Title
Computational tools for the analysis of high-throughput genome-scale sequence data

Author
Lopez, David Adrian

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational tools for the analysis of high-throughput genome-scale sequence data

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular Biology

by

David Adrian Lopez, Jr.

2016

ABSTRACT OF THE DISSERTATION

Computational tools for the analysis of high-throughput genome-scale sequence data

by

David Adrian Lopez, Jr.

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2016

Professor Matteo Pellegrini, Chair

As high-throughput sequence data becomes increasingly used in a variety of fields, there is a growing need for computational tools that facilitate analyzing and interpreting the sequence data to extract biological meaning. To date, several computational tools have been developed to analyze raw and processed sequence data in a number of contexts. However, many of these tools primarily focus on well-studied, reference organisms, and in some cases, such as the visualization of molecular signatures in expression data, there is a scarcity or complete absence of tools. Furthermore, the compendium of genome-scale data in publicly accessible databases can be leveraged to inform new studies. The focus of this dissertation is the development of computational tools and methods to analyze high-throughput genome-scale sequence data, as well as applications in mammalian, algal, and bacterial systems. Chapter 1 introduces the challenges of analyzing high-throughput sequence data. Chapter 2 presents the Signature Visualization Tool (SaVanT), a framework to visualize molecular signatures in user-

generated expression data on a sample-by-sample basis. This chapter demonstrates that SaVanT can use immune activation signatures to distinguish patients with different types of acute infections (influenza A and bacterial pneumonia), and determine the primary cell types underlying different leukemias (acute myeloid and acute lymphoblastic) and skin disorders. Chapter 3 describes the Algal Functional Annotation Tool, which biologically interprets large gene lists, such as those derived from differential expression experiments. This tool integrates data from several pathway, ontology, and protein domain databases and performs enrichment testing on gene lists for several algal genomes. Chapter 4 describes a survey of the *Chlamydomonas reinhardtii* transcriptome and methylome across various stages of its sexual life cycle. This chapter discusses the identification and function of 361 gamete-specific and 627 zygote-specific genes, the first base-resolution methylation map of *C. reinhardtii*, and the changes in chloroplast methylation throughout key stages of its life cycle. Chapter 5 presents a comparative genomics approach to identifying previously uncharacterized bacterial microcompartment (BMC) proteins. Based on genomic proximity of genes in 131 fully-sequenced bacterial genomes, this chapter describes new putative microcompartments and their function.

The dissertation of David Adrian Lopez, Jr. is approved.

Jason Ernst

Eleazar Eskin

Sabeeha Merchant

Todd O. Yeates

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2016

DEDICATION

This thesis is dedicated to my family, for their

constant love, support, and encouragement without which

this work would not have been possible

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Firstly, I thank my mentor, Matteo Pellegrini, for igniting the spark and for providing an environment of academic freedom, trust, and support to carry out my research projects. His approaches to research have been inspirational and have taught me how to frame and pursue scientific questions.

A special thanks to my committee members – Jason Ernst, Eleazar Eskin, Sabeeha Merchant, and Todd Yeates – for their constant guidance. In particular, I appreciate the years of collaboration and for pointing me toward several fellowship and training opportunities.

I also thank Tama Hasson for encouraging me to continue my studies, as well as providing extensive mentorship and support throughout my undergraduate career and the graduate school application process. The staff of the Undergraduate Research Center played a critical role in my success as an early researcher.

I am thankful to all of the members of the Pellegrini lab for the countless ideas, discussions, and fun times. I'd like to especially thank Shawn Cokus for his mentorship throughout my graduate career. The work in this dissertation would not have been possible without many, many collaborators. In particular, I'd like to thank Mark Hildebrand, Jesse Traller, James Umen, Takashi Hamaji, and Robert Modlin.

My funding sources have been instrumental in giving me the freedom to explore my research interests. I want to thank the Genome Analysis Training Program, the Eugene V. Cota Robles Fellowship, as well as the Fred and Judith Eiseling Doctoral Fellowship for funding the work throughout my graduate career.

Finally, but certainly not least, I'd like to thank my family for their love and support throughout the years. The emotional, financial, and intellectual support of my parents – David and Leticia Lopez – enabled me to take the path that has led me here. Their encouragement, compassion, and example have helped in ways I cannot describe in words. I thank my brother, Andy Lopez, for always encouraging me and being my ally. I also thank Leyla Naimi for the years of support and for being by my side to share all the wonderful memories. I couldn't have done it without all of you.

VITA

EDUCATION

University of California, Los Angeles Los Angeles, CA

Ph.D. Candidate, Molecular, Cell, and Developmental Biology 2011-2016
Advisor: Matteo Pellegrini
B.S., Molecular, Cell, and Developmental Biology 2006-2011

GRANTS AND AWARDS

Fred Eiserling and Judith Lengyel Doctoral Fellowship 2014-2016
University of California, Los Angeles

Human Genetics Genome Analysis Training Grant 2013-2015
National Institutes of Health (NIH) Training Grant (T32)

Eugene V. Cota-Robles Fellowship 2011-2012
University of California, Los Angeles

PUBLICATIONS

Lopez D., Hamaji T., Kropat J., De Hoff P., Morselli M., Rubbi L., Fitz-Gibbon S., Gallaher S.D., Merchant S.S., Umen J., Pellegrini M. (2015). Dynamic changes in the transcriptome and methylome of *Chlamydomonas reinhardtii* throughout its life cycle. *Plant Phys*. 169(4):2730-2743.

Orozco L.D., Morselli M., Rubbi L., Guo W., Go J., Shi H., Lopez D., Furlotte N.A., Bennett B.J., Farber C.R., Ghazalpour A., Zhang M.Q., Bahous R., Rozen R., Lusis A.J., Pellegrini M. (2015). Epigenome-wide association of liver methylation patterns and complex metabolic traits in mice. *Cell Metabolism*. 21(6):905-917.

Inkeles M.S., Scumpia P.O., Swindell W.R., Lopez D., Teles R.M., Graeber T.G., Meller S., Homey B., Elder J.T., Gilliet M., Modlin R.L., Pellegrini M. (2015). Comparison of molecular signatures from multiple skin diseases identifies mechanisms of immunopathogenesis. *Journal of Investigative Dermatology*. 135(1):151-159.

Chen P.Y., Montanini B., Liao W.W., Morselli M., Jaroszewicz A., Lopez D., Ottonello S., Pellegrini M. (2014). A comprehensive resource of genomic, epigenomic and transcriptomic sequencing data for the black truffle Tuber melanosporum. *Gigascience.* 3:25.

Montanini B., Chen P.Y., Morselli M., Jaroszewicz A., Lopez D., Martin F., Ottonello S., Pellegrini M. (2014) Non-exhaustive DNA methylation mediated transposon silencing in the black truffle genome, a complex fungal genome with massive repeat element content. *Genome Biology*. 15:411

Jorda J.*, Lopez D.*, Wheatley N.M., Yeates T.O. (2013). Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria. *Protein Science*. 22(2):179-95.

Molnár I., Lopez D., Wisecaver J.H., Devarenne T.P., Weiss T.L., Pellegrini M., Hackett J.D. (2012). Bio- crude transcriptomics: Gene discovery and metabolic network reconstruction for the biosynthesis of the terpenome of the hydrocarbon oil-producing green alga, *Botryococcus braunii* race B (Showa). *BMC Genomics*. 13:576.

Chodavarapu R.K., Feng S., Ding B., Simon S.A., Lopez D., Jia Y., Wang G.L., Meyers B.C., Jacobsen S.E., Pellegrini M. (2012). Transcriptome and methylome interactions in rice hybrids. *Proc Natl Acad Sci*. 109(30):12040-5.

Lopez D., Casero D., Cokus S.J., Merchant S.S., Pellegrini M. (2011). Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. *BMC Bioinformatics*. 12:282.

* Authors contributed equally

CONFERENCES AND INVITED TALKS

"Uniparental Chloroplast Inheritance in Chlamydomonas reinhardtii". Molecular Biology Institute Annual Retreat. University of California, Los Angeles. January 2014.

"Active degradation and selective methylation of chloroplast DNA contribute to uniparental chloroplast inheritance in Chlamydomonas reinhardtii". Academy of Computational Life Sciences Meeting. Imperial College, London, England. September 2013.

"Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data". Emerging Researchers National Conference. Washington, DC. February 2011.

Chapter 1:

Computational tools and methods to analyze
high-throughput sequence data

As high-throughput sequence data becomes increasingly used in a variety of fields (Reuter et al., 2015), there is a growing need for computational tools that facilitate the analysis and interpretation of sequence data to extract biological meaning. Illustrating the diversity of biological applications for high-throughput sequencing, its utility has moved beyond genome sequencing and measuring transcript abundance to also include surveys of protein binding (Furey, 2012), methylation profiling (Krueger et al., 2012), chromatin conformation (Stadhouders et al., 2013), replication and transcription activity (Ingolia et al., 2009), and large-scale spatial genome structure (van de Werken et al., 2012). Furthermore, as these studies become more common, the compendium of processed sequence data gets larger – the repository Gene Expression Omnibus (GEO) (Barrett et al., 2013) contains over one million expression profiles, MSigDB (Subramanian et al., 2005) catalogs more than ten thousand molecular signatures, and more than 150 million single nucleotide polymorphisms (SNPs) are described in dbSNP (Sherry et al., 2001).

The vast amount of publicly-accessible sequence data can be leveraged to extract additional biological information or re-use in new contexts (Rung and Brazma, 2013). For example, thousands of previously-published cell-specific gene expression profiles have been combined to identify genes specifically regulated in immunological cell types and skin diseases (Swindell et al., 2013). Classifiers trained on gene expression profiles have been used to determine the tissue of origin for metastatic tumors (Ojala et al., 2011), and as well as determine tissue-specific expression (Kohane and Valtchinov, 2012). However, sequence data that can be reutilized and used in new contexts is not limited to expression profiles. Genome sequence data, when taken in combination across hundreds

of genomes, can also provide key biological insights (Rogers and Gibbs, 2014). Similar approaches can also be applied to methylation data (McCarthy et al., 2014) and protein binding profiles (Ouma et al., 2015). However, the vast collection of data necessitates the development of computational tools, methods, and algorithms to efficiently interrogate and interpret existing data.

To date, several computational tools have been developed to analyze raw and processed sequence data in a number of contexts. For example, tools have been created to functionally interpret large lists of genes (or coordinates) derived from large, genome-scale experiments. Such tools include Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005), the Database for Annotation, Visualization, and Integrated Discovery (DAVID) (Huang da et al., 2009), Ingenuity (Kramer et al., 2014), Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010). Many of these tools are routinely used in differential gene expression analyses and ChIP-seq experiments. The general methodology of these tools is the computation of statistics that describe the enrichment or difference between the gene or coordinate set and a background set of genes, such as by performing hypergeometric (DAVID, Ingenuity, GREAT) or Kolmogorov–Smirnov (GSEA) tests.

However, even in such contexts, these tools primarily support widely-studied reference genomes that have been manually curated over a large span of time, such as the human and mouse genomes. Such tools do not generally support relatively newer genomes, such as those of green alga and diatoms, which hinders genomic analyses in these species. To this end, this dissertation describes the development of tools tailored to these organisms (Chapter 3).

Regardless of model species, there is a scarcity of tools to integrate other genome-scale data that may provide biological insights, such as molecular signatures. Signatures are collections of genes that are characteristic of a particular cell type, developmental time point, or disease state (Nilsson et al., 2009). Several large-scale efforts by consortia aiming to systematically profile cell types of interest have generated large compendiums of expression data from which molecular signatures can be generated. For example, the Immunological Genome Project (ImmGen) (Heng et al., 2008) has profiled 214 cell types important in immunology. Furthermore, thousands of expression profiles from GEO may also be used to generate signatures, and tens of thousands are catalogued in repositories such as MSigDB. However, despite their utility, determining the behavior of these signatures in newly-generated expression experiments is not straightforward. Chapter 2 of this dissertation describes SaVanT, a tool to visualize molecular signatures.

Lastly, advances in the cost-effectiveness of sequencing can be leveraged to design multifaceted approaches to understanding the biology of species such as algae and bacteria. For example, genome-wide transcript abundance data can be supplemented with the methylation status of each gene to create a multi-layered model. An example of a combinatorial approach is shown in Chapter 4, which describes a study that interrogates methylation, expression, and genomic variation in *Chlamydomonas reinhardtii* throughout its sexual life cycle. Another approach, which leverages the breadth of complete bacterial genome sequences, is used to identify proteins involved in the formation of bacterial microcompartments (see Chapter 5). Bacterial microcompartments are polyhedral structures present in bacterial cytosol that encapsulate distinct metabolic processes (Yeates et al., 2010).

References

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res **41,** D991-995.

Furey, T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet **13,** 840-852.

Heng, T.S., Painter, M.W., and Immunological Genome Project, C. (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol **9,** 1091-1094.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc **4,** 44-57.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science **324,** 218-223.

Kohane, I.S., and Valtchinov, V.I. (2012). Quantifying the white blood cell transcriptome as an accessible window to the multiorgan transcriptome. Bioinformatics **28,** 538-545.

Kramer, A., Green, J., Pollard, J., Jr., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics **30,** 523-530.

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012). DNA methylome analysis using short bisulfite sequencing data. Nat Methods **9,** 145-151.

McCarthy, N.S., Melton, P.E., Cadby, G., Yazar, S., Franchina, M., Moses, E.K., Mackey, D.A., and Hewitt, A.W. (2014). Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. BMC Genomics **15,** 981.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol **28,** 495-501.

Nilsson, R., Bjorkegren, J., and Tegner, J. (2009). On reliable discovery of molecular signatures. BMC Bioinformatics **10,** 38.

Ojala, K.A., Kilpinen, S.K., and Kallioniemi, O.P. (2011). Classification of unknown primary tumors with a data-driven method based on a large microarray reference database. Genome Med **3,** 63.

Ouma, W.Z., Mejia-Guerra, M.K., Yilmaz, A., Pareja-Tobes, P., Li, W., Doseff, A.I., and Grotewold, E. (2015). Important biological information uncovered in previously unaligned reads from chromatin immunoprecipitation experiments (ChIP-Seq). Sci Rep **5,** 8635.

Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-throughput sequencing technologies. Mol Cell **58,** 586-597.

Rogers, J., and Gibbs, R.A. (2014). Comparative primate genomics: emerging patterns of genome content and dynamics. Nat Rev Genet **15,** 347-359.

Rung, J., and Brazma, A. (2013). Reuse of public genome-wide gene expression data. Nat Rev Genet **14,** 89-99.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res **29,** 308-311.

Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R.J., Wendt, K.S., Grosveld, F., van Ijcken, W., and Soler, E. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc **8,** 509-524.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **102,** 15545-15550.

Swindell, W.R., Johnston, A., Voorhees, J.J., Elder, J.T., and Gudjonsson, J.E. (2013). Dissecting the psoriasis transcriptome: inflammatory- and cytokine-driven gene expression in lesions from 163 patients. BMC Genomics **14,** 527.

van de Werken, H.J., de Vree, P.J., Splinter, E., Holwerda, S.J., Klous, P., de Wit, E., and de Laat, W. (2012). 4C technology: protocols and data analysis. Methods Enzymol **513,** 89-112.

Yeates, T.O., Crowley, C.S., and Tanaka, S. (2010). Bacterial microcompartment organelles: protein shell structure and evolution. Annu Rev Biophys **39,** 185-205.

Chapter 2:

SaVanT -- a web-based tool for the sample-level visualization of molecular signatures in gene expression profiles

Abstract

Molecular signatures are collections of genes characteristic of a particular cell type, tissue, disease, or perturbation. Signatures can also be used to interpret expression profiles generated from heterogeneous samples. Large collections of gene signatures have been previously developed and catalogued in the MSigDB database. In addition, several consortia and large-scale projects have systematically profiled broad collections of purified primary cells, molecular perturbations of cell types, and tissues from specific diseases, and the specificity and breadth of these datasets can be leveraged to create additional molecular signatures. However, to date there are few tools that allow the visualization of individual signatures across large numbers of expression profiles. Signature visualization of individual samples allows, for example, the identification of patient subcategories a priori on the basis of well-defined molecular signatures. Here, we generate and compile 10,985 signatures (636 newly-generated and 10,349 previously available from MSigDB) and provide a web-based Signature Visualization Tool (SaVanT; http://pathways.mcdb.ucla.edu/savant), to visualize these signatures in user-generated expression data. We show that using SaVanT, immune activation signatures can distinguish patients with different types of acute infections (influenza A and bacterial pneumonia). Furthermore, SaVanT is able to identify the prominent signatures within each patient group, and identify the primary cell types underlying different leukemias (acute myeloid and acute lymphoblastic) and skin disorders. The development of SaVanT facilitates large-scale analysis of gene expression profiles on a patient-level basis to identify patient subphenotypes, or potential therapeutic target pathways.

Background

Molecular signatures are collections of genes with an associated biological interpretation. For example, signatures can be generated from genes associated with specific cell types, diseases, or perturbations of cells by stimulatory signals. Signatures are typically generated from expression experiments that identify genes upregulated in a specific subset of samples when compared to a much broader group. Once generated, these signatures can be used to provide insights into the composition of heterogeneous samples. Signatures can also be composed of genes specifically associated with a disease. For example, molecular signatures from breast cancer samples have identified subphenotypes indistinguishable by traditional histological analyses [1], which can in turn be used predict tumor invasiveness and inform patient treatment options.

Generally, the generation of molecular signatures involves the identification of a set of genes that are overexpressed in a subgroup of samples compared to the entire dataset. Several methods have been used to identify these genes, such as hierarchical clustering [2], machine learning [3], and neural networks [4]. In combination, these methods have led to the creation of thousands of molecular signatures and gene sets, which are compiled in established repositories such as MSigDB [5]. Furthermore, some signatures are manually curated for certain biochemically-determined pathways, such as REACTOME [6] and KEGG [7]. In general, the most popular current pathway enrichment tools, Ingenuity [8], GSEA [5], and DAVID [9], calculate enrichment of molecular signatures that have the highest statistical overlap with a gene list that the user has filtered by analysis of their expression study. By limiting the analysis to a single

gene list, all of the individual variation of each expression profile is lost and further subcategorization of patient groups based upon these signatures is not immediately possible. Therefore the utility of signatures is limited by a lack of tools that are able to visualize the actual expression level of the signature genes within each user-supplied individual expression profile.

Furthermore, the current repositories of signatures are not exhaustive, and their signatures can be supplemented by additional signatures generated from large studies. For example, several consortia and large-scale projects have collected expression data with the aim of systematically profiling, and in some cases generating molecular signatures for, a diverse group of cells, tissues, and diseases. These include collections of immune cell subsets [10-13], other primary and cultured cells [14, 15], tissue types [16, 17], cytokine-activated immune cells [18, 19], and skin diseases [20, 21]. Collectively, these projects have produced over 3,000 expression profiles for more than 600 cell and tissue types. The specificity and breadth of these expression experiments can be leveraged to create molecular signatures that are not currently represented in MSigDB that can then be used to interpret new datasets.

To overcome the limitations of existing tools, we have generated 636 new signatures from expression dataset collections and supplemented them with 10,349 signatures from MSigDB for a total of 10,985 signatures and have developed a web-based Signature Visualization Tool (SaVanT), to visualize these signatures in user-generated expression profiles. SaVanT is able to analyze user-supplied expression studies and visualize the

average gene expression of molecular signatures across each individual expression profile. Through several examples, we show that SaVanT can be used to distinguish inflammatory patterns found between patients with different acute infections, identify the neoplastic cell type in leukemia samples, and provide insights into the immune response of several skin diseases. Through the visualization of molecular signatures, SaVanT allows users to efficiently leverage existing biological knowledge to interpret transcriptomic experiments.

Results

Generation of New Signatures

To leverage the vast number of reference expression profile repositories and add to MSigDB, we generated new molecular signatures using publicly-available expression data retrieved from a collection of repositories and sources (Table 1). Normalized data was used where available from the original study, but in lieu of preprocessed data, frozen robust multiarray analysis (fRMA) [22] normalization was used. Samples corresponding to biological replicates were averaged at the probe level, and genes with multiple probes were represented by the probe with the highest average intensity across all samples. In total, 4,677 microarray profiles were retrieved to generate molecular signatures.

Molecular signatures were generated from expression data by computing genome-wide 'proportional median' (PM) values. PM values are calculated by dividing the intensity of a microarray probe in a particular sample by the median intensity of the same probe across all samples in the corresponding data series. Therefore, high PM values are assigned to genes that are highly expressed in a certain sample relative to the others. A molecular signature consists of the top genes ranked in order of descending PM values. PM values have been previously used to generate signatures for a variety of skin diseases and conditions [20]. We note that the signatures we generated are ranked lists, while the signatures of MSigDB are unranked collections of genes. Using this PM metric, 636 ranked molecular signatures were created. The signatures represent a diverse set of biological states, as a consequence of the variety of sources used: we generated signatures

for 158 tissue types, 277 cell types, 70 primary cells, 114 molecular perturbations, and 17 skin diseases.

To assess and validate the use of proportional median values to create molecular signatures, we have annotated the genes from two representative signatures generated from the Human Primary Cell Atlas: an adipocyte-specific signature and a keratinocyte-specific signature (Supplementary Tables 1 and 2). The annotations assigned to the top genes (by PM rank) are characteristic of the distinct biology underlying the samples. For example, the adipocyte signature contains genes required for fatty acid processing and metabolism (fatty acid binding protein 4 [FABP4]), lipogenic proteins (lipogenic protein 1/THRSP), regulatory genes (adipogenesis regulatory factor/C10orf116), as well as genes known to be uniquely expressed in adipocytes, such as adiponectin (ADIPOQ). Similarly, the keratinocyte signature contains several keratin genes (keratin 6AII, keratin 14I, keratin 2II), envelope proteins (small proline-rich protein 1A [SPRR1A]), and regulatory genes involved in keratinocyte differentiation and maintenance (keratinocyte differentiation-associated protein [KRTDAP]). The enrichment of adipocyte- and keratinocyte-related annotations for the top genes in each respective signature suggests that our PM values capture genes that are specifically representative of the cell type or state of interest.

Visualization of Molecular Signatures

In order to visualize molecular signatures across any expression data of interest, we have developed the Signature Visualization Tool (SaVanT). SaVanT is a web-accessible tool

that accepts matrices of gene expression data (i.e., from RNA-seq or microarray experiments) and produces a visual representation of the signatures across the submitted samples as an interactive heatmap. The key step in the SaVanT pipeline is to create a 'sample-signature' matrix whose columns are the input samples and the rows are the user-selected molecular signatures (Figure 1). Using the default settings, every cell in this matrix contains the average value of signature genes for a particular signature-sample combination. This average value is computed by looking up the top genes for the user-selected signature in the SaVanT database and subsequently averaging the values of these genes in a particular sample in the user-submitted data. The sample-signature matrix is displayed by SaVanT as an interactive heatmap that can be optionally clustered along its axes. Alternatively, the 'sample-signature' matrix can consist of sums instead of mean values, and can be converted to z-scores or filtered by minimum values.

In order to enhance the visualization of the 'sample-signature' matrix, several optional steps can be used to transform the user-uploaded data or the 'sample-signature' matrix (Figure 2). For example, to dampen the effects of the large dynamic ranges characteristic of RNA-seq data, expression values can be log-transformed, converted to ranks, as well as shown as the difference from the mean value of all the samples. Once the sample-signature matrix is computed, its values can be converted to z-scores. On the submission page, an interactive description of the steps to create the matrix is shown, reflecting the chosen parameters. Clustering of the sample-signature matrix can be performed using several distance metrics (Euclidean distance or Pearson correlation) as well as different linkage parameters. The heatmap produced by SaVanT is interactive, and additional

information (such as the sample-signature combination, p-values, and the matrix value) are shown as hover-over boxes.

Analyses of Example Datasets

To demonstrate the capabilities of SaVanT, we provide biologically-motivated examples using publicly-available datasets retrieved from GEO [23].

Cell type identification within tissue samples

SaVanT can be used to identify the relative abundance of cell types found within tissue samples. To demonstrate this capability, we retrieved samples from patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) (Figure 3A). Using a panel of signatures representing different hematopoietic cells, SaVanT produced heatmaps identifying the principal cell type in AML samples (monocytes) and ALL samples (B cells). Furthermore, the heatmap identifies one ALL sample that may be misclassified (the first sample in the heatmap), although we could not find supporting metadata to support this.

Discrimination of disease phenotypes

Often the main goals of expression studies of clinical samples are to distinguish between clinical phenotypes and to identify the molecular signatures that differ between phenotypes to provide insight on disease pathogenesis. To demonstrate the ability of SaVanT to accomplish both goals, expression data was retrieved from a study profiling expression of whole blood samples collected daily from 17 patients with either influenza

A or bacterial pneumonia [24]. The study found enrichment of interferon, cell cycle genes, apoptosis, DNA damage, B cell , CD4+ T helper cells, and neutrophils in bacterial versus influenza-induced pneumonia. We used SaVanT to visualize equivalent gene signatures from MSigDB or cell type-specific expression profiles on a per-patient basis, filtering for those signatures that are statistically different between bacterial pneumonia and influenza. The clustered heatmap produced by SaVanT separates the acute infection samples into two groups: the predominantly influenza cluster was characterized by higher signature values for type I interferon pathways, B cells, cell-cycle, DNA damage, and apoptosis. The bacterial pneumonia cluster was composed of 92% bacterial pneumonia samples, characterized by higher neutrophil signature values relative to influenza. Five other samples were clustered as outliers. In addition to identifying the main clusters between disease groups, the SaVanT analysis displays intra-disease differences in molecular and cellular pathways. For example, there are two different bacterial subclusters in which one group has higher B cell signatures with the other higher in neutrophils. Furthermore, upon examination of the influenza group we can see that the misclassified bacterial pneumonia samples still have higher neutrophil signatures, but also have high type 1 interferon signatures, potentially identifying the reason for misclassification and targeting for further investigation. If the group structure of the submitted samples is known, as in this case, the tag 'SAVANT_GROUP' can be included in the submitted expression matrix with integers designating group membership of the samples, which automatically runs an ANOVA analysis on the signature-sample matrix.

Dermatoses

Lastly, in order to illustrate the analyses of heterogenous tissue samples, we used expression data from a collection of skin diseases [20] and analyzed these using signatures for specific cell types found in the skin (Figure 3C). The predominant signature for most samples is that of keratinocytes, which illustrates that while our signature values cannot be interpreted as quantitative estimates of cell type fractions, a higher relative value does reflect that the underlying cell type is more abundant than those associated with other lower scoring signatures. Within these dermatoses we also find several samples that have weaker keratinocyte signatures, but higher values for other signatures (designated by blue boxes). For example, the macrophage signature is elevated in leprosy lesions (erythema nodosum leprosum, lepromatous leprosy, and reversal reaction), as would be expected from the presence of macrophages within the granulomas in these biopsies. Furthermore, signatures derived from hematopoietic cells are elevated in tissue samples from patients with Stevens Johnsons disease, which are collected from blister fluid, along with mucosis fungoides, a T cell neoplasm, and sarcoidosis, which also typically has abundant granulomas. Overall, these signatures help interpret the components of these skin biopsies, which may in large part underlie the differences in gene expression between them.

Discussion

SaVanT provides an interactive platform for compiling and visualizing molecular signatures in order to interpret user-submitted data. The newly-generated signatures that supplement MSigDB, leverage the specificity and depth of large expression studies to

capture the biology pertaining to specific diseases, cell types, and immune states. The functions of the top genes in our signatures often reflect signature-specific characteristics. The adipocyte and keratinocyte signatures as representative examples, with each of their top 10 genes reflecting specific association to the differentiated cell lineages.

Moreover, in addition to compiling a large database of signatures, we also provide a framework that enables their mining and visualization to help interpret user-supplied expression data. To this end, we provide very flexible options to enable different analyses. For example, signature values can be filtered to only display those above a certain threshold or clustered using a number of different options. Furthermore, SaVanT has been optimized to quickly scan more than 10,000 signatures in seconds, thus allowing all signatures to be computed against user-uploaded expression matrices.

The power of SaVanT is illustrated in the examples shown in Figure 3. The fundamental objectives of each analysis are distinct: identification of inflammatory states that differentiate two clinical presentation (Fig. 3A), identification of the neoplastic cell types in a liquid tumor (Fig. 3B), and gaining insights into the composition of heterogeneous biopsies (Fig. 3C). Viral infection, including influenza is characterized by a strong induction of a type I interferon antiviral response, composed of genes induced by interferon α and β stimulation [25, 26], and this is reflected in the signature-sample heatmap. The neoplastic cell type present in a leukemia can be seen using the signature-sample heatmap, and misclassified patients identified. Lastly, the cell type compositions of skin biopsies from a number of dermatoses can be determined from expression data.

In the future we plan to continue to develop SaVanT by adding more signatures, along with additional features that facilitate the interpretation of complex expression patterns.

Methods

Expression Data Retrieval and Processing

Microarray data was retrieved from Gene Expression Omnibus (GEO) for samples and series listed in Supplementary Table 2. Raw CEL files were processed using the 'affy' R package and were normalized using the 'frma' R package in conjunction with the respective 'frmavecs' package for the platform used. Intensities for multiprobe genes were taken from the probe with the highest mean expression across all samples. Samples annotated as biological replicates in the GEO series description were combined by taking the average value of probes in the replicates.

Signature Generation

'Proportional median' (PM) values were calculated by dividing the intensity of a probe in a particular sample by its median value across all samples. For PM calculations, datasets from different sources were considered independently (i.e., the denominator was composed of only samples within a certain series when calculating PMs for a sample within that series). PMs were calculated at the probe level, and PM values were subsequently associated with gene symbols using the platform-specific annotation tables from GEO.

Signature Visualization

In order to produce the sample-signature heatmap with SaVanT, the user-submitted expression matrix is processed by a series of scripts. Ambiguous values, such as those for gene symbols appearing multiple times in the user input, are resolved by taking the average value of all instances. Optional transformations (log-transformation and/or conversion to ranks, in that order) are performed on the input expression matrix, and the 'sample-signature matrix' is created by taking the average (or sum, optionally) of expression values for genes in every signature-sample pair. If conversion to z-scores is selected, the mean and standard deviation is computed for the entire 'sample-signature' matrix, which are used to convert the values to z-scores. Clustering is optionally performed by the R 'heatmaps.2' function of the 'gplots' package. The signature-sample matrix is displayed interactively using a modified version of the HighCharts JavaScript library.

