43,44). There are growing datasets in which *Prochlorococcus* [*sensu* (9)] and *Synechococcus* ecotype populations are identified in depth profiles (8,35), time series (17,56) and ocean transects (16,45). Population analyses are instrumental in identifying key environmental variables that have resulted in niche differentiation among *Prochlorococcus* (16), and can be used to guide physiological, metabolic and genomic analyses.

The vast quantity and complexity of genome and transcriptome data for microorganisms has become increasingly difficult to effectively describe solely through conventional publications and databases. This is especially true when dealing with model organisms where consistency and cross-referencing of annotations among growing numbers of strains and datasets is essential. It is also becoming clear that the interpretation of genomic and transcriptomic data is greatly enriched when framed within an environmental context—i.e. meta-data from the sites of isolation of the strains or metagenomes being studied (2,19,20,32). Because the marine habitat of *Prochlorococcus* contains smooth and measurable gradients, this type of meta-data is more accessible than that of more complex habitats. We created ProPortal to allow expedient access to data that has been curated and to allow direct correspondence between data of different types placed in the context of habitat. It was built to allow biologists to access the data *via* a web interface, and computational biologists or database developers to access the data *via* SQL statements and Python modules. While there are several excellent resources available to

explore and compare microbial genomes—e.g. CyanoBase (46), IMG (47) and MicrobesOnline (48), the unique strength of ProPortal is its comprehensive nature—including genomic, transcriptomic, metagenomic and population data from both domesticated and wild populations of cyanobacteria and phage.

IMPLEMENTATION

ProPortal is built and maintained with Django (<http://www.djangoproject.com/>), a high-level, Python-based web framework. Data visualization is implemented with the Google Chart Tools API (<http://code.google.com/apis/chart/>), Google Map Tools API (<http://code.google.com/apis/maps/>), Bioheatmap, a package in the Visualizations and Controls toolkit developed at the Institute of Systems Biology (<http://code.google.com/p/systemsbiology-visualizations/>) and GraphViz (49). The back end of ProPortal is a MySQL database server (version 4.1.22) compiled for redhat-linux-gnu. While the data presented in ProPortal is specific to *Prochlorococcus*, *Synechococcus* and cyanophage, the database framework is generalizable for any group working on microbial and phage genomes for which extensive metadata is available. We have provided access to our database schema (<http://proportal.mit.edu/schema/>) should other groups want to replicate the whole database or portions of it to maintain their own data.

DATA TYPES AND ACCESS

ProPortal is divided into four main sections for data access: Genomes, Microarrays, Population Dynamics and Metagenomes (Figure 1). The foundation of the database is the genomes and orthologous gene clusters, and we therefore describe these features in more detail below. The underlying database structure enables a user to ask questions that integrate different data types.

A representative page layout for *Prochlorococcus* genomes is shown in Figure 1. It includes a global map that shows the site of isolation for each cultured strain with a sequenced genome. Each teardrop is clickable, revealing the location and depth from which the strain was collected, as well as the ProPortal link to the genome. A navigation panel ‘Browse Site Data’ and a search panel ‘Search Data’ are available from most pages. The four sections under Browse Site Data are expandable, and users can always navigate back to the home page by clicking on the ProPortal icon or the tabs across the top of the screen. Users can do keyword searches for genes or orthologous gene clusters (termed ‘CyCOGs’ for cyanobacterial clusters and ‘PhCOGs’ for phage clusters) in the Search Data box. For example, a search for the protein PMM0370 would return the link for a single gene, in this case the putative substrate binding protein for the cyanate ABC transporter in *Prochlorococcus* MED4, whereas searching for ‘cyanate’ would return links to 17 genes from both *Prochlorococcus* and *Synechococcus*.

Prochlorococcus

The screenshot displays the ProPortal interface for *Prochlorococcus*. At the top, navigation links include Home, Press, Download, Chisholm Lab, and Contact Us. The main heading is 'Prochlorococcus'. Below this is a world map showing the distribution of *Prochlorococcus* strains. A call-out box for strain AS9601 provides details: Location: 19°N, 67°E, Depth: 50m, and a ProPortal link. The map includes a search bar and a dropdown menu set to 'Prochloro COGs'. Below the map are two interactive boxes: 'Browse Site Data' with expandable links for Genomes, Population Dynamics, Microarrays, and Metagenomes; and 'Search Data' with a search input field containing 'cyanate', a dropdown menu set to 'Prochloro COGs', and a 'submit' button.

