Enhancing heterologous expression in *Chlamydomonas reinhardtii* by transcript sequence optimization

Iddo Weiner¹,², Shimshi Atar², Shira Schweitzer¹, Haviva Eilenberg³, Yael Feldman¹, Meital Avitan¹,², Mor Blau³, Avihai Danon³, Tamir Tuller²,⁴ and Iftach Yacoby¹,*

¹The George S. Wise Faculty of Life Sciences, School of Plant Sciences and Food Security, Tel Aviv University, Ramat Aviv, Tel Aviv, 69978, Israel,
²Department of Biomedical Engineering, The Iby and Aladar Fleischman Faculty of Engineering, Tel Aviv University, Tel Aviv 6997801, Israel,
³Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot 76100, Israel, and
⁴The Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 6997801, Israel

Received 16 November 2017; revised 22 December 2017; accepted 4 January 2018.
*For correspondence (e-mail iftachy@tau.ac.il; tamirtul@post.tau.ac.il).
Iddo Weiner, Shimshi Atar and Shira Schweitzer are the authors who contributed equally.

**SUMMARY**

Various species of microalgae have recently emerged as promising host-organisms for use in biotechnology industries due to their unique properties. These include efficient conversion of sunlight into organic compounds, the ability to grow in extreme conditions and the occurrence of numerous post-translational modification pathways. However, the inability to obtain high levels of nuclear heterologous gene expression in microalgae hinders the development of the entire field. To overcome this limitation, we analyzed different sequence optimization algorithms while studying the effect of transcript sequence features on heterologous expression in the model microalga *Chlamydomonas reinhardtii*, whose genome consists of rare features such as a high GC content. Based on the analysis of genomic data, we created eight unique sequences coding for a synthetic ferredoxin–hydrogenase enzyme, used here as a reporter gene. Following *in silico* design, these synthetic genes were transformed into the *C. reinhardtii* nucleus, after which gene expression levels were measured. The empirical data, measured *in vivo* show a discrepancy of up to 65-fold between the different constructs. In this work we demonstrate how the combination of computational methods and our empirical results enable us to learn about the way gene expression is encoded in the *C. reinhardtii* transcripts. We describe the deleterious effect on overall expression of codons encoding for splicing signals. Subsequently, our analysis shows that utilization of a frequent subset of preferred codons results in elevated transcript levels, and that mRNA folding energy in the vicinity of translation initiation significantly affects gene expression.

**Keywords:** heterologous expression, sequence optimization, microalgae, *Chlamydomonas reinhardtii*, hydrogenase, gene expression.

**INTRODUCTION**

Microalgae are a diverse group of unicellular photosynthetic organisms found in a variety of habitats, from arid deserts to saline seawater to freshwater lakes and damp soils. Recently, various species of microalgae have emerged as promising host organisms for use in biotechnology industries due their unique features. These primarily include: (i) efficient conversion of sunlight to organic compounds – obviating the need for expensive fixed carbon in the growth medium, (ii) the ability of many microalgal species to grow in extreme conditions – minimizing costs related to keeping growth conditions sterile, and (iii) the occurrence of numerous post-translational modification pathways enabling proper maturation for a variety of proteins (León-Bañares et al., 2004; Walker et al., 2005; Yusuf, 2007; Markou et al., 2012; Scaife et al., 2015).

However, attempts to scale up the production of various biological products in microalgae-based technologies face several challenges. The most severe underlying problem, which is considered rate-limiting for the advancement of...
the microalgae-based industry, is the poor overall expression of functional heterologous genes (Fuhrmann et al., 1999; Schroda et al., 2000; Shao and Bock, 2008; Neupert et al., 2009; Rasala et al., 2012). Indeed, most existing large-scale microalgae-based production systems focus on producing substances that occur naturally in the host, such as beta-carotene (NBT), astaxanthin (Algategach), polysaccharides (Nikken Sohonsa) and unsaturated fatty acids (Subitec GmbH). While microalgae do offer an impressive variety of natural substances, novel technologies require the expression of heterologous genes – thus these techniques must be significantly improved for the whole industry to progress.

Recently, transformation protocols and genomic data have been accumulating for several microalgal species, with the most prominent one being *Chlamydomonas reinhardtii*. Effective *C. reinhardtii* transformation protocols exist (Yamano et al., 2013) and its genome has been sequenced completely (Merchant et al., 2007). Transcription abundance was measured in different strains and conditions (González-Ballester et al., 2010) and initial data from more advanced genomic analyses (of which data exist for very few eukaryotes), such as proteomics (Rolland et al., 2009) and ribosomal profiling (Zoschke et al., 2013), is starting to accumulate as well. Other advantages of using *C. reinhardtii* as a model microalga include its rapid cell growth and dominant haploid vegetative stage that simplifies the detection of mutants (Harris, 2001; Grossman et al., 2003).