Tables

Table 1

Expression Data Source	Reference	Platform	Normalization	# Signatures Generated
Human U133A/GNF1H Gene Atlas (BioGPS)	Su AI et al. (2004) *PNAS*	Affymetrix U133A/GNF1H	fRMA	84
Mouse MOE430 Gene Atlas (BioGPS)	Lattin JE et al. (2008) *Immunome Res.*	Affymetrix 430 2.0 Array	fRMA	94
Immunological Genome Project (ImmGen)	Heng TS et al. (2008) *Nature Immunology*	Affymetrix Gene 1.0 ST	Pre-processed	214
Human Cell Types (Swindell)	Swindell WR et al. (2013) *BMC Genomics*	Affymetrix Genome Plus 2.0	fRMA	24
Macrophage Activation	Xue J et al. (2014) *Immunity*	Illumina HumanHT-12 V3.0	Pre-processed	80
Primary Cell Atlas	Mabbott NA (2013) *BMC Genomics*	Affymetrix U133 Plus 2.0	fRMA	26
Skin Diseases (“DermDB”)	Inkeles MS et al. (2015) *J. Invest. Dermatol.*	Mixed	fRMA	23

Supplementary Table 1

Gene Symbol	PM Value	Gene Description	Notes
FABP4	577.23	Fatty Acid Binding Protein 4, Adipocyte	encodes the fatty acid binding protein found in adipocytes; FABPs roles include fatty acid uptake, transport, and metabolism.
ADIPOQ	563.24	Adiponectin, C1Q And Collagen Domain Containing	gene is expressed in adipose tissue exclusively; encoded protein circulates in the plasma and is involved with metabolic and hormonal processes
COL3A1	519.97	Collagen, Type III, Alpha 1	encodes the pro-alpha1 chains of type III collagen, a fibrillar collagen that is found in extensible connective tissues such as skin, lung, uterus, intestine and the vascular system
ADH1B	494.47	Alcohol Dehydrogenase 1B (Class I), Beta Polypeptide	a member of the alcohol dehydrogenase family; metabolize a wide variety of substrates, including ethanol, retinol, other aliphatic alcohols, hydroxysteroids, and lipid peroxidation products
THRSP	468.30	Thyroid Hormone Responsive; Lipogenic Protein 1	shown to be expressed in liver and adipocytes, particularly in lipomatous modules ; also found to be expressed in lipogenic breast cancers

RBP4	458.86	Retinol Binding Protein 4, Plasma
COL1A2	452.31	Collagen, Type I, Alpha 2
SRPX	350.80	Sushi-Repeat Containing Protein, X-Linked
TIMP4	338.46	TIMP Metallopeptidase Inhibitor 4
C10orf116	317.78	Adipogenesis Regulatory Factor

Supplementary Table 2

Gene Symbol	PM Value	Gene Description	Notes
SPRR1A	1238.97	Small Proline-Rich Protein 1A	cross-linked envelope protein of keratinocytes; first appears in the cell cytosol, but ultimately becomes cross-linked to membrane proteins by transglutaminase
KRTDAP	1186.22	Keratinocyte Differentiation-Associated Protein	may function in the regulation of keratinocyte differentiation and maintenance of stratified epithelia
KRT6A	1175.16	Keratin 6A, Type II	consist of basic or neutral proteins which are arranged in pairs of heterotypic keratin chains coexpressed during differentiation of simple and stratified epithelial tissues
KRT14	1173.01	Keratin 14, Type I	usually found as a heterotetramer with two keratin 5 molecules, a type II keratin; together they form the cytoskeleton of epithelial cells.
SPRR1B	1040.41	Small Proline-Rich Protein 1B	cross-linked envelope protein of keratinocytes; diseases associated with SPRR1B include epidermolytic hyperkeratosis
FLG	883.49	Filaggrin	
FLG2	880.08	Filaggrin Family Member 2	
DSC1	772.33	Desmocollin 1	
KRT2	678.45	Keratin 2, Type II	
DMKN	647.12	Dermokine	

Figures

Figure 1. Constructing 'Signature-Sample' Matrix From Expression Data

The SaVanT pipeline converts user-submitted expression data into a signature-sample matrix whose columns are the submitted samples and rows are the user-selected molecular signatures. By default (shown above), every cell in this matrix contains the average value of signature genes for a particular signature-sample combination. The breakdown for an example cell in the signature-sample matrix is shown in red. The matrix value is computed by looking up the genes in any given user-selected signature in the SaVanT database (middle panel) and subsequently averaging the values of these genes in a particular sample in the user-submitted data (left and right panels). Above, samples are designated with numbers, genes with letters, and signatures with Roman numerals.

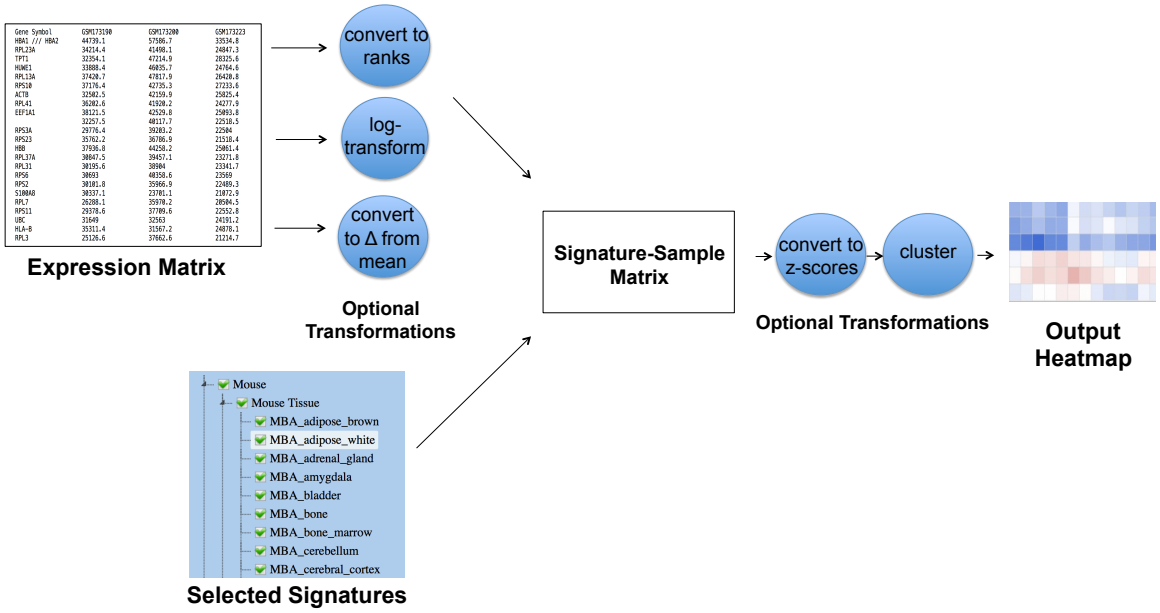


Figure 2. SaVanT Pipeline

In the first step, an expression matrix containing values for genes in several samples is optionally converted to ranked lists of genes in samples or log-transformed. The expression matrix is then converted into a signature-sample matrix as described in Figure 1 using the selected signatures. Optionally, the signature-sample matrix is converted to differences from mean values, converted to z-scores, and/or clustered to produce a final heatmap.

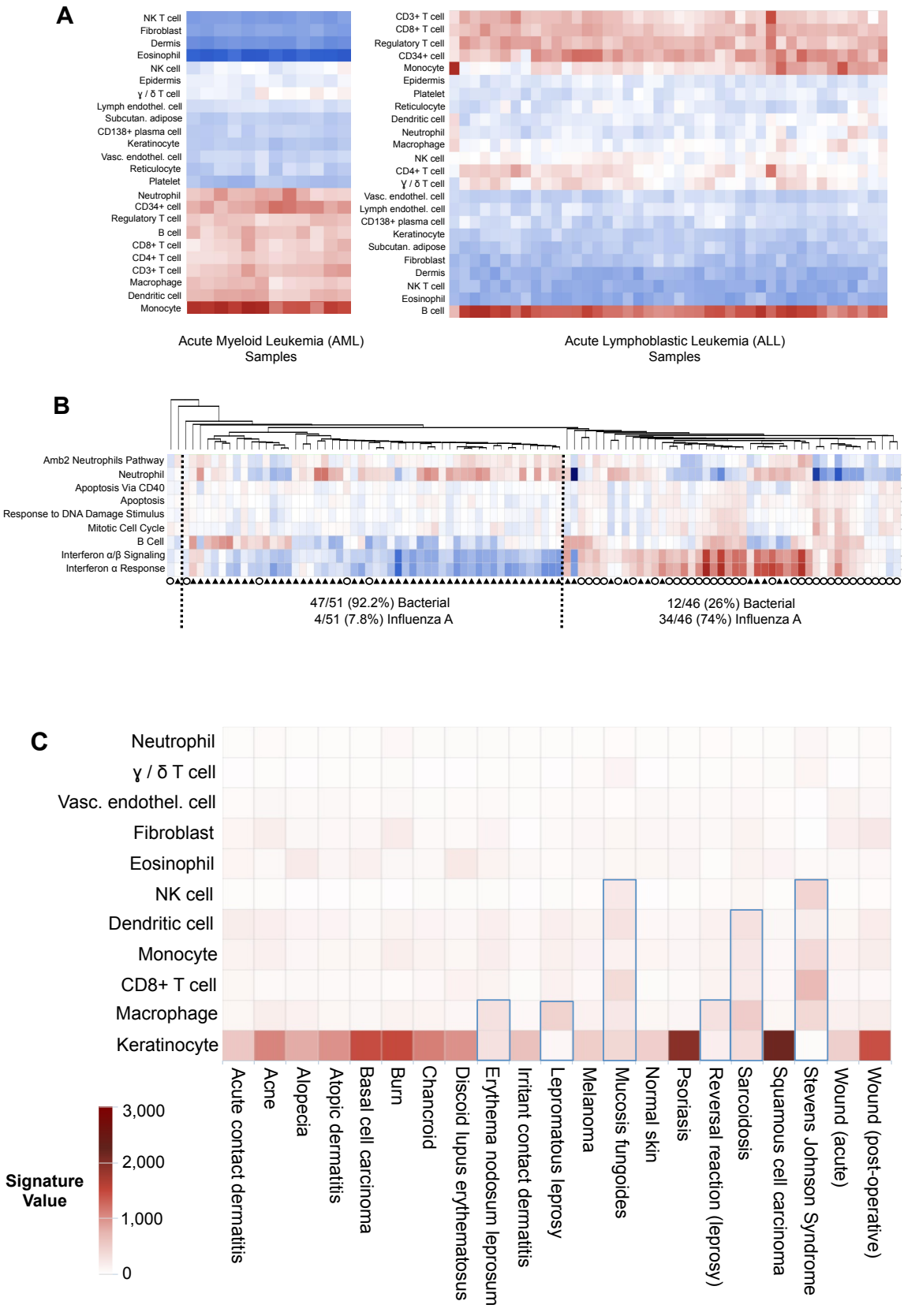

A
NK T cell
Fibroblast
Dermis
Eosinophil
NK cell
Epidermis
γ / δ T cell
Lymph endothel. cell
Subcutan. adipose
CD138+ plasma cell
Keratinocyte
Vasc. endothel. cell
Reticulocyte
Platelet
Neutrophil
CD34+ cell
Regulatory T cell
B cell
CD8+ T cell
CD4+ T cell
CD3+ T cell
Macrophage
Dendritic cell
Monocyte
Acute Myeloid Leukemia (AML) Samples
CD3+ T cell
CD8+ T cell
Regulatory T cell
CD34+ cell
Monocyte
Epidermis
Platelet
Reticulocyte
Dendritic cell
Neutrophil
Macrophage
NK cell
CD4+ T cell
γ / δ T cell
Vasc. endothel. cell
Lymph endothel. cell
CD138+ plasma cell
Keratinocyte
Subcutan. adipose
Fibroblast
Dermis
NK T cell
Eosinophil
B cell
Acute Lymphoblastic Leukemia (ALL) Samples
B
Amb2 Neutrophils Pathway
Neutrophil
Apoptosis Via CD40
Apoptosis
Response to DNA Damage Stimulus
Mitotic Cell Cycle
B Cell
Interferon α/β Signaling
Interferon α Response
47/51 (92.2%) Bacterial
4/51 (7.8%) Influenza A
12/46 (26%) Bacterial
34/46 (74%) Influenza A
C
Neutrophil
γ / δ T cell
Vasc. endothel. cell
Fibroblast
Eosinophil
NK cell
Dendritic cell
Monocyte
CD8+ T cell
Macrophage
Keratinocyte
Acute contact dermatitis
Acne
Alopecia
Atopic dermatitis
Basal cell carcinoma
Burn
Chancroid
Discoid lupus erythematosus
Erythema nodosum leprosum
Irritant contact dermatitis
Lepromatous leprosy
Melanoma
Mucosis fungoides
Normal skin
Psoriasis
Reversal reaction (leprosy)
Sarcoidosis
Squamous cell carcinoma
Stevens Johnson Syndrome
Wound (acute)
Wound (post-operative)
Signature Value
3,000
2,000
1,000
0

Figure 3. SaVanT Distinguishes Between Patients, Cell Types, and Underlying Biology

(A) SaVanT output for expression data from acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) patients. (B) SaVanT output for expression data from 99 patients with acute infections (either Influenza A or bacterial pneumonia). The infection type for each patient is represented by a hatched circle (Influenza A) or filled triangle (bacterial pneumonia). The numbers below each cluster quantify the proportion of infection types. (C) SaVanT output for expression data from different skin diseases.

References

1. Pedraza V, Gomez-Capilla JA, Escaramis G, Gomez C, Torne P, Rivera JM, Gil A, Araque P, Olea N, Estivill X, Farez-Vidal ME: **Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness.** *Cancer* 2010, **116:**486-496.

2. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M, et al: **Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value.** *PLoS Med* 2013, **10:**e1001453.

3. Bartsch G, Jr., Mitra AP, Mitra SA, Almal AA, Steven KE, Skinner DG, Fry DW, Lenehan PF, Worzel WP, Cote RJ: **Use of Artificial Intelligence and Machine Learning Algorithms with Gene Expression Profiling to Predict Recurrent Nonmuscle Invasive Urothelial Carcinoma of the Bladder.** *J Urol* 2016, **195:**493-498.

4. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7:**673-679.

5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.

6. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, et al: **The Reactome pathway Knowledgebase.** *Nucleic Acids Res* 2016, **44:**D481-487.

7. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic Acids Res* 2014, **42:**D199-205.

8. Kramer A, Green J, Pollard J, Jr., Tugendreich S: **Causal analysis approaches in Ingenuity Pathway Analysis.** *Bioinformatics* 2014, **30:**523-530.

9. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44-57.

10. Heng TS, Painter MW, Immunological Genome Project C: **The Immunological Genome Project: networks of gene expression in immune cells.** *Nat Immunol* 2008, **9:**1091-1094.

11. Kim CC, Lanier LL: **Beyond the transcriptome: completion of act one of the Immunological Genome Project.** *Curr Opin Immunol* 2013, **25:**593-597.

12. Shay T, Kang J: **Immunological Genome Project and systems immunology.** *Trends Immunol* 2013, **34:**602-609.

13. Xue J, Schmidt SV, Sander J, Draffehn A, Krebs W, Quester I, De Nardo D, Gohel TD, Emde M, Schmidleithner L, et al: **Transcriptome-based network analysis reveals a spectrum model of human macrophage activation.** *Immunity* 2014, **40:**274-288.

14. Mabbott NA, Baillie JK, Brown H, Freeman TC, Hume DA: **An expression atlas of human primary cells: inference of gene function from coexpression networks.** *BMC Genomics* 2013, **14:**632.

15. Swindell WR, Johnston A, Voorhees JJ, Elder JT, Gudjonsson JE: **Dissecting the psoriasis transcriptome: inflammatory- and cytokine-driven gene expression in lesions from 163 patients.** *BMC Genomics* 2013, **14:**527.

16. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101:**6062-6067.

17. Lattin JE, Schroder K, Su AI, Walker JR, Zhang J, Wiltshire T, Saijo K, Glass CK, Hume DA, Kellie S, Sweet MJ: **Expression analysis of G Protein-Coupled Receptors in mouse macrophages.** *Immunome Res* 2008, **4:**5.

18. Teles RM, Graeber TG, Krutzik SR, Montoya D, Schenk M, Lee DJ, Komisopoulou E, Kelly-Scumpia K, Chun R, Iyer SS, et al: **Type I interferon suppresses type II interferon-triggered human anti-mycobacterial responses.** *Science* 2013, **339:**1448-1453.

19. Montoya D, Inkeles MS, Liu PT, Realegeno S, Teles RM, Vaidya P, Munoz MA, Schenk M, Swindell WR, Chun R, et al: **IL-32 is a molecular marker of a host defense network in human tuberculosis.** *Sci Transl Med* 2014, **6:**250ra114.

20. Inkeles MS, Scumpia PO, Swindell WR, Lopez D, Teles RM, Graeber TG, Meller S, Homey B, Elder JT, Gilliet M, et al: **Comparison of molecular signatures from multiple skin diseases identifies mechanisms of immunopathogenesis.** *J Invest Dermatol* 2015, **135:**151-159.

21. Wong D, Kea B, Pesich R, Higgs BW, Zhu W, Brown P, Yao Y, Fiorentino D: **Interferon and biologic signatures in dermatomyositis skin: specificity and heterogeneity across diseases.** *PLoS One* 2012, **7:**e29161.

22. McCall MN, Bolstad BM, Irizarry RA: **Frozen robust multiarray analysis (fRMA).** *Biostatistics* 2010, **11:**242-253.

23. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al: **NCBI GEO: archive for functional genomics data sets--update.** *Nucleic Acids Res* 2013, **41:**D991-995.

24. Parnell GP, McLean AS, Booth DR, Armstrong NJ, Nalos M, Huang SJ, Manak J, Tang W, Tam OY, Chan S, Tang BM: **A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia.** *Crit Care* 2012, **16:**R157.

25. Randall RE, Goodbourn S: **Interferons and viruses: an interplay between induction, signalling, antiviral responses and virus countermeasures.** *J Gen Virol* 2008, **89:**1-47.

26. Haller O, Kochs G, Weber F: **The interferon response circuit: induction and suppression by pathogenic viruses.** *Virology* 2006, **344:**119-130.

Chapter 3:

Algal Functional Annotation Tool – a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data

Lopez et al. BMC Bioinformatics 2011, 12:282
http://www.biomedcentral.com/1471-2105/12/282

BMC Bioinformatics

DATABASE Open Access

Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data

David Lopez[1], David Casero[1], Shawn J Cokus[1], Sabeeha S Merchant[2,3] and Matteo Pellegrini[1,3*]

Abstract

Background: Progress in genome sequencing is proceeding at an exponential pace, and several new algal genomes are becoming available every year. One of the challenges facing the community is the association of protein sequences encoded in the genomes with biological function. While most genome assembly projects generate annotations for predicted protein sequences, they are usually limited and integrate functional terms from a limited number of databases. Another challenge is the use of annotations to interpret large lists of 'interesting' genes generated by genome-scale datasets. Previously, these gene lists had to be analyzed across several independent biological databases, often on a gene-by-gene basis. In contrast, several annotation databases, such as DAVID, integrate data from multiple functional databases and reveal underlying biological themes of large gene lists. While several such databases have been constructed for animals, none is currently available for the study of algae. Due to renewed interest in algae as potential sources of biofuels and the emergence of multiple algal genome sequences, a significant need has arisen for such a database to process the growing compendiums of algal genomic data.

Description: The Algal Functional Annotation Tool is a web-based comprehensive analysis suite integrating annotation data from several pathway, ontology, and protein family databases. The current version provides annotation for the model alga *Chlamydomonas reinhardtii*, and in the future will include additional genomes. The site allows users to interpret large gene lists by identifying associated functional terms, and their enrichment. Additionally, expression data for several experimental conditions were compiled and analyzed to provide an expression-based enrichment search. A tool to search for functionally-related genes based on gene expression across these conditions is also provided. Other features include dynamic visualization of genes on KEGG pathway maps and batch gene identifier conversion.

Conclusions: The Algal Functional Annotation Tool aims to provide an integrated data-mining environment for algal genomics by combining data from multiple annotation databases into a centralized tool. This site is designed to expedite the process of functional annotation and the interpretation of gene lists, such as those derived from high-throughput RNA-seq experiments. The tool is publicly available at http://pathways.mcdb.ucla.edu.

* Correspondence: matteop@mcdb.ucla.edu
[1]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA, USA
Full list of author information is available at the end of the article

Background

Next-generation sequencers are revolutionizing our ability to sequence the genomes of new algae efficiently and in a cost effective manner. Several assembly tools have been developed that take short read data and assemble it into large continuous fragments of DNA. Gene prediction tools are also available which identify coding structures within these fragments. The resulting transcripts can then be analyzed to generate predicted protein sequences. The function of these protein sequences are subsequently determined by searching for close homologs in protein databases and transferring the annotation between the two proteins. While some versions of the previously described data processing pipeline have become commonplace in genome projects, the resulting functional annotation is typically fairly minimal and includes only limited biological pathway information and protein structure annotation. In contrast, the integration of a variety of pathway, function and protein databases allows for the generation of much richer and more valuable annotations for each protein.

A second challenge is the use of these protein-level annotations to interpret the output of genome-scale profiling experiments. High-throughput genomic techniques, such as RNA-seq experiments, produce measurements of large numbers of genes relevant to the biological processes being studied. In order to interpret the biological relevance of these gene lists, which commonly range in size from hundreds to thousands of genes, the members must be functionally classified into biological pathways and cellular mechanisms. Traditionally, the genes within these lists are examined using independent annotation databases to assign functions and pathways. Several of these annotation databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [1], MetaCyc [2], and Pfam [3], include a rich set of functional data useful for these purposes.

However, presently researchers must explore these different knowledge bases separately, which requires a substantial amount of time and effort. Furthermore, without systematic integration of annotation data, it may be difficult to arrive at a cohesive biological picture. In addition, many of these annotation databases were designed to accommodate a single gene search, a methodology not optimal for functionally interpreting the large lists of genes derived from high-throughput genomic techniques. Thus, while modern genomic experiments generate data for many genes in parallel, their output must often still be analyzed on a gene-by-gene basis across different databases. This fragmented analysis approach presents a significant bottleneck in the pipeline of biological discovery.

One approach to solving this problem is integrating information from multiple annotation databases and providing access to the combined biological data from a single comprehensive portal that is equipped with the proper statistical foundations to effectively analyze large gene lists. For example, the DAVID database integrates information from several pathway, ontology, and protein family databases [4]. Similarly, Ingenuity Pathway Analysis (IPA) provides an integrated knowledge base derived from published literature for the human genome [5]. The integrated functional information and annotation terms are then assigned to lists of genes and for some analyses, enrichment tests are performed to determine which biological terms are overrepresented within the group of genes. By combining the information found in a number of knowledge bases and performing the analysis of lists of genes, these tools permit the efficient processing of high-throughput genomic experiments and thus expedite the process of biological discovery. However, most of these integrated databases have been developed for the analysis of well-annotated and thoroughly studied organisms, and are lacking for many newly genome-enabled organisms.

One large group of organisms for which integrated functional databases are lacking are the algae. The algae constitute a branch in the plant kingdom, although they form a polyphyletic group as they do not include all the descendants of their last common ancestor. As many as 10 algal genomes have been sequenced, including those of a red alga and several chlorophyte algae, with several more in the pipeline [6-11]. Algal genomic studies have provided insights into photosymbiosis, evolutionary relationships between the different species of algae, as well as their unique properties and adaptations. Recently, there has been a renewed interest in the study of algal biochemistry and biology for their potential use in the development of renewable biofuels [reviewed in [12]]. This has promoted the study of varied biochemical processes in diverse algae, such as hydrogen metabolism, fermentation, lipid biosynthesis, photosynthesis and nutrient assimilation [13-20]. One of the most studied algae is *Chlamydomonas reinhardtii*. It has a sequenced genome that has been assembled into large scaffolds that are placed on to chromosomes [6]. For many years, *Chlamydomonas* has served as a reference organism for the study of photosynthesis, photoreceptors, chloroplast biology and diseases involving flagellar dysfunction [21-25]. Its transcriptome has recently been profiled by RNA-seq experiments under various conditions of nutrient deprivation [[26,27], unpublished data (Castruita M., et al.)].

While *Chlamydomonas* has been extensively characterized experimentally, annotation of its genome is still

approximate. Although KEGG categorizes some *C. reinhardtii* gene models into biological pathways, other databases - such as Reactome [28] - do not directly provide information for proteins of this green alga. Complicating the analysis of *Chlamydomonas* genes is the fact that there are two assemblies of the genome in use (version 3 and version 4) and multiple sets of gene models have been developed that are catalogued under diverse identifiers: Joint Genome Institute (JGI) FM3.1 protein IDs for the version 3 assembly, and JGI version FM4 protein IDs and Augustus version 5 IDs for the version 4 assembly [11,29]. The differences between these assemblies are significant; for example, the version 3 assembly contains 1,557 continuous segments of sequence while the fourth version contains 88. Although the version 3 assembly is superseded by version 4, users presently access version 3 because of the richer user-based functional annotations. In addition, other sets of gene predictions have been generated using a variety of additional data, including ESTs and RNA-seq data, to more accurately delineate start and stop positions and improve upon existing gene models. One such gene prediction set is Augustus u10.2. As such, there are a variety of gene models between different assemblies being simultaneously used by researchers, presenting complications in genomics studies. To facilitate the analysis of *Chlamydomonas* genome-scale data, we developed the Algal Functional Annotation Tool, which provides a comprehensive analysis suite for functionally interpreting *C. reinhardtii* genes across all available protein identifiers. This web-based tool provides an integrative data-mining environment that assigns pathway, ontology, and protein family terms to proteins of *C. reinhardtii* and enables term enrichment analysis for lists of genes. Expression data for several experimental conditions are also integrated into the tool, allowing the determination of overrepresented differentially expressed conditions.

Table 1 List of annotation resources integrated into the Algal Functional Annotation Tool

Resource	URL	Reference
KEGG	http://www.genome.jp/kegg/	[1]
MetaCyc	http://www.metacyc.org/	[2]
Pfam	http://pfam.sanger.ac.uk	[3]
Reactome	http://www.reactome.org/	[28]
Panther	http://www.pantherdb.org/pathway	[30]
Gene Ontology	http://www.geneontology.org/	[31]
InterPro	http://www.ebi.ac.uk/interpro	[32]
MapMan Ontology	http://mapman.gabipd.org/	[33]
KOG	http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi	[35]

Primary databases used to functionally annotate gene models and integrated into the Algal Functional Annotation Tool.

Additionally, a gene similarity search tool allows for genes with similar expression patterns to be identified based on expression levels across these conditions.

Construction and Content

Integration of Multiple Annotation Databases

The Algal Functional Annotation Tool integrates annotation data from the biological knowledge bases listed in Table 1. Publically available flat files containing annotation data were downloaded and parsed for each individual resource. *Chlamydomonas reinhardtii* proteins were assigned KEGG pathway annotations by means of sequence similarity to proteins within the KEGG genes database [1]. MetaCyc [2], Reactome [28], and Panther [30] pathway annotations were assigned to *C. reinhardtii* proteins by sequence similarity to subsets of UniProt IDs annotated in each corresponding database. In all cases, sequence similarity was determined by BLAST. BLAST results were filtered to contain only best hits with an E-value < 1e-05.

Gene Ontology (GO) [31] terms were downloaded from the *Chlamydomonas reinhardtii* annotation provided by JGI. These GO terms were associated with their respective ancestors in the hierarchical ontology structure to include broader functional terms and provide a complete annotation set. Pfam domain annotations were assigned by direct search against protein domain signatures provided by Pfam. InterPro [32] and user-submitted manual annotations are based on those contained within JGI's annotation of the *C. reinhardtii* genome [11]. These methods were applied to four types of gene identifiers commonly used for *C. reinhardtii* proteins: JGI protein identifiers (versions 3 and 4) and Augustus gene models (versions 5 and 10.2). In total, over 12,600 unique functional annotation terms were assigned to 65,494 *C. reinhardtii* gene models spanning four different gene identifier types by these methods (Table 2). These assigned annotations may be explored for single genes using a built-in keyword search tool as well as an integrated annotation lookup tool which displays all annotations for a particular identifier.

Assignment of Annotation from *Arabidopsis thaliana*

To extend the terms associated with *C. reinhartdii* genes, functional terms were inferred by homology to the annotation set of the plant *Arabidopsis thaliana* (thale cress). Identification of orthologous proteins was based on sequence similarity and subsequent filtering of the results by retaining only mutual best hits between the two sets of protein sequences. The corresponding *Arabidopsis thaliana* annotation was used to supplement GO terms and was similarly expanded to contain term ancestry. The *A. thaliana* annotations of the MapMan Ontology [33] and MetaCyc Pathway database [2]

Table 2 Number of gene identifiers associated with annotation databases

Identifier Type	Total Gene IDs	KEGG	Reactome	Panther	Gene Ontology	MapMan	KOG	Pfam	InterPro
JGI v3.0	14598	5348	2740	1147	6563	5214	9139	7166	7532
JGI v4.0	16706	4232	1949	1085	7568	3171	9973	7305	8151
Augustus v5.0	16888	4686	2983	1673	4334	3160	5123	8202	5202
Augustus u10.2	17302	4583	3326	1913	6956	3892	8977	8691	7464

Number of *Chlamydomonas reinhardtii* identifiers with at least one functional annotation for each primary database, shown per identifier type.

were also used to provide more complete annotation coverage of the *C. reinhardii* genome.

Functional Term Enrichment Testing

The hypergeometric distribution is commonly used to determine the significance of functional term enrichment within a list of genes. In this test, the occurrence of a functional term within a gene list is compared to the background level of occurrence across all genes in the genome to determine the degree of enrichment. A p-value based on this test can be calculated from four parameters: (1) the number of genes within the list, (2) the frequency of a term within the gene list, (3) the total number of genes within the genome, and (4) the frequency of a term across all genes in the genome. This test effectively distinguishes truly overrepresented terms from those occurring at a high frequency across all genes in the genome and therefore within the gene list as well. The cumulative hypergeometric test assigns a p-value to each functional term associated with genes within a given list, and all functional terms are ranked by ascending p-value (i.e. by descending levels of enrichment). Huang et al. reviews the use of the hypergeometric test for functional term enrichment [34]. The Algal Functional Annotation Tool computes hypergeometric p-values using a Perl wrapper for the GNU Scientific Library cumulative hypergeometric function written in C to provide a quick and accurate implementation of this statistical test.

Dynamic Visualization of KEGG Pathway Maps

Individual pathway maps from KEGG provide information on protein localization within the cell, compartmentalization into different cellular components, or of reactions within a larger metabolic process. Visualization of proteins from gene lists onto pathway maps is useful for their interpretation. The Algal Functional Annotation Tool utilizes the publicly available KEGG application programming interface (API) for pathway highlighting. The information linking *C. reinhardtii* proteins to identifiers within the KEGG database is used to determine the subset of KEGG IDs within the supplied gene list associated with a particular pathway. The Algal Functional Annotation Tool also deduces which proteins within the pathway are located within the genome of *C. reinhardtii* but not found in the gene list and sends the corresponding identifiers to the KEGG API to be highlighted in a different background color. This API interface is implemented using the SOAP architecture for web applications.

Integration of Expression Data

The expression levels of *C. reinhardtii* genes have been experimentally characterized under numerous conditions using high-throughput methods such as RNA-seq [[26,27], unpublished data (Castruita M., et al.)]. These expression data were compiled and analyzed to determine which genes are over- and under-expressed in each experimental condition. The expression data was preprocessed to normalize the counts for uniquely mappable reads in any experiment. Genes exhibiting greater than a two-fold change in expression compared to average expression across all conditions with a Poisson cumulative p-value of less than 0.05 were considered differentially expressed. Using this data, *C. reinhardtii* genes were associated with conditions in which they were over- and under-expressed.

The compiled expression data was also analyzed to find functionally related genes based on their expression levels across the different experimental conditions [[26,27], unpublished data (Castruita M., et al.)]. Genes demonstrating low variance of expression across all samples were not considered. This analysis was performed for three representations of the expression data: absolute counts, log counts, and log ratios of expression. By this method, *C. reinhardtii* genes are each associated with 100 genes with the most similar expression patterns to determine potentially functionally related genes.

Gene Identifier Conversion

Due to the existence of several protein identifier types (FM3.1, FM4, Au5, Au10.2), different identifiers are associated with an individual protein within the *Chlamydomonas* genome. In order to extend annotations from one identifier type to another, matching protein identifiers are deduced by sequence similarity filtering for mutual best hits between identifiers using BLAST. Matching identifiers with 100% sequence coverage are kept, and the rest of the mutual best hits are filtered to include only those proteins with matches with at least

75% coverage. Potential ambiguities involving proteins similar to multiple other proteins are resolved by considering only the reciprocal best hit from the BLAST query in the opposite direction. The information derived by this analysis is used to convert gene identifiers between different types, which allows the Algal Annotation Tool to work with multiple protein identifier types.