Figure 1. Features of the ProPortal entry page for *Prochlorococcus* genomes. Blue markers on the map indicate the sites of isolation for *Prochlorococcus* strains with complete genomes. Each marker can be expanded to reveal a 'call-out' that shows the strain name, isolation coordinates and depth, as shown for the Arabian Sea isolate AS9601. The Browse Site Data and Search Data boxes shown under the map are available on most ProPortal pages for navigation, as are the tabs across the top of the page. The four Browse Site Data links are expandable; the Genomes expansion is shown here. Keyword searches can be used to find specific orthologous gene clusters (e.g. 'CyCOG2748') or clusters with a keyword in the annotation (e.g. 'cyanate'). Likewise, proteins can be searched using identifiers (i.e. PMM0370) or annotation keywords (e.g. 'cyanate').

Cyanobacterial and cyanophage genomes and orthologous gene clusters

Genome annotation requires updating as new genomes become available. Interpretation of experimental data (e.g. gene expression data) relies heavily on genome annotations, and we therefore created, and maintain and update orthologous gene cluster information exclusively for *Prochlorococcus*, *Synechococcus* and phage in ProPortal to give researchers access to the latest annotations. These annotations supplement those available through sites like IMG (47), Genbank (50) and Cyanobase (46) by enabling phylogenetic profiling—i.e. examining the distribution of orthologous gene clusters among genomes of strains belonging to different phylogenetic groups as determined using classical molecular taxonomic approaches. Users can search for genetic features common to phylogenetic groups of interest, for example high-light adapted *Prochlorococcus*. This information can be particularly useful for unannotated genes where distributions of genes among phylogenetic groups with known environmental distributions could provide clues to the relative importance of genes for different organisms. The information is available on the ProPortal cluster page and does not require BLASTing sequences to locate orthologs.

ProPortal currently contains 13 *Prochlorococcus*, 11 *Synechococcus* and 28 cyanophage genomes. Protein coding genes in each genome are linked to orthologous gene clusters, defined using both external databases like NCBI COGs (50) and with internal CyCOG designations for cyanobacteria and PhCOG designations for phage (3,37). There are many genes specific to *Prochlorococcus*, *Synechococcus* and cyanophage that, while conserved among these organisms, do not have external annotations. Each group of genomes has genes denoted 'core'; that is shared by all organisms, and genes that are 'flexible', or shared among some subset of genomes (3). Because phage and host do not share 'orthologs' that arise from vertical descent, we process phage and host sequences separately for ortholog analysis.

Pairwise orthologous relationships were first mapped in all *Prochlorococcus*, *Synechococcus* and cyanophage genomes as described previously (3,37). Briefly, orthologous genes were assigned using reciprocal best blastp scores ($e\text{-value} \leq 1e-5$) where the sequence identity was at least 35% and the sequence alignment length was at least 75% of the protein length of the shorter protein of the two compared. These criteria were established to avoid identifying only shorter conserved domains within larger proteins. Clusters of orthologous genes were built by transitively clustering orthologs together as follows:

if gene A and B are orthologs and gene B and C are orthologs, then genes A, B and C are clustered into an orthologous group. Orthologous members are denoted with the tag 'ortholog' under 'Cluster Evidence' on each gene cluster page.

To find divergent orthologs missed by the initial BLAST-based approach we next built Hidden Markov Model (HMM) profiles (51) by aligning each gene in an orthologous cluster using MUSCLE version 3.7 (52) with default parameters. We used the hmmbuild program from HMMER version 2.3.2 (<http://hmmer.janelia.org/>) to build HMM profiles from the resulting alignments. The program hmmsearch (also from HMMER version 2.3.2) was then used to search remaining, non-clustered protein sequences against these in-house HMM profiles. Singletons with homology ($e\text{-value} \leq 1e^{-5}$) to an HMM profile were added to that cluster. Genes added to the clusters by this method are denoted by the tag 'hmmscan' under 'Cluster Evidence' on each gene cluster page. All remaining singleton genes are then assigned to their own gene cluster. Importantly, some orthologous gene clusters have representatives in both cyanobacterial and cyanophage genomes. These shared phage/host genes are indicated with an 'associated gene clusters tag' on the gene cluster page. For example, phage and host versions of the phosphate transporter subunit *pstS* are represented by phage gene cluster PhCOG174 and the cyanobacterial gene cluster CyCOG3129.

The general gene-clustering method described above works well for single-copy cyanobacterial core genes. However, genes that have undergone loss, transfer, or duplication events yield more complicated clusters and would not be distinguished by this simple method. Furthermore, the high gene diversity in phage genomes also inhibits the construction of cohesive orthologous clusters. We therefore provide visualized reciprocal best-hit networks for each cluster to help users detect these potential complications. To study these events more carefully, we will build phylogenetic trees for orthologous gene clusters in a future update of ProPortal.

