Microbiology Monographs

Series Editor: Alexander Steinbüchel

Michael Hippler *Editor*

Chlamydomonas: Molecular Genetics and Physiology

Springer

Microbiology Monographs

Volume 30

Series editor
Alexander Steinbüchel
Münster, Germany

More information about this series at http://www.springer.com/series/7171

Michael Hippler
Editor

Chlamydomonas: Molecular Genetics and Physiology

Springer

Editor
Michael Hippler
Institute of Plant Biology and Biotechnology
Universität Münster
Münster, Germany

ISSN 1862-5576 ISSN 1862-5584 (electronic)
Microbiology Monographs
ISBN 978-3-319-66363-0 ISBN 978-3-319-66365-4 (eBook)
DOI 10.1007/978-3-319-66365-4

Library of Congress Control Number: 2017958132

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Genomics and Functional Genomics in *Chlamydomonas reinhardtii* 1
Ian K. Blaby and Crysten E. Blaby-Haas

Nuclear Transformation and Toolbox Development 27
Jan H. Mussgnug

Mitochondrial Bioenergetics Pathways in Chlamydomonas 59
Simon Massoz, Pierre Cardol, Diego González-Halphen, and Claire Remacle

Bioenergetic Pathways in the Chloroplast: Photosynthetic Electron Transfer 97
Philipp Gäbelein, Laura Mosebach, and Michael Hippler

Chlamydomonas: Bioenergetic Pathways—Regulation of Photosynthesis 135
Jun Minagawa

Chlamydomonas: Anoxic Acclimation and Signaling 155
Anja Hemschemeier

Chlamydomonas: Regulation Toward Metal Deficiencies 201
Eugen I. Urzica

Calcium-Dependent Signalling Processes in Chlamydomonas 233
Glen L. Wheeler

Chlamydomonas: The Eyespot 257
Mark D. Thompson, Telsa M. Mittelmeier, and Carol L. Dieckmann

Genomics and Functional Genomics in *Chlamydomonas reinhardtii*

Ian K. Blaby and Crysten E. Blaby-Haas

Abstract The availability of the *Chlamydomonas reinhardtii* nuclear genome sequence continues to enable researchers to address biological questions relevant to algae, land plants, and animals in unprecedented ways. As we continue to characterize and understand biological processes in *C. reinhardtii* and translate that knowledge to other systems, we are faced with the realization that many genes encode proteins without a defined function. The field of functional genomics aims to close this gap between genome sequence and protein function. Transcriptomes, proteomes, and phenomes can each provide layers of gene-specific functional data while supplying a global snapshot of cellular behavior under different conditions. Herein we present a brief history of functional genomics, the present status of the *C. reinhardtii* genome, how genome-wide experiments can aid in supplying protein function inferences, and provide an outlook for functional genomics in *C. reinhardtii*.

1 Introduction

The first bacterial genome sequence, completed in the mid-1990s (Fleischmann et al. 1995), was rapidly followed by a succession of milestone accomplishments [e.g., the first eukaryote in 1996 (Goffeau et al. 1996), the first plant genome in 2000 (Kaul et al. 2000), and the first mammalian genome in 2002 (Waterston et al. 2002)]. These achievements marked the beginning of a new era in biological sciences. The resulting post-genomic age has equipped researchers with the genetic blueprints for thousands of organisms. Unfortunately, however, it is not immediately clear how these blueprints translate into complex, thriving, and adaptable organisms. Indeed, a key observation made with early genome sequences, and subsequently echoed with each new genome, is that we have complete functional understanding for a small percentage of the encoded proteins [although this perhaps came as little surprise (Bork et al. 1992a, b)]. Increases in the speed of sequence acquisition, accuracy (due in large part to increased depth, or coverage, of

I.K. Blaby (✉) • C.E. Blaby-Haas
Biology Department, Brookhaven National Laboratory, 50 Bell Avenue, Building 463, Upton, NY 11973, USA
e-mail: iblaby@bnl.gov

M. Hippler (ed.), *Chlamydomonas: Molecular Genetics and Physiology*,
Microbiology Monographs 30, DOI 10.1007/978-3-319-66365-4_1

sequence), reductions in cost per genome, and the required sample size have resulted in the pace of genome sequencing far outstripping experimental validation of gene function. Consequently, biologists are presently working with a growing parts list without the knowledge as to what the pieces do, how they interact and how this results in a functioning cell.

