BNL-113704-2017-JA

**Genomics and functional
genomics in *Chlamydomonas reinhardtii***

Ian K. Blaby & Crysten E. Blaby-Haas

Submitted to Microbiology Monographs

March 21, 2017

Biology Department

Brookhaven National Laboratory

**U.S. Department of Energy
USDOE Office of Science (SC),
Biological and Environmental Research (BER) (SC-23)**

Notice: This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DE- SC0012704 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Genomics and functional genomics in *Chlamydomonas reinhardtii*

Ian K. Blaby & Crysten E. Blaby-Haas

Biology Department, Brookhaven National Laboratory, 50 Bell Avenue, Building 463 Upton, NY 11973

To whom correspondence should be addressed: Ian K Blaby; Biology Department, Brookhaven National Laboratory, 50 Bell Avenue, Building 463 Upton, NY 11973 USA; Tel.: 631-344-6037; Email: iblaby@bnl.gov

Table of contents

Abstract

1. Introduction

- 1.1 What is functional genomics?**
- 1.2 What is protein function?**
- 1.3 The functional annotation problem**
- 1.4 Why use functional genomics?**

2. *Chlamydomonas reinhardtii* genomics

- 2.1 The status of protein function annotation in *Chlamydomonas***
- 2.2 The *Chlamydomonas* nuclear genome and gene models**
- 2.3 Present and future directions of the *Chlamydomonas* nuclear genome**

3. Functional genomics in *Chlamydomonas*

- 3.1 Transcriptomics**
- 3.2 Proteomics**
- 3.3 Phenomics**
- 3.4 Other “omics” experimental types**
- 3.5 Databases and resources for *Chlamydomonas* genomics and functional genomics**

4 Outlook and future directions

- 4.1 Integrative approaches to analysis**
- 4.2 Comparative algal genomics**

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 90's (Fleischmann et al., 1995), was rapidly followed by a succession of milestone accomplishments (for example, first eukaryote in 1996 (Goffeau et al., 1996), 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., 1992b, a)). Increases in the speed of sequence acquisition, accuracy (due in large part to increased depth, or coverage, of 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 (Figure 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) (Figure 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-throughput (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 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 multi-faceted 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 in this way: 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; Altschul et al., 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 (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 relevant 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; Harris, 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 2,500 (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 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 2000's, 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; Haas et al., 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 6-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 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 Labadof et al inferred 611 alternative splice variants (Labadof 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 CreX.gY. In similar fashion to *Arabidopsis* loci identifiers (written as ATXG Y), Cre indicates the organism (*Chlamydomonas reinhardtii*), X denotes the chromosome/scaffold, and Y is a unique 6-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 were 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 an 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 8Mb 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 catalogue. 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 forwards 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% (2,388) 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 towards 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, with further revised builds planned for the future. *Chlamydomonas* loci IDs will henceforth remain 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 predicted protein sequence warrants further attention. Another study suggests exonic regions are not included in the present reference assembly genome, potentially indicating that the complete coding sequence (and consequently protein sequences) is not represented (Tulin and Cross, 2015). A further issue requiring attention is the omission of a number of gene models, possibly numbering on the order of a few hundred, that were represented in v4.3 but are not present in v5.5 (Blaby unpublished). It is presently unknown as to why these models were not picked up by the Augustus 11.6 algorithm as many are highly supported by ESTs/RNASeq datasets, and include, for example, *PSBW*, encoding the W protein of photosystem II.

3 Functional genomics in *Chlamydomonas*

Here, we define a functional genomics dataset as the cumulative results of a genome-wide experiment. Specifically, the data has to be linked directly to precise gene loci or proteins. Needless to say, a functional genomics experiment attempts to inform on every gene, transcript or protein in the cell. However, the extent to which a global investigation is truly genome-wide is dependent on the technique used. In this section, we discuss the three predominant types of

functional genomics studies in *Chlamydomonas*: mRNA quantification (transcriptomics), protein detection and quantification (proteomics), and mutant phenotype screening (phenomics). Transcriptomic methods (mainly microarray and RNA-Seq) enable high-throughput examination of gene expression, identifying genes whose mRNA abundance alters significantly between experimental conditions (environment) or strains (wild-type/mutant comparisons). In *Chlamydomonas*, RNA-seq, a deep-sequencing technology, comes the closest to providing quantitative data on every gene on the genome (theoretically informing on every gene model, assuming sequencing depth is sufficient). Proteomic methods can provide information on where proteins are in the cell, an indication of where they perform their function, and can provide information of the differential abundance of proteins between conditions or strains. Proteomic approaches have been able to inform on the presence of 3500 *Chlamydomonas* proteins in a single study, and differential abundance demonstrated for approaching 1000 proteins (Park et al., 2015). Phenomic methods traditionally rely on availability of a library of single-gene deletions or disruptions. The library is screened for mutant phenotypes that deviate from the parent strain used to construct the library (growth defect, colony morphology, chlorophyll content, *etc.*) when grown under precise conditions (such as high light, low nutrient availability, *etc.*). Efforts are being made to provide a sequenced insertional library (Li et al., 2016), which may enable functional genomics studies of mutant phenotypes. In the meantime, phenomic studies in *Chlamydomonas* are limited to forward genetic screens of un-sequenced libraries, isolation of mutants with desired phenotypes, and the subsequent identification of the site of mutation.

3.1 Transcriptomics

Transcriptomes offer a valuable tool in the functional genomics arsenal by providing a snapshot of transcript abundance under well-defined conditions for a large number of transcripts. Transcript abundance changes can be used in several ways to infer function. Higher mRNA abundance in one growth condition (for example low sulfur medium) compared to a second growth condition (such as sulfur-replete medium) can give an indication as to when the gene product is needed (growth during sulfur starvation). There are however several assumptions being made in this simplistic type of function deduction. It can be difficult to know if the observed transcript changes are due to the experimental variable (presence/absence of sulfur) or an indirect consequence of the variable (perhaps redox stress), leading to the question is the corresponding protein involved in sulfur metabolism or a downstream process? Also, in some cases, transcript abundance may be linked to a feedback response involving activity of the encoded protein during the experimental condition. For instance, under iron starvation, the abundances of transcripts encoding several iron-dependent proteins increase while the protein's abundance decreases; it appears that the increase in transcript abundance is responding to the reduced activity of these iron-dependent proteins during iron deprivation rather than a role for

these proteins in acclimating to poor iron nutrition (Urzica et al., 2012). The identification of co-expressed genes can also inform function using the “guilt-by-association” principle (Niehrs and Pollet, 1999), whereby the function of a gene is inferred by its co-expression with another gene of known function.