Some features for the analysis of orthologous gene clusters are illustrated in Figure 2. Each gene cluster has a page combining available DNA, protein, phylogenetic and expression data for a representative gene from that cluster. For example, from the annotation information for the cluster for the putative high light inducible (*hli*) gene *hli16*, part of the orthologous gene cluster CyCOG3918, a user can observe that this gene also has associated phage orthologous gene clusters (PhCOGs212, 213 and 494) (Figure 2A). Gene expression data over the day/night cycle, measured using custom Affymetrix arrays, is shown for the four representative *Prochlorococcus* MED4 genes representing the gene cluster (Figure 2B). ProPortal indicates that CyCOG3918 is present only in *Prochlorococcus* genomes and sometimes present in multiple copies in individual genomes (Figure 2C). Users can also view DNA and protein sequences for all members of the cluster, or alternatively pages for an individual gene can be accessed using the clickable network maps on each cluster page; genes are shown as light green circles and reciprocal best-hit relationships are indicated by lines

connecting the genes (Figure 2D). By clicking on a gene in the cluster's gene table (not shown) users can navigate to genes upstream and downstream of this cluster in an individual genome using the genome browser; as an example we show the *hli16* gene PMED4_09101 (Figure 2E).

ProPortal can also be used to report differences in expression of multiple *hli* gene families over the 24-h light-dark cycle. To do this, a user can search for 'hli' using the Prochloro COGs option in the search box, and find 16 different orthologous gene clusters. For example, the three *Prochlorococcus* MED4 *hli* genes represented by cluster CyCOG3919 all have similar expression patterns over the diel cycle. In contrast, of the four *Prochlorococcus* MED4 *hli* genes represented by the gene cluster CyCOG3918 (Figure 2A and B), one gene (PMED4_15861) has the opposite pattern of the other three genes. Individual genes are also linked to experiments identifying changes in gene expression in *Prochlorococcus* under perturbation conditions, as described in detail below.

Transcriptional response to stress

Microarray-based readouts of transcript levels in *Prochlorococcus* strains MED4 and MIT9313 exposed to various phosphate (18), nitrogen (40), iron (22) and ambient light conditions (41) have been integrated into ProPortal. Transcript data for changing O₂/CO₂ ratios (53) will soon be added. Datasets from other groups describing transcriptional response in *Synechococcus* (54,55) are not currently integrated, but could be in future releases. Under the microarray section users can choose a time point after the onset of a perturbation, for example phosphate starvation, and examine genes that are up and downregulated. After 12h of phosphate starvation, for example, the user would find that the most upregulated gene in *Prochlorococcus* MED4 is PMM0705 (PMED4_07791), encoding *phoB*, the response regulator of a conserved two-component phosphate sensing system. From ProPortal's page for PMED4_07791, the user can then easily identify the position of the gene in *Prochlorococcus* MED4.

Additionally, information on neighboring genes can be retrieved to examine genomic conservation surrounding a gene of interest using the genome browser (such as shown in Figure 2E for the *hli16* gene). For example, the two genes flanking *phoB* in *Prochlorococcus* MED4 are the *phoR* two-component sensor histidine kinase for phosphate sensing (PMED4_07801), found in 17 cyanobacterial genomes in ProPortal, and a putative potassium channel (PMED4_07781) found in all cyanobacterial genomes. A user could then explore the relative locations of these genes in other genomes using the browser for any genome of interest. Alternatively, a user could select the probeset representing this gene (MED4_ARR_0701_x_at) and find its expression patterns under conditions of nitrogen availability or exposure to different light levels. Finally, from this page users can also find up- and downregulated genes in the low light-adapted *Prochlorococcus* strain MIT9313 when available.

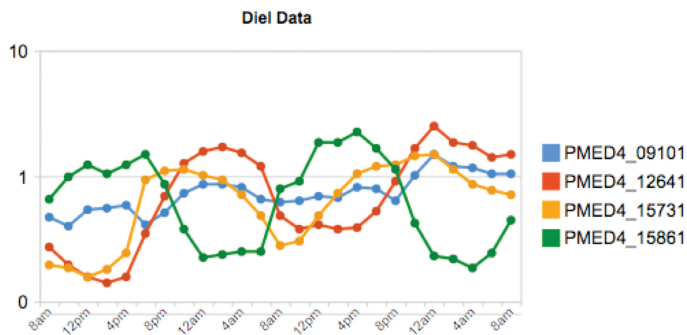
A CyCog 3918: possible high light inducible protein

[Protein Fasta](#) | [Transcript Fasta](#)