The main targets for heterologous gene expression are the nuclear and chloroplast genomes. While chloroplast expression was shown to be effective in some cases (Su et al., 2005; Rasala et al., 2010; Jones et al., 2013), nuclear expression contains unique features, including the eukaryotic maturation processes, stable genome transformation and optional secretion pathways (León-Bañares et al., 2004). Thus, it is necessary to obtain an efficient tool for nuclear heterologous gene expression.

Nuclear transformation is most commonly done by inserting linear DNA into the cell via electroporation (Yamano et al., 2013). The DNA strand is incorporated into the genome by non-homologous end joining, resulting in unpredicted locations of the new insert in the genome. This causes a ‘position effect’, in which transcription levels vary between different clones transformed with the same DNA fragment (León and Fernández, 2007). Furthermore, gene silencing is an additional factor that plays a role in hindering heterologous expression in microalgae. Transgene silencing in microalgae was shown to occur both at the transcriptional and post-transcriptional levels (Cerutti et al., 1997; Wu-Scharf et al., 2000; Shaver et al., 2010), and is most likely part of an ‘immune system’ protecting the cell against viral infections and transposable elements.

Several important attempts to elevate the expression of heterologous genes in the nucleus have been reported. The first inquiries focused on searching for strong promoters; the most notable ones found being the endogenous promoters of the *PsaD* (Fischer and Rochaix, 2001), *RBCS2* (Lumbrañas et al., 1998) and *HSP70A* (Schroda et al., 2000, 2002) native *C. reinhardtii* genes. In a recent work it has been shown that intelligently designed synthetic promoters could also support robust expression and outcompete native promoters in some cases (Scranton et al., 2016). Subsequently, integration of the native *RBCS2* introns in the promoter and coding sequence (CDS) was shown to drive higher expression (Lumbrañas et al., 1998; Eichler-Stahlberg et al., 2009).

The next rounds of improvement included the establishment of strong molecular toolkits. In one method the CDS of the gene of interest is fused to the zeocin resistance *shbleh* gene, with a self-cleaving peptide between them. Thus, the gene, which is separated from the resistance marker during translation, enjoys overall high expression driven by selection (Rasala et al., 2012). Another beneficial tool is the versatile pOptimized vector which contains different modules for facilitating nuclear expression (Lauersen et al., 2015). In recent reports this vector was used to achieve better levels of valuable recombinant proteins (Lauersen et al., 2016).

Although these notable advances have supported a significant improvement in the field of heterologous expression in microalgae, the need for a holistic gene-expression platform remains. Some important codon selection techniques have been reported (Fuhrmann et al., 1999, 2004; Shao and Bock, 2008; Lauersen et al., 2015), but the greater field of transcript sequence optimization has not been thoroughly investigated and constitutes a missing piece in the current system comprising effective promoters and strong expression toolkits. *In silico* models for efficient design of synthetic sequences could play a key role in optimizing gene expression in microalgae.

To address these needs, we studied gene expression features in *C. reinhardtii* and incorporated them into different sequence optimization algorithms, thus creating different versions of a synthetic ferredoxin-hydrogenase (fd-hyd) enzyme (Yacoby et al., 2011; Eilenberg et al., 2016), used here as a reporter gene. This enzyme was built by fusing endogenous ferredoxin to a hydrogenase with a truncated signal peptide by a synthetic linker. Throughout this work we kept the exact same amino acid (AA) sequence of fd-hyd, making only synonymous changes in the CDS and slight modifications of the untranslated regions (UTRs). Nevertheless, we observed high variability in both protein and transcript abundances resulting from these changes. Ultimately, we show here that a properly constructed synthetic gene drives significantly higher heterologous expression.
RESULTS

Conserved sequence features

To detect conserved sequence features in *C. reinhardtii* related to gene expression we studied a reference set of highly expressed genes from the analysis of two independent RNA sequencing (RNA-seq) measurements (González-Ballester et al., 2010; Blau & Danon). We found strong selection for high mRNA folding energy (i.e. loosely bound structures) in the translation initiation site compared with randomized mRNA sequences (Figure 1a). This observation matches similar global trends reported across an array of organisms (Gu et al., 2010; Tuller et al., 2010b).

In a separate analysis, conserved UTR base compositions were detected by aligning all the reference genes to the start or stop codons (Figure 1b). A similar analysis performed on the intron edges yielded the 5′ and 3′ splice sites (Figure 1c). Additionally, codon usage bias (CUB) was determined (Figure S1 in the online Supporting Information). To measure the effect of these features on *C. reinhardtii* gene expression we created eight separate variants (Table 1) all coding for the same fd-hyd reporter gene (Figure 2).