Web-Based Interface and Updates

The web interface of the Algal Functional Annotation Tool consists of a set of portals that give access to the different types of analyses available. Results are shown within expandable/collapsible HTML tables that display annotation information along with the statistical results of the analysis. When expanded, the results table shows which gene identifiers contain a specific annotation along with further information regarding matching gene identifiers and BLAST E-values. Updates to the Algal Functional Annotation Tool are semi-automated using a set of Perl scripts that parse and process updated flat files from the various integrated annotation databases at regular intervals. Currently, functional data from the primary annotation databases is set to be updated every 4 months.

Utility and Discussion

Comprehensive, Integrated Data-Mining Environment

The Algal Functional Annotation Tool is composed of three main components - functional term enrichment tests (which are separated by type), a batch gene identifier conversion tool, and a gene similarity search tool. A 'Quick Start' analysis is provided from the front page, featuring enrichment analysis using a sample set of databases containing the richest set of annotations (Figure 1). From any page, the sidebar provides access to the 'Quick Start' function of the tool.

Numerous other enrichment analyses - including enrichment using pathway, ontology, protein family, or differential expression data - are available within the Algal Functional Annotation Tool. Enrichment results are always sorted by hypergeometric p-value and whenever possible contain links to the primary database's entry for that annotation or to the protein page of the gene identifier. The number of hits to a certain annotation term are also displayed alongside the p-value, and results may always be expanded to show additional details, such as the specific gene IDs within the list matching a certain annotation (Figure 2). These results

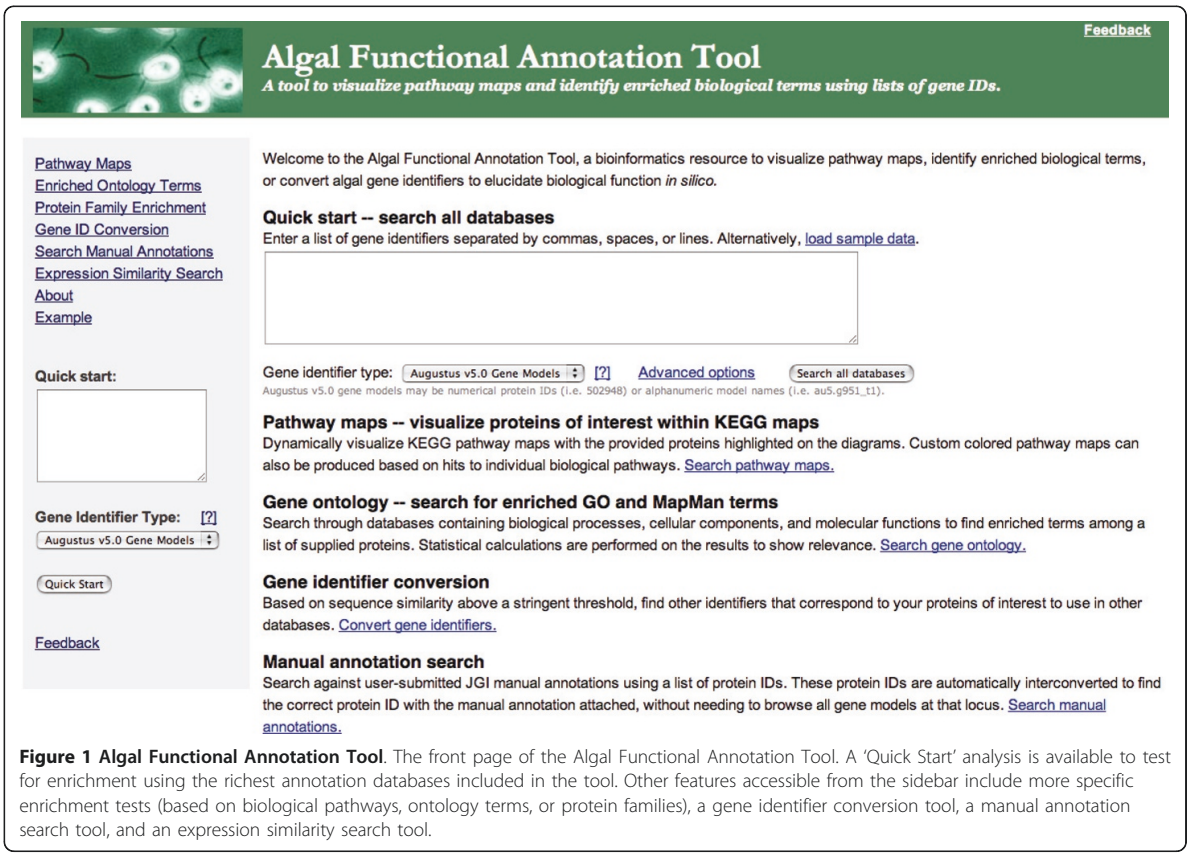
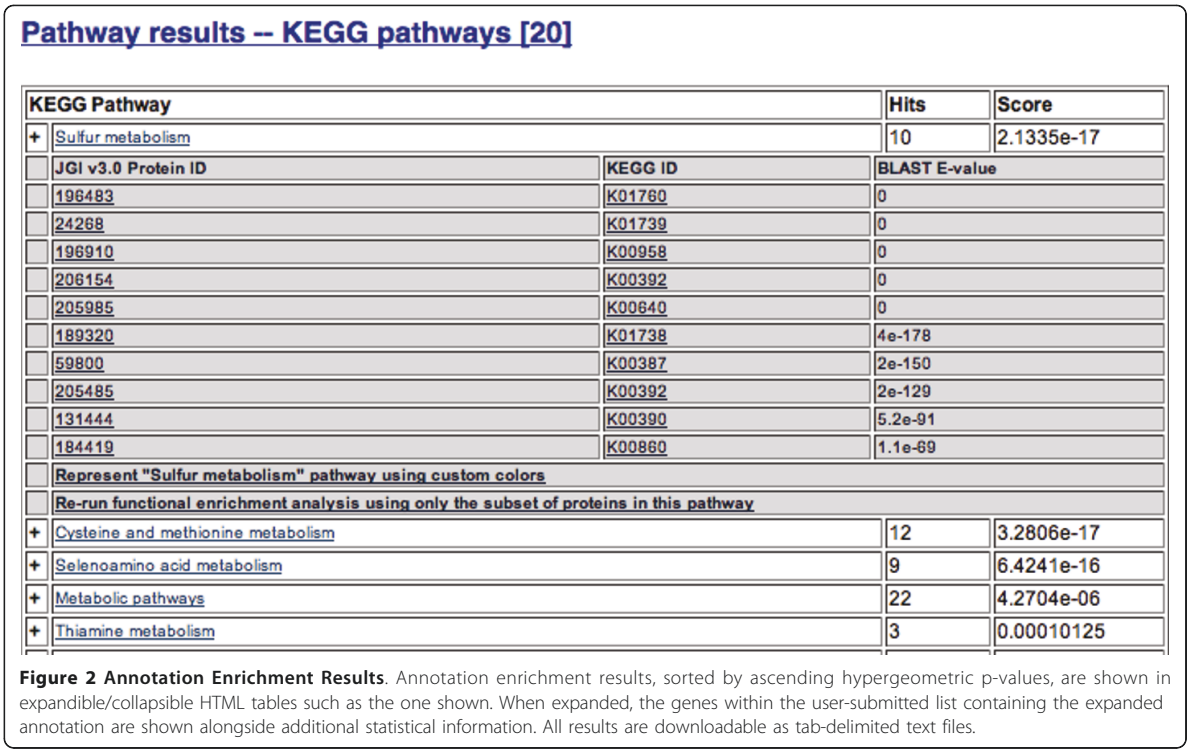


Figure 1 Algal Functional Annotation Tool. The front page of the Algal Functional Annotation Tool. A 'Quick Start' analysis is available to test for enrichment using the richest annotation databases included in the tool. Other features accessible from the sidebar include more specific enrichment tests (based on biological pathways, ontology terms, or protein families), a gene identifier conversion tool, a manual annotation search tool, and an expression similarity search tool.



Pathway results -- KEGG pathways [20]

	KEGG Pathway		Hits	Score
+	Sulfur metabolism		10	2.1335e-17
	JGI v3.0 Protein ID	KEGG ID	BLAST E-value	
	196483	K01760	0	
	24268	K01739	0	
	196910	K00958	0	
	206154	K00392	0	
	205985	K00640	0	
	189320	K01738	4e-178	
	59800	K00387	2e-150	
	205485	K00392	2e-129	
	131444	K00390	5.2e-91	
	184419	K00860	1.1e-69	
	Represent "Sulfur metabolism" pathway using custom colors			
	Re-run functional enrichment analysis using only the subset of proteins in this pathway			
+	Cysteine and methionine metabolism		12	3.2806e-17
+	Selenoamino acid metabolism		9	6.4241e-16
+	Metabolic pathways		22	4.2704e-06
+	Thiamine metabolism		3	0.00010125

Figure 2 Annotation Enrichment Results. Annotation enrichment results, sorted by ascending hypergeometric p-values, are shown in expandible/collapsible HTML tables such as the one shown. When expanded, the genes within the user-submitted list containing the expanded annotation are shown alongside additional statistical information. All results are downloadable as tab-delimited text files.

are downloadable as tab-delimited text files which may then be further analyzed or used in conjunction with other databases.

Dynamic visualization of KEGG pathway maps may be accessed from the results table for KEGG pathway enrichment by clicking on any pathway name. The proteins in the list that are members of the particular biological pathway will appear in red, while those proteins existing in *Chlamyomonas reinhardtii* but not in the list appear in green (Figure 3). Alternatively, by expanding the pathway results and following the link at the bottom, the user may select a custom color scheme for visualizing the proteins on pathway maps. These custom color schemes may be designed on a gene-by-gene basis (choosing colors individually for genes) or in a group-by-group fashion (such as choosing a color for those proteins found within the organism but not in the gene list).

A list of genes may also be converted into a list of gene identifiers of another type. This feature allows easy transformation of gene IDs into corresponding models for use in other databases that may have additional annotation information. Additionally, the resulting list of gene identifiers may be used as a new starting point for enrichment analysis. Because of the different annotations associated with other gene identifier types (albeit of the same proteins), enrichment results using a converted set of gene IDs may yield new biological information.

The gene similarity search tool, the third component of the Algal Functional Annotation Tool, accepts single genes and returns functionally related genes (based on gene expression across different experimental conditions) using user-specified distance metrics and thresholds. Presently, functionally related genes may be determined using correlation distance based on absolute counts, log counts, or log ratios of expression. The results page shows the original query gene at the top in gray and any resulting genes, sorted by similarity, are shown below the query gene (Figure 4). A colormap based on gene expression is generated for the different genes across the conditions, and this colormap may be changed to display absolute expression, log expression, or log ratios of expression. The distance between any gene and the original query gene is displayed by hovering the mouse over the gene identifier of interest. Quantitative expression data (e.g. absolute counts) are provided for each experiment by hovering over the colormap. Whenever a description of a gene is available, this is displayed when hovering over the gene identifier as well. Links to external databases (e.g. JGI, KEGG) providing more information about the genes are provided with the results.

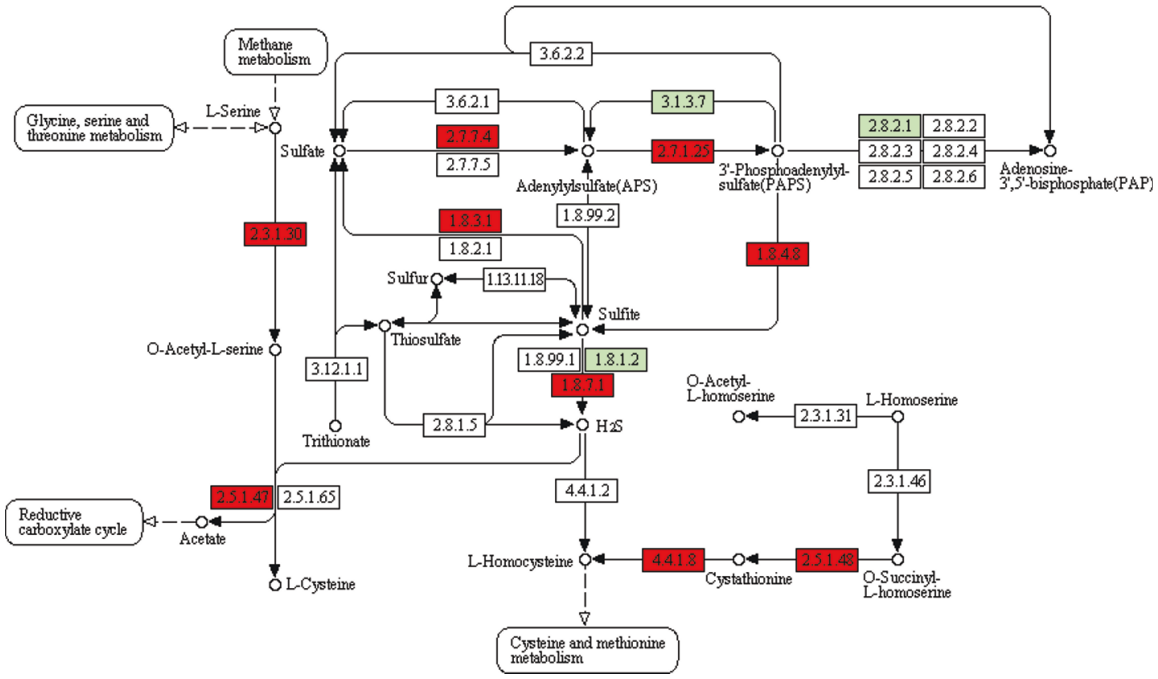


Figure 3 Dynamic Visualization of Gene Lists onto KEGG Pathway Maps. Dynamic KEGG pathway maps may be visualized to show the different proteins within a user-submitted gene list. Shown is the 'Sulfur Metabolism' dynamic pathway with the matching proteins submitted highlighted in red. In this example, the submitted gene list is drawn from literature characterizing *Chlamydomonas* under sulfur-deprived conditions [26].

Ability to Re-Run Analysis for Subsets of Genes

Once a gene list is supplied and enrichment results have been returned, a subset of genes corresponding to those that contain a particular annotation may be isolated and re-run through the tool to be analyzed as a separate, smaller gene list. This allows users to select a particularly interesting group of functionally related genes and isolate them to see if they are also enriched for other functional terms. This also allows the user to prune large gene lists into more focused lists of functionally similar genes and removing some of the inherent noise associated with high-throughput experimental techniques and their resulting gene lists. This feature of the tool may be accessed by expanding the enrichment results of a particular annotation and selecting to re-run the analysis using only that subset of proteins. From this step, users may select which database types to query for enrichment (e.g. pathway, ontology, protein family).

Expanded Annotation Coverage

The methods described to compensate for the incomplete annotation coverage of *Chlamydomonas reinhardtii* genes resulted in the addition of a vast number of unique annotations to the genome. While there is a strong overlap between pre-existing annotations and those assigned by inference, many new terms have also been added. The annotations derived by orthology, however, are not mixed with the annotations attained directly to decrease the possibility of false positive associations of functional terms that may distort the analysis, and to permit a comparison with the functional terms derived directly from the *Chlamydomonas* annotation.

Example - Sulfur-Related Genes

Using a filtered list of *C. reinhardtii* genes derived from transcriptome sequencing of the green alga under sulfur-depleted conditions [26], the Algal Functional Annotation Tool found enrichment for annotations related to sulfur metabolism, cysteine and methionine metabolism, and sulfur compound biosynthesis. For each annotation, the results may be expanded to reveal the genes containing that particular annotation. Furthermore, there is significant overlap between terms directly assigned to *C. reinhardtii* proteins and those inferred from *A. thaliana* orthology. Visualization of the sulfur metabolism KEGG pathway shows that a majority of the enzymes involved

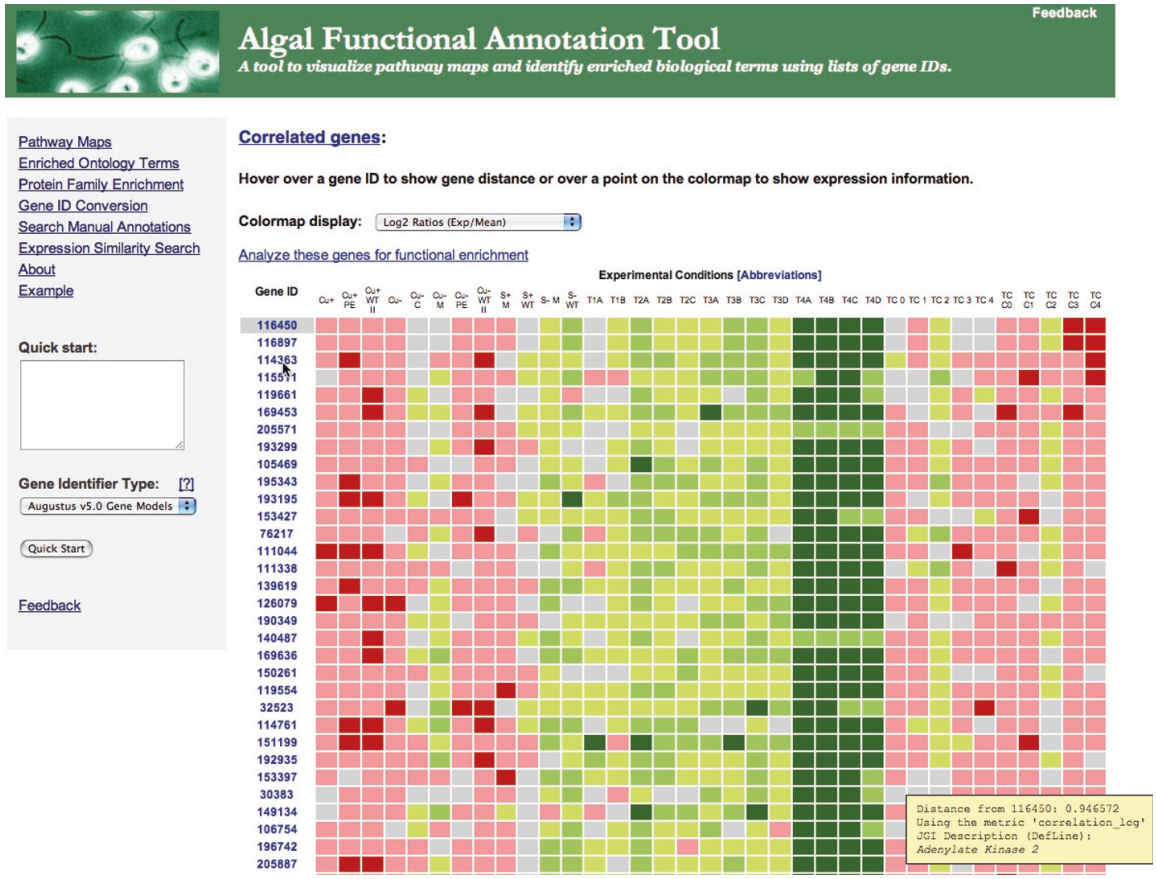


Figure 4 Expression Similarity Search Tool Results. An example of the results from the Gene Similarity Search Tool. Pairwise distances between resulting genes and the submitted gene are shown in the lower right corner when the mouse hovers over a gene of interest. Whenever applicable, a short description of the resulting gene is also shown when hovering over a gene. Expression data is shown when hovering over a point of the colormap.

in this biological process is in the sample list, and the reactions they catalyze may be seen on the pathway map. The results for any enrichment analysis may be downloaded as a tab-delimited text file. Taking a gene found to be associated with the KEGG pathway 'Sulfur metabolism' by this enrichment analysis (JGI v. 3 ID 206154) as a starting input into the gene similarity search tool, the genes corresponding to sulfate transporter, methionine synthase reductase, and cysteine dioxygenase were found within the top 15 results using the correlation metric between log counts.

Future Directions

As with all tools that integrate data from multiple external sources, the power of analysis using the Algal Functional Annotation Tool is ultimately limited by the quality of the annotations within the primary databases. With the steady growth of knowledge in these annotation databases, the utility of the analyses provided is expected to increase in the future as more biological associations are assigned to genes. Additionally, as *Chlamydomonas reinhardtii* genes continue to be experimentally characterized, the assignment of manual annotations will also fill in the gaps left by automated annotation assignment and thus expand the annotation coverage throughout the genome, further improving the results generated by our portal. Lastly, the extensible nature of the Algal Functional Annotation Tool will allow us to add other algal organisms in the future using the same platform so that genomic data from other algal model organisms may be analyzed in a similar fashion as that currently available for *Chlamydomonas reinhardtii*.

Conclusions

The Algal Functional Annotation Tool is intended as a comprehensive analysis tool to elucidate biological

meaning from gene lists derived from high-throughput experimental techniques. Annotation sets from a number of biological databases have been pre-processed and assigned to gene identifiers of the green alga *Chlamydomonas reinhardtii*, and this annotation data may be explored in multiple ways, including the use of enrichment tests designed for large gene lists. Furthermore, the site enables the visualization of proteins within pathway maps. Using several methods, such as inferring annotations from orthologous proteins of other organisms, the initially sparse annotation coverage of *C. reinhardtii* is alleviated, allowing for a more effective functional term enrichment analysis. Other functions of the tool include a batch gene identifier conversion tool and a manual annotation search tool. Lastly, similar genes based on expression across several conditions may be explored using the gene similarity search tool.

Availability and Requirements

Project name: Algal Functional Annotation Tool

- Public web service: http://pathways.mcdb.ucla.edu; Free and no registration.
- Programming language: Perl/CGI
- Database: MySQL
- Software License: GNU General Public License

List of Abbreviations Used

API: Application Programming Interface; BLAST: Basic Local Alignment Search Tool; CGI: Common Gateway Interface; DAVID: Database for Annotation, Visualization, and Integrated Discovery; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; JGI: Joint Genome Institute; SOAP: Simple Object Access Protocol.

Acknowledgements and Funding

We acknowledge funding of this work by the US Department of Energy under contract DE-EE0003046 awarded to the National Alliance for Advanced Biofuels and Bioproducts.

Author details

[1]Department of Molecular, Cell, and Developmental Biology, University of California, Los Angeles, CA, USA. [2]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA. [3]Institute of Genomics and Proteomics, University of California, Los Angeles, CA, USA.

Authors' contributions

MP conceived the analysis and main features of the tool. DL wrote and tested the code, constructed the annotation database, designed the user interface, and wrote the initial draft of the manuscript. SC provided the implementation of the hypergeometric distribution function. DC provided Pfam data and compiled the expression data. SM provided access to the expression data and tested the tool. All authors read, edited and approved the final manuscript.

Received: 8 February 2011 Accepted: 12 July 2011
Published: 12 July 2011

References

1. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, , **38 Database:** D355-360.
2. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2010, , **38 Database:** D473-479.
3. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2010, , **38 Database:** D211-222.
4. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4(1)**:44-57.
5. **Ingenuity Pathway Analysis (IPA), Ingenuity Systems.** [http://www.ingenuity.com].
6. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Marechal-Drouard L, *et al*: **The Chlamydomonas genome reveals the evolution of key animal and plant functions.** *Science* 2007, **318(5848)**:245-250.
7. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S, Partensky F, Degroeve S, Echeynie S, Cooke R, *et al*: **Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features.** *Proc Natl Acad Sci USA* 2006, **103(31)**:11647-11652.
8. Palenik B, Grimwood J, Aerts A, Rouze P, Salamov A, Putnam N, Dupont C, Jorgensen R, Derelle E, Rombauts S, *et al*: **The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation.** *Proc Natl Acad Sci USA* 2007, **104(18)**:7705-7710.
9. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV, *et al*: **Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas.** *Science* 2009, **324(5924)**:268-272.
10. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S, Pangilinan J, Polle J, *et al*: **The Chlorella variabilis NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex.** *Plant Cell* 2010, **22(9)**:2943-2955.
11. **Chlamydomonas reinhardtii v4.0, Joint Genome Institute.** [http://genome.jgi-psf.org/Chlre4/].
12. Rupprecht J: **From systems biology to fuel–Chlamydomonas reinhardtii as a model for a systems biology approach to improve biohydrogen production.** *J Biotechnol* 2009, **142(1)**:10-20.
13. Grossman AR, Croft M, Gladyshev VN, Merchant SS, Posewitz MC, Prochnik S, Spalding MH: **Novel metabolism in Chlamydomonas through the lens of genomics.** *Curr Opin Plant Biol* 2007, **10(2)**:190-198.
14. Beer LL, Boyd ES, Peters JW, Posewitz MC: **Engineering algae for biohydrogen and biofuel production.** *Curr Opin Biotechnol* 2009, **20(3)**:264-271.
15. Ghirardi ML, Dubini A, Yu J, Maness PC: **Photobiological hydrogen-producing systems.** *Chem Soc Rev* 2009, **38(1)**:52-61.
16. Hemschemeier A, Melis A, Happe T: **Analytical approaches to photobiological hydrogen production in unicellular green algae.** *Photosynth Res* 2009.
17. Finazzi G, Moreau H, Bowler C: **Genomic insights into photosynthesis in eukaryotic phytoplankton.** *Trends Plant Sci* 2010, **15(10)**:565-572.
18. Kruse O, Hankamer B: **Microalgal hydrogen production.** *Curr Opin Biotechnol* 2010, **21(3)**:238-243.
19. Scott SA, Davey MP, Dennis JS, Horst I, Howe CJ, Lea-Smith DJ, Smith AG: **Biodiesel from algae: challenges and prospects.** *Curr Opin Biotechnol* 2010, **21(3)**:277-286.
20. Radakovits R, Jinkerson RE, Darzins A, Posewitz MC: **Genetic engineering of algae for enhanced biofuel production.** *Eukaryot Cell* 2010, **9(4)**:486-501.
21. Eberhard S, Finazzi G, Wollman FA: **The dynamics of photosynthesis.** *Annu Rev Genet* 2008, **42**:463-515.
22. Rochaix JD: **Genetics of the biogenesis and dynamics of the photosynthetic machinery in eukaryotes.** *Plant Cell* 2004, **16(7)**:1650-1660.
23. Harris EH: **Chlamydomonas as a model organism.** *Annu Rev Plant Physiol Plant Mol Biol* 2001, **52**:363-406.
24. Marshall WF: **Basal bodies platforms for building cilia.** *Curr Top Dev Biol* 2008, **85**:1-22.
25. Scholey JM, Anderson KV: **Intraflagellar transport and cilium-based signaling.** *Cell* 2006, **125(3)**:439-442.
26. Gonzalez-Ballester D, Casero D, Cokus S, Pellegrini M, Merchant SS, Grossman AR: **RNA-seq analysis of sulfur-deprived Chlamydomonas cells reveals aspects of acclimation critical for cell survival.** *Plant Cell* 2010, **22(6)**:2058-2084.

27. Miller R, Wu G, Deshpande RR, Vieler A, Gartner K, Li X, Moellering ER, Zauner S, Cornish AJ, Liu B, *et al*: **Changes in transcript abundance in Chlamydomonas reinhardtii following nitrogen deprivation predict diversion of metabolism.** *Plant Physiol* 2010, **154(4)**:1737-1752.
28. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, *et al*: **Reactome knowledgebase of human biological pathways and processes.** *Nucleic Acids Res* 2009, , **37 Database:** D619-622.
29. **Chlamydomonas reinhardtii v3.0, Joint Genome Institute.** [http://genome.jgi-psf.org/Chlre3/].
30. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Res* 2003, **13(9)**:2129-2141.
31. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1)**:25-29.
32. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, , **37 Database:** D211-215.
33. Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M: **MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes.** *Plant J* 2004, **37(6)**:914-939.
34. Huang da W, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37(1)**:1-13.
35. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al*: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.

doi:10.1186/1471-2105-12-282
Cite this article as: Lopez *et al.*: Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. *BMC Bioinformatics* 2011 **12**:282.

Chapter 4:

Dynamic changes in the transcriptome and methylome of *Chlamydomonas reinhardtii* throughout its life cycle

Dynamic Changes in the Transcriptome and Methylome of *Chlamydomonas reinhardtii* throughout Its Life Cycle[1]

David Lopez, Takashi Hamaji, Janette Kropat, Peter De Hoff[2], Marco Morselli, Liudmilla Rubbi, Sorel Fitz-Gibbon, Sean D. Gallaher, Sabeeha S. Merchant, James Umen, and Matteo Pellegrini*

Molecular Biology Institute (D.L.), Department of Molecular, Cell, and Developmental Biology (D.L., M.M., L.R., S.F.-G., M.P.), Department of Chemistry and Biochemistry (J.K., S.F.-G., S.D.G., S.S.M.), and Institute for Genomics and Proteomics (S.S.M., M.P.), University of California, Los Angeles, California 90095; Donald Danforth Plant Science Center, St. Louis, Missouri 63132 (T.H., J.U.); and Salk Institute for Biological Studies, La Jolla, California 92037 (P.D.H.)

ORCID IDs: 0000-0001-9363-9329 (D.L.); 0000-0002-2801-2148 (T.H.); 0000-0001-8952-3286 (J.K.); 0000-0002-0163-7535 (P.D.H.); 0000-0003-3351-5791 (M.M.); 0000-0002-9236-202X (L.R.); 0000-0001-7090-5719 (S.F.-G.); 0000-0002-9773-6051 (S.D.G.); 0000-0002-2594-509X (S.S.M.); 0000-0003-4094-9045 (J.U.); 0000-0001-9355-9564 (M.P.).

The green alga *Chlamydomonas reinhardtii* undergoes gametogenesis and mating upon nitrogen starvation. While the steps involved in its sexual reproductive cycle have been extensively characterized, the genome-wide transcriptional and epigenetic changes underlying different life cycle stages have yet to be fully described. Here, we performed transcriptome and methylome sequencing to quantify expression and DNA methylation from vegetative and gametic cells of each mating type and from zygotes. We identified 361 gametic genes with mating type-specific expression patterns and 627 genes that are specifically induced in zygotes; furthermore, these sex-related gene sets were enriched for secretory pathway and alga-specific genes. We also examined the *C. reinhardtii* nuclear methylation map with base-level resolution at different life cycle stages. Despite having low global levels of nuclear methylation, we detected 23 hypermethylated loci in gene-poor, repeat-rich regions. We observed mating type-specific differences in chloroplast DNA methylation levels in plus versus minus mating type gametes followed by chloroplast DNA hypermethylation in zygotes. Lastly, we examined the expression of candidate DNA methyltransferases and found three, *DMT1a*, *DMT1b*, and *DMT4*, that are differentially expressed during the life cycle and are candidate DNA methylases. The expression and methylation data we present provide insight into cell type-specific transcriptional and epigenetic programs during key stages of the *C. reinhardtii* life cycle.

Chlamydomonas reinhardtii is a unicellular, biflagellate species of green alga found primarily in fresh water and soil (Harris et al., 2009). *C. reinhardtii* is an important reference organism for diverse eukaryotic cellular and metabolic processes, including photosynthetic biology (Rochaix, 2001), flagellar function and biogenesis (Silflow and Lefebvre, 2001), nutrient homeostasis (Grossman, 2000; Merchant et al., 2006; Glaesener et al., 2013), and sexual cycles (Goodenough et al., 2007). The nuclear and chloroplast genomes of *C. reinhardtii* have been fully sequenced, enabling genomic and epigenomic analyses (Maul et al., 2002; Merchant et al., 2007). The approximately 112-Mb haploid *C. reinhardtii* nuclear genome comprises 17 chromosomes. The circular chloroplast DNA (cpDNA) genome is 203 kb and present in 80 to 100 copies per cell that are organized into eight to 10 nucleoprotein complexes called nucleoids, which are distributed through the stroma.