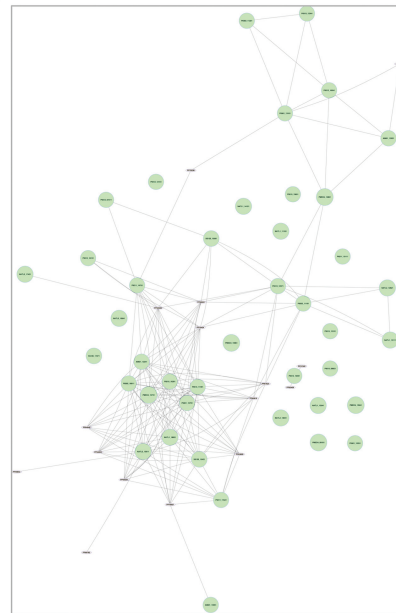
CyCog3918 Annotation Info

Gene name: None
 Cluster size: 42
 Synonyms:
 Associated gene clusters: [PhCOG212](#), [PhCOG213](#), [PhCOG494](#),

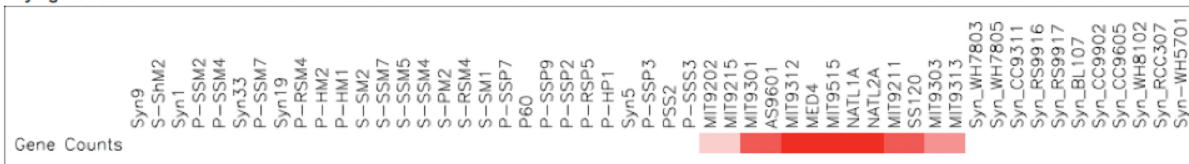
B



D



C Phylogenetic Profile



E

Protein Info: PMED4_09101 possible high light inducible protein (hli16)

Location: 773693 - 773800 (-) on scaffold: NC_005072

Protein Length: 35 aa

MED4 Mini Browser

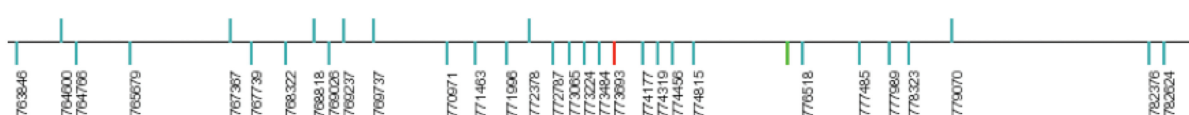


Figure 2. The orthologous gene cluster page for CyCOG3918, the putative high light-inducible protein *hli16*. (A) Annotations, including associated phage orthologous gene clusters. (B) Expression of four *Prochlorococcus* MED4 versions of this gene over the day/night cycle. (C) The phylogenetic profile indicates that this gene is found only in *Prochlorococcus* genomes, and that some genomes have multiple copies of this gene (red squares; color intensity proportional to number of copies). (D) An interactive network map of all genes (light green circles) in this cluster identifies the reciprocal best-hit relationships among them. Links to individual genes are presented in table format on the cluster page but are omitted from this figure for clarity. (E) A portion of the gene page for PMED4_09101 (*hli16*), a putative high-light inducible protein. The genome browser for MED4 is shown, with this gene highlighted in red.

Population dynamics in the wild

The Population Dynamics link in ProPortal is in its infancy; as a starting point we have incorporated the pioneering time series data from DuRand *et al.* (56), which describes the dynamics of *Synechococcus* and *Prochlorococcus* populations as a function of depth and time at the Bermuda Atlantic Time-series Study (BATS) station in the Sargasso Sea from 1990 to 1994. This data is available for download, but can also be visualized for different time intervals and at depths selected by the user. We are in the process of integrating a more recent 5-year time series of *Prochlorococcus* ecotypes comparing depth profiles and abundance in both the Pacific and Atlantic Oceans and associated ancillary data (17), and others will soon follow.

Metagenomics

Because of their global distribution, *Prochlorococcus*, *Synechococcus* and cyanophage are well represented in many metagenomic datasets (20,36,57,58). ProPortal currently integrates extensive metagenomic data available from the GOS database (32), a valuable resource for understanding the ecology, diversity and biogeography of these groups. For example, the identification of *Prochlorococcus*-like reads carrying nitrate assimilation genes in the GOS database expanded the metabolic repertoire of *Prochlorococcus* in ways not evident from cultured strains (34). More generally, the combination of a vast metagenomic database and genomes from diverse cultured strains allows one to explore the biogeography of different genomic variants of the group.