Transcriptomics in *Chlamydomonas* can be traced back to the first DNA micro-arrays containing probes to detect the differential abundance of ~3000 assembled ESTs (Figure 2) (Im et al., 2003). These early transcriptomes provided data on the influence of several environmental conditions on transcript abundance, including light and CO₂ (Im et al., 2003), sulfur nutrition (Zhang et al., 2004), and photooxidative stress (Ledford et al., 2004). As the EST library grew (Asamizu et al., 2004), DNA microarrays grew to measure over 10,000 ESTs. After the release of the draft *Chlamydomonas* nuclear genome in the early 2000’s, genomic information in addition to ESTs was used in the design of microarrays (Eberhard et al., 2006). The first “whole genome” microarray was published in 2005 and based on gene models associated with draft v2 of the nuclear genome (Stolc et al., 2005). Subsequent microarrays have been designed and employed as more accurate genome sequence drafts and accompanying gene models were released, but in large part, microarray technology has been supplanted by quantitative cDNA sequencing (i.e. RNA-seq).

Reductions in cost and sample size, fast turnaround and streamlined bioinformatic applications (reviewed in (Van Verk et al., 2013)), have resulted in a swelling stream of *Chlamydomonas* RNA-seq datasets (see Chapter 4 of this volume). Furthermore, RNA-Seq datasets adapt to changes in gene models, since the raw sequenced reads can be realigned should updates occur subsequent to the experiment being performed. The types of experiments published over the past 5 years reflect the status of *Chlamydomonas* as a popular microbial experimental system for studying photosynthesis, nutrition, cilia, and abiotic stresses.

3.2 Proteomics

One must remember that transcriptome data only gives an indication of transcriptional and post-transcriptional regulation of mRNA abundance. Many proteins in the cell, however, are regulated at the post-translational level either through post-translational modifications and/or degradation. The detection and quantification of these condition-specific changes requires analysis of proteins. The history of proteomics in *Chlamydomonas* is similar to that of transcriptomics: the number of proteins identified and quantified in a typical experiment has increased with time largely due to availability of the genome sequence (for identification of peptides/proteins) and technological advancements (Rolland et al., 2009).

To gain a functional genomics insight, differential abundance measurements from a quantitative proteomics study utilize the same logic as a transcriptomics study. The assumption is that proteins that are more abundant in a test condition versus standard growth conditions are

involved in acclimating to that test condition thus inferring a functional role. Several studies have been performed in *Chlamydomonas* where researchers quantify changes in the transcriptome and proteome. In general, RNA-Seq and microarrays are able to quantify transcripts from an order of magnitude more genes than proteins using shotgun proteomics (Hsieh et al., 2013; Mettler et al., 2014; Schmollinger et al., 2014).

Unlike transcriptomics, proteomics can inform function by cataloguing the location of proteins within the cell. In this case, proteins are identified after purification of the target compartment or structure. MS-based identification has been performed on proteins present in (and from a functional genomics perspective assumed to carry out a biological role in) the chloroplast (Allmer et al., 2006; Terashima et al., 2010; Bienvenut et al., 2011; Terashima et al., 2011), the eyespot (Schmidt et al., 2006; Wagner et al., 2008; Eitzinger et al., 2015), flagella (Keller et al., 2005; Pazour et al., 2005; Diener et al., 2015), mitochondria (Atteia et al., 2009), nucleus (Winck et al., 2012), and lipid bodies (Moellering and Benning, 2010; Nguyen et al., 2011)

3.3 Phenomics

For geneticists, a condition-specific phenotype resulting from the disruption of a single gene represents the Holy Grail in understanding the function of that disrupted gene. Given this, it comes as little surprise that for many organisms the creation of an indexed, genome-wide mutant library (typically defined as containing independent disruptions in 85% of predicted genes (Carpenter and Sabatini, 2004)) follows shortly after attainment of the genome sequence (Winzeler et al., 1999; Alonso et al., 2003; Baba et al., 2006). With such a library in hand, researchers begin screening the library under various growth conditions and cataloguing the measured phenotypes associated with each disrupted locus. HTP phenotypic profiling enables systematic genotype-to-phenotype relations to be generated. This approach differs from the classical forward genetic screen, where researchers are seeking the genetic basis for a particular phenotype. With reverse genetics, researchers are seeking phenotypes associated with each genotype.

Typically, a large number of parameters are collected, either sequentially or simultaneously, to generate comprehensive datasets for each mutant strain's phenome. Similar logic as with gene co-expression can be applied: if two strains deleted for different genes display similar phenotypes under the same conditions, there may be a functional linkage. This field is presently in its infancy in *Chlamydomonas*, although it is primed to come online in the near future as the availability of known site-of-insertion mutant libraries become available (Li et al., 2016).

Presently, *Chlamydomonas* genetics has a rich history of forward genetic screens. In these screens, a library of mutants with unknown genotypes is screened. The disrupted locus is identified for only those mutants with the desired phenotype. With the availability of HTP sequencing, the identification of these loci has become easier. For instance, a recently published

large-scale study screened roughly 49,000 insertional mutants (of unknown genotype) for growth defects under 12 growth conditions and identified the site of insertion for 439 mutants.

3.4 Other “omics” experimental types

Other global studies, such as metabolomes and lipidomes, have been conducted in *Chlamydomonas*. These investigations can illuminate changes within the cell on the systems level. From a functional genomics perspective of linking function to specific proteins, the data generated from these types of experiments can be useful when viewed in the context of other datasets (such as those described in this section) that can ascribe each observation to a specific loci. Several studies in *Chlamydomonas* have taken this multi-“omic” approach to understanding how *Chlamydomonas* acclimates to different growth conditions, such as rapamycin treatment (Kleessen et al., 2015), nutrient deprivation (Schmollinger et al., 2014; Park et al., 2015), oxygen status (Subramanian et al., 2014), temperature (Légeret et al., 2016) and light (Mettler et al., 2014).