Unintended splicing and its deleterious effect in heterologous transcripts

When planning a synthetic sequence *de novo*, unintended splice sites could be formulated within the CDS. To evaluate the effect of such cases on heterologous gene expression, we constructed two genes coding for the same fd-hyd enzyme. These base CDSs were built by either mimicking the natural CUB in our reference set (construct 2) or selecting the most frequent codon for each AA (construct 4). Subsequently, the conserved 5′ and 3′ UTR nucleotide (NT) patterns (Figure 1b) were added to the edges of these constructs. These base sequences were next scanned for subsequences resembling splice sites. Such subsequences were synonymously mutated to lower spliceosome affinity. This methodology yielded sequence 3, which was 98% identical to sequence 2, and sequence 5, which was 92% identical to sequence 2 (Figure 3a, c).

---

**Figure 1.** Conserved sequence features in *Chlamydomonas reinhardtii*. (a) mRNA folding energy profile of original highly expressed genes (black) compared with the mean profile over 1000 randomizations of these genes (green). For each window the result of the empirical statistical test is shown. (b) Nucleotide bias at the edges of the 5′ (left) and 3′ (right) untranslated regions closest to the coding sequence. The upper panels show the distribution of nucleotides per position, where the orange circles indicate strongly preferred nucleotides which were incorporated into the template of the synthetic genes. The lower panels depict the corresponding Shannon entropy profiles. (c) A sequence logo for the 5′ (left) and 3′ (right) splicing signals.
These four constructs were cloned into the psl18 vector in which transcription is driven by the strong constitutive light-dependent psaD promoter (Fischer and Rochaix, 2001). The vectors were transformed into the nucleus of a C. reinhardtii mutant with negligible hydrogenase activity (hydA1,2) (Meuser et al., 2012) and screened using the Rhodobacter capsulatus high-throughput system (Wecker et al., 2011; Wecker and Ghirardi, 2014; Eilenberg et al., 2016) for detection of hydrogen-producing clones, which signify successful heterologous expression (Figure S2b–e).

The positive clones were isolated and grown in standard conditions to mid-log phase. The cells were next concentrated and anaerobically induced to drive protein synthesis and hydrogenase maturation (Happe and Naber, 1993). Protein abundance was measured using the methyl violo- gen protein quantification assay (Eilenberg et al., 2016), which sensitively detects the abundance of fd-hyd enzymes in a cell lysate (see Figure S3 and Experimental Procedures for more details). In order to average out the noise introduced by the ‘position effect’ we isolated a pool of individual positive clones from each group, measured each of their protein abundances separately and used the overall mean and variance for statistical comparison between different groups. The results show that groups 3 and 5, in which the splicing signals were erased, dramatically outperform groups 2 and 4, respectively (Figure 3b, d).

The codon composition of a gene affects its overall expression in various ways, including transcription rate (Edri et al., 2014; Zhou et al., 2016), mRNA stability (Hafner et al., 2010; Presnyak et al., 2015; Harigaya et al., 2016; Mishima and Tomari, 2016), translation initiation (Tuller and Zur, 2015) and the efficiency of translation elongation.

<table>
<thead>
<tr>
<th>Number</th>
<th>Codon selection method</th>
<th>Optimized UTR edges</th>
<th>Splicing sites deleted</th>
<th>Optimized mRNA folding energy around START site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WT codons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Mimicking the reference CUB (shown in Figure S1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Mimicking the reference CUB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Most frequent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Most frequent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ChimeraMap (Zur and Tuller, 2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Outsourcing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Outsourcing</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

UTR, untranslated region; WT, wild type; CUB, codon usage bias.

Table 1 A list of the different constructs analyzed in this work, including the optimization features carried out on each of them.

Figure 2. Schematic illustration of the base construct for the ferredoxin–linker–hydrogenase enzyme, used here as a reporter gene. [Colour figure can be viewed at wileyonlinelibrary.com].
and ribosome allocation (Tuller, Carmi, et al., 2010; Tuller, Waldman, et al., 2010; Ben-Yehezkel et al., 2015; Dana and Tuller, 2015). To better understand the mechanisms responsible for the typical expression received with each codon selection method, we selected three representative clones from each group and quantified the abundance of each fd-hyd transcript by quantitative PCR (qPCR). Total RNA was isolated from all selected strains (RNeasy Plant Mini Kit, Qiagen 74903) immediately after growth and induction. For each strain, 2 μg of RNA was reverse-transcribed (High Capacity cDNA Reverse Transcription Kit, ABI 4368814) to form a cDNA pool. Real-time qPCR was carried out to measure the abundance of fd-hyd mRNA in all selected clones. The obtained reads were fitted to a standard curve calculated for serial dilutions of the relevant clean plasmid. Subsequently, these values were normalized by the absolute abundance of the stably expressed cblp reference gene (Jokel et al., 2015). It has been previously shown that this normalization step yields accurate and comparable transcript quantities (Lu et al., 2012).