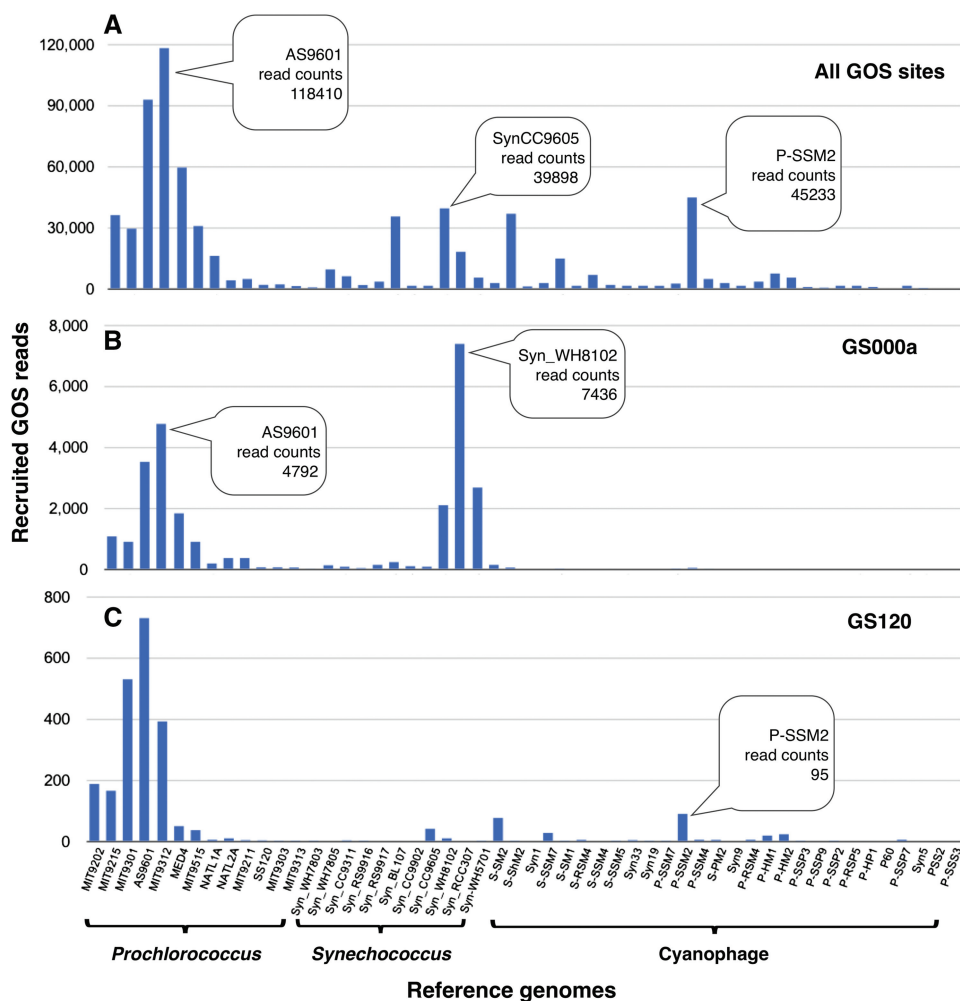


Figure 3. Recruited GOS metagenomic DNA fragments (un-normalized ‘raw reads’) to cultured *Prochlorococcus*, *Synechococcus* and cyanophage genomes in ProPortal. Approximately 10 million reads from the GOS database (32) were recruited to the 52 genomes in ProPortal. Recruited reads from (A) all GOS sites, (B) a Sargasso Sea site (GS000a) and (C) an Indian Ocean site (GS120). Each bar in the plot is clickable in ProPortal, revealing the number of reads recruited to a particular reference genome. For example, the call-outs in (A) show the read counts recruiting to the genomes of the Arabian Sea *Prochlorococcus* strain AS9601, *Synechococcus* strain CC9605 and the cyanophage P-SSM2.

To this end, approximately 10 million GOS reads downloaded from the CAMERA database (59) were recruited to 52 *Prochlorococcus*, *Synechococcus* and cyanophage genomes [*sensu* (32)] using the following parameters: `-p blastn -r 5 -q -4 -e 1e-4 -z 10000000000 -F ‘m L’ -X 150 -U T -m 9`. This initial BLAST recruited 3 315 149 GOS reads (out of 9 893 120 total) to *Prochlorococcus*, *Synechococcus* and cyanophage genomes. After filtering reads by requiring an alignment length $\geq 50\%$ of the read length, 1 972 428 reads remained. Most GOS reads also have an associated paired-end read, which enables greater confidence in organism assignment because both reads are from the same cloned sequence. After examining paired-ends, 728 441 reads remained. It was not required that paired-ends align best to the same genome, rather they were required to have a best hit to the same organism group, that is, either *Prochlorococcus*, *Synechococcus* or cyanophage.