1.1 What Is Functional Genomics?

Fortunately, a key development that came with the first genome sequences was the birth of functional genomics (Hieter and Boguski 1997). Functional genomics lies at the interface of two revelations: first, each genome encodes a plethora of proteins with unknown or poorly defined function; second, each genome provides a list of those proteins enabling the scale-up of experiments to simultaneously inform on the function of as many of them as possible (Fig. 1). Determining protein function plays a large part in modern biology generally, but what distinguishes functional genomics from other fields is that protein function is informed by genome-wide studies, such as transcriptomics and proteomics, and high-throughput screening including protein-protein interactions, global protein localization efforts, and mutant library phenotyping (phenomics) (Fig. 1).

As of yet, no single functional genomics approach has come close to deciphering the function of every protein in a given genome. Instead each experiment provides a single, global (but condition-specific) snapshot. The field therefore also encompasses computational analyses that endeavor to establish bridges between the mountains of high-throughout (HTP) genome-wide experimental data and protein function. Integration of functional genomics experiments and other genome-based analyses such as comparative genomics can generate functional inferences. Functional genomics can overlap significantly with systems biology, since both fields employ global approaches to experimentation. Accordingly, systems-wide studies can be highly informative to

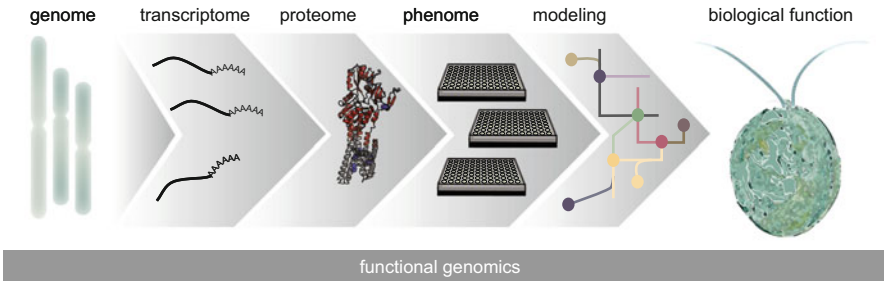

Fig. 1 The genome provides the raw sequence to which genes can be mapped. The function of proteins encoded by those genes can be informed through functional genomics experiments that relate biological information such as changes in transcript or protein abundance and phenotypes to each locus. In order to maximize the value of genome-wide studies, computational analyses and modeling are needed to gain a systems-level understanding of biological function from these experimental pieces

protein function where data is accumulated and interpreted from a holistic perspective. Thereupon, functional genomics and systems biology can form the basis of, and inform the design of, metabolic engineering and synthetic biology approaches.

1.2 What Is Protein Function?

Function is a multifaceted description of a protein's role in the cell. This description has both biochemical and biological elements. For an enzyme, the function would require understanding the reaction(s) catalyzed, the pathway or process the protein participates in, the location of the protein within the cell, how the gene and encoded protein are regulated, and what cofactors are required for activity. As experimental evidence for each of these attributes is provided, the annotation approaches a "gold-standard" level. By virtue of the detailed level of genetic/biochemical characterization to achieve this status, these experiments tend to be performed in laboratories with expertise in the process or pathway under examination, adding an additional level of robustness to the resulting functional annotations.

1.3 The Functional Annotation Problem

Traditionally, functional characterization of proteins has largely been performed one protein at a time. As the number of uncharacterized proteins (individual proteins and protein families) increases with each new genome sequence, the detailed but time-consuming characterization of one protein at a time is an impracticable approach to understanding the function of every protein in every organism. A compromise has been to transfer, or propagate, experimentally determined functions (called functional annotations) to non-experimentally characterized proteins employing sequence similarity. In an age where protein sequences are at our fingertips and bioinformatics play an increasingly large part in biological research, automated functional annotations abound in online databases. It is often difficult to know where annotations originated (i.e., by whom and in which organism the experimental validation was performed) and how reliable they are.