3.5 Databases and resources for *Chlamydomonas* genomics and functional genomics

A growing number of online resources are available to curate, disseminate and interpret *Chlamydomonas* genomics datasets, as well as catalogs for resource distribution (such as strains and vectors). These tools range from data repositories to sophisticated platforms enabling integrative analysis of different data types, including comparative analyses with other organisms. We have summarized a number of the available tools, their uses and the genome assembly/gene model version they presently employ (Table 4). Ideally, the most recent gene models should be used since these represent the most accurate predicted protein sequences. However, many tools offering unique modes of analysis utilize legacy gene model versions, and model generation conversion is necessary.

4 Outlook and future directions

4.1 Integrative approaches to analysis

Since it is unlikely that a single genome-wide technique will definitively enable function to be assigned to proteins, a key concept of functional genomics is the need to assimilate and exploit each piece of evidence relating to any given gene to generate function inferences. As more compatible inferences are made, confidence is gained and function predictions can be made. These predictions can be experimentally tested by applying appropriate targeted genetic and/or biochemical analysis to generate “gold-standard” annotations. As new global datasets are generated, and newer technologies (for instance Chip-Seq or ribosome-profiling) are applied, the integration of data becomes increasingly powerful, as the likelihood of any given gene displaying differential behavior relative to background increases. In any case, analysis of datasets in

isolation may result in incorrect conclusions due to potential false positives and negatives. Consequently an integrative approach to analysis is important (Vidal, 2001; Ge et al., 2003; Joyce and Palsson, 2006) although the inherent difficulties in integrating datasets have also been highlighted (Palsson and Zengler, 2010). Specifically, although large-scale repositories exist for specific data-types (such as SRA and GEO (Edgar et al., 2002; Leinonen et al., 2011)), no system is in place for depositing all datasets. Furthermore experimental variables, such as precise genotypes, lighting conditions, media recipes, and temperature are likely to vary to a greater or lesser extent between studies even when performed within the same laboratory. Stringent record keeping and documentation of experimental conditions are therefore critical, but can be overlooked.

Several online tools catering to *Chlamydomonas* do present multiple data types (Table 4), in particular, expression and co-expression data coupled with genomics. Additional integration of datasets, including proteomes, phenotype and other genome-wide studies will aid these analyses. More recently, big-data approaches are being considered to deploy machine learning techniques to gain maximal knowledge from datasets (Ma et al., 2014).

4.2 Comparative algal genomics

Recently, two *C. reinhardtii* genome comparison studies have offered an appreciation for genome diversity of both commonly used laboratory strains and natural variation in wild isolates (Flowers et al., 2015; Gallaher et al., 2015). Both investigations compared their re-sequenced genomes to the published reference genome of strain CC-503, and identified significant genotypic diversity between strains. Although the relatively small number of strains analyzed in these two investigations prevented GWAS-style linkage of genotype to phenotype, a combination of sequencing more genomes, for example of field isolates, and phenomics approaches of all sequenced strains may help to elucidate these relations. At least to our knowledge, no nuclear genome for any *Chlamydomonas* besides *C. reinhardtii* is presently available, although the mitochondrial genome of *C. eugametos* has been sequenced (Denovan-Wright et al., 1998). As more *Chlamydomonas* genomes become available, the quality of their assembly and critically gene model accuracy improves, and -omic methods are applied to them, the ability to confidently apply the principles outlined herein become more powerful.

Algae comprise a large, diverse group of organisms. Although *Chlamydomonas* remains by far the most studied, and the most “genomically advanced”, a growing number of algal genomes are now available enabling genomes to be interpreted comparatively. Logically, where clear orthologs exist between organisms, indicating vertical evolutionary descent, functional inferences and annotations can be legitimately transferred with higher confidence than simple similarity based methods. Furthermore, this data transfer can be contextualized by the lifestyle,

growth requirements and habitat of the respective organisms as well as by experimentally gathered data.

It is important to note that application of the approaches discussed here only provide functional inferences, and that experimental validation is crucial to be certain of such inferences, and should be considered a requisite before very high confidence (i.e. gold-standard) annotations can be applied. Experimental procedures available in the *Chlamydomonas* molecular toolbox for performing this next step have been reviewed recently (Jinkerson and Jonikas, 2015; Mussnug, 2015) and are updated here in Chapter 2 – Nuclear transformation and toolbox development.

Acknowledgments

IKB and CEB-H are supported by the Office of Biological and Environmental Research of the Department Of Energy.

Tables

Annotation type ^a	Number of genes	Fraction (of primary transcripts) ^a
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%

Table 1 Gene annotation status in *Chlamydomonas*

^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)

Genome	Number of scaffolds	Total	Gene model	Number of transcripts (alternate
--------	---------------------	-------	------------	----------------------------------

assemble (release date)	(of which chromosomes)	sequence Mb(% gaps)	version (release date)	transcripts)
3 (2006)	1557 (na)	120.2 (12.5%)	JGI v3	15143 (82)
4 (2008)	88 (17)	112.3 (7.5)	JGI v4	16709 (0)
			Aug u5	15818 (1070)
			Aug u9	15935 (0)
			JGI v4.3	17114 (0)
5 (2012)	54 (17)	111.1 (3.6%)	JGI v5.3.1	17737 (1789)
			JGI v5.5	17741 (1785)

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

	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			

Table 3 Codon usage of highly expressed genes in Chlamydomonas

Tool/database	Location	Chlamydomonas Genome assembly/gene model generation ^a	Purpose for using			
			Genomic/functional genomics data repository	Functional genomics analysis abilities	Comparative genomics abilities ^b	Tools for validating functional analysis
Phytozome (Goodstein et al., 2012)	http://www.phytozome.net	v5/v5.5	Primary repository of Chlamydomonas genome/gene models. Can bulk-download data as well as obtaining correspondence tables to translate gene model IDs.	Contains user validated annotations, and PFAM, Panther and GO predicted annotations. Gene coexpression tool utilizing some transcriptomes.	Genomic data for > 40 plants/photosynthetic organisms	
UCLA algal genomics portal	http://genomes.mcdb.ucla.edu/	v3-v5/ (transcriptomes presently available mapped to v4/v4.3)	Chlamydomonas genome browser, based on the UCSC browser (Kent et al., 2002). Can bulk download some genomic data	Multiple transcriptomes available to view in genome browser.	Can compare relative mRNA abundances of genes in different conditions	
Algal Annotation Tool (Lopez et al., 2011)	http://pathways.mcdb.ucla.edu/algal/index.html	v3-v5	Batch conversion of gene identifiers.	Bulk annotation prediction via Kegg, MapMan, GO, Panther, Metacyc. Gene coexpression tool utilizing some transcriptomes.		
NCBI	http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins	v3/JGIv3	Presently the only online resource to query Chlamydomonas using PsiBlast ^c (Altschul et al., 1997)			
Predalgo (Tardif et al., 2012)	https://giavap-genomes.ibpc.fr/cgi-bin/predalgodb.perl?page=main	N/A (dependent upon user provided sequence)		Green algal-specific protein localization predictions.		
Plant Transcription	http://plntfdb.bio.uni-	v4/v4.3		Prediction of all	Database covers 20	