The GOS recruitment data can be explored in several ways. The bar graphs provided on the ProPortal website

(Figure 3) report the number of reads assigned to homologous regions in currently available genomes. Reporting raw read counts is intended to answer simple questions such as ‘Is this genome or genomic region represented in GOS, and if so, at what sites?’ Users can download this data and should then normalize the read counts to both genome size and the total number of reads for each site if they want to compare the abundances of different genomes at different sites, for example. Analysis of the complete GOS dataset with the ProPortal pipeline indicates that the most abundant raw reads (un-normalized) within *Prochlorococcus*, *Synechococcus* and cyanophage are those with highest sequence similarity to strains AS9601, CC9605 and P-SSM2 respectively. *Prochlorococcus* reads are the most abundant raw reads within this triad overall (Figure 3A). One can also query individual sites in GOS and see how these relative abundances change at different sites in the ocean. For example, at site GS000a in the Sargasso Sea, reads most similar to *Synechococcus* WH8102 are most abundant (Figure 3B).

In contrast, in the Indian Ocean sites GS120 and GS121 *Prochlorococcus* AS9601-like sequences are most abundant and *Prochlorococcus* genomes, as a collection, recruit far more reads than do *Synechococcus* (Figure 3C).

In the future, users will be able to download all reads with hits to orthologous gene clusters directly from the orthologous gene cluster pages. This will enable users to more easily track how gene abundances change across different GOS sites and tie the abundances of specific genes or pathways to environmental metadata. At the moment, users can, however, query a region of a genome of interest to locate all reads recruited to this region and their associated site metadata. This is useful, for example, to identify how conserved particular sets of syntenic genes are in the wild.

FUTURE DEVELOPMENT

Aside from the improvements highlighted above, incorporating new genomes and metagenomic datasets is our first priority; we anticipate adding more than 30 new genomes from cultured strains in the next year in addition to approximately 100 single cell genomes from wild *Prochlorococcus*. As part of updates to the orthologous gene clusters, designations of 'core' and 'flexible' will be added to each cluster to aid in phylogenetic analysis. As RNASeq becomes a standard protocol for transcriptome analysis, this data will be incorporated into ProPortal to allow users to compare expression patterns among diverse strains. Population distribution data in the oceans will be expanded greatly, and an entry point for physiological data—temperature and light optima for growth—will be added. Finally, a future goal is to enable automated community submissions to the resource.

ACKNOWLEDGEMENTS

The authors would like to thank past and present members of the Chisholm lab for their contributions to the site, Mark Breidenbach (Berkeley), Jessica Weidemier Thompson, Steven Biller and Paul Berube (MIT) for helpful suggestions on the manuscript, Stuart Levine and the MIT BioMicro Center for their support, and the MIT Darwin Project for computational facilities.

FUNDING

The National Science Foundation Biological Oceanography Section and the National Science Foundation Center for Microbial Oceanography Research and Education (C-MORE) (grant numbers OCE-0425602 and EF0424599), US Department of Energy-GTL (grant numbers DE-FG02-02ER63445, DE-FG02-08ER64516 and DE-FG02-07ER64506) to S.W.C.; the Gordon and Betty Moore Foundation (Grant Award Letter #495.01). Funding for open access charge: The Gordon and Betty Moore Foundation.

Conflict of interest statement. None declared.