The logic of most automated protein functional annotation is based on the evolutionary concepts of orthology and ancestral relationships; a function may be shared among members of a protein family. Such methods are routinely implemented using BLAST (Altschul et al. 1990, 1997). While some estimates for reliability of similarity-based approaches have been determined (Tian and Skolnick 2003), this mechanism of annotation is generally confounded by the fact that function may not be conserved between even highly similar sequences. Conversely, proteins with dissimilar sequences may have the same function (for example, due to convergent evolution). Furthermore, as a consequence of runaway annotation propagation from genome to genome, 60% of database-deposited functional annotations are estimated to be incorrect for some protein families (Schnoes et al. 2009).

At some point, the genome being automatically annotated is so phylogenetically distant from the organism in which the original function-determining experiment was actually performed that the annotation becomes less relevant and ultimately invalid. This problem is particularly acute for organisms that are evolutionarily distant from very well experimentally characterized model organisms. Automated, structured classifications comprising strictly controlled vocabularies, such as those provided by the gene ontology (GO), and plant-specific tools such as MapMan (Thimm et al. 2004) and Mercator (Lohse et al. 2014) mitigate this annotation error to some extent but generally fail to link an annotation to a specific piece of evidence (Rhee et al. 2008).

1.4 Why Use Functional Genomics?

Each type of functional genomics experiment provides a snapshot of some functional data for many, if not all, genes in an organism at relatively little expense (in terms of time and finance). Although highly detailed protein characterization necessitates expert experimental validation, functional genomics can quickly provide testable hypotheses that drastically reduce the time at the bench it takes for detailed characterization of protein function.

2 *Chlamydomonas reinhardtii* Genomics

2.1 The Status of Protein Function Annotation in Chlamydomonas

A wide range of biological phenomena is studied in *C. reinhardtii* (herein Chlamydomonas), resulting in a wealth of information about Chlamydomonas and by extension plants and animals (Harris 2001, 2009; Goodenough 2015). The Chlamydomonas nuclear genome encodes roughly 19,500 proteins (17,741 loci with 19,526 putative transcripts in JGI 5.5/Phytozome 10 gene model version). The sequence of each protein is predicted from a gene model, which is derived from a combination of prediction tools and sequenced transcript data (such as expressed sequence tags (ESTs), full-length cDNAs, and RNA-Seq reads). Some of these genes are possibly pseudogenes, but it is expected that the number of non-expressed genes is less than 2500 (Zones et al. 2015). All other genes perform some role ensuring that each Chlamydomonas cell survives to the next generation.

Yet, despite intensive study for more than half a century, the role of many genes in Chlamydomonas is unknown (Table 1). This situation is not uncommon. The first eukaryotic genome sequenced, belonging to *Saccharomyces cerevisiae* (Goffeau et al. 1996) and by far the most characterized eukaryotic genome to date, contains 30% of genes that are not associated with a biological process and 40% that are not associated

Table 1 Gene annotation status in Chlamydomonas

Annotation type[a]	Number of genes	Fraction (of primary transcripts[b]) (%)
Primary gene symbol[c]	1815	10.2
Defline and/or description[c]	7834	44.2
PFAM domain[c]	9242	52.1
GO ontology[c]	6636	37.4
MapMan ontology[d]	5685	32.0
Loci that have none of the above	5699	32.1

[a]Determined using v5.5 models, data collected March 2016
[b]Since alternate transcripts (i.e., splice variants) are almost entirely predicted, only primary transcripts are analyzed here
[c]Data sourced from Phytozome (Goodstein et al. 2012)
[d]Thimm et al. (2004)

with a molecular function even though nearly 80% of genes are linked with some form of experimental evidence (as of 24 March 2016; http://www.yeastgenome.org/genomesnapshot). One can imagine that the state of gene function knowledge in more complex eukaryotes such as plants, which have many more genes, is by comparison limited (Chlamydomonas has nearly 10,000 more genes than *S. cerevisiae*, although there is likely some functional redundancy due to gene duplications).