factor database (Pérez-Rodríguez et al., 2010)	potsdam.de/v3.0/index.php?sp_id=CRE4			genome encoded transcription factors	photosynthetic organisms	
PLANTTFDB (Jin et al., 2014)	http://planttfdb.cbi.pku.edu.cn/	v4/v4.3		Predicted genome encoded transcription factors	Database covers 83 photosynthetic species	
cis-regulatory site prediction tool (Ding et al., 2012)	http://hulab.ucf.edu/research/projects/Microalgae/sdcree/motifcomb.html	v4/JGIv4.3		Gene regulatory element prediction		
algaePath (Zheng et al., 2014)	http://algaepath.itps.ncku.edu.tw/algae_path/home.html	v5/v5.3.1		Maps mRNA abundance from a number of transcriptomes to KEGG pathway maps	Analysis can also be performed on <i>Neodesmus</i> sp. UTEX 2219-4	
ALCOdb (Aoki et al., 2016)	http://alcodb.jp/	v5/v5.5		Microalgal gene coexpression networks	Identifies orthologs in <i>Cyanidioschyzon merloae</i> , for which (microarray-based) coexpression networks are also present	
ChlamyChem (Alfred et al., 2012)	http://chlamychem.utoronto.ca/ChlamyChem/download.php			Phenotypic and chemical screen resource		
ChlamyCyc5 (May et al., 2009)	http://pmn.plantcyc.org/organism-summary?object=CHLAMY	v3/JGIv3				
PlantSEED (Seaver et al., 2014)	http://bioseed.mcs.anl.gov/~seaver/FIG/seedviewer.cgi?page=PlantSEED	v3/JGIv3			Comparative platform embodying the “subsystem” approach for genome annotation (Overbeek et al., 2005)	
picoPlaza (Vandepoele et al., 2010)	http://bioinformatics.psb.ugent.be/plaza/versions/pico-	v4/JGI4.3			Comparative platform for genomics analysis of	

2013)	plaza/				Chlamydomonas	
Plantgdb (Duvick et al., 2008)	http://www.plantgdb.org/	v4/JGIv4.3	Genome repository and browser		Comparative genomics platform for >100 plants, genome browser for 14 species	
ChlamyNET (Romero-Campero et al., 2016)	http://viridiplantae.ibvf.csic.es/ChlamyNet/	v5/v5.5		Chlamydomonas gene coexpression networks, putative transcription factor binding site predictions		
Chlamydomonas Resource Center	http://chlamycollection.org/					Distributes > 5000 catalogued Chlamydomonas strains, in addition to phenotyped and sequenced mutant collections (Dent et al., 2015; Li et al., 2016). Also stocks plasmids, cDNA libraries, kits etc.
ChlamyStation	http://chlamystation.free.fr/					Paris (IBPC) Collection of photosynthesis mutants.

Table 4 Summary of data repositories and online tools enabling genomics and functional genomics in Chlamydomonas

^a as of March 2016

^b defined here as a tool that can directly comparatively analyze data between i) organisms or ii) conditions within an organism

^c Later versions of the genome/gene models can be queried by running PsiBLAST on a local terminal; source code is available at https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download

References:

- Alfred, S.E., Surendra, A., Le, C., Lin, K., Mok, A., Wallace, I.M., Proctor, M., Urbanus, M.L., Giaever, G., and Nislow, C. (2012). A phenotypic screening platform to identify small molecule modulators of *Chlamydomonas reinhardtii* growth, motility and photosynthesis. *Genome Biol* 13, R105.
- Allmer, J., Naumann, B., Markert, C., Zhang, M., and Hippler, M. (2006). Mass spectrometric genomic data mining: Novel insights into bioenergetic pathways in *Chlamydomonas reinhardtii*. *Proteomics* 6, 6207-6220.
- Alonso, J.M., Stepanova, A.N., Lisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C.C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D.E., Marchand, T., Risseuw, E., Brogden, D., Zeko, A., Crosby, W.L., Berry, C.C., and Ecker, J.R. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* 301, 653-657.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.
- Aoki, Y., Okamura, Y., Ohta, H., Kinoshita, K., and Obayashi, T. (2016). ALCodb: Gene Coexpression Database for Microalgae. *Plant Cell Physiol* 57, e3.
- Asamizu, E., Nakamura, Y., Miura, K., Fukuzawa, H., Fujiwara, S., Hirano, M., Iwamoto, K., Matsuda, Y., Minagawa, J., Shimogawara, K., Takahashi, Y., and Tabata, S. (2004). Establishment of publicly available cDNA material and information resource of *Chlamydomonas reinhardtii* (Chlorophyta) to facilitate gene function analysis. *Phycologia* 43, 722-726.
- Atteia, A., Adrait, A., Brugière, S., Tardif, M., van Lis, R., Deusch, O., Dagan, T., Kuhn, L., Gontero, B., Martin, W., Garin, J., Joyard, J., and Rolland, N. (2009). A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor. *Mol Biol Evol* 26, 1533-1548.
- Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., and Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* 2, 2006.0008.
- Bienvenut, W.V., Espagne, C., Martinez, A., Majeran, W., Valot, B., Zivy, M., Vallon, O., Adam, Z., Meinel, T., and Giglione, C. (2011). Dynamics of post-translational modifications and protein stability in the stroma of *Chlamydomonas reinhardtii* chloroplasts. *Proteomics* 11, 1734-1750.
- Bino, R.J., Hall, R.D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B.J., Mendes, P., Roessner-Tunali, U., Beale, M.H., Trethewey, R.N., Lange, B.M., Wurtele, E.S.,