Like many unicellular eukaryotes, *C. reinhardtii* has a biphasic life cycle where haploid cells can reproduce vegetatively by mitotic division or, alternatively, undergo a sexual cycle. Vegetative cells can propagate indefinitely when provided with nutrients and light. Upon nitrogen starvation, however, cells stop dividing and differentiate into gametes whose mating type (plus or minus) is determined genetically by an approximately 300-kb mating type locus on chromosome 6, with two haplotypes, *MT+* and *MT−* (Umen, 2011; De Hoff et al.,

[1] This work was supported by the National Institutes of Health (grant nos. R24 GM092473 and R01 GM078376 and T32 Training Fellowship in Genome Analysis no. 5T32HG002536–13 to D.L.); by the Office of Science (Biological and Environmental Research), U.S. Department of Energy (grant no. DE–FC02–02ER63421); by the Eugene V. Cota-Robles Fellowship and the Fred Eiserling and Judith Lengyel Doctoral Fellowship (to D.L.); and by the Japan Society for the Promotion of Science (Postdoctoral Fellowship for Research Abroad no. 26–495 to T.H.).

[2] Present address: Synthetic Genomics, 11149 North Torrey Pines Road, La Jolla, CA 92037.

* Address correspondence to matteop@mcdb.ucla.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Matteo Pellegrini (matteop@mcdb.ucla.edu).

M.P., J.U., and S.S.M. designed the experiment; J.K., M.M., L.R., and P.D.H. collected the samples and created the sequence libraries in preparation for sequencing; D.L. and T.H. aligned the sequence data and performed all bioinformatic analyses, with the exception of the analysis of the genomic data and identification of structural variants, which were carried out by S.F.-G. and S.D.G.; D.L., J.U., M.P., and S.S.M. wrote the article.

www.plantphysiol.org/cgi/doi/10.1104/pp.15.00861

2013). Gametes express a set of mating-related proteins that are different between minus and plus cells and that allow cells of opposite mating type to recognize each other and fuse to form a quadriflagellate zygote. Upon fertilization, the heterodimeric KNOX/BELL-type homeodomain proteins gamete-specific minus (GSM1) and gamete-specific plus (GSP1) initiate a zygote-specific developmental program that includes flagellar resorption, fusion of organelles including nuclei and chloroplasts, destruction of *MT*− cpDNA, and secretion of a thick, environment-resistant cell wall that protects the zygospore from cold, desiccation, and other environmental stresses (Cavalier-Smith, 1976; Catt, 1979; Grief et al., 1987; Brawley and Johnson, 1992; Goodenough et al., 2007; Lee et al., 2008). Upon return to favorable conditions of light and nutrients, zygospores undergo meiosis to produce four haploid progeny (two *MT*+ and two *MT*−) that can reenter the vegetative life cycle. While nuclear loci segregate in a Mendelian pattern of 2:2, both chloroplast and mitochondrial genomes are inherited uniparentally, with cpDNA inherited from the *MT*+ parent and mitochondrial DNA from the *MT*− parent (Nakamura, 2010; Nishimura, 2010).

While previous high-throughput expression studies have focused on the transcriptional programs underlying processes such as nutrient deprivation (Nguyen et al., 2008; González-Ballester et al., 2010; Toepel et al., 2011, 2013; Schmollinger et al., 2014), environmental responses (Simon et al., 2008, 2013; Matsuo et al., 2011; Fang et al., 2012), flagellar biogenesis (Albee et al., 2013), lipid accumulation (Miller et al., 2010; Boyle et al., 2012; Lv et al., 2013), and diurnal rhythms (Idoine et al., 2014; Panchy et al., 2014), only a few studies have explored the genome-wide transcriptional and epigenetic changes associated with the sexual cycle (Kubo et al., 2008; Ning et al., 2013; Aoyama et al., 2014). Several genes expressed in the early zygote, termed *EZY* genes, have predicted functions related to cell wall production, vesicular transport, and secretion (Ferris and Goodenough, 1987; Ferris et al., 2002; Kubo et al., 2008). A separate analysis of zygospore transcripts following light-induced germination revealed the up-regulation of photosynthetic and Met synthesis pathways (Aoyama et al., 2014).

DNA methylation studies have also been conducted on both the nuclear and chloroplast genomes (Hattman et al., 1978; Dyer, 1982). cpDNA methylation has been studied more extensively and shows dramatic changes in 5-methylcytosine (5meC) content at different stages of the *C. reinhardtii* life cycle. Vegetative cells have low levels of 5meC in cpDNA, while gametes show a substantial increase within cpDNA (12% 5meC in *MT*+ gamete cells and 4% in *MT*–; Royer and Sager, 1979; Feng and Chiang, 1984). In zygotes, *MT*− cpDNA is eliminated while *MT*+ cpDNA becomes hypermethylated. While differential cpDNA methylation was once thought to be part of a restriction-methylation system regulating uniparental inheritance (Burton et al., 1979), this model is unlikely, since loss of cpDNA methylation in *MT*+ cells does not result in its destruction in zygotes (Umen and Goodenough, 2001), and ectopic methylation of *MT*− cpDNA does not spare it from destruction (Bolen et al., 1982). However, previous studies are consistent with a role for 5meC in promoting cpDNA replication upon zygote germination, which can influence the amount of residual *MT*− cpDNA that is inherited by exceptional progeny (Umen and Goodenough, 2001; Nishiyama et al., 2004). An alternative proposed mechanism involves the digestion of *MT*− cpDNA by differentially localized or activated nucleases that are methylation insensitive early in zygote development before chloroplast fusion (Nishimura et al., 2002).

Several methyltransferase enzymes that modify cpDNA have been investigated biochemically (Sano et al., 1981), and one candidate chloroplast methyltransfersase gene has been cloned (Nishiyama et al., 2002, 2004). Since that time, the genome sequence of *C. reinhardtii* has become available (Merchant et al., 2007) and extensively annotated (Blaby et al., 2014) so that a comprehensive identification of genes encoding DNA methyltransferases can be undertaken.

Few studies have focused on the role of nuclear cytosine methylation in *C. reinhardtii*, but previous work has shown that induced silencing of nuclear transgenes does not correlate with transgene cytosine methylation levels, leaving open the question of what role cytosine methylation plays in chromatin structure and gene expression in *C. reinhardtii* (Cerutti et al., 1997). To date, the absence of detailed methylation maps has precluded a clear view of methylation patterns in the nuclear and chloroplast genomes of *C. reinhardtii*.

Here, we have performed RNA sequencing (RNA-seq)-based transcriptome analysis and bisulfite DNA sequencing of *C. reinhardtii* at different life cycle stages. We identify sex- and mating-related changes in gene expression, including genes that are preferentially expressed in gametes of each mating type or in zygotes. We generated a high-resolution map of nuclear and chloroplast cytosine methylation during the life cycle and identified candidate DNA methyltransferases whose expression profiles correlate with dynamic changes in cpDNA methylation patterns.

RESULTS

To quantify nuclear gene expression and 5-cytosine DNA methylation throughout the *C. reinhardtii* life cycle, we collected RNA and DNA samples at various stages for RNA and whole-genome bisulfite sequencing (Fig. 1). Samples were collected from vegetative and gametic cells of both mating types and zygotic stages to enable multiple comparisons. The design of the experiment was for matched DNA and RNA samples, but the RNA protocol was modified to avoid degradation.

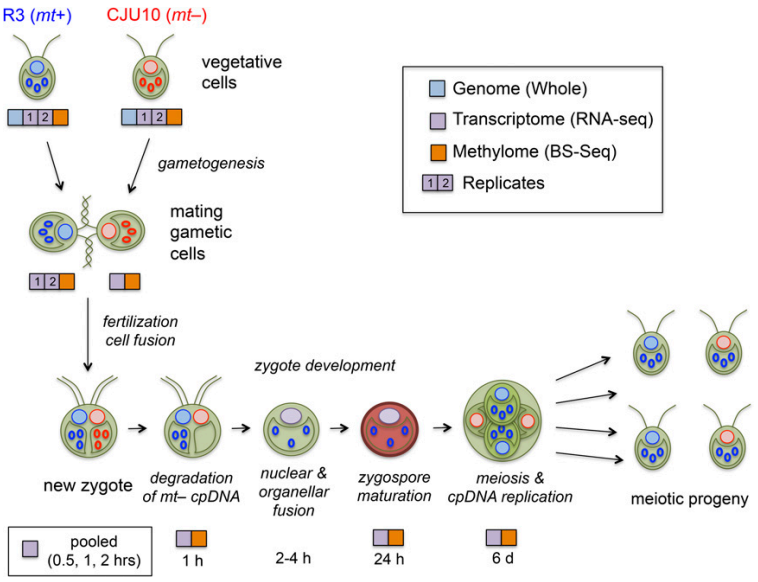


Figure 1. *C. reinhardtii* sexual life cycle and sequenced samples. Vegetative *C. reinhardtii* cells of each mating type (*MT+* and *MT−*) can be induced to undergo gametogenesis by nitrogen starvation. Gametes of opposite mating type recognize each other through flagellar adhesion and fuse to form a diploid zygote. During zygote maturation, *MT−* cpDNA is eliminated, flagella are resorbed, and a thick zygote cell wall forms. Upon return to nitrogen and light, zygospores undergo meiosis to form four haploid progeny (two of each mating type, all containing uniparentally inherited *MT+* parental cpDNA) that reenter the vegetative cycle. Colored boxes designate samples and material sequenced. Blue and red unfilled circles represent *MT+* and *MT–* chloroplast genomes, respectively. Filled pink and light blue circles represent nuclear DNA from *MT+* and *MT–* strains, respectively. BS-Seq, Bisulfite sequencing.

We note that for transcriptome studies, our vegetative samples were prepared in a different manner from those used for typical nitrogen-starvation studies. We grew all cultures to saturation on solid agar medium and then resuspended cells at high density (approximately 2×10^7 mL^{-1}) under illumination in liquid medium (high-salt medium) with or without nitrogen for approximately 3 to 5 h. Under these conditions, neither culture grew measurably, but the cultures without nitrogen expressed the gametic program and efficiently mated at greater than 90% efficiency, while the cultures with nitrogen could not mate and did not express gametic marker genes (see below). The advantage of this procedure is that it minimizes the differences between the plus and minus nitrogen cultures that would normally be attributable to growth rate differences and thereby allows more reliable identification of mating-related gene expression. For DNA methylation studies, the vegetative and gametic samples were obtained from growing nitrogen-replete and nitrogen-starved samples, respectively, as described in "Materials and Methods." Actively growing cultures were used for studies of methylation in vegetative cells, since agar plate-grown cells are already partially gametic and would require additional rounds of division in order to remove preexisting methylation.

Sequence Polymorphisms between R3 (*MT+*) and CJU10 (*MT–*) Parental Strains

As a prelude to transcriptome and methylome analyses, we cataloged the genetic differences (single-nucleotide variants, insertions, and deletions) between our two parental strains using genome resequencing (Fig. 2A; Gallaher et al., 2015). In total, the two strains differ by 0.16% in their nuclear genomes, and most of these differences are single-nucleotide variants. Of all the variants, 98.2% are localized to two regions, one of length 2.2 Mb on chromosome 17 and one of 2 Mb encompassing the mating locus on chromosome 6 (Fig. 2B), where mating-type haplotype differences have been observed previously (De Hoff et al., 2013). Additionally, the chloroplast genome of the CJU10 strain contains an insertion adjacent to the *ATP synthase subunit betaaadA* (*atpB*) gene of a spectinomycin resistance marker (*aminoglycoside-3′-adenylyltransferase* [*aadA*]) flanked by the *Rubisco large subunit* (*rbcL*) 5′ promoter and *photosystem II CP43* (*psbC*) 3′ untranslated region (Fig. 2C; Goldschmidt-Clermont, 1991; Umen and Goodenough, 2001).

Gamete and Zygote-Specific Genes

Following the quantitation of RNA-seq data for all of our samples, we identified genes with mating type- and zygote-specific expression patterns using a series of filters to screen for genes matching the expression patterns of known gametic and zygotic genes. We required that mating type-specific genes be expressed in gametes at least 4-fold higher than in vegetative cells of the same mating type and at least 10-fold higher than in gametes or vegetative cells of the opposite mating type. For zygote-specific genes, we required that expression be at least 4-fold higher in early zygotes than in any other sample. Using these criteria, we identified 293 and 68 genes whose expression is specific to plus and minus gametes, respectively, and 627 genes whose expression is specific to zygotes (Fig. 3A; Supplemental Table S1). Genes whose expression is known to be gamete or zygote specific, including *GSM1* (minus gametes), *SAG1* (plus gametes), and *EZY1* (early zygotes), were

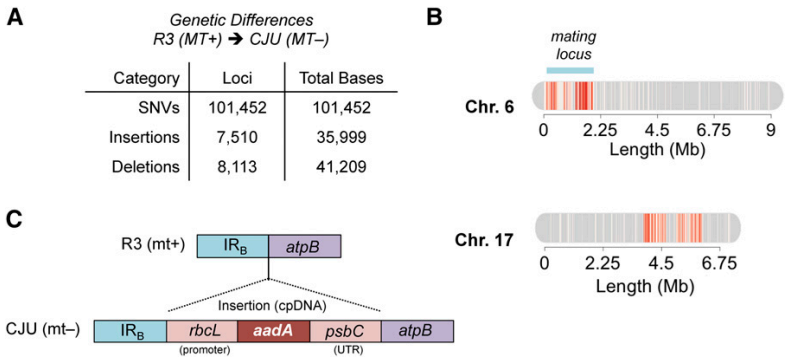


Figure 2. Genetic differences between R3 (*MT+*) and CJU10 (*MT−*) parental strains. A, Genetic differences, categorized by variant type (single-nucleotide variants [SNVs], insertions, and deletions), are shown with the total number of variant loci and total bases. B, Large-scale view of chromosomes 6 and 17, where 98.5% of the identified variants are located. Locations of sequence differences are shown in red. The region of chromosome 6 containing the mating locus is denoted by the blue bar. C, Schematic representation of the antibiotic resistance transgene insertion in CJU10 cpDNA. The *aadA* gene, along with the *rbcL* 5′ promoter and the *psbC* 3′ untranslated region (UTR), is inserted between inverted repeat region B (IR_B) and the endogenous *atpB* gene.

retained by our filters (Armbrust et al., 1993; Kurvari et al., 1998; Ferris et al., 2002, 2005; Lee et al., 2008). Of the 32 previously described zygotic genes (Matters and Goodenough, 1992; Armbrust et al., 1993; Uchida et al., 1993, 1999; Kuriyama et al., 1999; Suzuki et al., 2000; Ferris et al., 2002; Kubo et al., 2008), 25 were in our zygote data set, while the remaining seven (*EZY6*, *EZY15*, *EZY16*, *EZY17*, *EZY21*, *EZY23*, and *Lysozome 1A*) were found to have significant expression in other samples and, therefore, were excluded from the zygote-specific set (Supplemental Tables S1 and S2).

We functionally classified the predicted proteins encoded by genes whose expression was plus gamete specific, minus gamete specific, or early zygote specific. Protein localization prediction revealed the enrichment of putative secretory proteins in plus-gamete (*MT+*) and zygote-specific sets (Fig. 3B). We also assessed the conservation and phylogenetic distribution for homologs of each protein in a set of nested phylogenetic domains that encompass different taxonomic levels from cellular organisms (prokarya, archaea, and eukarya) to the single species level of *C. reinhardtii* (Fig. 3C). Zygote and gamete expression groups were enriched for *C. reinhardtii*-specific genes and/or alga-specific genes. In addition, gametic genes were underrepresented for more widely conserved genes (i.e. those with homologs outside of chlorophyte algae). Consistent with these findings, the gametic and zygotic gene lists were also depleted to various extents for GreenCut2 and CiliaCut genes, whose members are associated with conserved photosynthetic and flagella/basal body functions (Merchant et al., 2007; Karpowicz et al., 2011; Heinnickel and Grossman, 2013; Fig. 4, A and B); however, the overall low numbers of genes in these categories precluded obtaining a significant statistical result in all cases but one. Gene Ontology, Kyoto Encyclopedia of Genes and Genomes, and MapMan (Kanehisa and Goto, 2000; Harris et al., 2004; Thimm et al., 2004) classification of predicted proteins encoded by gametic and zygotic genes was performed using the Algal Functional Annotation Tool (Lopez et al., 2011) but had limited utility because of the large number of nonconserved proteins in these groups and incomplete annotations. Nonetheless, we found significant enrichment in two MapMan categories for zygotic genes, cell wall and transport, both of which may relate to the requirement for new cell wall biosynthesis in zygotes (Fig. 4, C and D). Indeed, examination of manually curated early zygotic gene annotations revealed numerous cell wall-related protein-coding genes as described below.

Volvocine cell walls are composed primarily of glycosylated hydroxyproline-rich glycoproteins (HRGPs) that enter the secretory pathway and are exported to the extracellular space where they coassemble (Woessner et al., 1994). The thick and environment-resistant cell walls of zygospores are formed by a specialized set of HRGPs that are synthesized shortly after fertilization (Minami and Goodenough, 1978; Catt, 1979; Grief et al., 1987). Manual annotation and inspection of zygotic up-regulated genes verified the MapMan ontology assignments of cell wall and transport categories as described in Supplemental Table S1. At least 57 zygotic genes are predicted HRGPs or have putative cell wall biogenesis-related functions that include secretion, glycosylation, and metabolism of nucleotide sugars (e.g. UDP-Glc 4-epimerase, pyrophosphorylase, dehydrogenase, dTDP-6-deoxy-L-lyxo-4-hexulose reductase related, and exotosin-like glycosyltransferase) or sugar metabolite transport (e.g. ATP-binding cassette transporter, triose phosphate transporter, and UDP-GlcNAc transporter). Among these were some previously identified early zygotic genes as noted in Supplemental Table S1 (*EZY4*, *EZY11/UDP-glucose:protein transglucosylase1* [*UPT1*; also known as *UPTG1*]/*EZY12/UDP-glucose*

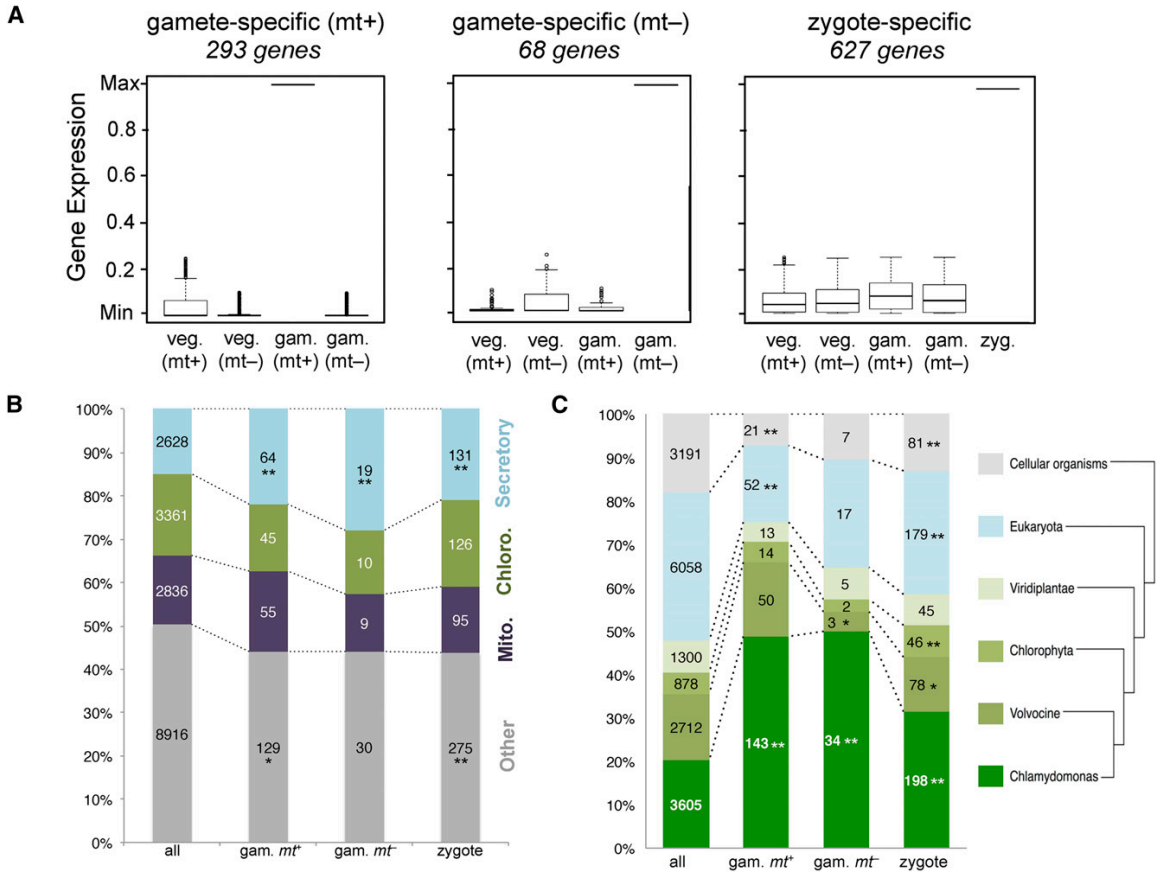


Figure 3. Gamete- and zygote-specific genes. A, Box and whisker plots of gene expression profiles for gamete- and zygote-specific genes. Values are plotted relative to the maximum expression value of each of the genes. A total of 293 and 68 genes were expressed specifically in *MT+* and *MT–* gametes, respectively. A total of 627 genes were expressed specifically in zygotes. B, Localization predictions for proteins encoded by gamete-specific (gam. *mt+* and gam. *mt−*) and zygote-specific genes, compared with all predicted proteins in the *C. reinhardtii* proteome (all). Data are plotted as the fraction predicted for each of four compartments, with total numbers indicated within each graph portion: secretory in blue; chloroplast (Chloro.) in green; mitochondria (Mito.) in purple; and other in gray. Values with asterisks are significantly different from the total proteome distribution (*, $P < 0.05$ and **, $P < 0.01$). C, Predicted protein groups as described in B are plotted according to the taxonomic distribution of encoded proteins from each category. From bottom (*C. reinhardtii*-specific proteins; dark green) to top (all cellular organisms; gray) are increasingly broad phylogenetic distributions. Samples with asterisks indicate significant enrichment or depletion in a taxonomic category relative to the distribution of all proteins (*, $P < 0.05$; and **, $P < 0.01$).

dehydrogenase1, *EZY14*/*triose phosphate transporter14*, and *EZY22*; Kubo et al., 2008).

Besides cell wall and secretory pathway genes, we also noted in the zygote gene set predicted functions that may be related to other zygotic processes, including elimination of *MT–* cpDNA, zygotic cpDNA methylation, and packaging for long-term dormancy, nuclear fusion (karyogamy), chloroplast fusion, and flagellar resorption (Goodenough et al., 2007). These annotations include predicted chloroplast-targeted DNA-binding proteins such as a DNA recombination protein A homolog and predicted nuclease (*EZY19*/Cre07.g314650), chloroplast-targeted DNA methyltransferases (*DMT1A*, *DMT1B*, and *DMT4*; discussed below), and a chloroplast-targeted dynamin (*EZY8*/Cre06.g25065) that may be involved in chloroplast fusion (Kubo et al., 2008). Predicted nucleus-targeted zygotic proteins include several DNA-binding transcription factors (*EZY18*/Cre02.g091550, *Regeneration Protein A* [*RegA*]*/RlsA-like protein7*/Cre14.g617200, and *zygote-specific transcription factor 1A*/Cre17.g719200) such as RLS7 that contains a SAND domain (Duncan et al., 2006, 2007) related to RegA, a repressor of germ cell fate in *Volvox carteri* (Kirk et al., 1999), and several types of chromatin-related proteins (Cre03.g184900, Cre08.g367000, Cre08.g400200, Cre09.g401812, and *histone H1*/Cre13.g567450) that may be involved in nuclear DNA packaging in preparation for zygospore dormancy. Minutes after fertilization and

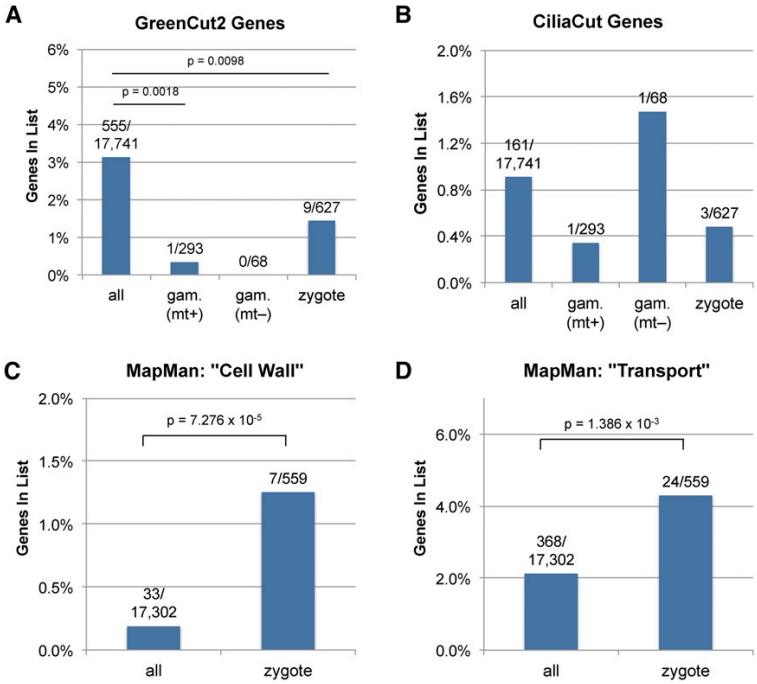


Figure 4. Functional annotation of proteins encoded by gamete- and zygote-specific genes. Functional annotations for gamete- and zygote-specific gene lists are shown as bar plots showing the fraction of total proteins in each group with the specified annotation. The total number of genes in each annotation category versus the number of genes with the described annotation are shown. Significant differences between all genes and gamete- or zygote-specific genes were calculated with the hypergeometric test, and significant *P* values are shown. A, GreenCut2 genes (Karpowicz et al., 2011; Heinnickel and Grossman, 2013). B, CiliaCut genes (Merchant et al., 2007). C, MapMan cell wall-related genes (Thimm et al., 2004). D, MapMan transport-related genes.

prior to flagellar resorption, the four basal bodies and flagella of newly formed zygotes (two from each parent) move to a single apical location via an unknown mechanism. One possible participant in this process could be the striated fiber protein SF-assemblin (Cre07.g332950), a zygote up-regulated gene whose protein product in vegetative cells associates with rootlet microtubules that are proximal to basal bodies and are thought to play a role in the organization of the rootlet structure (Lechtreck et al., 2002). Two other cytoskeletal proteins whose genes are up-regulated in zygotes are a flagella-associated protein of unknown function, FAP79 (Cre04.g217908), and the flagella length regulatory protein LF5/FAP279 (Cre12.g538300; Tam et al., 2013), which may be related to flagellar resorption that begins 2 or 3 h after fertilization. Lastly, we identified a gene for a predicted secreted trypsin-related protease (Cre06.g287750), which could contribute to the rapid postfertilization degradation of gametic plasma membrane surface proteins such as fusion protein1 and generative cell specific1, a process that is thought to restrict polygamy (fusion between more than two gametes; Liu et al., 2010).

DNA Methylation of the Nuclear Genome

We conducted bisulfite sequencing of vegetative, gametic, and zygotic samples to generate DNA methylation profiles. The nuclear genome had an average per-site CG methylation of less than 0.75% in all samples, and this level of methylation did not differ significantly between plus and minus strains or at different life cycle stages (Fig. 5A). However, CG methylation densities greater than 80% were identified for 23 loci that ranged in size between 10 and 22 kb (Fig. 5B; Supplemental Table S3). The highly methylated regions are enriched for repeats, and their overall protein-coding gene densities are significantly lower than average, although there are still genes in these regions (Fig. 5C). In addition, one example where methylation was strain specific is shown in Figure 5D, although most hypermethylated sites did not have strain-specific methylation patterns. The expression levels of genes overlapping hypermethylated loci are not strongly correlated with degree of methylation (Supplemental Table S4).

Chloroplast Methylation Changes during the Life Cycle

In contrast to the relatively stable pattern of cytosine methylation in nuclear DNA, the *C. reinhardtii* chloroplast genome underwent dynamic changes in cytosine methylation throughout the life cycle (Fig. 6A). In the vegetative stage, global per-cytosine methylation was less than 2% for both mating types for all cytosine contexts (CG, CHG, and CHH). After gametogenesis, cpDNA methylation increased in a mating type-dependent manner. *MT*+ gametes had an average of approximately 10% per-site CG methylation, while *MT*− gametes had an average of approximately 3%. A large increase in 5meC was observed for all sequence contexts during zygote development, with 54% (CG) methylation

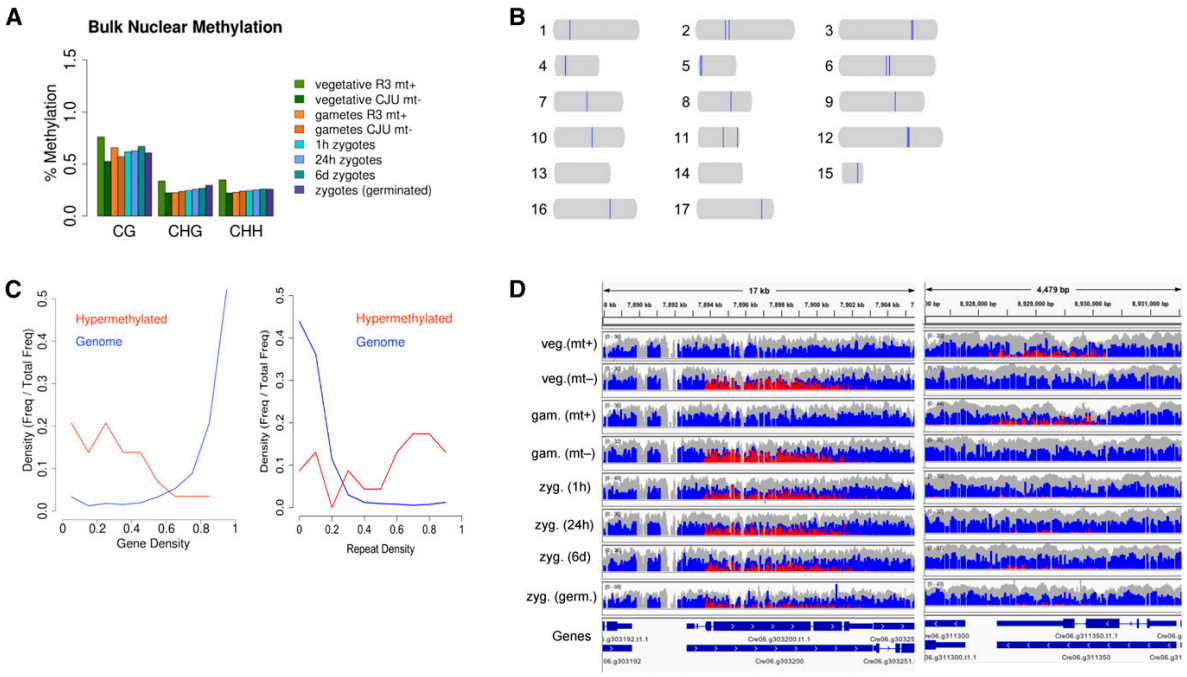


Figure 5. Nuclear methylation at different *C. reinhardtii* life cycle stages. A, Bulk averages of 5meC in the nuclear genome for each sample and categorized by the fraction of 5meC in each of three sequence contexts (CG, CHG, and CHH). B, Scaled representations of the 17 *C. reinhardtii* chromosomes, with 23 hypermethylated regions larger than 10 kb shaded in blue. C, Plots comparing gene and repeat densities of hypermethylated regions relative to the entire genome. D, Genome browser display of nuclear cytosine methylation frequency at two representative hypermethylated loci from chromosome 6. Two representative examples of strain-specific hypermethylated loci are shown side by side, with a gene track below for genomic context. In the coverage plot, blue represents unmethylated cytosines (converted by bisulfite treatment), red represents methylated cytosines (unconverted), and gray represents total coverage (including reads from the opposite strand that do not provide methylation information).

reached by 24 h. Methylation levels began to drop during germination to an average per-site CG methylation level of 45%. CHG and CHH methylation levels were lower than CG methylation levels in all cases, and this difference was particularly notable in germinated zygotes, where CHH and CHG methylation dropped from their peak zygotic levels much faster than CG methylation did.