To a certain extent experimental characterization in other organisms can provide some functional information. For instance, about 10,000 Chlamydomonas proteins have at least one PFAM domain (Table 1; however, some of these domains are not associated with a known function). A combination of sequence similarity, more stringent bioinformatics analyses, and experimental evidence has led to annotation of approximately 10% of predicted Chlamydomonas transcripts (Table 1). Making inroads into deciphering complete function of the remaining 90% of genes and placing these genes in a biological context requires functional genomics approaches.

2.2 *The Chlamydomonas Nuclear Genome and Gene Models*

Since many techniques in functional genomics rely on accurate gene models and protein sequences, we start our discussion of functional genomics in Chlamydomonas with a prelude on structural genomics. Raw genome sequence indicates little of biological function. A critical step prior to assigning and defining functional annotations, which describe the role of an encoded protein, is to physically annotate a genome (otherwise known as structural genomics). Physical annotations define the structural attributes of the genome, such as the number and size of chromosomes/scaffolds, positions (i.e., coordinates) of genes on chromosomes, intron/exon junctions, transcription start/stop sites, and translation start/stop sites. Since the accuracy of the predicted protein sequence is dependent upon the quality of the raw genome sequence and accuracy of the gene models, it is crucial for the physical annotations to be of a very high standard before functional annotations can be reliably applied. Fortunately, as a result of numerous efforts integrating

new computational algorithms and the incorporation of experimental datasets, the quality of the Chlamydomonas genome and its gene models are of a high level, as described herein.

After initiation of the Chlamydomonas genome project in the early 2000s and subsequent to two early releases, the draft Chlamydomonas genome (v3) was published in 2007 (Merchant et al. 2007), achieving an averaged 13x coverage by Sanger-based sequencing of 2.1 million reads (Table 2). The draft genome assembly was annotated with the aid of de novo gene prediction tools. This suite of tools comprised statistical models, which computationally scanned the genome for occurrences of putative open reading frames (as determined by translational start and stop sites) and possible intron/exon splice junctions. These models incorporated homology data of translated sequence to known proteins/conserved domains and were supported in conjunction with ~0.25 million ESTs (Haas et al. 2003, 2008; Holt and Yandell 2011). These methods resulted in the JGI v3 generation of gene models, released in 2005 (Table 2). The collection consisted of 15,143 predicted transcripts, of which 82 had manually annotated alternate splice forms, and was deposited in NCBI databases under accession number ABCN01000000 (Merchant et al. 2007). For nomenclature purposes each transcript was assigned a unique six-digit ID. Concurrently with the JGI-led effort, an independent investigation was performed in which a collection of >2000 cDNAs was used as a training set for ab initio gene identification. This study resulted in the GreenGenie2 catalog of gene models (Kwan et al. 2009). As an indication of robustness of these two gene model collections, most (78%) putative transcripts overlapped between the two collections, although many were unique to either the GreenGenie2 (23%) or JGI v3 (22%) collections, suggesting additional work was necessary to capture the complete coding sequence.

A series of successive generations incorporating new methods, data, and algorithms have built upon these early releases resulting in improvements to both the genome assembly and the gene models (Blaby et al. 2014). These post-publication updates started with targeted Sanger-based sequencing of regions with low quality in combination with manual attempts at gap closing. The new sequence attained from these approaches was assisted by a genetic map (Rymarquis et al. 2005), aiding the positioning of scaffolds onto chromosomes. These approaches enabled

Table 2 History of Chlamydomonas genome assemblies and generations of gene models