- and Sumner, L.W. (2004). Potential of metabolomics as a functional genomics tool. *Trends Plant Sci* 9, 418-425.
- Blaby, I.K., Blaby-Haas, C.E., Tourasse, N., Hom, E.F., Lopez, D., Aksoy, M., Grossman, A., Umen, J., Dutcher, S., Porter, M., King, S., Witman, G.B., Stanke, M., Harris, E.H., Goodstein, D., Grimwood, J., Schmutz, J., Vallon, O., Merchant, S.S., and Prochnik, S. (2014). The *Chlamydomonas* genome project: a decade on. *Trends Plant Sci* 19, 672-680.
- Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E. (1992a). What's in a genome? *Nature* 358, 287.
- Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E. (1992b). Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Protein Sci* 1, 1677-1690.
- Boyle, N.R., Page, M.D., Liu, B., Blaby, I.K., Casero, D., Kropat, J., Cokus, S., Hong-Hermesdorf, A., Shaw, J., Karpowicz, S.J., Gallaher, S., Johnson, S., Benning, C., Pellegrini, M., Grossman, A., and Merchant, S.S. (2012). Three acyltransferases and a nitrogen responsive regulator are implicated in nitrogen starvation-induced triacylglycerol accumulation in *Chlamydomonas*. *J. Biol. Chem.* 287, 15811-15825.
- Carpenter, A.E., and Sabatini, D.M. (2004). Systematic genome-wide screens of gene function. *Nat Rev Genet* 5, 11-22.
- Castruita, M., Casero, D., Karpowicz, S.J., Kropat, J., Vieler, A., Hsieh, S.I., Yan, W., Cokus, S., Loo, J.a., Benning, C., Pellegrini, M., and Merchant, S.S. (2011). Systems Biology Approach in *Chlamydomonas* Reveals Connections between Copper Nutrition and Multiple Metabolic Steps. *Plant cell*, 1-21.
- Cross, F.R. (2015). Tying Down Loose Ends in the *Chlamydomonas* genome: Functional Significance of Abundant Upstream Open Reading Frames. G3 (Bethesda).
- Denovan-Wright, E.M., Nedelcu, A.M., and Lee, R.W. (1998). Complete sequence of the mitochondrial DNA of *Chlamydomonas eugametos*. *Plant Mol Biol* 36, 285-295.
- Dent, R.M., Sharifi, M.N., Malnoë, A., Haglund, C., Calderon, R.H., Wakao, S., and Niyogi, K.K. (2015). Large-scale insertional mutagenesis of *Chlamydomonas* supports phylogenomic functional prediction of photosynthetic genes and analysis of classical acetate-requiring mutants. *Plant J* 82, 337-351.
- Diener, D.R., Lupetti, P., and Rosenbaum, J.L. (2015). Proteomic analysis of isolated ciliary transition zones reveals the presence of ESCRT proteins. *Curr Biol* 25, 379-384.
- Ding, J., Li, X., and Hu, H. (2012). Systematic prediction of cis-regulatory elements in the *Chlamydomonas reinhardtii* genome using comparative genomics. *Plant Physiol* 160, 613-623.
- Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M.D., Lawrence, C.J., Lushbough, C., and Brendel, V. (2008). PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res* 36, D959-965.
- Eberhard, S., Jain, M., Im, C.S., Pollock, S., Shrager, J., Lin, Y., Peek, A.S., and Grossman, A.R. (2006). Generation of an oligonucleotide array for analysis of gene expression in *Chlamydomonas reinhardtii*. *Curr Genet* 49, 106-124.

- Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30, 207-210.
- Eitzinger, N., Wagner, V., Weisheit, W., Geimer, S., Boness, D., Kreimer, G., and Mittag, M. (2015). Proteomic Analysis of a Fraction with Intact Eyespots of *Chlamydomonas reinhardtii* and Assignment of Protein Methylation. *Front Plant Sci* 6, 1085.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496-512.
- Flowers, J.M., Hazzouri, K.M., Pham, G.M., Rosas, U., Bahmani, T., Khraiwesh, B., Nelson, D.R., Jijakli, K., Abdrabu, R., Harris, E.H., Lefebvre, P.A., Hom, E.F., Salehi-Ashtiani, K., and Purugganan, M.D. (2015). Whole-Genome Resequencing Reveals Extensive Natural Variation in the Model Green Alga *Chlamydomonas reinhardtii*. *Plant Cell* 27, 2353-2369.
- Gallaher, S.D., Fitz-Gibbon, S.T., Glaesener, A.G., Pellegrini, M., and Merchant, S.S. (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.
- Ge, H., Walhout, A.J., and Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* 19, 551-560.
- Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. (1996). Life with 6000 genes. *Science* 274, 546, 563-547.
- Goodenough, U. (2015). Historical perspective on Chlamydomonas as a model for basic research: 1950-1970. *Plant J* 82, 365-369.
- Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D.S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40, D1178-1186.
- Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 9, R7.
- Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., and White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31, 5654-5666.
- Harris, E. (2009). The Chlamydomonas Sourcebook, 2nd edition ((San Diego, CA: Academic Press)).
- Harris, E.H. (2001). Chlamydomonas as a model organism. *Annu Rev Plant Physiol Plant Mol Biol* 52, 363-406.
- Hieter, P., and Boguski, M. (1997). Functional genomics: it's all how you read it. *Science* 278, 601-602.

- Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491.
- Hsieh, S.I., Castruita, M., Malasarn, D., Urzica, E., Erde, J., Page, M.D., Yamasaki, H., Casero, D., Pellegrini, M., Merchant, S.S., and Loo, J.A. (2013). The proteome of copper, iron, zinc, and manganese micronutrient deficiency in *Chlamydomonas reinhardtii*. *Mol Cell Proteomics* 12, 65-86.
- Im, C.S., Zhang, Z., Shrager, J., Chang, C.W., and Grossman, A.R. (2003). Analysis of light and CO₂ regulation in *Chlamydomonas reinhardtii* using genome-wide approaches. *Photosynth Res* 75, 111-125.
- Kaul, S., Koo, H.L., Jenkins, J., Rizzo, M., Rooney, T., Tallon, L.J., Feldblyum, T., Nierman, W., Benito, M.I., Lin, X. and Town, C.D. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796-815.
- Jin, J., Zhang, H., Kong, L., Gao, G., and Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* 42, D1182-1187.
- Jinkerson, R.E., and Jonikas, M.C. (2015). Molecular techniques to interrogate and edit the *Chlamydomonas* nuclear genome. *Plant J* 82, 393-412.
- Joyce, A.R., and Palsson, B. (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 7, 198-210.
- Keller, L.C., Romijn, E.P., Zamora, I., Yates, J.R., and Marshall, W.F. (2005). Proteomic analysis of isolated *chlamydomonas* centrioles reveals orthologs of ciliary-disease genes. *Curr Biol* 15, 1090-1098.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res* 12, 996-1006.
- Kleessen, S., Irgang, S., Klie, S., Giavalisco, P., and Nikoloski, Z. (2015). Integration of transcriptomics and metabolomics data specifies the metabolic response of *Chlamydomonas* to rapamycin treatment. *Plant J* 81, 822-835.
- Kwan, A.L., Li, L., Kulp, D.C., Dutcher, S.K., and Stormo, G.D. (2009). Improving gene-finding in *Chlamydomonas reinhardtii*: GreenGenie2. *BMC Genomics* 10, 210.
- Labadorf, A., Link, A., Rogers, M.F., Thomas, J., Reddy, A.S., and Ben-Hur, A. (2010). Genome-wide analysis of alternative splicing in *Chlamydomonas reinhardtii*. *BMC Genomics* 11, 114.
- Ledford, H.K., Baroli, I., Shin, J.W., Fischer, B.B., Eggen, R.I., and Niyogi, K.K. (2004). Comparative profiling of lipid-soluble antioxidants and transcripts reveals two phases of photo-oxidative stress in a xanthophyll-deficient mutant of *Chlamydomonas reinhardtii*. *Mol Genet Genomics* 272, 470-479.
- Leinonen, R., Sugawara, H., Shumway, M., and Collaboration, I.N.S.D. (2011). The sequence read archive. *Nucleic Acids Res* 39, D19-21.
- Li, X., Zhang, R., Patena, W., Gang, S.S., Blum, S.R., Ivanova, N., Yue, R., Robertson, J.M., Lefebvre, P., Fitz-Gibbon, S.T., Grossman, A.R., and Jonikas, M.C. (2016). An indexed, mapped mutant library enables reverse genetics studies of biological processes in *Chlamydomonas reinhardtii*. *Plant Cell*.

- Lilly, J.W., Maul, J.E., and Stern, D.B. (2002). The *Chlamydomonas reinhardtii* organellar genomes respond transcriptionally and post-transcriptionally to abiotic stimuli. *Plant Cell* 14, 2681-2706.
- Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A.R., Stitt, M., and Usadel, B. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ* 37, 1250-1258.
- Lopez, D., Casero, D., Cokus, S.J., Merchant, S.S., and 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, 282-282.
- Légeret, B., Schulz-Raffelt, M., Nguyen, H.M., Auroy, P., Beisson, F., Peltier, G., Blanc, G., and Li-Beisson, Y. (2016). Lipidomic and transcriptomic analyses of *Chlamydomonas reinhardtii* under heat stress unveil a direct route for the conversion of membrane lipids into storage lipids. *Plant Cell Environ* 39, 834-847.
- Ma, C., Zhang, H.H., and Wang, X. (2014). Machine learning for Big Data analytics in plants. *Trends Plant Sci* 19, 798-808.
- Malasarn, D., Kropat, J., Hsieh, S.I., Finazzi, G., Casero, D., Loo, J.a., Pellegrini, M., Wollman, F.-A., and Merchant, S.S. (2013). Zinc deficiency impacts CO₂ assimilation and disrupts copper homeostasis in *Chlamydomonas reinhardtii*. *The Journal of biological chemistry* 288, 10672-10683.
- May, P., Christian, J.O., Kempa, S., and Walther, D. (2009). ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii*. *BMC Genomics* 10, 209.
- Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., Marshall, W.F., Qu, L.-H., Nelson, D.R., Sanderfoot, A.a., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S.M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.-L., Cognat, V., Croft, M.T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P.a., Lemaire, S.D., Lobanov, A.V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J.V., Moseley, J., Napoli, C., Nedelcu, A.M., Niyogi, K., Novoselov, S.V., Paulsen, I.T., Pazour, G., Purton, S., Ral, J.-P., Riaño-Pachón, D.M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S.L., Allmer, J., Balk, J., Bisova, K., Chen, C.-J., Elias, M., Gendler, K., Hauser, C., Lamb, M.R., Ledford, H., Long, J.C., Minagawa, J., Page, M.D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A.M., Yang, P., Ball, S., Bowler, C., Dieckmann, C.L., Gladyshev, V.N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R.T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y.W., Jhaveri, J., Luo, Y., Martínez, D., Ngau, W.C.A., Otilar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I.V., Rokhsar, D.S., and Grossman, A.R. (2007). The *Chlamydomonas*

- genome reveals the evolution of key animal and plant functions. *Science* (New York, N.Y.) 318, 245-250.
- Mettler, T., Mühlhaus, T., Hemme, D., Schöttler, M.A., Rupprecht, J., Idoine, A., Veyel, D., Pal, S.K., Yaneva-Roder, L., Winck, F.V., Sommer, F., Vosloh, D., Seiwert, B., Erban, A., Burgos, A., Arvidsson, S., Schönfelder, S., Arnold, A., Günther, M., Krause, U., Lohse, M., Kopka, J., Nikoloski, Z., Mueller-Roeber, B., Willmitzer, L., Bock, R., Schroda, M., and Stitt, M. (2014). Systems Analysis of the Response of Photosynthesis, Metabolism, and Growth to an Increase in Irradiance in the Photosynthetic Model Organism *Chlamydomonas reinhardtii*. *Plant Cell* 26, 2310-2350.
- Moellering, E.R., and Benning, C. (2010). RNA interference silencing of a major lipid droplet protein affects lipid droplet size in *Chlamydomonas reinhardtii*. *Eukaryotic cell* 9, 97-106.
- Mussnug, J.H. (2015). Genetic tools and techniques for *Chlamydomonas reinhardtii*. *Appl Microbiol Biotechnol* 99, 5407-5418.
- Nguyen, H.M., Baudet, M., Cuiné, S., Adriano, J.M., Barthe, D., Billon, E., Bruley, C., Beisson, F., Peltier, G., Ferro, M., and Li-Beisson, Y. (2011). Proteomic profiling of oil bodies isolated from the unicellular green microalga *Chlamydomonas reinhardtii*: with focus on proteins involved in lipid metabolism. *Proteomics* 11, 4266-4273.
- Niehrs, C., and Pollet, N. (1999). Synexpression groups in eukaryotes. *Nature* 402, 483-487.
- Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E.D., Gerdes, S., Glass, E.M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A.C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G.D., Rodionov, D.A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33, 5691-5702.
- Palsson, B., and Zengler, K. (2010). The challenges of integrating multi-omic data sets. *Nat Chem Biol* 6, 787-789.
- Park, J.J., Wang, H., Gargouri, M., Deshpande, R.R., Skepper, J.N., Holguin, F.O., Juergens, M.T., Shachar-Hill, Y., Hicks, L.M., and Gang, D.R. (2015). The response of *Chlamydomonas reinhardtii* to nitrogen deprivation: a systems biology analysis. *Plant J* 81, 611-624.
- Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. *J Cell Biol* 170, 103-113.
- Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L.K., Hellsten, U., Chapman, J., Simakov, O., Rensing, S.A., Terry, A., Pangilinan, J., Kapitonov, V., Jurka, J., Salamov, A., Shapiro, H., Schmutz, J., Grimwood, J., Lindquist, E., Lucas, S., Grigoriev, I.V., Schmitt, R., Kirk, D., and Rokhsar, D.S. (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* 329, 223-226.