For all life cycle stages and mating types, methylation of cpDNA is uniformly distributed in the chloroplast genome without bias toward genes or other sequence features (Fig. 6B). With the exception of two large inverted repeats in the chloroplast sequence, where methylation cannot be measured accurately with short reads, average methylation levels tabulated in 1-kb intervals rarely deviated more than 5% from the global average, suggesting that no specific regions of the chloroplast genome are targeted for methylation.

DNA Methyltransferases

In order to further explore the mechanisms responsible for the dynamic patterns of DNA methylation, we identified candidate cytosine methyltransferases from the predicted *C. reinhardtii* proteome based on the presence of predicted DNA cytosine methylase domains (see "Materials and Methods"). We found a total of six candidate methyltransferases with different domain architectures (Supplemental Fig. S1), including a homolog of *V. carteri* MET1, a putative nuclear methyltransferase (Babinger et al., 2007). As described below, three predicted methyltransferases, *DMT1a*, *DMT1b*, and *DMT4*, had expression patterns and/or predicted localization sequences suggesting a role in cpDNA methylation. Each of these predicted proteins has DNA methylase as well as bromo-adjacent homology (BAH) domains. *DMT1a*, *DMT1b*, and *DMT4* were all expressed in gametes and showed the highest levels of expression in zygotes, coinciding with elevated levels of methylation in the zygotic samples (Fig. 7). *DMT1a* and *DMT1b* are near the mating locus in its telomere-proximal domain and encode highly similar paralogs. Both genes had higher expression in *MT+* than in *MT−* gametes, a pattern that matches the methylation bias seen in gametic cpDNA from the two mating types (Fig. 6A). *DMT1a* and *DMT1b* sequences have been described

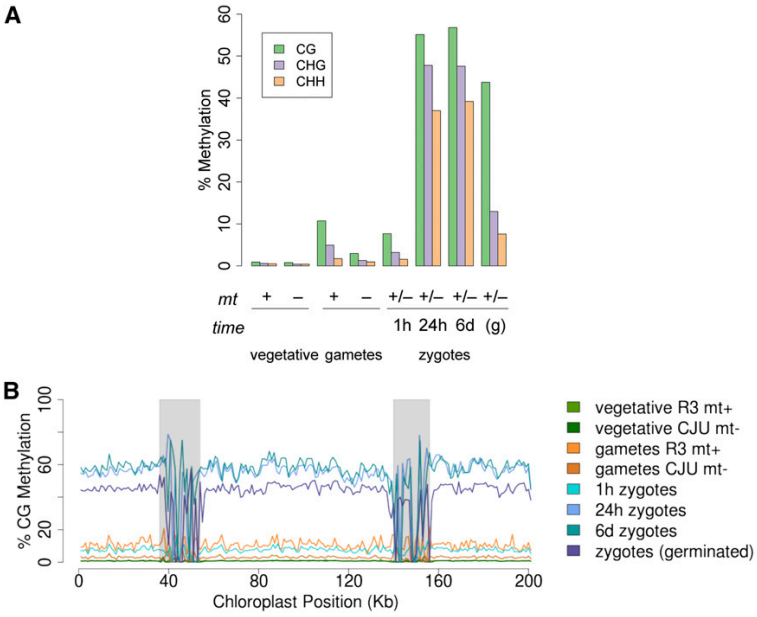


Figure 6. Chloroplast genome methylation at different life cycle stages. A, Bulk cytosine methylation frequencies for cpDNA at different life cycle stages plotted for each methylation context. The mating type of each sample (+, −, or diploid +/−) and time of zygote development and germination (g) are shown below the graph. B, Plot of average CG methylation frequency for each sample in 1-kb bins across the chloroplast genome. The two inverted repeat regions are shaded in gray. Plots are color coded according to the legend at right.

previously as a single gene (Nishiyama et al., 2002), but the published cDNA sequence is a hybrid with 5′ sequences derived from *DMT1a* and 3′ sequences from *DMT1b* (National Center for Biotechnology Information [NCBI] gene identifiers 5722229 and 5722231). Consistent with subcellular targeting predictions (Fig. 7), the DMT1a presequence directs chloroplast localization (Nishiyama et al., 2002). While the *MT+* strain copy of *DMT1a* appears to be intact, a survey of structural variants derived from genome resequencing data led to the identification of a point insertion in exon 10 of the *MT−* strain copy of *DMT1a* leading to a frame shift and premature termination before the methyltransferase domains (Fig. 8B). This insertion was observed in *MT–* transcriptome data. Furthermore, this point insertion is found in the genome of 12 out of 13 *MT–* strains resequenced by Gallaher et al. (2015). No variants predicted to be deleterious were found within the *DMT1b* gene. Targeting predictions of DMT1b suggest that it is mitochondria localized. Although the mitochondrial genome is largely devoid of cytosine methylation (1%–2% global per-cytosine methylation), the zygote sample at 24 h shows evidence of methylation (approximately 13% global per-cytosine methylation; Supplemental Fig. S2). *DMT4* is also predicted to encode a chloroplast-targeted cytosine methyltransferase and is one of the 361 genes identified with a strong zygotic expression pattern (Fig. 3A), consistent with a possible participation of DMT4 in zygotic cpDNA hypermethylation.

DISCUSSION

The dynamic changes in gamete- and zygote-specific mRNA abundance and DNA methylation presented in this work provide a framework for understanding cell differentiation during the *C. reinhardtii* sexual life cycle. A previous study of plus and minus gamete-specific

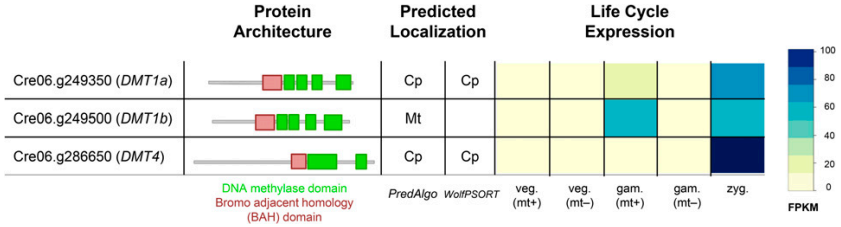


Figure 7. Candidate chloroplast methyltransferases in *C. reinhardtii*. The protein domain structure of each candidate methyltransferase is shown schematically (green = DNA methylase domain and red = BAH domain). Predicted localizations from PredAlgo and WolfPSORT (Cp = chloroplast and Mt = mitochondria) are shown alongside the log-transformed normalized expression level of each candidate from different RNA-seq samples. FPKM, Fragments per kilobase of transcript per million mapped reads.

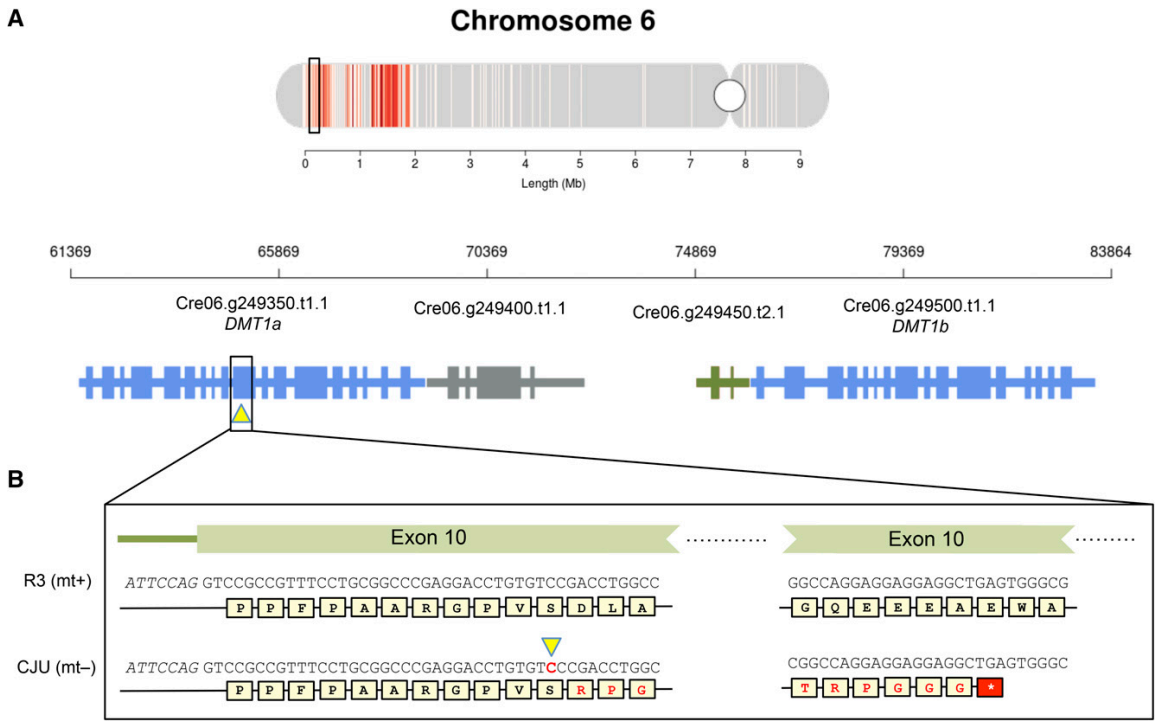


Figure 8. Single-base insertion variant in *DMT1a* leads to a premature stop codon. A, The *DMT1* locus on chromosome 6 containing *DMT1a* and *DMT1b* is shown below its context in the entire chromosome. B, The diagram at top shows an enlarged chromosomal region containing the *DMT1* locus with genomic coordinates and adjacent genes. The position of the insertion variant in *DMT1a* is indicated by the yellow triangle. Below is an expanded view of exon 10 from *DMT1a* showing the single-base insertion in the *MT–* strain (CJU10) and the altered protein-coding sequence and premature stop codon caused by the cytosine insertion, shown in boldface and marked with the yellow triangle.

genes focused on cell type genes whose expression was up-regulated during the process of mating (Ning et al., 2013). However, direct comparison with the previous data is confounded by differences in annotation between genome versions used to define transcripts as well as in the different clustering criteria used in the two studies. Here, we made use of culture conditions that were designed to suppress the differential transcript abundance signal resulting from growing versus nongrowing nitrogen-starved cultures to help identify gamete-specific transcripts. Our method identified known gametic genes whose expression is mating type limited (expressed preferentially in plus versus minus gametes or vice versa). However, because we used nitrogen resupply of stationary phase cultures to create nongrowing vegetative samples, we may have missed highly stable gametic transcripts that did not turn over after nitrogen addition. Nonetheless, nitrogen resupply for several hours was completely effective at suppressing mating, so the transcriptome differences we were able to identify in our gametic versus vegetative samples are likely tied to gametogenesis-related functions and perhaps less relevant for other nitrogen-starvation responses such as neutral lipid accumulation.

Sex-related genes evolve rapidly and, therefore, are expected to appear younger and have a more restricted phylogenetic distribution than other genes (Swanson and Vacquier, 2002). Indeed, using phylogenomic profiling, we found an enrichment of *C. reinhardtii*-specific genes belonging to the plus or minus gamete up-regulated categories and zygote up-regulated category (Fig. 3C). This finding underscores the importance of species-specific and clade-specific genes as potential drivers of cell type specialization related to sex and speciation. Although the functions of these genes are difficult to predict, since they have no homologs outside of *C. reinhardtii* or volvocine algae, we did find enrichment for secretory pathway targeting signals within the plus gamete and zygote predicted proteins (Fig. 3B), which could indicate a role for sex-related proteins in the plasma membrane, cell wall, or extracellular space. Manual annotation of zygotic genes showed that their predicted functions match processes such as glycosylation and transport that are associated with cell wall formation. In addition, we identified or confirmed the expression of zygotic genes that may be associated with other poorly understood differentiation processes, including chloroplast fusion, *MT–* cpDNA

elimination, DNA methylation, and cytoskeletal remodeling (Supplemental Table S1). Deeper investigation of these genes may yield insights into the cell biology of zygote differentiation that may have parallels in other zygospore-forming algae (Brawley and Johnson, 1992) and even in plants where pollen cells must undergo a similar process of dormancy and reactivation when exposed to appropriate conditions (Brown and Lemmon, 2011). Early zygotes in *C. reinhardtii* also undergo dramatic changes in their cytoskeleton. Unlike the case in animals, where paternal but not maternal centrioles are contributed to zygotes (Avidor-Reiss et al., 2015), in *C. reinhardtii*, both parental basal bodies (structurally similar to centrioles) and flagella are retained initially in zygotes but eventually are disassembled and rebuilt during germination (Cavalier-Smith, 1976). Nonetheless, investigation of conserved zygotically up-regulated cytoskeletal proteins such as LF5 and SF-assemblin may shed light on more general mechanisms involved in controlling the dynamic behavior of flagella/cilia and basal bodies during cellular remodeling and life cycle transitions.

DNA Methylation Changes during the Life Cycle

In contrast to many organisms in which a large fraction of the nuclear genome is methylated (Smith and Meissner, 2013; Bestor et al., 2015), we found relatively low levels of cytosine methylation in the nuclear genome of *C. reinhardtii*, consistent with previous surveys that were not as high resolution as reported here (Hattman et al., 1978; Feng et al., 2010). Interestingly, we did identify 23 hypermethylated loci larger than 10 kb (Supplemental Table S3) that tended to occur within gene-poor, repeat-rich regions of the genome. However, the mechanism that leads to the methylation of these loci remains unknown. Previous studies on the silencing of transgenes in *C. reinhardtii* found that inserted transgene repeats were frequently methylated, but they reported little correlation between methylation and gene expression (Cerutti et al., 1997). On the other hand, in the related colonial alga *V. carteri*, where methylcytosine frequency is slightly higher than in *C. reinhardtii* (1.1% versus approximately 0.75%), nuclear methylation did appear to be associated with gene silencing (Babinger et al., 2001, 2007). Interestingly, *V. carteri* also has a more repeat-rich genome than *C. reinhardtii* with many active transposons (Miller et al., 1993; Ueki and Nishii, 2008; Prochnik et al., 2010) and may have retained or evolved higher levels of nuclear methylation activity than *C. reinhardtii* to suppress transposon activity.

The low frequency of cytosine methylation in the nuclear genome contrasts with the dynamic and abundant cytosine methylation in the chloroplast genome. The overall patterns we observed are in agreement with previous findings that vegetative cpDNA from both mating types had low levels of cytosine methylation and that gametes showed elevated levels, with *MT+* cpDNA having a 2- or 3-fold higher frequency of methylcytosine compared with *MT–* cpDNA (Royer and Sager, 1979; Dyer, 1982). We observed hypermethylation in zygotes, which has also been seen previously. Our study extended these earlier results by examining genome-wide methylation in cpDNA at single-base resolution. Unlike the methylation of nuclear DNA, which was mostly restricted to a few loci, cpDNA cytosine methylation was uniformly distributed across all regions and sequence feature types (genic, intergenic, repeats, exons, introns, etc.). This finding has implications for the function of cpDNA methylation and the enzymes that are responsible for methylating cpDNA (elaborated in the next section). Previous models of cpDNA inheritance invoked a methylation-restriction system similar to that in prokaryotes where sequence-specific methyltransferases protect *MT+* cpDNA from site-specific restriction endonucleases. However, the non-sequence-specific distribution of methylcytosines we observed and the modest differences between levels of *MT+* and *MT–* cpDNA methylation in gametes are not consistent with a methylation-restriction mechanism. Moreover, if cpDNA methylation were related to methylation and restriction, it would be most effective if it were established at the gametic stage of the life cycle. Instead, the majority of cpDNA methylation occurs in zygotes and is pervasive genome wide. The purpose of this massive methylation is unknown but is likely tied to the packaging and protection of cpDNA in zygospores, where it may be dormant for many years before germination (Brawley and Johnson, 1992).

DNA Methyltransferases

Our survey of candidate methyltransferases revealed two candidates with expression patterns and predicted chloroplast targeting sequences that make them likely responsible for cpDNA methylation: *DMT1a* and *DMT4*. Both genes have detectable expression in gametes and zygotes (Fig. 7). *DMT1a* and *DMT1b* are physically and genetically linked to the mating locus and, furthermore, the *MT–* linked copy of *DMT1a* has a point mutation that is predicted to generate a truncated, nonfunctional protein. Although they are linked to the mating locus, the *DMT1a/b* locus could still recombine with the mating-type locus or be subject to gene conversion, possibly leading to the creation of strains where the levels of gametic cpDNA in *MT+* and *MT–* parents are equivalent. Such strains, if they could be isolated, would be useful for testing the significance of differential gametic cpDNA methylation and the contribution of the *DMT1a* gene to gametic cpDNA methylation levels.

Although the predicted *DMT1*- and *DMT4*-encoded methyltransferases have accessory BAH domains that are implicated in protein-protein interactions and targeting to specific loci (Callebaut et al., 1999; Yang and Xu, 2013), the nonspecific pattern of cpDNA methylation we observed and attribute to the predicted *DMT1a*

and *DMT4* proteins suggests that their targeting is not sequence specific. The BAH domains in these proteins may serve other functions, such as general targeting of the methyltransferase enzymes to chloroplast nucleoids. The lack of sequence specificity predicted for *DMT1a* and *DMT4* may also prove useful for biotechnology applications that require nonspecific DNA cytosine methylation activity.

MATERIALS AND METHODS

Sample Generation for Bisulfite-Treated DNA

Chlamydomonas reinhardtii strains R3 (CC-620 R3 NM *MT+*) and CJU10 (Umen and Goodenough, 2001) were grown in high-salt medium to concentrations of 3.4×10^6 and 3.5×10^6 cells mL^{-1}, respectively. A total of 100 mL of each strain culture was used for DNA and RNA isolation as vegetative samples. Cells were collected by centrifugation (3,500 rpm for 5 min), resuspended in high-salt medium without nitrogen, and after 15 h, 40 mL of each strain culture was collected as gametic samples. Gametes were mixed and checked for mating efficiency (85% efficient), and after 1 h, a 40-mL sample was collected, corresponding to the 1-h zygote sample. Mixed gametes were split into two flasks: (1) cells in the first were collected after 24 h, corresponding to the 24-h zygote sample; (2) cells in the second were resuspended in water and plated with high-salt medium, incubated in the light for 24 h, incubated for 5 d in the dark, and resuspended in Tris-EDTA-NaCl (TEN) buffer with 0.2% (v/v) Nonidet P-40, at which time half of the culture was collected as 6-d zygote samples. The remaining half was transferred to Tris-acetate phosphate medium, incubated in the light for 24 h, and collected as the germinated zygote sample.

DNA Isolation and Purification for Bisulfite Treatment

Cells were resuspended in 4 mL of TEN buffer, with the exception of 24-h and 6-d zygotes. The 24-h and 6-d zygotes were resuspended in 10 mL of TEN, mixed for 1 min, and repelleted (repeated three times) followed by resuspension in 4 mL of TEN. A total of 400 μL of 20% (w/v) SDS and 400 μL of 20% (v/v) sarkosyl was added to each sample. For 24-h and 6-d zytores, 4 mL of zirconium beads was also added followed by vortexing for 5 min. A total of 200 μL of pronase E solution (10 mg mL^{-1}) was then added, and samples were incubated at 37°C for 30 min. A total of 2.5 mL of phenol (10 mM Tris-Cl, pH 8, saturated) was added, followed by 2.5 mL of chloroform:isoamylalcohol (24:1). The phases were separated by centrifugation (5,000 rpm for 10 min) at 10°C, and the upper phase (4.5 mL) was transferred into 9 mL of 100% ethanol and incubated overnight at −20°C. Nucleic acids were collected by centrifugation, resuspended in 1 mL of 10 mM Tris-Cl, pH 8, and 100 μL of RNase (5 mg mL^{-1}) was added. After RNase treatment, samples were extracted again with 1 mL of phenol:chloroform:isoamylalcohol (25:24:1), and DNA was precipitated with 0.3 M sodium acetate and 70% (v/v) isopropanol for 30 min at room temperature.

DNA Library Preparation

A total of 500 ng of purified genomic DNA was sheared by sonication with Covaris S2 to generate DNA fragments spanning from 100- to 400-bp size range. Library preparation was carried out using NEBNext DNA Library Prep Master Mix (Set for Illumina; New England Biolabs; catalog no. E6040) according to the manufacturer's instructions with minor modifications. The ligation was performed using Illumina TruSeq Adapters (catalog no. 15025064), and DNA size selection (200- to 400-bp range) was carried out with AMPure XP beads (Beckman Coulter) prior to bisulfite conversion using the EZ DNA Methylation-Lightning Kit (Zymo; catalog no. D5030). The bisulfite-treated DNA was amplified using Illumina Primer Cocktail Mix (catalog no. 15027084) and MyTaq Mix (Bioline; catalog no. BIO-25045) according to the following program: 98°C for 2 min; 12 cycles of 98°C for 15 s, 60°C for 30 s, and 72°C for 30 s; and then 72°C for 5 min.

RNA Library Preparation

Total RNA isolation for RNA-seq analysis was performed as described previously (De Hoff et al., 2013) with additional DNase treatment (2 units of Roche RNase-free DNase per 110 μg of total RNA, 37°C for 20 min) before final Qiagen RNeasy column purification.

Genomic Variation

Genomic reads were aligned using Burrows-Wheeler Aligner version 0.6.2 (Li and Durbin, 2010), with default parameters, to the version 5.0 assembly of the *C. reinhardtii* CC-503 genome (Merchant et al., 2007). After removing duplicates with Picard MarkDuplicates (http://broadinstitute.github.io/picard/), we applied Genome Analyzer Tool Kit (McKenna et al., 2010) base quality score recalibration, insertion/deletion realignment, and small variant discovery (DePristo et al., 2011). This was followed by hard filtering of variants with extensive manual calibration guided by inspections in Integrative Genomics Viewer (Thorvaldsdóttir et al., 2013).

RNA-seq Analysis

RNA-seq data were aligned to the *C. reinhardtii* genome (assembly version 5.5; Phytozome version 10.3 gene annotation) with TopHat version 2.0.10 using the annotation to guide spliced alignment. Default parameters were kept, with the exception of constraining intron lengths to less than 5 kb. Expression levels were quantified using Cufflinks version 2.2.1 to compute fragments per kilobase of transcript per million mapped reads.

Identification of Gamete- and Zygote-Specific Genes

Gamete- and zygote-specific genes were identified by applying a series of filters to the fragments per kilobase of transcript per million mapped reads data generated as described above. We defined gamete-specific genes for each mating type as those that have expression in all samples (excluding the zygote) that is less than 10% of the expression level in the gamete of that mating type. Zygote-specific genes were defined as those whose expression in all other samples was less than 25% of the expression level in the zygote. Expression values for samples with replicates were averaged.

Bisulfite-Treated DNA Sequencing Analysis

Raw sequence data were demultiplexed using standard Illumina barcode indices and checked for quality using FastQC (version 0.10.1). Bisulfite-converted sequences were aligned to the *C. reinhardtii* nuclear genome (November 2011 assembly) and chloroplast genome using the BS-Seeker2 alignment pipeline version 2.0.5 (Guo et al., 2013). Default whole-genome bisulfite alignment parameters were chosen with the following exceptions: Bowtie2 was used as the aligner, and local alignments were enabled. Methylation levels were called for cytosines covered by at least four reads. Sequence data have been deposited in NCBI's Short Read Archive under the accession numbers SRR2051057, SRR2051058, SRR2051059, SRR2051060, SRR2051061, SRR2051062, SRR2051063, and SRR2051065.

Identification of Candidate Methyltransferases

To identify candidate DNA methyltransferases, protein sequences of *C. reinhardtii* (Phytozome version 10 gene annotation) were scanned against the Pfam-A database (release 27) using the stand-alone PfamScan scripts provided by the Wellcome Trust Sanger Institute. The presence of the C-5 cytosine-specific DNA methylase domain (PF00145) was used as a criterion for assignment as a cytosine methyltransferase.

Sequence data have been deposited in NCBI's Short Read Archive under the accession numbers SRR2051057, SRR2051058, SRR2051059, SRR2051060, SRR2051061, SRR2051062, SRR2051063, and SRR2051065.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Candidate methyltransferases in *C. reinhardtii*.

Supplemental Figure S2. Methylation patterns for *C. reinhardtii* mitochondrial DNA.

Supplemental Table S1. Annotation of gamete and zygote specific genes.

Supplemental Table S2. Expression values for gamete and zygote specific genes.

Supplemental Table S3. Coordinates of hypermethylated regions in nuclear genome with repeat data.

Supplemental Table S4. Methylation and expression data for genes in regions with strain specific methylation patterns.

Received June 9, 2015; accepted October 7, 2015; published October 8, 2015.

LITERATURE CITED

Albee AJ, Kwan AL, Lin H, Granas D, Stormo GD, Dutcher SK (2013) Identification of cilia genes that affect cell-cycle progression using whole-genome transcriptome analysis in Chlamydomonas reinhardtii. G3 (Bethesda) **3:** 979–991

Aoyama H, Saitoh S, Kuroiwa T, Nakamura S (2014) Comparative analysis of zygospore transcripts during early germination in Chlamydomonas reinhardtii. J Plant Physiol **171:** 1685–1692

Armbrust EV, Ferris PJ, Goodenough UW (1993) A mating type-linked gene cluster expressed in Chlamydomonas zygotes participates in the uniparental inheritance of the chloroplast genome. Cell **74:** 801–811

Avidor-Reiss T, Khire A, Fishman EL, Jo KH (2015) Atypical centrioles during sexual reproduction. Front Cell Dev Biol **3:** 21

Babinger P, Kobl I, Mages W, Schmitt R (2001) A link between DNA methylation and epigenetic silencing in transgenic Volvox carteri. Nucleic Acids Res **29:** 1261–1271

Babinger P, Völkl R, Cakstina I, Maftei A, Schmitt R (2007) Maintenance DNA methyltransferase (Met1) and silencing of CpG-methylated foreign DNA in Volvox carteri. Plant Mol Biol **63:** 325–336

Bestor TH, Edwards JR, Boulard M (2015) Notes on the role of dynamic DNA methylation in mammalian development. Proc Natl Acad Sci USA **112:** 6796–6799

Blaby IK, Blaby-Haas CE, Tourasse N, Hom EF, Lopez D, Aksoy M, Grossman A, Umen J, Dutcher S, Porter M, et al (2014) The Chlamydomonas genome project: a decade on. Trends Plant Sci **19:** 672–680

Bolen PL, Grant DM, Swinton D, Boynton JE, Gillham NW (1982) Extensive methylation of chloroplast DNA by a nuclear gene mutation does not affect chloroplast gene transmission in Chlamydomonas. Cell **28:** 335–343

Boyle NR, Page MD, Liu B, Blaby IK, Casero D, Kropat J, Cokus SJ, Hong-Hermesdorf A, Shaw J, Karpowicz SJ, et al (2012) Three acyltransferases and nitrogen-responsive regulator are implicated in nitrogen starvation-induced triacylglycerol accumulation in Chlamydomonas. J Biol Chem **287:** 15811–15825

Brawley SH, Johnson LE (1992) Gametogenesis, gametes and zygotes: an ecological perspective on sexual reproduction in the algae. Br Phycol J **27:** 233–252

Brown RC, Lemmon BE (2011) Spores before sporophytes: hypothesizing the origin of sporogenesis at the algal-plant transition. New Phytol **190:** 875–881

Burton WG, Grabowy CT, Sager R (1979) Role of methylation in the modification and restriction of chloroplast DNA in Chlamydomonas. Proc Natl Acad Sci USA **76:** 1390–1394

Callebaut I, Courvalin JC, Mornon JP (1999) The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. FEBS Lett **446:** 189–193

Catt JW (1979) Isolation and chemical composition of the zygospore cell wall of Chlamydomonas reinhardtii. Plant Sci Lett **15:** 69–74

Cavalier-Smith T (1976) Electron microscopy of zygospore formation in Chlamydomonas reinhardtii. Protoplasma **87:** 297–315

Cerutti H, Johnson AM, Gillham NW, Boynton JE (1997) Epigenetic silencing of a foreign gene in nuclear transformants of *Chlamydomonas*. Plant Cell **9:** 925–945

De Hoff PL, Ferris P, Olson BJ, Miyagi A, Geng S, Umen JG (2013) Species and population level molecular profiling reveals cryptic recombination and emergent asymmetry in the dimorphic mating locus of C. reinhardtii. PLoS Genet **9:** e1003724

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43:** 491–498

Duncan L, Nishii I, Harryman A, Buckley S, Howard A, Friedman NR, Miller SM (2007) The VARL gene family and the evolutionary origins of the master cell-type regulatory gene, regA, in Volvox carteri. J Mol Evol **65:** 1–11

Duncan L, Nishii I, Howard A, Kirk D, Miller SM (2006) Orthologs and paralogs of regA, a master cell-type regulatory gene in Volvox carteri. Curr Genet **50:** 61–72

Dyer TA (1982) Methylation of chloroplast DNA in Chlamydomonas. Nature **298:** 422–423

Fang W, Si Y, Douglass S, Casero D, Merchant SS, Pellegrini M, Ladunga I, Liu P, Spalding MH (2012) Transcriptome-wide changes in *Chlamydomonas reinhardtii* gene expression regulated by carbon dioxide and the CO_2-concentrating mechanism regulator CIA5/CCM1. Plant Cell **24:** 1876–1893

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al (2010) Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci USA **107:** 8689–8694

Feng TY, Chiang KS (1984) The persistence of maternal inheritance in Chlamydomonas despite hypomethylation of chloroplast DNA induced by inhibitors. Proc Natl Acad Sci USA **81:** 3438–3442

Ferris PJ, Armbrust EV, Goodenough UW (2002) Genetic structure of the mating-type locus of Chlamydomonas reinhardtii. Genetics **160:** 181–200

Ferris PJ, Goodenough UW (1987) Transcription of novel genes, including a gene linked to the mating-type locus, induced by Chlamydomonas fertilization. Mol Cell Biol **7:** 2360–2366

Ferris PJ, Waffenschmidt S, Umen JG, Lin H, Lee JH, Ishida K, Kubo T, Lau J, Goodenough UW (2005) Plus and minus sexual agglutinins from *Chlamydomonas reinhardtii*. Plant Cell **17:** 597–615

Gallaher SD, Fitz-Gibbon ST, Glaesener AG, Pellegrini M, Merchant SS (2015) Chlamydomonas genome resource for laboratory strains reveals a mosaic of sequence variation, identifies true strain histories, and enables strain-specific studies. Plant Cell **27:** 2335–2352

Glaesener AG, Merchant SS, Blaby-Haas CE (2013) Iron economy in Chlamydomonas reinhardtii. Front Plant Sci **4:** 337