Genome assemble (release date)	Number of scaffolds (of which chromosomes)	Total sequence Mb (% gaps)	Gene model version (release date)	Number of transcripts (alternate transcripts)
3 (2006)	1557 (na)	120.2 (12.5%)	JGI v3	15,143 (82)
4 (2008)	88 (17)	112.3 (7.5)	JGI v4	16,709 (0)
			Aug u5	15,818 (1070)
			Aug u9	15,935 (0)
			JGI v4.3	17,114 (0)
5 (2012)	54 (17)	111.1 (3.6%)	JGI v5.3.1	17,737 (1789)
			JGI v5.5	17,741 (1785)

significant forward strides to be made; the total number of scaffolds reduced from 1557 in v3 to 88 in the v4 assembly, providing the basis for several rounds of gene model improvements. The first generation of gene models made available for assembly v4, known as JGI v4, constituted 16,709 transcripts. Although this collection of models did not officially support alternate transcripts, analysis by Labadorf et al. inferred 611 alternative splice variants (Labadorf et al. 2010). Subsequent updates were made possible by incremental improvements in the Augustus algorithm, which supported inclusion of EST data (Stanke et al. 2008), and proteomic data (Specht et al. 2011). Consequently, these model collections, released between 2008 and 2012, saw the predicted transcript number increase from 15,818 with the Aug5 release to 17,114 with the Aug10.2 release (Table 1). This latter release, known as JGI v4.3, benefited significantly from the advent of second-generation sequencing technologies, and the inclusion of >6 million ESTs sequenced on the 454 platform, as well as homology to the then-recently sequenced *Volvox carteri* genome (Prochnik et al. 2010). The adoption of Augustus for gene model identification also marked the transition to the now-standardized (and permanent) system for Chlamydomonas loci ID, written in the format Cre*X*.g*Y*. In similar fashion to Arabidopsis loci identifiers (written as AT*X*G*Y*), Cre indicates the organism (*Chlamydomonas reinhardtii*), *X* denotes the chromosome/scaffold, and *Y* is a unique six-digit identifier, which proceeds in numerical order from the "beginning" of chromosome 1 to the "end" of the final scaffold. Initially this identifier increased in increments of 50, thus allowing for the incorporation of new genes without the need for all subsequent loci ID to be reshuffled. JGI v4.3 represented the final gene model version to be based upon the v4 genome.

Work on the next (and, presently, the most up-to-date) revisions to both the nuclear genome sequence and gene models, v5 and v5.5, respectively, took full advantage of advances in second-generation sequencing technologies and was released in mid-2012. Efforts were made to close remaining gaps in the genome sequence by construction of new genomic fragment libraries of differing insert sizes [as a means of compromising between achieving some sequence (with small inserts) and sequencing across repetitive sequence (with long inserts)] with a combination of Sanger and 454-based technologies. These endeavors resulted in reducing the total number of scaffolds from 88 in v4 to 54 in v5 (Table 2). In its present state (v5), the genome totals 111.1 Mbp, averages 65% GC, and comprises 17 chromosomes plus additional 37 repeat-rich non-anchored scaffolds. These chromosomes and scaffolds range ~1–10 Mbp and 0.1–0.8 Mbp, respectively. Common metrics used to describe the quality of a draft genome are the N50 and L50 statistics. N50 describes the smallest number of sequenced fragments required to make 50% of the complete assembly, and L50 is the size of the smallest fragment within the fragments comprising N50. Thus, a more complete genome will have a higher N50 and a lower L50 than a less complete genome of the same size. As work on the Chlamydomonas genome has progressed, the N50 has reduced from 24 to 7, and the L50 has increased from 1.7 to 8 Mb between v3 and v5.

Building upon the JGI v4 gene model inventory, extensive revisions, incorporating more than a billion reads from 59 independent RNA-Seq experiments, including paired reads and stranded libraries (enabling direction and strand of the model to be

determined), were incorporated into an updated Augustus algorithm (Aug11.6) and resulted in the v5.3.1 gene model catalog. During this transition, many complex revisions to gene models were made, such as splitting a single-gene model into multiple models and fusing several models into one (causing additional complications if a named gene is split; in such cases, the name migrated with the conserved domain). Consequently, 28% of loci could not be easily mapped forward and were provisioned with a temporary loci ID in the format g.*X*, where *X* provided a unique placeholder ID. Another, relatively minor, update released in 2014 (v5.5) amended these temporary loci by employing a more robust technique to map ~74% of loci from v4.3 resulting in gene models v5.5. Approximately 15% of gene models were given a new locus as a result of movement between chromosomes/scaffolds. The remaining ~12% could not easily be mapped from the v4 assembly to v5 and were also provided with a new locus. Minor changes were also made to a handful of gene models, with four loci each splitting into two in going from v5.3.1 to v5.5.