- Pérez-Rodríguez, P., Riaño-Pachón, D.M., Corrêa, L.G., Rensing, S.A., Kersten, B., and Mueller-Roeber, B. (2010). PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res* 38, D822-827.
- Rhee, S.Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat Rev Genet* 9, 509-515.
- Rolland, N., Atteia, A., Decottignies, P., Garin, J., Hippler, M., Kreimer, G., Lemaire, S.D., Mittag, M., and Wagner, V. (2009). *Chlamydomonas* proteomics. *Curr Opin Microbiol* 12, 285-291.
- Romero-Campero, F.J., Perez-Hurtado, I., Lucas-Reina, E., Romero, J.M., and Valverde, F. (2016). ChlamyNET: a *Chlamydomonas* gene co-expression network reveals global properties of the transcriptome and the early setup of key co-expression patterns in the green lineage. *BMC Genomics* 17, 227.
- Rymarquis, L.A., Handley, J.M., Thomas, M., and Stern, D.B. (2005). Beyond complementation. Map-based cloning in *Chlamydomonas reinhardtii*. *Plant Physiol* 137, 557-566.
- Schmidt, M., Gessner, G., Luff, M., Heiland, I., Wagner, V., Kaminski, M., Geimer, S., Eitzinger, N., Reissenweber, T., Voytsekh, O., Fiedler, M., Mittag, M., and Kreimer, G. (2006). Proteomic analysis of the eyespot of *Chlamydomonas reinhardtii* provides novel insights into its components and tactic movements. *Plant Cell* 18, 1908-1930.
- Schmollinger, S., Mühlhaus, T., Boyle, N.R., Blaby, I.K., Casero, D., Mettler, T., Moseley, J.L., Kropat, J., Sommer, F., Strenkert, D., Hemme, D., Pellegrini, M., Grossman, A.R., Stitt, M., Schroda, M., and Merchant, S.S. (2014). Nitrogen-Sparing Mechanisms in *Chlamydomonas* Affect the Transcriptome, the Proteome, and Photosynthetic Metabolism. *Plant Cell*.
- Schnoes, A.M., Brown, S.D., Dodevski, I., and Babbitt, P.C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS computational biology* 5, e1000605-e1000605.
- Seaver, S.M., Gerdes, S., Frelin, O., Lerma-Ortiz, C., Bradbury, L.M., Zallot, R., Hasnain, G., Niehaus, T.D., El Yacoubi, B., Pasternak, S., Olson, R., Pusch, G., Overbeek, R., Stevens, R., de Crécy-Lagard, V., Ware, D., Hanson, A.D., and Henry, C.S. (2014). High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. *Proc Natl Acad Sci U S A* 111, 9645-9650.
- Specht, M., Stanke, M., Terashima, M., Naumann-Busch, B., Janssen, I., Höhner, R., Hom, E.F., Liang, C., and Hippler, M. (2011). Concerted action of the new Genomic Peptide Finder and AUGUSTUS allows for automated proteogenomic annotation of the *Chlamydomonas reinhardtii* genome. *Proteomics* 11, 1814-1823.
- Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637-644.
- Stolc, V., Samanta, M.P., Tongprasit, W., and Marshall, W.F. (2005). Genome-wide transcriptional analysis of flagellar regeneration in *Chlamydomonas reinhardtii* identifies orthologs of ciliary disease genes. *Proc Natl Acad Sci U S A* 102, 3703-3707.

- Subramanian, V., Dubini, A., Astling, D.P., Laurens, L.M., Old, W.M., Grossman, A.R., Posewitz, M.C., and Seibert, M. (2014). Profiling *Chlamydomonas* metabolism under dark, anoxic H₂-producing conditions using a combined proteomic, transcriptomic, and metabolomic approach. *J Proteome Res* 13, 5431-5451.
- Tardif, M., Atteia, A., Specht, M., Cogne, G., Rolland, N., Brugière, S., Hippler, M., Ferro, M., Bruley, C., Peltier, G., Vallon, O., and Cournac, L. (2012). PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol Biol Evol* 29, 3625-3639.
- Terashima, M., Specht, M., and Hippler, M. (2011). The chloroplast proteome: a survey from the *Chlamydomonas reinhardtii* perspective with a focus on distinctive features. *Curr Genet* 57, 151-168.
- Terashima, M., Specht, M., Naumann, B., and Hippler, M. (2010). Characterizing the anaerobic response of *Chlamydomonas reinhardtii* by quantitative proteomics. *Mol Cell Proteomics* 9, 1514-1532.
- Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.a., Rhee, S.Y., and Stitt, M. (2004). Mapman: a User-Driven Tool To Display Genomics Data Sets Onto Diagrams of Metabolic Pathways and Other Biological Processes. *The Plant Journal* 37, 914-939.
- Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 333, 863-882.
- Tulin, F., and Cross, F. (2015). Patching holes in the *Chlamydomonas* genome (bioRxiv).
- Urzica, E.I., Casero, D., Yamasaki, H., Hsieh, S.I., Adler, L.N., Karpowicz, S.J., Blaby-Haas, C.E., Clarke, S.G., Loo, J.A., Pellegrini, M., and Merchant, S.S. (2012). Systems and Trans-System Level Analysis Identifies Conserved Iron Deficiency Responses in the Plant Lineage. *The Plant cell*, 1-29.
- Van Verk, M.C., Hickman, R., Pieterse, C.M., and Van Wees, S.C. (2013). RNA-Seq: revelation of the messengers. *Trends Plant Sci* 18, 175-179.
- Vandepoele, K., Van Bel, M., Richard, G., Van Landeghem, S., Verhelst, B., Moreau, H., Van de Peer, Y., Grimsley, N., and Piganeau, G. (2013). pico-PLAZA, a genome database of microbial photosynthetic eukaryotes. *Environ Microbiol* 15, 2147-2153.
- Vidal, M. (2001). A biological atlas of functional maps. *Cell* 104, 333-339.
- Wagner, V., Ullmann, K., Mollwo, A., Kaminski, M., Mittag, M., and Kreimer, G. (2008). The phosphoproteome of a *Chlamydomonas reinhardtii* eyespot fraction includes key proteins of the light signaling pathway. *Plant Physiol* 146, 772-788.
- Wakao, S., Chin, B.L., Ledford, H.K., Dent, R.M., Casero, D., Pellegrini, M., Merchant, S.S., and Niyogi, K.K. (2014). Phosphoprotein SAK1 is a regulator of acclimation to singlet oxygen in *Chlamydomonas reinhardtii*. *Elife* 3, e02286.
- Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M.R., Brown, D.G., Brown, S.D., Bult, C., Burton, J., Butler, J., Campbell, R.D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A.T., Church, D.M., Clamp, M., Clee, C., Collins, F.S.,