Goldschmidt-Clermont M (1991) Transgenic expression of aminoglycoside adenine transferase in the chloroplast: a selectable marker of site-directed transformation of Chlamydomonas. Nucleic Acids Res **19:** 4083–4089

González-Ballester D, Casero D, Cokus S, Pellegrini M, Merchant SS, Grossman AR (2010) RNA-seq analysis of sulfur-deprived *Chlamydomonas* cells reveals aspects of acclimation critical for cell survival. Plant Cell **22:** 2058–2084

Goodenough U, Lin H, Lee JH (2007) Sex determination in Chlamydomonas. Semin Cell Dev Biol **18:** 350–361

Grief C, O'Neill MA, Shaw PJ (1987) The zygote cell wall of Chlamydomonas reinhardtii: a structural, chemical and immunological approach. Planta **170:** 433–445

Grossman A (2000) Acclimation of Chlamydomonas reinhardtii to its nutrient environment. Protist **151:** 201–224

Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, Chen PY, Pellegrini M (2013) BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics **14:** 774

Harris EH, Stern DB, Witman G, editors (2009) The Chlamydomonas Sourcebook. Elsevier/Academic Press, Amsterdam

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res **32:** D258–D261

Hattman S, Kenny C, Berger L, Pratt K (1978) Comparative study of DNA methylation in three unicellular eucaryotes. J Bacteriol **135:** 1156–1157

Heinnickel ML, Grossman AR (2013) The GreenCut: re-evaluation of physiological role of previously studied proteins and potential novel protein functions. Photosynth Res **116:** 427–436

Idoine AD, Boulouis A, Rupprecht J, Bock R (2014) The diurnal logic of the expression of the chloroplast genome in Chlamydomonas reinhardtii. PLoS One **9:** e108760

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res **28:** 27–30

Karpowicz SJ, Prochnik SE, Grossman AR, Merchant SS (2011) The GreenCut2 resource, a phylogenomically derived inventory of proteins specific to the plant lineage. J Biol Chem **286:** 21427–21439

Kirk MM, Stark K, Miller SM, Müller W, Taillon BE, Gruber H, Schmitt R, Kirk DL (1999) regA, a Volvox gene that plays a central role in germsoma differentiation, encodes a novel regulatory protein. Development **126:** 639–647

Kubo T, Abe J, Oyamada T, Ohnishi M, Fukuzawa H, Matsuda Y, Saito T (2008) Characterization of novel genes induced by sexual adhesion and gamete fusion and of their transcriptional regulation in Chlamydomonas reinhardtii. Plant Cell Physiol **49:** 981–993

Kuriyama H, Takano H, Suzuki L, Uchida H, Kawano S, Kuroiwa H, Kuroiwa T (1999) Characterization of *Chlamydomonas reinhardtii* zygote-specific cDNAs that encode novel proteins containing ankyrin repeats and WW domains. Plant Physiol **119:** 873–884

Kurvari V, Grishin NV, Snell WJ (1998) A gamete-specific, sex-limited homeodomain protein in Chlamydomonas. J Cell Biol **143:** 1971–1980

Lechtreck KF, Rostmann J, Grunow A (2002) Analysis of Chlamydomonas SF-assemblin by GFP tagging and expression of antisense constructs. J Cell Sci **115:** 1511–1522

Lee JH, Lin H, Joo S, Goodenough U (2008) Early sexual origins of homeoprotein heterodimerization and evolution of the plant KNOX/BELL family. Cell **133:** 829–840

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2 **6:** 589–595

Liu Y, Misamore MJ, Snell WJ (2010) Membrane fusion triggers rapid degradation of two gamete-specific, fusion-essential proteins in a membrane block to polygamy in Chlamydomonas. Development **137:** 1473–1481

Lopez D, Casero D, Cokus SJ, Merchant SS, Pellegrini M (2011) Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. BMC Bioinformatics **12:** 282

Lv H, Qu G, Qi X, Lu L, Tian C, Ma Y (2013) Transcriptome analysis of Chlamydomonas reinhardtii during the process of lipid accumulation. Genomics **101:** 229–237

Matsuo M, Hachisu R, Tabata S, Fukuzawa H, Obokata J (2011) Transcriptome analysis of respiration-responsive genes in Chlamydomonas reinhardtii: mitochondrial retrograde signaling coordinates the genes for cell proliferation with energy-producing metabolism. Plant Cell Physiol **52:** 333–343

Matters GL, Goodenough UW (1992) A gene/pseudogene tandem duplication encodes a cysteine-rich protein expressed during zygote development in Chlamydomonas reinhardtii. Mol Gen Genet **232:** 81–88

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB (2002) The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell **14:** 2659–2679

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res **20:** 1297–1303

Merchant SS, Allen MD, Kropat J, Moseley JL, Long JC, Tottey S, Terauchi AM (2006) Between a rock and a hard place: trace element nutrition in Chlamydomonas. Biochim Biophys Acta **1763:** 578–594

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, et al (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318:** 245–250

Miller R, Wu G, Deshpande RR, Vieler A, Gärtner K, Li X, Moellering ER, Zäuner S, Cornish AJ, Liu B, et al (2010) Changes in transcript abundance in *Chlamydomonas reinhardtii* following nitrogen deprivation predict diversion of metabolism. Plant Physiol **154:** 1737–1752

Miller SM, Schmitt R, Kirk DL (1993) *Jordan*, an active *Volvox* transposable element similar to higher plant transposons. Plant Cell **5:** 1125–1138

Minami SA, Goodenough UW (1978) Novel glycopolypeptide synthesis induced by gametic cell fusion in Chlamydomonas reinhardtii. J Cell Biol **77:** 165–181

Nakamura S (2010) Paternal inheritance of mitochondria in Chlamydomonas. J Plant Res **123:** 163–170

Nguyen AV, Thomas-Hall SR, Malnoë A, Timmins M, Mussgnug JH, Rupprecht J, Kruse O, Hankamer B, Schenk PM (2008) Transcriptome for photobiological hydrogen production induced by sulfur deprivation in the green alga Chlamydomonas reinhardtii. Eukaryot Cell **7:** 1965–1979

Ning J, Otto TD, Pfander C, Schwach F, Brochet M, Bushell E, Goulding D, Sanders M, Lefebvre PA, Pei J, et al (2013) Comparative genomics in Chlamydomonas and Plasmodium identifies an ancient nuclear envelope protein family essential for sexual reproduction in protists, fungi, plants, and vertebrates. Genes Dev **27:** 1198–1215

Nishimura Y (2010) Uniparental inheritance of cpDNA and the genetic control of sexual differentiation in Chlamydomonas reinhardtii. J Plant Res **123:** 149–162

Nishimura Y, Misumi O, Kato K, Inada N, Higashiyama T, Momoyama Y, Kuroiwa T (2002) An mt(+) gamete-specific nuclease that targets mt(−) chloroplasts during sexual reproduction in C. reinhardtii. Genes Dev **16:** 1116–1128

Nishiyama R, Ito M, Yamaguchi Y, Koizumi N, Sano H (2002) A chloroplast-resident DNA methyltransferase is responsible for hypermethylation of chloroplast genes in Chlamydomonas maternal gametes. Proc Natl Acad Sci USA **99:** 5925–5930

Nishiyama R, Wada Y, Mibu M, Yamaguchi Y, Shimogawara K, Sano H (2004) Role of a nonselective de novo DNA methyltransferase in maternal inheritance of chloroplast genes in the green alga, Chlamydomonas reinhardtii. Genetics **168:** 809–816

Panchy N, Wu G, Newton L, Tsai CH, Chen J, Benning C, Farré EM, Shiu SH (2014) Prevalence, evolution, and cis-regulation of diel transcription in Chlamydomonas reinhardtii. G3 (Bethesda) **4:** 2461–2471

Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, Ferris P, Kuo A, Mitros T, Fritz-Laylin LK, et al (2010) Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. Science **329:** 223–226

Rochaix JD (2001) Assembly, function, and dynamics of the photosynthetic machinery in *Chlamydomonas reinhardtii*. Plant Physiol **127:** 1394–1398

Royer HD, Sager R (1979) Methylation of chloroplast DNAs in the life cycle of Chlamydomonas. Proc Natl Acad Sci USA **76:** 5794–5798

Sano H, Grabowy C, Sager R (1981) Differential activity of DNA methyltransferase in the life cycle of Chlamydomonas reinhardtii. Proc Natl Acad Sci USA **78:** 3118–3122

Schmollinger S, Mühlhaus T, Boyle NR, Blaby IK, Casero D, Mettler T, Moseley JL, Kropat J, Sommer F, Strenkert D, et al (2014) Nitrogen-sparing mechanisms in *Chlamydomonas* affect the transcriptome, the proteome, and photosynthetic metabolism. Plant Cell **26:** 1410–1435

Silflow CD, Lefebvre PA (2001) Assembly and motility of eukaryotic cilia and flagella: lessons from *Chlamydomonas reinhardtii*. Plant Physiol **127:** 1500–1507

Simon DF, Descombes P, Zerges W, Wilkinson KJ (2008) Global expression profiling of Chlamydomonas reinhardtii exposed to trace levels of free cadmium. Environ Toxicol Chem **27:** 1668–1675

Simon DF, Domingos RF, Hauser C, Hutchins CM, Zerges W, Wilkinson KJ (2013) Transcriptome sequencing (RNA-seq) analysis of the effects of metal nanoparticle exposure on the transcriptome of Chlamydomonas reinhardtii. Appl Environ Microbiol **79:** 4774–4785

Smith ZD, Meissner A (2013) DNA methylation: roles in mammalian development. Nat Rev Genet **14:** 204–220

Suzuki L, Woessner JP, Uchida H, Kuroiwa H, Yuasa Y, Waffenschmidt S, Goodenough UW, Kuroiwa T (2000) Zygote-specific protein with hydroxyproline-rich glycoprotein domains and lectin-like domains involved in the assembly of the cell wall of Chlamydomonas reinhardtii (Chlorophyta). J Phycol **36:** 571–583

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nat Rev Genet **3:** 137–144

Tam LW, Ranum PT, Lefebvre PA (2013) CDKL5 regulates flagellar length and localizes to the base of the flagella in Chlamydomonas. Mol Biol Cell **24:** 588–600

Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M (2004). MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J **37:** 914–939

Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform **14:** 178–192

Toepel J, Albaum SP, Arvidsson S, Goesmann A, la Russa M, Rogge K, Kruse O (2011) Construction and evaluation of a whole genome microarray of Chlamydomonas reinhardtii. BMC Genomics **12:** 579

Toepel J, Illmer-Kephalides M, Jaenicke S, Straube J, May P, Goesmann A, Kruse O (2013) New insights into Chlamydomonas reinhardtii hydrogen production processes by combined microarray/RNA-seq transcriptomics. Plant Biotechnol J **11:** 717–733

Uchida H, Kawano S, Sato N, Kuroiwa T (1993) Isolation and characterization of novel genes which are expressed during the very early stage of

zygote formation in Chlamydomonas reinhardtii. Curr Genet **24:** 296–300

Uchida H, Suzuki L, Anai T, Doi K, Takano H, Yamashita H, Oka T, Kawano S, Tomizawa KI, Kawazu T, et al (1999) A pair of invertedly repeated genes in Chlamydomonas reinhardtii encodes a zygote-specific protein whose expression is UV-sensitive. Curr Genet **36:** 232–240

Ueki N, Nishii I (2008) Idaten is a new cold-inducible transposon of Volvox carteri that can be used for tagging developmentally important genes. Genetics **180:** 1343–1353

Umen JG (2011) Evolution of sex and mating loci: an expanded view from volvocine algae. Curr Opin Microbiol **14:** 634–641

Umen JG, Goodenough UW (2001) Chloroplast DNA methylation and inheritance in Chlamydomonas. Genes Dev **15:** 2585–2597

Woessner JP, Molendijk AJ, van Egmond P, Klis FM, Goodenough UW, Haring MA (1994) Domain conservation in several volvocalean cell wall proteins. Plant Mol Biol **26:** 947–960

Yang N, Xu RM (2013) Structure and function of the BAH domain in chromatin biology. Crit Rev Biochem Mol Biol **48:** 211–221

Chapter 5:

Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria

THE PROTEIN SOCIETY

Using comparative genomics to uncover new kinds of protein-based metabolic organelles in bacteria

Julien Jorda,[1] David Lopez,[1] Nicole M. Wheatley,[2] and Todd O. Yeates[1,2,3]*

[1]UCLA-DOE Institute for Genomics and Proteomics, 611 Charles Young Dr. East, Los Angeles, California 90095
[2]UCLA Molecular Biology Institute, 611 Charles Young Dr. East, Los Angeles, California 90095
[3]Department of Chemistry and Biochemistry, University of California, Los Angeles, 611 Charles Young Dr. East, Los Angeles, California 90095

Received 12 September 2012; Revised 9 November 2012; Accepted 12 November 2012
DOI: 10.1002/pro.2196
Published online 27 November 2012 proteinscience.org

Abstract: Bacterial microcompartment (MCP) organelles are cytosolic, polyhedral structures consisting of a thin protein shell and a series of encapsulated, sequentially acting enzymes. To date, different microcompartments carrying out three distinct types of metabolic processes have been characterized experimentally in various bacteria. In the present work, we use comparative genomics to explore the existence of yet uncharacterized microcompartments encapsulating a broader set of metabolic pathways. A clustering approach was used to group together enzymes that show a strong tendency to be encoded in chromosomal proximity to each other while also being near genes for microcompartment shell proteins. The results uncover new types of putative microcompartments, including one that appears to encapsulate B_{12}-independent, glycyl radical-based degradation of 1,2-propanediol, and another potentially involved in amino alcohol metabolism in mycobacteria. Preliminary experiments show that an unusual shell protein encoded within the glycyl radical-based microcompartment binds an iron-sulfur cluster, hinting at complex mechanisms in this uncharacterized system. In addition, an examination of the computed microcompartment clusters suggests the existence of specific functional variations within certain types of MCPs, including the alpha carboxysome and the glycyl radical-based microcompartment. The findings lead to a deeper understanding of bacterial microcompartments and the pathways they sequester.

Keywords: microcompartment; carboxysome; bacterial organelle; metabolic pathways; glycyl radical enzymes

Introduction

Over the last few decades, the general view that bacteria have simple internal structures has changed. Electron microscopy investigations have demonstrated that bacteria produce a wide variety of intracellular inclusions.[1,2] The discovery and subsequent isolation of one particular class of polyhedrally shaped bodies dates back almost 40 years.[3] These giant proteinaceous bodies, called bacterial microcompartments (hereafter referred to as MCPs) are typically 80–150 nm in diameter and consist of a set of interior enzymes surrounded by a thin protein shell reminiscent of a viral capsid.[4–8] MCPs have been proposed to serve diverse functional roles:

Abbreviations: BMC, bacterial microcompartment shell protein; Eut, ethanolamine utilization; Etu,ethanol utilization; Grp, glycyl radical-based 1,2-propanediol utilization; MCP, microcompartment; PCC, pairwise correlation coefficient; Pdu, propanediol utilization.

Additional Supporting Information may be found in the online version of this article.

Julien Jorda and David Lopez contributed equally to this work.

Grant sponsor: NIH; Grant number: R01AI081146; Grant sponsor: Ruth L. Kirschstein National Research Service; Grant number: GM007185.

*Correspondence to: Todd O. Yeates, Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095-1569. E-mail: yeates@mbi.ulca.edu

improving flux through key metabolic pathways,[9] sequestering cytotoxic or volatile intermediates in a pathway,[10,11] and protecting the encapsulating enzymes from exposure to competing or reactive molecules (e.g., O_2),[12] all while allowing passage of substrates and products across the shell. Biochemical and structural studies have revealed microcompartments to be mechanistically complex entities, warranting their classification as organelles (Fig. 1).

The enzymes and metabolic pathways encapsulated by microcompartments are diverse, allowing the delineation of a few distinct classes of MCPs.[7] The founding member is the carboxysome, present in cyanobacteria and some chemoautotrophs.[3,13] The carboxysome houses two enzymes: RuBisCO (a low efficiency enzyme essential to autotrophic fixation of carbon dioxide) and carbonic anhydrase [Fig. 1(B)]. The catalytic efficiency of RuBisCO is improved by having its CO_2 substrate produced by carbonic anhydrase inside the MCP, where its escape might be retarded by the shell.[14,15] Two carboxysome subtypes (alpha and beta) are delineated by their partially distinct protein components; they are distributed along phylogenetic lines within chemoautotrophs (alpha only) and cyanobacteria (alpha or beta). Biochemical and genetic studies have been conducted on two other microcompartments: the Pdu microcompartment of enteropathic *Salmonella enterica*[16–18] and the Eut microcompartment of *Salmonella* (also found in some strains of *Escherichia coli*) (refs. 11,19,20). These MCPs metabolize 1,2-propanediol and ethanolamine, respectively [Fig. 1(B)].

In contrast to the metabolic variations presented by different MCPs, the proteins that self-assemble to form the outer shell are homologous across the disparate functional types. MCP shells are composed mainly by proteins bearing one or sometimes two tandem bacterial microcompartment (or BMC) domains, identified first by Shively and coworkers.[13] We refer to these major shell components as BMC shell proteins. Bacterial microcompartments themselves are sometimes referred to as BMCs, but in this paper we refer to microcompartments as MCPs to avoid confusing the compartment with its main structural proteins. In each MCP, a few (three to seven) different paralogs of the BMC shell protein assemble, from a few thousand copies in total, to form the shell (Supporting Information Fig. S1). Crystallographic studies have given insight into the significance of the conserved BMC domain and how it relates to microcompartment shell organization as a whole.[4,5,8,21,22] In particular, structures of several BMC shell proteins have revealed that they generally assemble as cyclic homohexamers, which pack side-by-side to build a molecular layer[4,5,21,23] (Supporting Information Fig. S1). The center of each hexamer is typically perforated by a narrow pore along the sixfold axis of symmetry; these pores are

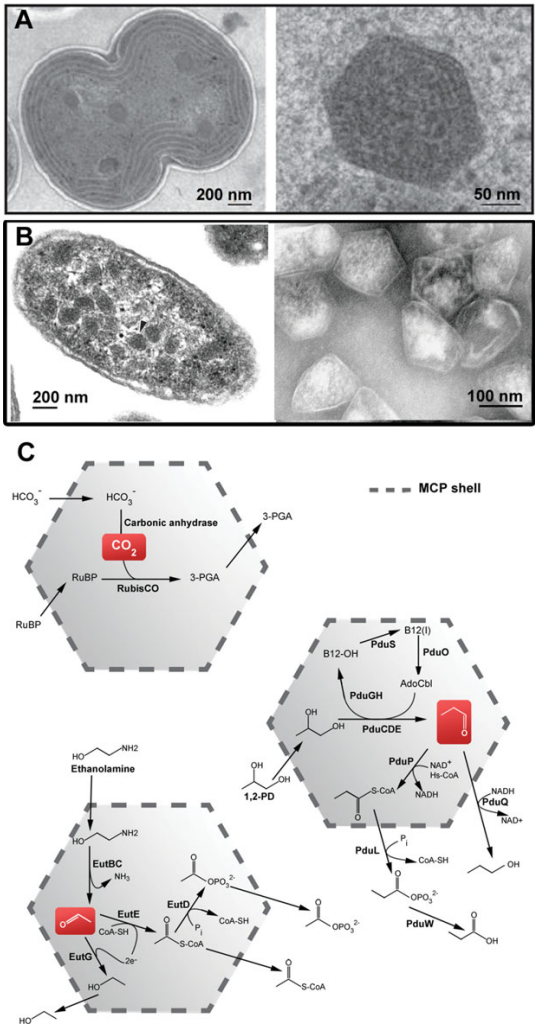


Figure 1. Morphology of MCPs and models of their sequestered pathways. (A) Thin-section EM of a dividing cell of the cyanobacterium *Synechocystis* sp. PCC6803 (left) along with an enlargement of a single carboxysome (right, courtesy of Wim Vermaas). (B) Electron micrographs of a thin-section of *Salmonella enterica* serovar Tyhpimurium LT2 (left; Reproduced from Ref. [96], with permission from Thomas Bobik) and purified Pdu microcompartments (right; Reproduced from Ref. [8], with permission from Thomas Bobik). (C) Models for CO_2 fixation and 1,2-propanediol and ethanolamine metabolism in the carboxysome, Pdu and Eut microcompartments, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

believed to provide the conduits for substrates and products to cross the shell.[4,23–25] Another gene family (ccmL/csoS4/eutn/pduN), which is distinct from the BMC family, is typically present as well and is believed to code for minor vertex proteins in the shell.[23] Genomic analyses have shown that the encapsulated enzymes and the BMC shell proteins are frequently encoded within the same operon,

consistent with the idea that expression and assembly of the distinct MCP components occur in a coordinated fashion.[7] Mechanisms underlying the targeting of enzymes to their correct destination in MCPs have been partially clarified. Experiments show that, for a few cases, short N-terminal sequence extensions are necessary and sufficient to target enzymes to the MCP interior or lumen.[26–30] In another case, a C-terminal region has been shown to be important for targeting an interior protein.[31]

Genomic studies offer prospects for new discoveries related to MCPs. A current search of sequence databases for proteins bearing a BMC domain indicates that microcompartments are distributed across approximately 17% (265 out of 1568) of the fully sequenced bacterial genomes currently available. Several comparative genomics analyses have suggested the existence of other types of MCPs besides the three currently studied,[4,8,21,22] but clear metabolic models have not been developed. The ever-expanding body of genomic sequences, combined with the tendency of BMC shell proteins to be encoded in proximity to the enzymes they encapsulate, suggests bioinformatics strategies for predicting the existence of novel metabolic schemes within MCPs. In this study, we sought to identify potentially novel MCP pathways by searching for the co-occurrence of groups of enzyme-encoding genes in MCP operons. The approach is based on the idea that groups of enzymes that tend to occur together within individual MCP operons are likely to represent encapsulated pathways. Here, we describe how our method recapitulates the metabolic pathways hosted in known MCPs, while also uncovering MCPs that are novel or that represent variations on previously studied types.

Results

MCP operons were examined across the 113 fully sequenced bacterial genomes where BMC shell proteins could be identified. Based on their identified PFAM domains and annotations from the KEGG Orthology system, proteins encoded by genes proximal to BMC shell genes were collapsed into distinct Protein Functional Groups intended to represent unique cellular functions. In order to cluster these into disjoint sets that might each represent one type of MCP, we evaluated in a full pairwise fashion the tendency of every pair of Protein Functional Groups to co-occur within individual MCP operons. Statistical tests were applied throughout the procedure to maximize the likelihood of producing biologically meaningful results (see Methods and Fig. 2).

Our genomic context-based approach produced 10 candidate metabolic clusters containing between two and 13 proteins and enzymes (Fig. 3). Protein Functional Groups are represented as nodes, with two nodes being connected by a line or edge when

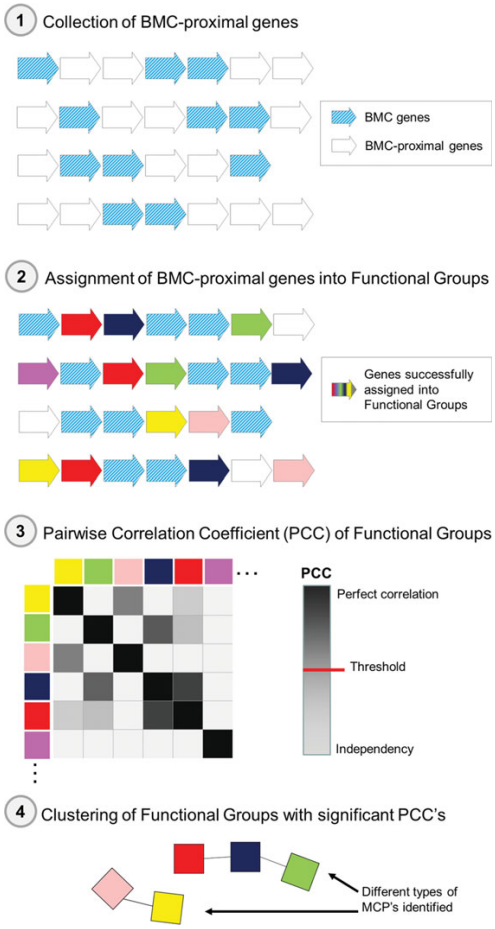


Figure 2. Schematic for identifying pairs of proteins or enzyme families that tend to co-occur in the context of microcompartment (MCP) operons. As step 1, BMC-proximal genes and their encoded proteins are collected following our operational definition of a MCP operon (see Methods). In step 2, to gain clarity and statistical power, the BMC-proximal genes are assigned to Protein Functional Groups, grouping together similar protein sequences where possible. In step 3, the co-occurrence tendency is evaluated by a pairwise correlation coefficient (PCC) for every pair of Functional Groups. In the last step, after applying statistical confidence tests, strongly linked Protein Functional Groups are clustered together. Different clusters identify MCPs with distinct metabolic functions. The scheme shown is a simplification; application to real genomic data leads to more MCP types and more proteins per cluster.

their tendency to co-occur was judged to be statistically significant. Among the 10 clusters, four were consistent with well-characterized, canonical MCPs: carboxysomes of the alpha and beta type, along with the Pdu and Eut microcompartments. Most of the enzymes known to participate in MCP function were found to be effectively clustered, in some cases together with other unanticipated Protein Functional Groups. Of particular interest, a number of poorly

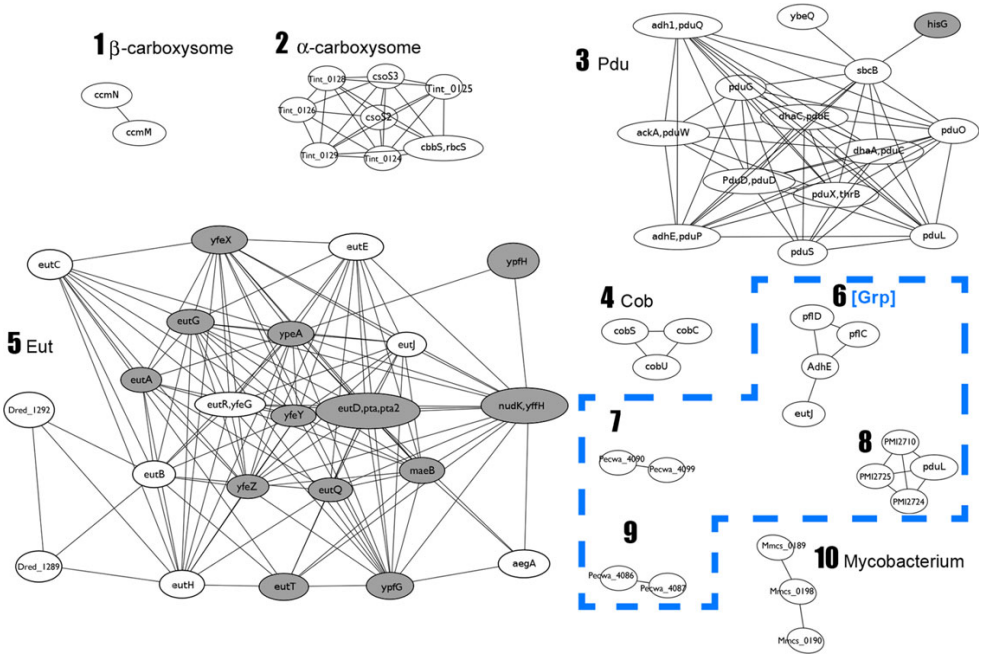


Figure 3. Clusters of proteins and enzymes predicted by a computational approach to constitute distinct kinds of microcompartments. Clusters have been numbered from 1 to 10. Each Protein Functional Group is represented as a node (plain white for strong nodes, light gray for weak nodes) and a significant correlation between two nodes is represented as an edge. Protein Functional Groups are labeled with their corresponding gene name when consistent annotations are available across the different species. Clusters 1, 2, 3, and 5 are related to the canonical microcompartments (beta and alpha carboxysomes, and Pdu and Eut microcompartments, respectively). Cluster 4 is related to the *cob* operon, which upon closer inspection is seen to relate to the Pdu MCP. Clusters 6–9 relate to a presumptive MCP for glycyl radical-based propanediol utilization (which we name Grp), along with variations under which it appears in different species. Cluster 10 identifies a potential MCP in mycobacteria that could involve amino alcohol metabolism. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

characterized or previously uncharacterized MCP types emerged from the analysis. The features of these MCPs are discussed below.

Carboxysomes and Pdu and Eut gene clusters

For the clusters corresponding to previously characterized MCP types, we analyzed the results for the presence of (1) enzymes well-established to be involved in that MCP, and (2) unexpected enzymes that could provide new insights into how these MCPs might function in diverse microbes.

Beta-carboxysome. Cluster 1 represents the beta carboxysome (Fig. 3). A total of nine cyanobacterial genomes form the basis for cluster 1. The beta carboxysome is unusual compared to the alpha carboxysome and the Eut and Pdu MCPs in that its component genes are typically distributed across multiple genomic regions rather than residing in a single operon.[5,8,12,23] This is seen in the clustering results, which identify strong connections only between two proteins in the beta carboxysome, namely CcmM and CcmN; genes for the RuBisCO large and small subunits, which perform the key CO_2 fixing reaction, do not co-occur as strongly with BMC shell proteins within the genomes of cyanobacteria that produce beta carboxysomes. CcmM and CcmN play important roles in organizing the enzymes and shell proteins of the beta carboxysome through specific protein–protein interactions,[31,32] and CcmM has been shown to carry redox-sensitive carbonic anhydrase activity.[33]

Alpha-carboxysome. Cluster 2 represents the alpha carboxysome. Eight Protein Functional Groups are clustered (Fig. 3). These include three proteins well-established to be part of the alpha carboxysome: the carbonic anhydrase CsoS3 (or CsoSCA), the RubisCO small subunit, and the CsoS2 protein, whose function remains enigmatic.[14,21,34] The correlations between these three Protein Functional Groups were high, based on their co-occurrence across 14 organisms included here (see Supporting Information). These results are consistent with the canonical genomic organization reported in the literature for the alpha carboxysome[12,14,34] and with the essential features of CO_2 fixation.

Five additional Protein Functional Groups are also found clustered with the alpha-carboxysome:

bacterioferritin, pterin-4a-carbinolamine dehydratase homologues, an ATP-binding protein homologous to CbbQ, along with its activation protein (CbbO), and cobyrinic acid a,c-diamide synthase (CobQ). The presence of these noncanonical genes appears to arise from their co-occurrence in the vicinity of the BMC shell gene csoS1D, which was recently confirmed to be a bona fide carboxysome shell protein.[35] The tendency of CsoS1D to be encoded near some of the proteins identified here was reported in an earlier analysis,[35] but the biological relevance of these BMC-proximal genes has not been clarified yet.

The CsoS1D protein[24] belongs to a group of BMC shell proteins that each contain two tandem BMC domains, and which assemble as pseudo-hexameric trimers.[4,25,36,37] Tandem domain BMC shell proteins from different kinds of MCPs have been shown to support both open and closed pore conformations.[4,25,36] In the case of the Eut MCP, it was proposed that the open conformation in the tandem BMC protein EutL is required for transporting bulky cofactors across the shell for their use by the encapsulated ethanolamine utilizing enzymes. The core enzymes of the carboxysome—carbonic anhydrase and RuBisCO—do not require cofactors, leaving the purpose of a large pore in CsoS1D unexplained. We propose here, based on the tight genomic association of CsoS1D with various complex enzymes, that this shell protein likely supports the transport of molecules—possibly including bulky cofactors—for as-yet undefined reactions that might occur within variant forms of the alpha carboxysome in diverse bacteria.