As it currently stands in this most recent iteration, the Chlamydomonas reference genome encodes 17,741 transcripts plus 1785 alternate transcripts resulting from splice variants. The coding sequence has an average GC skew of 68% although there is significant deviation between genes. More than 10% (2388) of predicted transcripts have a GC content of $\geq$75%; astonishingly, a handful (66) contain $\geq$80% GC. At the other end of the scale, 138 contain $\leq$55%. These outliers may represent recent acquisitions by horizontal gene transfer and have potentially not yet ameliorated to the average GC content. On the whole, it appears that genes containing a slight skew toward high GC content are more highly expressed, at least at the level of mRNA. By ranking genes by mRNA abundance (i.e., reads per kb per million bases; RPKM) and identifying the top 200 transcripts conserved in five RNA-Seq experiments (Castruita et al. 2011; Boyle et al. 2012; Urzica et al. 2012; Malasarn et al. 2013; Wakao et al. 2014), the codon usage of genes highly expressed in diverse conditions was determined (Table 3). The average GC content in these genes was 64%, with 88% GC at the third base. This codon usage data may aid in efforts for high expression of genes in metabolic engineering.

2.3 *Present and Future Directions for the Chlamydomonas Nuclear Genome*

As with all draft genomes, work continues on both the assembly (continued gap closing) and gene models. Chlamydomonas loci IDs are now static (or will deviate only marginally to allow for transfer of loci as required for future updates). However, the complicated history and the number of studies conducted using earlier generations of gene models necessitate the need to translate between versions. Correspondence tables are available to download for this purpose (http://phytozome.jgi.doe.gov/pz/portal.html). Several studies have highlighted potential directions for future exploration, including a recent genomic analysis suggesting that a fraction of gene models could be extended at the N-terminus (Cross 2015). Although the analysis is entirely of an informatic nature, the implications this may have on

Table 3 Codon usage of highly expressed genes in Chlamydomonas

	Codon	aa	%	Codon	aa	%	Codon	aa	%	Codon	aa	%	
U	UUU	F	0.33	UCU	S	0.46	UAU	Y	0.20	UGU	C	0.06	**U**
	UUC	F	3.18	UCC	S	1.80	UAC	Y	2.51	UGC	C	1.50	**C**
	UUA	L	0.05	UCA	S	0.11	UAA	STOP	0.31	UGA	STOP	0.45	**A**
	UUG	L	0.14	UCG	S	1.83	UAG	STOP	0.33	UGG	W	1.08	**G**
C	CUU	L	0.30	CCU	P	0.60	CAU	H	0.14	CGU	R	0.69	**U**
	CUC	L	1.01	CCC	P	3.34	CAC	H	1.70	CGC	R	4.75	**C**
	CUA	L	0.10	CCA	P	0.16	CAA	Q	0.16	CGA	R	0.05	**A**
	CUG	L	6.25	CCG	P	1.00	CAG	Q	3.31	CGG	R	0.59	**G**
A	AUU	I	1.15	ACU	T	0.64	AAU	N	0.13	AGU	S	0.13	**U**
	AUC	I	3.05	ACC	T	3.50	AAC	N	3.30	AGC	S	1.75	**C**
	AUA	I	0.04	ACA	T	0.14	AAA	K	0.08	AGA	R	0.03	**A**
	AUG	M	2.79	ACG	T	0.78	AAG	K	7.37	AGG	R	0.19	**G**
G	GUU	V	0.68	GCU	A	2.13	GAU	D	0.69	GGU	G	1.43	**U**
	GUC	V	2.32	GCC	A	6.12	GAC	D	4.21	GGC	G	6.29	**C**
	GUA	V	0.06	GCA	A	0.33	GAA	E	0.09	GGA	G	0.17	**A**
	GUG	V	4.61	GCG	A	2.10	GAG	E	5.68	GGG	G	0.28	**G**
	U			**C**			**A**			**G**			