REFERENCES

- Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V., Duprat, S., Galperin, M.Y., Koonin, E.V., Le Gall, F. *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl Acad. Sci. USA*, **100**, 10020–10025.
- Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., Hess, W.R. *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, **424**, 1042–1047.
- Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferreira, S., Johnson, J. *et al.* (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet.*, **3**, e231.
- Zinser, E.R., Lindell, D., Johnson, Z.I., Futschik, M.E., Steglich, C., Coleman, M.L., Wright, M.A., Rector, T., Steen, R., McNulty, N. *et al.* (2009) Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. *PLoS ONE*, **4**, e5135.
- Garcia-Fernandez, J.M., de Marsac, N.T. and Diez, J. (2004) Streamlined regulation and gene loss as adaptive mechanisms in *Prochlorococcus* for optimized nitrogen utilization in oligotrophic environments. *Microbiol. Mol. Biol. Rev.*, **68**, 630–638.
- Partensky, F., Hess, W.R. and Vaultot, D. (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.*, **63**, 106–127.
- Partensky, F. and Garczarek, L. (2010) *Prochlorococcus*: advantages and limits of minimalism. *Ann. Rev. Mar. Sci.*, **2**, 305–331.
- Ahlgren, N.A., Rocap, G. and Chisholm, S.W. (2006) Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. *Environ. Microbiol.*, **8**, 441–454.
- Rocap, G., Distel, D.L., Waterbury, J.B. and Chisholm, S.W. (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.*, **68**, 1180–1191.
- Lavin, P., Gonzalez, B., Santibanez, J.F., Scanlan, D.J. and Ulloa, O. (2010) Novel lineages of *Prochlorococcus* thrive within the oxygen minimum zone of the eastern tropical South Pacific. *Environ. Microbiol. Rep.*, **2**, 728–738.
- West, N.J., Lebaron, P., Strutton, P.G. and Suzuki, M.T. (2011) A novel clade of *Prochlorococcus* found in high nutrient low chlorophyll waters in the South and Equatorial Pacific Ocean. *ISME J.*, **5**, 933–944.
- Follows, M.J., Dutkiewicz, S., Grant, S. and Chisholm, S.W. (2007) Emergent biogeography of microbial communities in a model ocean. *Science*, **315**, 1843–1846.
- Moore, L.R., Rocap, G. and Chisholm, S.W. (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature*, **393**, 464–467.
- Moore, L.R. and Chisholm, S.W. (1999) Photophysiology of the marine cyanobacterium *Prochlorococcus*: Ecotypic differences among cultured isolates. *Limnol. Oceanogr.*, **44**, 628–638.
- Zinser, E.R., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D. and Chisholm, S.W. (2007) Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol. Oceanogr.*, **52**, 2205–2220.
- Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M. and Chisholm, S.W. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science*, **311**, 1737–1740.
- Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R. and Chisholm, S.W. (2010) Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J.*, **4**, 1252–1264.
- Martiny, A.C., Coleman, M.L. and Chisholm, S.W. (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc. Natl Acad. Sci. USA*, **103**, 12552–12557.
- Martiny, A.C., Huang, Y. and Li, W. (2009) Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ. Microbiol.*, **11**, 1340–1347.