- Cook, L.L., Copley, R.R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K.D., Deri, J., Dermitzakis, E.T., Dewey, C., Dickens, N.J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D.M., Eddy, S.R., Elnitski, L., Emes, R.D., Eswara, P., Eyra, E., Felsenfeld, A., Fewell, G.A., Flicek, P., Foley, K., Frankel, W.N., Fulton, L.A., Fulton, R.S., Furey, T.S., Gage, D., Gibbs, R.A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T.A., Green, E.D., Gregory, S., Guigó, R., Guyer, M., Hardison, R.C., Haussler, D., Hayashizaki, Y., Hillier, L.W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D.B., Johnson, L.S., Jones, M., Jones, T.A., Joy, A., Kamal, M., Karlsson, E.K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W.J., Kirby, A., Kolbe, D.L., Korf, I., Kucherlapati, R.S., Kulbokas, E.J., Kulp, D., Landers, T., Leger, J.P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D.R., Mardis, E.R., Matthews, L., Mauceli, E., Mayer, J.H., McCarthy, M., McCombie, W.R., McLaren, S., McLay, K., McPherson, J.D., Meldrim, J., Meredith, B., Mesirov, J.P., Miller, W., Miner, T.L., Mongin, E., Montgomery, K.T., Morgan, M., Mott, R., Mullikin, J.C., Muzny, D.M., Nash, W.E., Nelson, J.O., Nhan, M.N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M.J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K.H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C.S., Poliakov, A., Ponce, T.C., Ponting, C.P., Potter, S., Quail, M., Reymond, A., Roe, B.A., Roskin, K.M., Rubin, E.M., Rust, A.G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M.S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J.B., Slater, G., Smit, A., Smith, D.R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J.P., Von Niederhausern, A.C., Wade, C.M., Wall, M., Weber, R.J., Weiss, R.B., Wendl, M.C., West, A.P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R.K., Winter, E., Worley, K.C., Wyman, D., Yang, S., Yang, S.P., Zdobnov, E.M., Zody, M.C., Lander, E.S., and Consortium, M.G.S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562.
- Winck, F.V., Riaño-Pachón, D.M., Sommer, F., Rupprecht, J., and Mueller-Roeber, B. (2012). The nuclear proteome of the green alga *Chlamydomonas reinhardtii*. *Proteomics* 12, 95-100.
- Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D.J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J.L., Riles, L., Roberts, C.J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R.K., Véronneau, S., Voet, M., Volckaert, G., Ward, T.R., Wysocki, R., Yen, G.S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., and Davis, R.W. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901-906.

- Zhang, Z., Shrager, J., Jain, M., Chang, C.W., Vallon, O., and Grossman, A.R. (2004). Insights into the survival of *Chlamydomonas reinhardtii* during sulfur starvation based on microarray analysis of gene expression. *Eukaryot Cell* 3, 1331-1348.
- Zheng, H.Q., Chiang-Hsieh, Y.F., Chien, C.H., Hsu, B.K., Liu, T.L., Chen, C.N., and Chang, W.C. (2014). AlgaePath: comprehensive analysis of metabolic pathways using transcript abundance data from next-generation sequencing in green algae. *BMC Genomics* 15, 196.
- Zones, J.M., Blaby, I.K., Merchant, S.S., and Umen, J.G. (2015). High-Resolution Profiling of a Synchronized Diurnal Transcriptome from *Chlamydomonas reinhardtii* Reveals Continuous Cell and Metabolic Differentiation. *Plant Cell* 27, 2743-2769.

Figure legends

Figure 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.

Figure 2 Progress of genomics and transcriptomics in *Chlamydomonas*. The green line represents number of nuclear-encoded protein sequences in Genbank starting at 1990. Because of the redundancy caused by duplicated entries, protein entries from each year were mapped to JGI 5.5. If the sequence could not be mapped forward, it was not included in the graph. The number of entries in Genbank significantly increased with deposition of predicted proteins associated with the genome publication (Merchant et al., 2007), with only a handful of new sequences deposited subsequently. Presently, only about 60% of *Chlamydomonas* proteins in the Phytozome 11 database are represented in Genbank, and the gene models associated with many of those deposited in 2007 have changed as a result of more accurate genome assemblies and gene model prediction algorithms. Grey bars indicate the number of gene models released to the public in the indicated years that map onto the v5.5 gene models. The purple line represents the progress of transcriptomics in *Chlamydomonas*; the number of “gene-specific” probes reported for each microarray is given, but it should be noted that there are potential redundancies and artifacts due to the limited availability of accurate gene models when many of these microarrays were designed. Starting in 2002 (Lilly et al., 2002) with a microarray that included probes for 51 nuclear genes, transcriptomics of the nuclear genome progressively increased in scope over the next decade. As EST libraries were made public followed by the first drafts of the nuclear genome, the size of

microarrays continued to increase, while the publication of *Chlamydomonas* RNA-Seq experiments began in 2010. Although the present release of JGI gene models includes alternative transcripts, most RNA-Seq analyses only include mapping to the primary transcripts.

genome

transcriptome

proteome

phenome

modeling

biological function

functional genomics