The identities of the above five noncanonical Protein Functional Groups that appear linked to the alpha carboxysome allow for speculation regarding possible functional variations within this type of MCP. Two of the associated Protein Functional Groups, CbbQ and its activation protein, modulate the conformation and activity of RuBisCO, inducing a twofold increase in its V_{max}.[38] CbbQ relies on ATP for activity, raising the prospect of nucleotide transport across the alpha carboxysome shell. A third protein, bacterioferritin, is known to form molecular cages for iron storage. Its functional connection to the alpha carboxysome, if any, is unknown. The remaining two Protein Functional Groups identified here have functions likely related to cofactor synthesis. One is CobQ, which uses glutamine or ammonia as a substrate in a reaction for synthesis of cobalamin,[39] a cofactor used in other MCPs but not previously associated with the carboxysome. The involvement of ammonia is intriguing in view of the enzyme's relatively high Km (26–200 μM) for this substrate,[40] along with the well-established confinement or channeling of ammonia between enzymes in other metabolic contexts.[41,42] The possibility that ammonia could be used for a reaction inside a functionally extended form of the alpha-carboxysome constitutes a hypothesis for future testing.

The final Protein Functional Group strongly co-occurring with the canonical proteins of the alpha-carboxysome appears by sequence analysis to be distantly related to pterin-4 alpha-carbinolamine dehydratase (PCD), though key catalytic residues believed to play a role in PCD activity[43] are not obviously preserved by sequence alignments to the enzymes co-occurring with the alpha-carboxysome genes. PCD is involved in tetrahydrobiopterin recycling, where it catalyzes the reversible conversion of 4a-hydroxytetrahydrobiopterin to dihydrobiopterin.[44] Only tentative connections can be suggested between these Protein Functional Groups. For example, we note that tetrahydrobiopterin is often seen as a cofactor for aromatic amino acid hydroxylases,[45] which also require non-heme iron for activity;[46] this would provide a potential link to bacterioferritin. Also intriguing is the observation that some ferritin-like proteins are encapsulated within a different kind of protein compartment, considerably smaller than an MCP, known as the encapsulin nanocompartment.[47] Although clear functional relationships between these additional proteins found to be associated with alpha carboxysome operons cannot be established at the present time, recent experimental studies confirm that bacterioferritin and the enzyme homologous to PCD are in fact upregulated in concert with the canonical carboxysome genes.[48]

Pdu microcompartment. Cluster 3 represents the Pdu MCP (Fig. 3). It contains all the enzymes known to operate in the 1,2-propanediol utilization pathway (Fig. 1), with the exception of PduH and PduV. These enzymes share identical domains with other clustered enzymes (PduH with PduD and PduV with EutP respectively) and thus were missed by our approach due to the inability in each case to segregate the homologous enzymes into two distinct groups. The Protein Functional Group annotated as *hisG* was represented as a weak node due to its tendency to occur at the margins of the *pdu* operon.

Cluster 4 contains a few enzymes involved in synthesizing the cobalamin cofactor (Fig. 3). This cluster of *cob* genes was not automatically joined to the canonical Pdu MCP (cluster 3) by our analysis, but the two clusters appear to be functionally linked. Experimental studies show that the *pdu* and *cob* operons are both tightly regulated by the PocR protein, and that propanediol degradation is dependent on cobalamin, B_{12}.[18,49] In most cases we observe that the *cob* genes are adjacent or peripheral to the *pdu* operon and not interspersed with the BMC shell proteins. Thus in our analysis the correlations between these distinct clusters of Protein Functional Groups were not significant enough to merge the Pdu and Cob pathways into a single cluster.

Likewise, there are no experimental data tying these particular cobalamin synthesis reactions directly to the Pdu MCP. Nonetheless, B_{12} is a required cofactor for 1,2-propanediol degradation and there are a few bacterial species where the genomic arrangement is distinct, and suggestive of a closer relationship. The cobU cobC and cobS genes are used to synthesize the lower ligand of B_{12}, suggesting that lower ligand synthesis may be limiting for B_{12} production in some environments. Similarly, the PduX gene often found near the end of the pdu operon in enterica bacteria is also used for lower ligand synthesis.[50]

Eut microcompartment. Cluster 5 represents the ethanolamine utilization (Eut) MCP. The proteins typically encoded by that operon are clustered by our method. Some additional proteins, more weakly connected, are also identified, including two genes coding for a sensor histidine kinase and a response regulator. Indeed, it has been previously established that among the species associated to the Eut microcompartment, some of them present an extended version of the canonical operon and embed a two-component signal transduction system: a histidine kinase and its response regulator, referred to as EutW and EutV, respectively.[19] In vitro assays showed that ethanolamine induces a 15-fold increase in the rate of autophosphorylation of EutW, followed by the activation of EutV through phospho-transfer.[19,51] Reciprocally, a closer look at the 17 organisms featuring this variant of the *eut* operon showed that the eutR gene is absent. The latter is known to regulate the *eut* operon in response to ethanolamine and adenosylcobalamin (AdoCbl).[52] The EutVW and EutR regulatory systems appear to exist in mutually exclusive species that use Eut MCPs. The observed dichotomy appears to be largely phylogenetic; EutV and EutW are found mainly in the Firmicutes while EutR is found only in the Enterobacteriaceae.

We note that the putative microcompartment for ethanol utilization (Etu) discussed by Heldt *et al.*[53] gets grouped with the Eut cluster by our automatic clustering. This is because the latter pathway includes just two Protein Functional Groups, and these are also found in the ethanolamine utilization pathway (namely EutG and EutE). The operation of this presumptive Etu MCP has not been clarified yet by experimental studies, though physiological considerations suggest that it may be involved in converting ethanol to acetyl-CoA.[54]

Protein clusters representing new presumptive MCP types

Five additional small clusters, besides the clusters clearly related to the canonical MCPs discussed above, are identified in our analysis. They highlight systems that have not yet been characterized in detail (Fig. 3). Four of these clusters (6–9) are weakly interconnected at a level not sufficient for them to be automatically joined by our approach—they appear to represent variations within a complex type of MCP. Finally, a 10th cluster appears to represent a distinct entity. Findings related to these two presumptive MCP types—clusters 6–9 and cluster 10— are discussed below.

A putative glycyl radical-based MCP. Among several enzymes identified in cluster 6, one of particular note belongs to a diverse family of glycyl radical enzymes. The identification (in *Vibrio furnissi* M1) of BMC shell genes interspersed with an enzyme from this family was discussed by Wackett *et al.*[55] Based on sequence similarity, this enzyme has been previously annotated as a pyruvate formate lyase. However, the sequence similarity is low, and other considerations discussed subsequently argue that this enzyme, and the MCP that harbors it, most likely utilizes 1,2-propanediol rather than pyruvate, in a B_{12}-independent pathway. We report here that these enzymes are encoded in a genomic pattern substantially conserved across more than 20 species examined, suggesting the existence of a broadly distributed class of microcompartment apart from the better-known MCPs.

Cluster 7 consists of two co-occurring Functional Groups: an enzyme similar to the C-terminal domain of PduO (an ATP/cobalamin adenosyltransferase involved in vitamin B_{12} synthesis in the Pdu pathway), and a MIP family channel protein known to transport small neutral metabolites across the membrane.[56] Cluster 8 includes two drug resistance proteins along with two regulatory proteins and a phosphotransacylase (a PduL homolog). Strikingly, an analysis of the operons supporting clusters 7 and 8 showed that the enzymes identified in cluster 6 were also present, though our algorithmic approach did not automatically detect connections between cluster 6 and either cluster 7 or 8. The Protein Functional Groups represented by clusters 7 and 8 are only sometimes present in the larger set of operons that contain cluster 6 as the conserved core (Fig. 4). This is consistent with the failure of our automatic procedure to connect clusters 7 and 8 to cluster 6; they appear to represent specific compositional variations.

Clusters 7 and 8 contain a number of proteins or enzymes without obvious relationships to glycyl radical-based metabolism. The MIP channel protein from cluster 7 could be responsible for importing the substrate or substrates of this MCP, but the role of the enzyme similar to the PduO C-terminus is unclear. Among the five organisms found to have these two genes, three belong to the set of more than 20 where the cluster 7 genes occur together with the core enzymes of cluster 6. When present, these genes are found dispersed between multiple BMC shell protein genes (Fig. 4). Cluster 8,

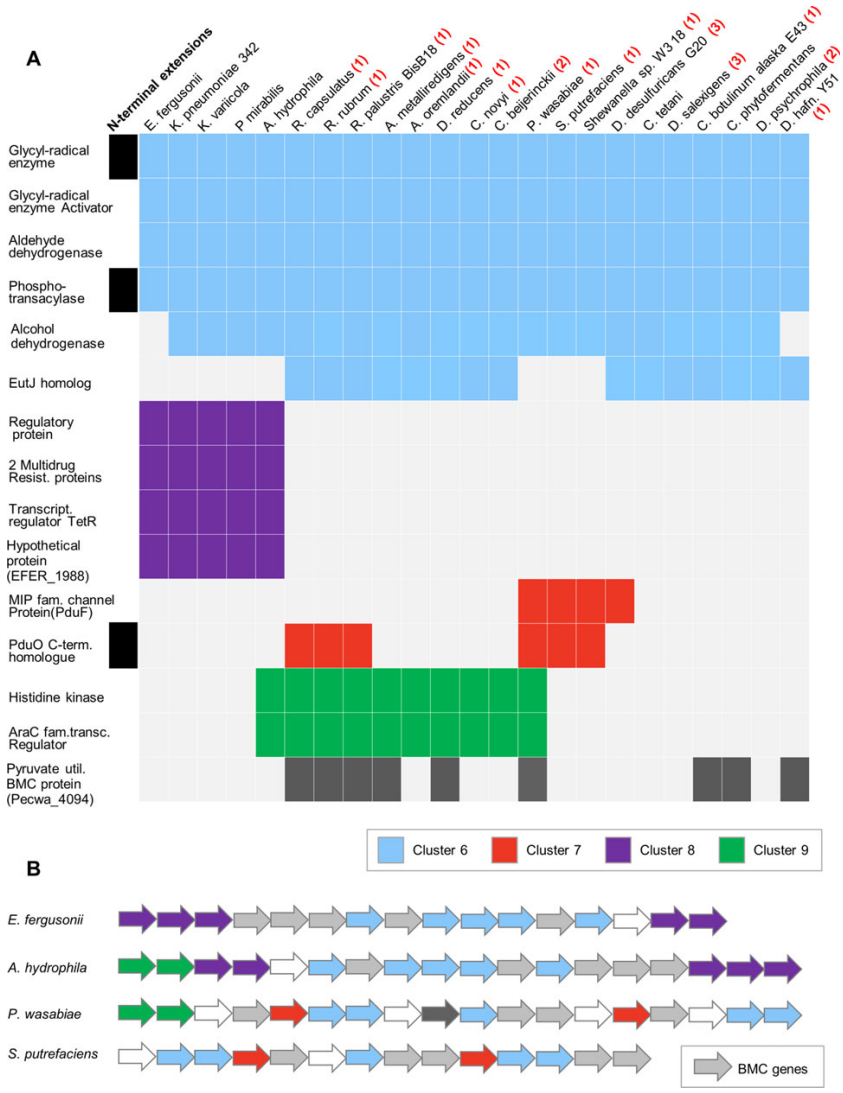


Figure 4. Phylogenetic profile of the BMC-proximal proteins and enzymes from the presumptive glycyl radical-based propanediol utilization (Grp) microcompartment operon across 23 bacterial species. (A) The 23 microorganisms are extracted from the analysis of cluster 6 (Fig. 3). Each colored block refers to the presence of a given protein (on the left) in the operon featured by a given organism (on the top). Cluster 6, corresponding to the core enzymes of the pathway, is depicted in light blue while supplemental enzymes from clusters 7, 8, and 9 are shown in red, purple and green respectively. A black block next to a protein name indicates the presence of a presumptive N-terminal targeting extension in this protein compared to homologues not involved in microcompartments, as analyzed following previously described methods.[27] The last row represents the profile of a divergent BMC shell protein, apparently specific to the Grp MCP, which was subjected to initial experimental characterization (see text). Next to each organism name, the number of tandem BMC proteins present in the operon is enclosed in red brackets. (B) Examples of gene organization in species featuring the Grp operon. BMC shell genes are colored in light gray, while the genes clustered by our approach are colored consistently with their corresponding clusters depicted in panel A. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

composed of four different Protein Functional Groups, was present in five different species (Fig. 4). These proteins appear to be associated with drug resistance mechanisms: two transcription regulators with the canonical helix-turn-helix DNA-binding motif found in antibiotic-resistance repressors like TetR,[57] two paralogs of a drug resistance protein, and one yet uncharacterized protein conserved across the five species. Genes from this group were sometimes found flanking BMC shell genes at both upstream and downstream positions within a genome (Fig. 4). It is presently unclear what functional connections might relate the glycyl radical-based enzyme cluster (6) to clusters 7 and 8.

Cluster 9, whose genes also occasionally occur with those of cluster 6, involves two Protein Functional Groups similar to those found in the expanded version of the *eut* operon: a histidine kinase sensor and a response regulator receiver. These genes occur in the operon containing the glycyl radical-based degradation enzyme in about half of the species examined, but their frequent location at the upstream end of the operon, along with their absence from the other half of the species, caused them to cluster separately.

Diverse MCP operons encoding a glycyl radical enzyme have been noted in the literature, and a just-published survey (see Supporting Information in Ref. 31) suggests the potential for multiple types of MCP based on those operons. In the present study, the diverse MCP operons belonging to clusters 6–9 are seen to share a conserved set of enzymes (Fig. 4). We therefore group them for the present as a single distinct type of MCP, though functional variations are evident.

A presumptive MCP in mycobacteria. Cluster 10 identifies another distinct type of MCP operon present in four organisms (*Mycobacterium smegmatis*, *Mycobacterium sp. MCS*, *Mycobacterium gilvum*, and *Mycobacterium vanbaalenii*) and containing at least three Protein Functional Groups (Fig. 3): an aminotransferase, a short chain dehydrogenase similar to amino alcohol dehydrogenase, and a GnTR family transcriptional regulator that Vindal *et al*.[58] have described as belonging to the FadR/HutC subfamily, whose members are known to bind ligands such as oxidized substrates related to amino acid metabolism or long chain fatty acids. Recently, genetic analysis of an operon coding for a similar MCP in *Rhodococcus erythropolis* MAK154 highlighted that amino alcohol dehydrogenase expression was repressed by a GntR transcriptional regulator.[59] That repression was relieved in the presence of 1-amino-2-propanol, which is the substrate of the amino alcohol dehydrogenase. After visual examination of the four operons, we found other enzymes missed by our automatic approach, which allowed for manual improvement of the functional predictions. These additional enzymes include an amino acid permease and an aldehyde dehydrogenase. Beyond a presumptive connection to amino alcohol metabolism, a more specific functional role for this putative MCP, which was also listed among prospective MCPs by Kinney *et al*.,[31] cannot be offered at this time.

Genomic characterization of a putative glycyl radical-based propanediol utilization (Grp) MCP

As noted above, the enzymes of cluster 6 exhibit a strongly conserved pattern of co-occurrence with BMC shell proteins across many bacteria, covering 23 species from the Firmicutes and Proteobacteria phyla.[55] The key enzyme from cluster 6 exhibits low but recognizable sequence similarity to pyruvate formate lyase from *Escherichia coli* (21% amino acid sequence identity), and glycerol dehydratase from *Clostridium butyricum* (33% identity). Both of those enzymes belong to a broad family of glycyl-radical enzymes.[60–62] Experimental studies demonstrate that a pyruvate-formate lyase activase enzyme, which converts Gly734 (G-H) into a glycine radical (G*) in the active site of pyruvate formate lyase, is necessary for activating the latter, and for triggering the reaction in anaerobic conditions.[63] Glycyl radical enzymes generally require such activators to modify a critical residue in their active site.[61,62,64] A glycyl-radical enzyme activase is indeed present in cluster 6; its occurrence is perfectly correlated with the presence of the glycyl-radical enzyme across the genomes in our analysis (Fig. 4). The activase is known to use S-adenosyl methionine as a substrate to generate the required radical,[65,66] and a closer look at the genomic context showed the presence in some cases of a S-adenosylmethionine synthase gene (Supporting Information).

Pyruvate formate lyases contain two adjacent cysteine residues within their active sites, which are believed to be important in radical transfer.[56] Other types of glycyl radical enzymes, such as B_{12} independent glycerol dehydratase, ribonucleotide reductases, and 1,2-propanediol dehydratases, contain only one of those cysteine residues. Sequence alignments indicate that the glycyl radical enzymes in our cluster 6 MCP contain only one cysteine. This, coupled with the closer similarity of the cluster 6 glycyl radical enzyme to a B_{12}-independent glycerol dehydratase, calls into question its annotation as a pyruvate formate lyase. Likewise, experimental studies highlighted that two of the organisms represented in cluster 6, *Proteus mirabilis* and *Escherichia fergusonii* are not able to ferment glycerol,[67] suggesting that glycerol is not the primary substrate for this type of MCP. Conversely, it has been shown that the enzyme annotated as glycerol dehydratase from *Clostridium butyricum* is also able to catalyze the dehydration of 1,2-propanediol to propionaldehyde.[68] Further evidence that 1,2-propanediol is the primary substrate in these systems comes from the upregulation of microcompartment genes when 1,2-propanediol is metabolized anaerobically in *Roseburia inulinivorans*.[69] The 1,2-propanediol arises from the anaerobic degradation of fucose, followed by conversion to propionaldehyde by a B_{12}-independent glycyl-radical enzyme in *R. inulinivorans* that is highly similar to the glycerol dehydratase in *Clostridium butyricum*.[68] Consistent with this theme, one of the cluster 6 microcompartments identified in *Clostridium phytofermentans* (Supporting Information) has been shown to be involved in fucose/rhamnose degradation (GSM333252 data from the Gene

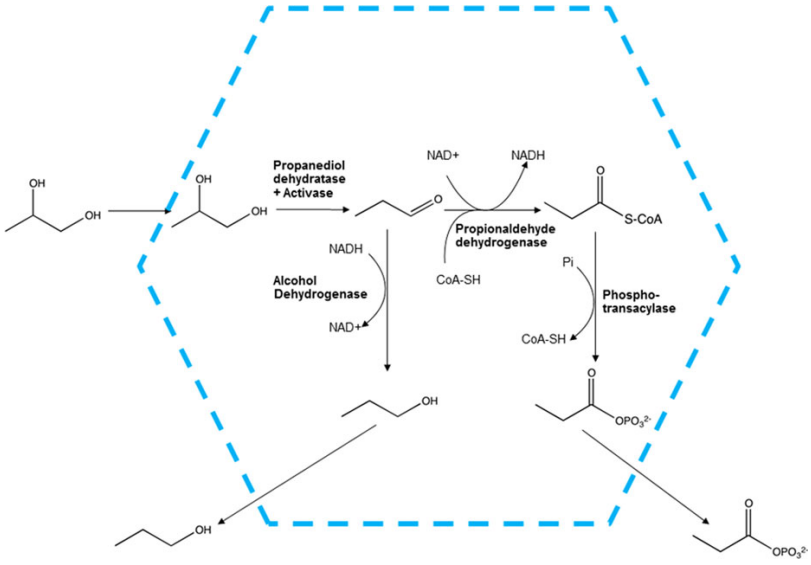


Figure 5. Proposed model for metabolism in the Grp microcompartment. Similar to the Pdu microcompartment, the expected final products include propanol and propionyl-phosphate. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Expression Omnibus, submitted by J.L. Blanchard). Based on these observations, we surmise that the glycyl-radical enzymes in these MCPs act as B_{12}-independent 1,2-propanediol dehydratases. In line with previous naming schemes, we refer to these systems as glycyl radical-based propanediol utilization or Grp MCPs.

Following the initial enzymatic step, similarities are evident between the Grp MCP and the B_{12}-dependent Pdu (and Eut) MCPs. Cluster 6 contains two Protein Functional Groups related to enzymes known to be targeted to the Eut or Pdu microcompartments: the aldehyde dehydrogenases EutE/PduP and EutJ, which plays a possible role of chaperone. Other enzymes believed to operate in the Eut or Pdu MCP are not seen in the glycyl radical-based cluster 6. However, an enzyme closely related in sequence to the phosphotransacylase PduL (whose substrate is propionyl-CoA) is present,[70] further supporting the hypothesis of a pathway beginning with 1,2-propanediol as the initial MCP substrate. The PduL-like Functional Group is not automatically placed with cluster 6 in our approach, mainly because of its wide occurrence in other genomic contexts, including those represented by cluster 3 (Pdu). This situation highlights a limitation of our approach, which seeks to divide MCPs based on their use of distinct enzymes. A similar situation was seen for the Protein Functional Group representing alcohol dehydrogenases. The alcohol dehydrogenase functional group clusters automatically with the Eut MCP (Fig. 3, cluster 5, EutG), but genes for alcohol dehydrogenases also co-occur systematically with the 1,2-propanediol metabolizing enzymes of cluster 6 (Fig. 4).

Modeling a glycyl radical-based propanediol utilization pathway

Looking at the substrates and products of the enzymes identified computationally (cluster 6), and also by subsequent visual inspection of the corresponding operons, one can logically combine the reactions in a pathway that would use 1,2-propanediol as a main metabolic substrate. As a first step in modeling a reaction pathway, we used the KEGG webserver[71] to identify candidate pathways that might involve a glycyl radical-based propanediol dehydratase enzyme. The most likely candidate was the glycerolipid pathway (K00128), leading to two products: propionylphosphate and 1-propanol. By mapping the enzymes identified in our cluster 6 onto the glycerolipid metabolism pathway, we found that they could carry out a reaction sequence similar to the one found in Pdu. The putative 1,2-propanediol dehydratase in the Grp MCP would play a role analogous to the PduCDE enzyme in the B_{12}-dependent Pdu MCP; both produce propionaldehyde. This intermediate would then be converted to propionyl-CoA and propanol by the sequential actions of aldehyde dehydrogenase and alcohol dehydrogenase, both found in cluster 6. Based on similarities with the characterized Pdu pathway, the PduL homolog can be integrated into the reaction scheme to carry out the transacylation between propionyl-CoA and propionyl-phosphate. From these reactions, we can draw a

prospective pathway likely to occur within the Grp MCP (Fig. 5).

Detection of N-terminal extensions in the glycyl radical-based propanediol utilization enzymes

Recent experiments have demonstrated that some enzymes in the B_{12}-dependent Pdu system are targeted to the MCP interior via the presence of short N-terminal extensions in their sequences.[27,30] Bioinformatics studies suggest that an equivalent mechanism might exist in other MCPs as well.[27,28,31] Here, we asked whether special N-terminal extensions might be evident in the enzymes predicted to be associated with the proposed Grp MCP (Fig. 4). Consistent with our previous calculations, we found that the phosphotransacylase and the putative glycyl radical 1,2-propanediol dehydratase both present special N-terminal extensions.[27] Additionally, the enzyme similar to the PduO C-terminus in cluster 7 was also detected as having such an extension. These observations support a model that places those enzymes within (or physically bound to) the MCP.

The approach we used to identify potential targeting signals is based, not on recognition of any particular sequence motif, but on the presence of extended sequences at the termini.[27] In addition, following recent work,[31] we attempted to identify potentially conserved sequence motifs in the enzymes of the Grp MCP operons that might match other established targeting sequences. In our examination we did not judge matches to potential targeting motifs as being statistically significant enough to make clear predictions about other targeting mechanisms in this MCP.

Shell proteins of the Grp MCP

In parallel with our analysis of Protein Functional Groups from cluster 6 and their genomic organization (Fig. 4), we examined the BMC genes that would presumably form the shell of the Grp MCP. In a typical bacterial genome, an MCP operon for glycyl radical-based propanediol utilization contains about four distinct BMC shell protein paralogs. According to current models for MCP architecture,[5,6,25] a few thousand copies of these proteins would assemble into hexagonally-packed arrays in forming the surface of the shell, with pores allowing for molecular transport (Supporting Information Fig. S1). Also in keeping with other MCP operons, the operons for Grp MCPs code for a gene from the *ccmL/csoS4/pduN/eutN* family, which are presumed to code for minor (e.g., vertex) proteins in the shell.[23]

Structural studies of shell proteins from different MCPs have revealed interesting variations that have been ascribed to different functional requirements of the distinct shells. The presumptive BMC shell proteins for the Grp MCP reveal additional variations. For instance, the BMC shell proteins encoded by Sputw3181_0423 in *Shewanella putrefaciens* and Pecwa_4089 in *Pectobacterium wasabiae* exhibit long C-terminal tails extending beyond the conserved BMC domain; these segments are predicted by IUPRED[72] to be disordered. Short, flexible extensions of about 10–20 residues have been noted in previous structural studies,[21,23,73] but the extensions observed here are unusually long, ranging from 80 to 100 residues. Also unusual is the presence of a BMC shell protein (present in eight of the species examined) that is especially divergent from those previously characterized. The occurrence of this shell protein across various species has been integrated into Figure 4. Another interesting feature is the abundance of tandem BMC proteins in the Grp MCP operons of some bacteria, such as *Desulfovibrio salexigens* and *Desulfovibrio desulfuricans* G20, which exhibit up to three tandem BMC proteins (Fig. 4). The apparently novelty of the shell proteins in the Grp MCP is consistent with the expectation that its metabolic features will be different in substantial ways from those explored so far.

As an initial step in characterizing the protein components of a Grp MCP, we overexpressed the unusually divergent shell protein noted above, Pecwa_4094 from *Pectobacterium wasabiae*, in *E. coli* and purified it to homogeneity. As judged by gel filtration, the protein appears to assemble into a homooligomer consistent with a hexamer, by analogy to previously studied BMC shell proteins. Surprisingly, the purified protein was brown in color and an absorption spectrum showed broad peaks at 330 and 420 nm, consistent with a partially oxidized iron-sulfur cluster (Supporting Information Fig. S2). An iron-sulfur cluster has been proposed to occupy the pore of a previously characterized tandem BMC shell protein, PduT, in the Pdu MCP.[25,74,75] The unusual shell protein from *Pectobacterium wasabiae* reported here is the first single-domain BMC to be identified as a metalloprotein. Further biochemical and structural studies on components of this system will be required to clarify how the shell proteins of the Grp MCP modulate its function.

Discussion

Since the discovery of the carboxysome and the Pdu and Eut MCPs, important clues about MCP composition and function have come from examining sequence data and genomic organization.[12,76–79] In this study, we have taken those ideas further with an algorithmic approach aimed at classifying new and varied types of MCPs, relying on patterns across more than a hundred bacterial genomes in which BMC shell proteins can be found. Because each type of MCP houses a group of enzymes, a central element of our strategy was to try to automatically group together genes that co-occur strongly in the vicinity of BMC shell proteins. The resulting

clustering strategy was generally effective in detecting genomic patterns and organizing the data in a functionally relevant form. A minor challenge arose from the fact that different MCPs actually contain some overlapping enzyme activities (e.g., alcohol dehydrogenases), whereas the clustering approach naturally seeks to generate separate enzyme groupings. Nonetheless, with some manual analysis to mitigate such challenges, the method was able to classify two types of uncharacterized MCPs, while also illuminating potential variations within previously characterized types, most notably the alpha carboxysome.

Our analysis suggests a model for a microcompartment for B_{12}-independent, glycyl radical-based propanediol utilization, for which we have introduced the name Grp. Multiple lines of reasoning support the proposed operation of such a microcompartment. An operon-like organization is evident, with genes for enzymes dispersed among multiple BMC shell genes. The enzymes can be sequentially connected in a pathway for 1,2-propanediol metabolism that resembles the one that occurs in the Pdu MCP, with the key distinction being the initial reaction, which involves a glycyl-radical enzyme and its activase in the Grp MCP, instead of a B_{12} cofactor. Further experimental studies will be required to test whether MCPs of the Grp type might be able to metabolize a wider range of substrates, perhaps with different specificities in different bacteria. Finally, multiple enzymes encoded in the operon reveal special N-terminal extensions; this has been established as a key mechanism for targeting enzymes to MCPs.[27–29]

The encapsulated enzymes and metabolic intermediates provide clues regarding biological advantages that could be offered by the Grp microcompartment. One of the pathway intermediates is the cytotoxic propionaldehyde; retaining such intermediates is recognized as a key role for MCPs.[10,80] The reactivity of the key glycyl radical enzyme offers another clue about potential roles for this MCP. The presence of a glycyl radical in the activated state of this enzyme renders it sensitive to oxygen, via oxygen-mediated cleavage of the polypeptide backbone. This property is general to glycyl-radical enzymes, confining their existence to strictly anaerobic conditions.[62] The suggestion that the Grp MCP could protect the glycyl-radical enzyme from destructive oxygen exposure would parallel similar ideas in other MCPs; enzymes in the Pdu, Eut, and carboxysome systems are all sensitive to either damage by or competition with molecular oxygen.[60]

In addition to the core pathway illustrated for propanediol utilization in the Grp MCP, several genomic variations were found across the species examined. A few distinct groups of additional proteins were sometimes present, with different groups appearing in different bacteria (Fig. 4). The existence of three types of extensions could be postulated from the gene clusters automatically identified. The first extension, involving a histidine kinase and a regulatory sensor, is reminiscent of an analogous pair observed in the Eut MCP (and also present in the proposed Etu system). It occurs in about half of the Grp operons identified here. The similarity suggests that this ubiquitous phosphorylation-based signal transduction mechanism also regulates the Grp system, perhaps in response to propanediol. Another variant of the Grp operon appears in several bacteria in the form of two transcriptional regulators homologous to the tetR family of repressors at the upstream end of the operon, and two genes coding for small multidrug resistance (SMR) family proteins at the downstream end. It seems unlikely that the drug resistance proteins are sequestered in the microcompartment lumen since their primary role is to export small molecules extracellularly.[81] Moreover, their structural properties as transmembrane efflux proteins seem incompatible with the structural requirements of a microcompartment shell. Drug resistance elements are known to be prone to horizontal gene transfer, and their presence in this case may be explained by a necessity to be under the influence of the promoter controlling the expression of the Grp operon. This extended Grp operon occurs almost exclusively in enteropathic bacteria. In another variational form, we identified two proteins homologous to proteins usually found in the Pdu operon, PduF and PduO. PduF is a channel protein transporting small metabolites, potentially facilitating 1,2-propanediol diffusion in this case.[82] A last variant, highlighted by J. L Blanchard in *Roseburia inulinovorans*[69] (but not automatically identified in our bioinformatics analysis), appears to be used for anaerobic fucose and rhamnose degradation. In this variation, which eluded our automatic analysis owing to its relative rarity, the Grp operon is extended by the presence of an aldolase, which is required in a multistep conversion of 6-carbon sugars to 1,2-propanediol, most likely before entering the MCP. Expression profiling data showed that, in the presence of fucose or rhamnose in anaerobic conditions, the aldolase and 12 other genes matching our definition of the Grp operon are indeed found in the top 20 upregulated genes in *Roseburia*.

The genomic variations observed in the Grp system suggest that MCPs providing core metabolic functions can be used differently or modified in distinct ways in diverse bacteria. The existence of a Eut operon variant with a signal transduction system, and alpha carboxysome operons extended by various additional genes in the vicinity of the specialized CsoS1D shell gene,[35] are consistent with this general view. For the latter, we speculate that extended functions beyond the canonical alpha carboxysome activities (i.e., CO_2 fixation) would require

the transport of larger molecules, perhaps cofactors, and the specialized shell protein CsoS1D could serve those functions with its larger pore. What additional functions might extend the core CO_2 fixing reactions have not been articulated yet, but preliminary observations implicate bacterioferritin and a homolog of a pterin recycling enzyme in this type of MCP.