20. Coleman, M.L. and Chisholm, S.W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc. Natl Acad. Sci. USA*, **107**, 18634–18639.
21. Moore, L.R., Post, A.F., Roco, G. and Chisholm, S.W. (2002) Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol. Oceanogr.*, **47**, 989–996.
22. Thompson, A.W., Huang, K., Saito, M.A. and Chisholm, S.W. (2011) Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. *ISME J.*, **5**, 1580–1594.
23. Sullivan, M.B., Waterbury, J.B. and Chisholm, S.W. (2003) Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature*, **424**, 1047–1051.
24. Avrani, S., Wurtzel, O., Sharon, I., Sorek, R. and Lindell, D. (2011) Genomic island variability facilitates *Prochlorococcus*-virus coexistence. *Nature*, **474**, 604–608.
25. Sher, D., Thompson, J.W., Kashtan, N., Croal, L. and Chisholm, S.W. (2011) Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. *ISME J.*, **5**, 1125–1132.
26. Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., Paulsen, I.T., de Marsac, N.T., Wincker, P., Dossat, C. *et al.* (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol.*, **9**, R90.
27. Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E.E., McCarren, J. *et al.* (2003) The genome of a motile marine *Synechococcus*. *Nature*, **424**, 1037–1042.
28. Palenik, B., Ren, Q., Dupont, C.L., Myers, G.S., Heidelberg, J.F., Badger, J.H., Madupu, R., Nelson, W.C., Brinkac, L.M., Dodson, R.J. *et al.* (2006) Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proc. Natl Acad. Sci. USA*, **103**, 13555–13559.
29. Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., Post, A.F., Hagemann, M., Paulsen, I. and Partensky, F. (2009) Ecological genomics of marine picocyanobacteria. *Microbiol. Mol. Biol. Rev.*, **73**, 249–299.
30. Luo, H., Friedman, R., Tang, J. and Hughes, A.L. (2011) Genome reduction by deletion of paralogs in the marine Cyanobacterium *Prochlorococcus*. *Mol. Biol. Evol.*, **28**, 2751–2760.
31. Yooseph, S., Nealson, K.H., Rusch, D.B., McCrow, J.P., Dupont, C.L., Kim, M., Johnson, J., Montgomery, R., Ferreira, S., Beeson, K. *et al.* (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, **468**, 60–66.
32. Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.
33. Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., Andrews-Pfannkoch, C., Fadrosch, D., Miller, C.S., Sutton, G. *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, **3**, e1456.
34. Martiny, A.C., Kathuria, S. and Berube, P.M. (2009) Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc. Natl Acad. Sci. USA*, **106**, 10787–10792.
35. West, N.J. and Scanlan, D.J. (1999) Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl. Environ. Microbiol.*, **65**, 2585–2591.
36. Rusch, D.B., Martiny, A.C., Dupont, C.L., Halpern, A.L. and Venter, J.C. (2010) Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc. Natl Acad. Sci. USA*, **107**, 16184–16189.
37. Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigle, P.R., Defrancesco, A.S., Kern, S.E., Thompson, L.R., Young, S. *et al.* (2010) Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.*, **12**, 3035–3056.
38. Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F. and Chisholm, S.W. (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl Acad. Sci. USA*, **101**, 11013–11018.
39. Millard, A.D., Zwirgmaier, K., Downey, M.J., Mann, N.H. and Scanlan, D.J. (2009) Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. *Environ. Microbiol.*, **11**, 2370–2387.
40. Tolonen, A.C., Aach, J., Lindell, D., Johnson, Z.I., Rector, T., Steen, R., Church, G.M. and Chisholm, S.W. (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol. Syst. Biol.*, **2**, 53.
41. Steglich, C., Futschik, M., Rector, T., Steen, R. and Chisholm, S.W. (2006) Genome-wide analysis of light sensing in *Prochlorococcus*. *J. Bacteriol.*, **188**, 7796–7806.
42. Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T., Kettler, G., Sullivan, M.B., Steen, R., Hess, W.R. *et al.* (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature*, **449**, 83–86.
43. Bouman, H.A., Ulloa, O., Scanlan, D.J., Zwirgmaier, K., Li, W.K., Platt, T., Stuart, V., Barlow, R., Leth, O., Clementson, L. *et al.* (2006) Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. *Science*, **312**, 918–921.
44. Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J. and Chisholm, S.W. (2006) *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl Environ. Microbiol.*, **72**, 723–732.
45. Zwirgmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F., Massana, R., Ulloa, O. and Scanlan, D.J. (2008) Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.*, **10**, 147–161.
46. Nakao, M., Okamoto, S., Kohara, M., Fujishiro, T., Fujisawa, T., Sato, S., Tabata, S., Kaneko, T. and Nakamura, Y. (2010) CyanoBase: the cyanobacteria genome database update 2010. *Nucleic Acids Res.*, **38**, D379–D381.
47. Markowitz, V.M., Szeto, E., Palaniappan, K., Grechkin, Y., Chu, K., Chen, I.M., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K. *et al.* (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.*, **36**, D528–D533.
48. Dehal, P.S., Joachimiak, M.P., Price, M.N., Bates, J.T., Baumohl, J.K., Chivian, D., Friedland, G.D., Huang, K.H., Keller, K., Novichkov, P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
49. Ellson, J., Gansner, E.R., Koutsofos, E., North, S.C. and Woodhull, G. (2003) Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools. In: Junger, M. and Mutzel, P. (eds), *Graph Drawing Software*. Springer, Berlin, Heidelberg, NY, pp. 127–148.
50. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
51. Eddy, S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
52. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
53. Bagby, S. (2009) Life in a drop of water. *Ph.D. Thesis*. Massachusetts Institute of Technology Department of Biology.
54. Tetu, S.G., Brahamsha, B., Johnson, D.A., Tai, V., Phillipy, K., Palenik, B. and Paulsen, I.T. (2009) Microarray analysis of phosphate regulation in the marine cyanobacterium *Synechococcus* sp. WH8102. *ISME J.*, **3**, 835–849.

55. Tai,V., Paulsen,I.T., Phillippy,K., Johnson,D.A. and Palenik,B. (2009) Whole-genome microarray analyses of *Synechococcus-Vibrio* interactions. *Environ. Microbiol.*, **11**, 2698–2709.
56. DuRand,M.D., Olson,R.J. and Chisholm,S.W. (2001) Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Res. Part II: Topical Studies Oceanogr.*, **48**, 1983–2003.
57. Ghai,R., Martin-Cuadrado,A.B., Molto,A.G., Heredia,I.G., Cabrera,R., Martin,J., Verdu,M., Deschamps,P., Moreira,D., Lopez-Garcia,P. *et al.* (2010) Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.*, **4**, 1154–1166.
58. DeLong,E.F., Preston,C.M., Mincer,T., Rich,V., Hallam,S.J., Frigaard,N.U., Martinez,A., Sullivan,M.B., Edwards,R., Brito,B.R. *et al.* (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, **311**, 496–503.
59. Sun,S., Chen,J., Li,W., Altintas,I., Lin,A., Peltier,S., Stocks,K., Allen,E.E., Ellisman,M., Grethe,J. *et al.* (2011) Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.*, **39**, D546–D551.