Finally, another type of presumptive microcompartment was identified, specific to *Mycobacterium* species. Genes for a group of about four proteins or enzymes are found interspersed with typically two BMC shell protein genes as well as a gene for the minor (vertex) shell protein. The enzymes suggest potential involvement in amino alcohol metabolism. Two of them have been shown to be involved in utilizing amino alcohols such as 1-amino-2-propanol, while other types of proteins occurring in these operons include a class III aminotransferase, an amino acid permease-associated protein, an aminoglycoside phosphotransferase, and a protein of unknown function. By analogy to other MCP systems, the amino alcohol dehydrogenase is likely to represent the key first reaction in some encapsulated pathway. The repression of this enzyme by the GntR transcriptional regulator would be lifted in the presence of the amino alcohol substrate, leading to the expression of structural shell proteins and the enzymes to be encapsulated. The permease, while unlikely to be part of the MCP structure itself, could facilitate uptake of an amino acid or amino alcohol substrate. The structural similarity of 1-amino-2-propanol to ethanolamine raises the possibility that the mycobacterial MCP discussed here could be similar to the Eut microcompartment. However, the presence of distinct groups of enzymes supports a separate classification for this presumptive MCP. Among the distinctive enzymes appearing to be associated with this MCP, the aminoglycoside phosphotransferases are bacterial antibiotic resistance proteins, conferring resistance to many aminoglycosides,[83] while aminotransferases have been presumed to play a role in aminoglycoside antibiotics biosynthesis.[86,87] This suggests potentially interesting connections between this novel MCP and mycobacterial persistence, a dormant phase of host infection sometimes lasting decades,[88] though we note that an MCP of this type is not found in the *M. tuberculosis* genome. Finally, sequence comparisons (using reciprocal Blast searches) between the putative MCP operons across different mycobacterium species highlighted one well-conserved protein (MSMEG_0274 as in *Mycobacterium smegmatis*) whose function is presently unknown.

The automatic computational analysis of MCP types presented here does not successfully identify all the MCP types proposed in the recent literature. Owing to the statistical criteria applied, the computational approach overlooks potential MCP types that occur in only a few instances across the genomic data. The failure to automatically classify the Etu MCP was discussed above. In addition, in a few select bacteria Kinney *et al*.[31] describe potential MCP types based on fuculose aldolase as a key encapsulated enzyme; the purpose of such an MCP could be to sequester the lactaldehyde intermediate. Indeed, we note that the fuculose aldolase in these operons carries an extended N-terminal domain that could be involved in targeting it to the MCP. An alternate sequence feature in the C-terminal region of fuculose aldolase has been indicated as a likely targeting signal by Kinney *et al*.[31]

In summary, our bioinformatics approach has allowed us to more clearly articulate the diversity of MCPs in up-to-date sequenced genomes, and to glean new insights from the organization of their underlying operons. Combining these findings with other recent analyses of MCP operons in the literature, we can assemble a census of all the microcompartment types and variants currently supported by genomic data (Fig. 6). As always with genomic/bioinformatics approaches, the lack of functional annotations for many genes leads to challenges and limitations. Exploring those uncharacterized proteins could be fruitful. Our initial experimental investigation of an unusual shell protein from the proposed Grp MCP supports that view. Meanwhile, further bioinformatics studies are likely to add additional discoveries, for example, using different algorithmic approaches, or using larger data sets as additional genomes are sequenced. Somewhat different approaches may be required to gain insights into systems like the beta carboxsome, where the genomic organization of MCP components is more fragmented. Bearing in mind that bioinformatics studies are only predictive, experimental investigations will be necessary to unravel the biological functions of the new microcompartment types and variations reported here, and to more fully understand the mechanisms by which they operate.

Material and Methods

Operational definition of an MCP operon and BMC-proximal genes

To circumvent the problem of defining true operons across hundreds of microbes, many without experimentally characterized regulatory signatures, we adopted a statistical view of MCP operons. We considered genes encoded within a certain number of open reading frames (upstream or downstream) of one or several paralagous BMC shell genes as potentially belonging to an MCP operon. The general idea of using conserved (bacterial) chromosomal proximity as an indicator of functional linkage has been explored widely and with good success in previous bioinformatics studies.[87–91] In our application, proteins whose genes satisfy the chromosomal proximity requirement in multiple genomes are statistically likely to be part

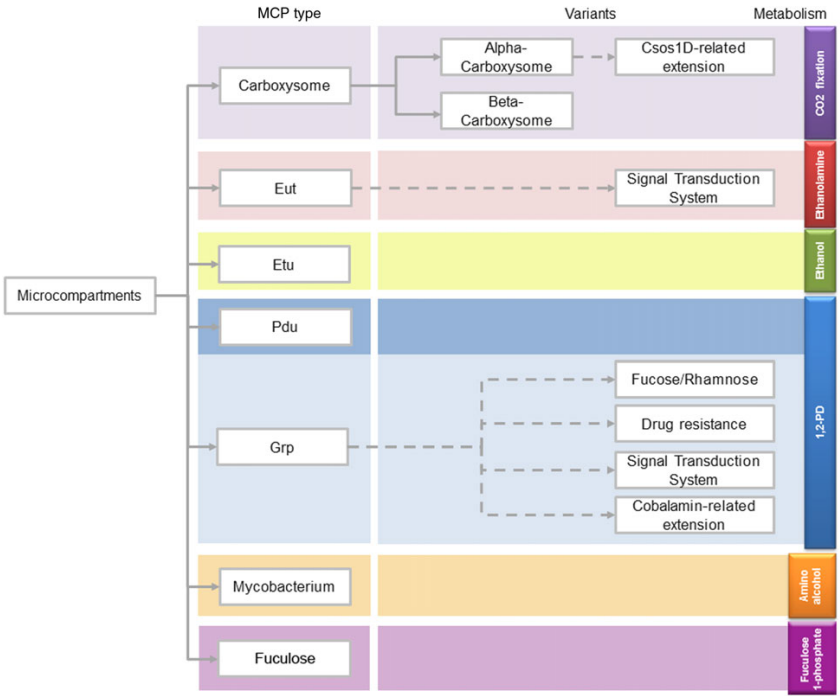


Figure 6. An updated classification of MCPs including presumptive types identified computationally. Microcompartments are divided here into seven main classes according to the core enzymes and pathways confined in their lumen; the two carboxysome subtypes are separated here on the basis of their partially distinct compositions. For some of the MCPs, their core functions appear to be augmented by the presence of additional groups of proteins or enzymes; some of these may be directly involved with MCP function while others could be more peripheral (e.g., regulatory). These extensions suggest the existence of more complex or diverse MCP variants. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of an encapsulated pathway. The strength of this assertion depends on the maximum number of ORFs allowed between a gene in question and a BMC gene; we elected not to consider actual physical intergenic distance or direction of transcription. We refer to this vicinity metric as v. In our analysis, setting v to a maximum value of five yielded the most consistent results in subsequent analyses.

Extracting BMC-proximal proteins

BMC genes were collected from fully sequenced bacterial genomes by scanning for the InterPro profile of the BMC domain (IPR000249)[92] against the UniProt database (release of September 2011) and mapping the resulting hits onto their corresponding gene names. Of all the fully sequenced bacterial genomes present in Uniprot, 113 genomes were identified as carrying BMC genes; no BMC genes were identified in the archaea. Following our operational definition of an MCP operon, the BMC-proximal genes (i.e., those within v genes of a BMC shell gene) were retrieved from the EBI integr8 database and subsequently mapped to their UniProt ID.[95] These constituted our starting set of BMC-proximal proteins. For $v = 5$, the dataset gathered a total of 3120 BMC-proximal protein sequences.

Classifying BMC-proximal protein into functional groups

A first assignment of BMC-proximal protein sequences into internally homologous protein families (or groups) was based on Pfam HMM queries of our database using the *hmmer* package.[94,95] Only those hits reporting an *E*-value lower than 10^{-4} were considered. Proteins exhibiting the same combination of domains were collapsed into the same group. Of the 3120 proteins, 2616 could be collapsed down to 759 protein groups. We judged that some of these groups contained somewhat divergent sequences, with possibly distinct metabolic functions. Therefore, to increase the sensitivity of our classification, a complementary scan of the dataset searching for KEGG Orthology annotations[71] allowed a subdivision of some of these groups into smaller ones, leading to a total number of 807 groups, referred to as Protein Functional Groups in our analysis.

Pairwise correlation coefficients between protein functional groups

Each possible pair of Protein Functional Groups was examined to see if they tended strongly to occur together within individual MCP operons (Fig. 2). To obtain a correlation coefficient for each Protein

Functional Group pair (A,B) we filled a 2×2 contingency table according to the number of operons exhibiting the following combinations: containing both A and B, containing A without B, containing B without A, and lacking both A and B. A pairwise correlation coefficient (PCC) was obtained by assigning an ordered pair value [(A,B) = (1,1), (1,0), (0,1), or (0,0)] for each operon according to the presence or absence of A and B, and then calculating the Pearson correlation coefficient. This measure ranges from +1 to −1, with the extremes indicating a perfect correlation or a perfect anticorrelation respectively, and 0 indicating independence between the two Protein Functional Groups.

Clustering BMC-proximal functional groups based on pairwise correlations

BMC-proximal functional groups were clustered using a graph-based strategy. In this scheme, each Protein Functional Group is a node, and a correlation between a pair of Protein Functional Groups, if it satisfies our statistical tests, is an edge between those two nodes (Fig. 2). We applied multiple statistical criteria to reduce the likelihood of including spuriously identified proteins and linkages in the final analysis. From the list of the Protein Functional Group pairs, we filtered out those combinations having a PCC lower than 0.5, a threshold below which we considered the correlations likely to be functionally insignificant. We applied two additional statistical tests to assess the reliability of both the edges and nodes. For edge validation, we computed a double-tail Fischer's exact test to cope with unevenly distributed values in the contingency tables that could lead to spurious correlations. In other words, we calculated a *P*-value for an edge between Protein Functional Groups A and B, corresponding to the probability of having B occur by chance at least as many times as actually observed in a set of operons that already have A, under the null hypothesis that both groups occur independently. Edges that had a *P*-value higher than a given threshold—set to 10^{-4} in our study—were discarded from the list, along with any subsequently unconnected nodes. This filtering step had the effect of discarding from further consideration Protein Functional Groups that did not occur enough times to allow statistically significant inferences to be made. Out of a starting set of 807 Protein Functional Groups, 64 survived this filtering step. A final statistical analysis aimed to estimate the strength of the nodes upon variation of our operon definition. Indeed, the edges connected to a specific node can be affected through varying the parameter v, since the BMC proximal proteins dataset is built upon its value. For a given node, we analyzed the variation of the set of other nodes to which it was connected when setting v from 5 to 10 by a dependent paired Student's *t*-test (confidence interval equal to 0.95), where the null hypothesis is true if both sample means are not significantly different. Nodes for which the test yielded a value more significant than 10^{-5} were designated as strong, while values inferior to that limit were synonymous for weak nodes. We used the remaining pairs as a basis for clustering.

The final list of Protein Functional Groups and their linkages were clustered using GraphViz(version 2.x) with the "neato" layout (Fig. 3). A node was named either by choosing a consensus string from genomic annotations, when available, or after one of the gene names associated with the node when a consensus could not be established.

Cloning, expression, and purification of the BMC shell protein Pecwa_4094

A codon-optimized version of the Pecwa_4094 gene was synthesized by assembly PCR to include an N-terminal hexa-histidine tag. The gene was cloned into pET22b+ vector via NdeI and XhoI restriction sites. Pecwa_4094 was expressed in BL21(DE3) cells in Luria-Broth at 37°C, shaking at 225 rpm, for 3 h. Cells were pelleted, frozen and stored at −20°C. The cells were resuspended in 50 m*M* Tris pH 7.6, 300 m*M* NaCl with protease inhibitors and sonicated until lysed. Lysate was centrifuged at 16,500 rpm to separate soluble and insoluble fractions for 30 min. The soluble fraction was filtered through a 0.22 μ*M* filter before being loaded onto a 5 mL HiTrap Ni Column at room temperature. Protein was eluted in one step with 50 m*M* Tris pH 7.6, 300 m*M* NaCl, 300 m*M* imidazole pH 8.

Acknowledgments

The authors thank Thomas Bobik, Rob Gunsalus, and Michael Thompson for helpful comments and for critical reading of the manuscript. We also thank Joan Valentine, Kevin Barnese, James Liao and Jenny Takasumi for assistance with anaerobic experiments, and Danny Gidaniyan for assistance in protein expression.

References

1. Gitai Z (2005) The new bacterial cell biology: moving parts and subcellular architecture. Cell 120:577–586.
2. Jensen GJ, Briegel A (2007) How electron cryotomography is opening a new window onto prokaryotic ultrastructure. Curr Opin Struct Biol 17:260–267.
3. Shively JM, Ball F, Brown DH, Saunders RE (1973) Functional Organelles in Prokaryotes - Polyhedral Inclusions (Carboxysomes) of Thiobacillus Neapolitanus. Science 182:584–586.
4. Kerfeld CA, Heinhorst S, Cannon GC (2010) Bacterial microcompartments. Annu Rev Microbiol 64:391–408.
5. Yeates TO, Thompson MC, Bobik TA (2011) The protein shells of bacterial microcompartment organelles. Curr Opin Struct Biol 21:223–231.

6. Yeates TO, Tsai Y, Tanaka S, Sawaya MR, Kerfeld CA (2007) Self-assembly in the carboxysome: a viral capsid-like protein shell in bacterial cells. Biochem Soc Trans 35:508–511.
7. Bobik TA (2006) Polyhedral organelles compartmenting bacterial metabolic processes. Appl Microbiol Biotechnol 70:517–525.
8. Yeates TO, Kerfeld CA, Heinhorst S, Cannon GC, Shively JM (2008) Protein-based organelles in bacteria: carboxysomes and related microcompartments. Nat Rev Microbiol 6:681–691.
9. Raushel FM, Thoden JB, Holden HM (2003) Enzymes with molecular tunnels. Acc Chem Res 36:539–548.
10. Sampson EM, Bobik TA (2008) Microcompartments for B12-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate. J Bacteriol 190:2966–2971.
11. Penrod JT, Roth JR (2006) Conserving a volatile metabolite: a role for carboxysome-like organelles in Salmonella enterica. J Bacteriol 188:2865–2874.
12. Cannon GC, Bradburne CE, Aldrich HC, Baker SH, Heinhorst S, Shively JM (2001) Microcompartments in prokaryotes: carboxysomes and related polyhedra. Appl Environ Microbiol 67:5351–5361.
13. English RS, Lorbach SC, Qin X, Shively JM (1994) Isolation and characterization of a carboxysome shell gene from Thiobacillus neapolitanus. Mol Microbiol 12: 647–654.
14. Heinhorst S, Williams EB, Cai F, Murin CD, Shively JM, Cannon GC (2006) Characterization of the carboxysomal carbonic anhydrase CsoSCA from Halothiobacillus neapolitanus. J Bacteriol 188:8087–8094.
15. Cannon GC, Heinhorst S, Kerfeld CA (2010) Carboxysomal carbonic anhydrases: Structure and role in microbial CO_2 fixation. Biochim Biophys Acta 1804: 382–392.
16. Bobik TA, Havemann GD, Busch RJ, Williams DS, Aldrich HC (1999) The propanediol utilization (pdu) operon of Salmonella enterica serovar Typhimurium LT2 includes genes necessary for formation of polyhedral organelles involved in coenzyme B(12)-dependent 1, 2-propanediol degradation. J Bacteriol 181:5967–5975.
17. Bobik TA, Xu Y, Jeter RM, Otto KE, Roth JR (1997) Propanediol utilization genes (pdu) of Salmonella typhimurium: three genes for the propanediol dehydratase. J Bacteriol 179:6633–6639.
18. Chen P, Ailion M, Bobik T, Stormo G, Roth J (1995) Five promoters integrate control of the cob/pdu regulon in Salmonella typhimurium. J Bacteriol 177: 5401–5410.
19. Garsin DA (2010) Ethanolamine utilization in bacterial pathogens: roles and regulation. Nat Rev Microbiol 8: 290–295.
20. Stojiljkovic I, Baumler AJ, Heffron F (1995) Ethanolamine utilization in Salmonella typhimurium: nucleotide sequence, protein expression, and mutational analysis of the cchA cchB eutE eutJ eutG eutH gene cluster. J Bacteriol 177:1357–1366.
21. Yeates TO, Crowley CS, Tanaka S (2010) Bacterial microcompartment organelles: protein shell structure and evolution. Annu Rev Biophys 39:185–205.
22. Cheng S, Liu Y, Crowley CS, Yeates TO, Bobik TA (2008) Bacterial microcompartments: their properties and paradoxes. Bioessays 30:1084–1095.
23. Tanaka S, Sawaya MR, Phillips M, Yeates TO (2009) Insights from multiple structures of the shell proteins from the beta-carboxysome. Protein Sci 18:108–120.
24. Klein MG, Zwart P, Bagby SC, Cai F, Chisholm SW, Heinhorst S, Cannon GC, Kerfeld CA (2009) Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. J Mol Biol 392:319–333.
25. Crowley CS, Cascio D, Sawaya MR, Kopstein JS, Bobik TA, Yeates TO (2010) Structural insight into the mechanisms of transport across the Salmonella enterica Pdu microcompartment shell. J Biol Chem 285: 37838–37846.
26. Choudhary S, Quin MB, Sanders MA, Johnson ET, Schmidt-Dannert C Engineered protein nano-compartments for targeted enzyme localization. PLoS One 7: e33342.
27. Fan C, Cheng S, Liu Y, Escobar CM, Crowley CS, Jefferson RE, Yeates TO, Bobik TA (2010) Short N-terminal sequences package proteins into bacterial microcompartments. Proc Natl Acad Sci U S A 107: 7509–7514.
28. Heinhorst S, Cannon GC (2010) Addressing microbial organelles: a short peptide directs enzymes to the interior. Proc Natl Acad Sci U S A 107:7627–7628.
29. Parsons JB, Frank S, Bhella D, Liang M, Prentice MB, Mulvihill DP, Warren MJ (2010) Synthesis of empty bacterial microcompartments, directed organelle protein incorporation, and evidence of filament-associated organelle movement. Mol Cell 38:305–315.
30. Fan C, Bobik TA (2011) The N-terminal region of the medium subunit (PduD) packages adenosylcobalamin-dependent diol dehydratase (PduCDE) into the Pdu microcompartment. J Bacteriol 193:5623–5628.
31. Kinney JN, Salmeen A, Cai F, Kerfeld CA (2012) Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly. J Biol Chem 287:17729–17736.
32. Cot SS, So AK, Espie GS (2008) A multiprotein bicarbonate dehydration complex essential to carboxysome function in cyanobacteria. J Bacteriol 190:936–945.
33. Pena KL, Castel SE, de Araujo C, Espie GS, Kimber MS (2010) Structural basis of the oxidative activation of the carboxysomal gamma-carbonic anhydrase, CcmM. Proc Natl Acad Sci U S A 107:2455–2460.
34. Baker SH, Lorbach SC, Rodriguez-Buey M, Williams DS, Aldrich HC, Shively JM (1999) The correlation of the gene csoS2 of the carboxysome operon with two polypeptides of the carboxysome in thiobacillus neapolitanus. Arch Microbiol 172:233–239.
35. Roberts EW, Cai F, Kerfeld CA, Cannon GC, Heinhorst S (2011) Isolation and characterization of the Prochlorococcus carboxysome reveal the presence of the novel shell protein CsoS1D. J Bacteriol 194:787–795.
36. Tanaka S, Kerfeld CA, Sawaya MR, Cai F, Heinhorst S, Cannon GC, Yeates TO (2008) Atomic-level models of the bacterial carboxysome shell. Science 319: 1083–1086.
37. Sagermann M, Ohtaki A, Nikolakakis K (2009) Crystal structure of the EutL shell protein of the ethanolamine ammonia lyase microcompartment. Proc Natl Acad Sci U S A 106:8883–8887.
38. Hayashi NR, Arai H, Kodama T, Igarashi Y (1997) The novel genes, cbbQ and cbbO, located downstream from the RubisCO genes of Pseudomonas hydrogenothermophila, affect the conformational states and activity of RubisCO. Biochem Biophys Res Commun 241:565–569.
39. Galperin MY, Grishin NV (2000) The synthetase domains of cobalamin biosynthesis amidotransferases cobB and cobQ belong to a new family of ATP-dependent amidoligases, related to dethiobiotin synthetase. Proteins 41:238–247.

40. Fresquet V, Williams L, Raushel FM (2004) Mechanism of cobyrinic acid a,c-diamide synthetase from Salmonella typhimurium LT2. Biochemistry 43:10619–10627.
41. Floquet N, Mouilleron S, Daher R, Maigret B, Badet B, Badet-Denisot MA (2007) Ammonia channeling in bacterial glucosamine-6-phosphate synthase (Glms): molecular dynamics simulations and kinetic studies of protein mutants. FEBS Lett 581:2981–2987.
42. Holden HM, Thoden JB, Raushel FM (1999) Carbamoyl phosphate synthetase: an amazing biochemical odyssey from substrate to product. Cell Mol Life Sci 56: 507–522.
43. Naponelli V, Noiriel A, Ziemak MJ, Beverley SM, Lye LF, Plume AM, Botella JR, Loizeau K, Ravanel S, Rebeille F, de Crecy-Lagard V, Hanson AD (2008) Phylogenomic and functional analysis of pterin-4a-carbinolamine dehydratase family (COG2154) proteins in plants and microorganisms. Plant Physiol 146: 1515–1527.
44. Hauer CR, Rebrin I, Thony B, Neuheiser F, Curtius HC, Hunziker P, Blau N, Ghisla S, Heizmann CW (1993) Phenylalanine hydroxylase-stimulating protein/pterin-4 alpha-carbinolamine dehydratase from rat and human liver. Purification, characterization, and complete amino acid sequence. J Biol Chem 268: 4828–4831.
45. Fitzpatrick PF (2000) The aromatic amino acid hydroxylases. Adv Enzymol Relat Areas Mol Biol 74:235–294.
46. Fitzpatrick PF (2003) Mechanism of aromatic amino acid hydroxylation. Biochemistry 42:14083–14091.
47. Sutter M, Boehringer D, Gutmann S, Gunther S, Prangishvili D, Loessner MJ, Stetter KO, Weber-Ban E, Ban N (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. Nat Struct Mol Biol 15:939–947.
48. Dobrinski KP, Enkemann SA, Yoder SJ, Haller E, Scott KM (2012) Transcriptional response of the sulfur chemolithoautotroph Thiomicrospira crunogena to dissolved inorganic carbon limitation. J Bacteriol 194: 2074–2081.
49. Ailion M, Roth JR (1997) Repression of the cob operon of Salmonella typhimurium by adenosylcobalamin is influenced by mutations in the pdu operon. J Bacteriol 179:6084–6091.
50. Fan C, Bobik TA (2008) The PduX enzyme of Salmonella enterica is an L-threonine kinase used for coenzyme B12 synthesis. J Biol Chem 283:11322–11329.
51. Del Papa MF, Perego M (2008) Ethanolamine activates a sensor histidine kinase regulating its utilization in Enterococcus faecalis. J Bacteriol 190:7147–7156.
52. Roof DM, Roth JR (1992) Autogenous regulation of ethanolamine utilization by a transcriptional activator of the eut operon in Salmonella typhimurium. J Bacteriol 174:6634–6643.
53. Heldt D, Frank S, Seyedarabi A, Ladikis D, Parsons JB, Warren MJ, Pickersgill RW (2009) Structure of a trimeric bacterial microcompartment shell protein, EtuB, associated with ethanol utilization in Clostridium kluyveri. Biochem J 423:199–207.
54. Seedorf H, Fricke WF, Veith B, Bruggemann H, Liesegang H, Strittmatter A, Miethke M, Buckel W, Hinderberger J, Li F, Hagemeier C, Thauer RK, Gottschalk G (2008) The genome of Clostridium kluyveri, a strict anaerobe with unique metabolic features. Proc Natl Acad Sci U S A 105:2128–2133.
55. Wackett LP, Frias JA, Seffernick JL, Sukovich DJ, Cameron SM (2007) Genomic and biochemical studies demonstrating the absence of an alkane-producing phenotype in Vibrio furnissii M1. Appl Environ Microbiol 73:7192–7198.
56. Reizer J, Reizer A, Saier MH, Jr. (1993) The MIP family of integral membrane channel proteins: sequence comparisons, evolutionary relationships, reconstructed pathway of evolution, and proposed functional differentiation of the two repeated halves of the proteins. Crit Rev Biochem Mol Biol 28:235–257.
57. Ramos JL, Martinez-Bueno M, Molina-Henares AJ, Teran W, Watanabe K, Zhang X, Gallegos MT, Brennan R, Tobes R (2005) The TetR family of transcriptional repressors. Microbiol Mol Biol Rev 69:326–356.
58. Vindal V, Suma K, Ranjan A (2007) GntR family of regulators in Mycobacterium smegmatis: a sequence and structure based characterization. BMC Genomics 8: 289.
59. Urano N, Kataoka M, Ishige T, Kita S, Sakamoto K, Shimizu S (2010) Genetic analysis around aminoalcohol dehydrogenase gene of Rhodococcus erythropolis MAK154: a putative GntR transcription factor in transcriptional regulation. Appl Microbiol Biotechnol 89: 739–746.
60. Frey PA, Hegeman AD, Reed GH (2006) Free radical mechanisms in enzymology. Chem Rev 106:3302–3316.
61. Selmer T, Pierik AJ, Heider J (2005) New glycyl radical enzymes catalysing key metabolic steps in anaerobic bacteria. Biol Chem 386:981–988.
62. Buckel W, Golding BT (2006) Radical enzymes in anaerobes. Annu Rev Microbiol 60:27–49.
63. Wagner AF, Frey M, Neugebauer FA, Schafer W, Knappe J (1992) The free radical in pyruvate formate-lyase is located on glycine-734. Proc Natl Acad Sci U S A 89:996–1000.
64. Vey JL, Yang J, Li M, Broderick WE, Broderick JB, Drennan CL (2008) Structural basis for glycyl radical formation by pyruvate formate-lyase activating enzyme. Proc Natl Acad Sci U S A 105:16137–16141.
65. Lehtio L, Goldman A (2004) The pyruvate formate lyase family: sequences, structures and activation. Protein Eng Des Sel 17:545–552.
66. Vey JL, Drennan CL Structural insights into radical generation by the radical SAM superfamily. Chem Rev 111:2487–2506.
67. Bouvet OM, Lenormand P, Ageron E, Grimont PA (1995) Taxonomic diversity of anaerobic glycerol dissimilation in the Enterobacteriaceae. Res Microbiol 146:279–290.
68. O'Brien JR, Raynaud C, Croux C, Girbal L, Soucaille P, Lanzilotta WN (2004) Insight into the mechanism of the B12-independent glycerol dehydratase from Clostridium butyricum: preliminary biochemical and structural characterization. Biochemistry 43:4635–4645.
69. Scott KP, Martin JC, Campbell G, Mayer CD, Flint HJ (2006) Whole-genome transcription profiling reveals genes up-regulated by growth on fucose in the human gut bacterium "Roseburia inulinivorans". J Bacteriol 188:4340–4349.
70. Liu Y, Leal NA, Sampson EM, Johnson CL, Havemann GD, Bobik TA (2007) PduL is an evolutionarily distinct phosphotransacylase involved in B12-dependent 1,2-propanediol degradation by Salmonella enterica serovar typhimurium LT2. J Bacteriol 189:1589–1596.
71. Mao X, Cai T, Olyarchuk JG, Wei L (2005) Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics 21:3787–3793.
72. Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) IUPred: web server for the prediction of intrinsically

unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434.
73. Kerfeld CA, Sawaya MR, Tanaka S, Nguyen CV, Phillips M, Beeby M, Yeates TO (2005) Protein structures forming the shell of primitive bacterial organelles. Science 309:936–938.
74. Pang A, Warren MJ, Pickersgill RW (2011) Structure of PduT, a trimeric bacterial microcompartment protein with a 4Fe-4S cluster-binding site. Acta Crystallogr D Biol Crystallogr 67:91–96.
75. Parsons JB, Dinesh SD, Deery E, Leech HK, Brindley AA, Heldt D, Frank S, Smales CM, Lunsdorf H, Rambach A, Gass MH, Bleloch A, McClean KJ, Munro AW, Rigby SE, Warren MJ, Prentice MB (2008) Biochemical and structural insights into bacterial organelle form and biogenesis. J Biol Chem 283:14366–14375.
76. Kofoid E, Rappleye C, Stojiljkovic I, Roth J (1999) The 17-gene ethanolamine (eut) operon of Salmonella typhimurium encodes five homologues of carboxysome shell proteins. J Bacteriol 181:5317–5329.
77. Cannon GC, Baker SH, Soyer F, Johnson DR, Bradburne CE, Mehlman JL, Davies PS, Jiang QL, Heinhorst S, Shively JM (2003) Organization of carboxysome genes in the thiobacilli. Curr Microbiol 46:115–119.
78. Badger MR, Price GD (2003) CO2 concentrating mechanisms in cyanobacteria: molecular components, their diversity and evolution. J Exp Bot 54:609–622.
79. Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–1860.
80. Brinsmade SR, Paldon T, Escalante-Semerena JC (2005) Minimal functions and physiological conditions required for growth of salmonella enterica on ethanolamine in the absence of the metabolosome. J Bacteriol 187:8039–8046.
81. Bay DC, Rommens KL, Turner RJ (2008) Small multidrug resistance proteins: a multidrug transporter family that continues to grow. Biochim Biophys Acta 1778: 1814–1838.
82. Chen P, Andersson DI, Roth JR (1994) The control region of the pdu/cob regulon in Salmonella typhimurium. J Bacteriol 176:5474–5482.
83. Kim C, Mobashery S (2005) Phosphoryl transfer by aminoglycoside 3'-phosphotransferases and manifestation of antibiotic resistance. Bioorg Chem 33:149–158.
84. Aoki R, Nagaya A, Arakawa S, Kato C, Tamegai H (2008) Identification and diversity of putative aminoglycoside-biosynthetic aminotransferase genes from deep-sea environmental DNA. Biosci Biotechnol Biochem 72:1388–1393.
85. Nagaya A, Takeyama S, Tamegai H (2005) Identification of aminotransferase genes for biosynthesis of aminoglycoside antibiotics from soil DNA. Biosci Biotechnol Biochem 69:1388–1393.
86. Daniel J, Deb C, Dubey VS, Sirakova TD, Abomoelak B, Morbidoni HR, Kolattukudy PE (2004) Induction of a novel class of diacylglycerol acyltransferases and triacylglycerol accumulation in Mycobacterium tuberculosis as it goes into a dormancy-like state in culture. J Bacteriol 186:5017–5030.
87. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96: 2896–2901.
88. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402:83–86.
89. Kyrpides NC, Ouzounis CA, Iliopoulos I, Vonstein V, Overbeek R (2000) Analysis of the Thermotoga maritima genome combining a variety of sequence similarity and genome context tools. Nucleic Acids Res 28: 4573–4576.
90. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5:R35.
91. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28:3442–3444.
92. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH, Yeats C (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37:211–215.
93. Mulder NJ, Kersey P, Pruess M, Apweiler R (2008) In silico characterization of proteins: UniProt, InterPro and Integr8. Mol Biotechnol 38:165–177.
94. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763.
95. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:29–37.
96. Crowley CS, Sawaya MR, Bobik TA, Yeates TO (2008) Structure of the PduU shell protein from the Pdu microcompartment of Salmonella. Structure 16: 1324–1332.