This analysis revealed an interesting phenomenon: the hierarchy observed at the protein level was maintained at the transcript level (Figure 4b), suggesting a correlation between codon composition and transcript abundance in C. reinhardtii. Curiously, while group 5 showed a remarkably high level of mRNA abundance, the transcript levels of the WT construct were below the detection rate, most likely due to a combination of extremely low transcript abundance and the inherent problem of primer non-specificity in this WT sequence.

Improving an outsource sequence by optimizing mRNA folding energy in the vicinity of translation initiation

To examine sequence optimization carried out by outsourcing and to compare it with our methods, an optimized fd-hyd gene was ordered from a company that offers C. reinhardtii gene optimization (but does not disclose its methods for doing this). This sequence (construct 7) was scanned for potential sub-optimal regions based on our previous findings. We built a new fd-hyd version (construct 8) in which we added the START and STOP contexts described above (Figure 1b). Additionally, we made 10 base substitutions just downstream of the START codon (Figure 5a). These synonymous mutations were made in order to increase mRNA folding energy in the vicinity of translation initiation, based on the observed selection for high free energy in this area (Figure 1a).

Both constructs were cloned and transformed into the nuclei of hydA1,2 mutant cells, then screened for H2 production (Figure S2g, h) and positive clones were isolated. Protein abundance measurements revealed a significant 2.4-fold increment in overall heterologous expression in
group 8 clones (Figure 5b). Subsequent transcript quantification of two clones from different groups but with similar protein levels (Figure 5c) revealed that the group 7 clone required higher levels of mRNA in order to produce the same amount of protein as its counterpart from group 8 (Figure 5d). These results emphasize the importance of loosely bound mRNA structures in the proximity of the START codon for effective translation initiation in *C. reinhardtii*.

**DISCUSSION**

The difficulty of achieving high levels of nuclear heterologous gene expression in microalgae is hindering the progress of both research and microalgae-based technologies. In this work we studied different aspects of gene expression in the model microalga *C. reinhardtii*, focusing on issues related to the design of synthetic genes. Using a fd-hyd reporter gene system, comprising a high-throughput biological screening system, an efficient methyl viologen-based protein quantification technique and absolute qPCR, we were able to analyze the performance of different sequence optimization algorithms and distinguish key principles for designing heterologous genes and for understanding gene expression in microalgae.

**Unintended splicing and its deleterious effect in heterologous transcripts**

*Chlamydomonas reinhardtii* has an intron-rich genome. The alga has a high number of introns per gene (about 8.5), frequent events of alternative splicing, typically long introns (about 370 bp) and only a small group of intron-less genes (around 8%) (Merchant *et al.*, 2007; Labadorf *et al.*, 2010). All these characteristics set it apart from most single-cell eukaryotes and suggest that the process of splicing plays an important role in post-transcriptional regulation. By comparing constructs in which the problem of spliceosome binding to unintended CDS loci was not treated with their optimized counterparts, we show the deleterious effect of these incidents on overall protein levels. Splicing is a complex process, and proper splicing requires a combination of splicing signals. The expression of a transcript containing splice sites could be hampered in several ways, including the incision of the mRNA, disruption of transcription, delay of transport through the nuclear pore complex and disruption of ribosome binding or elongation. A qPCR analysis on three representative clones from groups 2, 3, 4 and 5 (Figure S4) suggested that the mechanism responsible for the discrepancy in overall expression (Figure 3b, d) might not be the same for different constructs. Subsequently, since incisions may have occurred along the transcript, interpretation of these results might be impeded by the sensitivity of the qPCR method to primer selection.

**Codon selection**

Codons affect the overall expression of a gene in various ways. Codons are most commonly associated with translation, although it is well established and has been shown in various eukaryotes that a gene’s codon composition also

![Figure 4. The effect of codon selection methods on gene expression.](image-url)
regulates transcription efficiency and transcript stability; codon combinations form binding sites for transcription factors (Stergachis et al., 2013; Sullivan et al., 2014) and also play an important role in determining transcription elongation (Edri et al., 2014; Zhou et al., 2016). Mature transcripts containing certain subsequences can be targeted for downregulation by microRNA binding (Hafner et al., 2010; Presnyak et al., 2015; Harigaya et al., 2016; Mishima and Tomari, 2016) or other degradation processes (Duan et al., 2003). In this work we observed significant differences in overall expression of a transgene depending on the manner in which its codons were selected (Figure 4a). By examining a representative subset of each group, we have shown that a major cause for these differences is varying transcript levels (Figure 4b). A closer look at this result raises an interesting hypothesis: construct 5 utilizes only the most frequent codons by default, though splicing deletion and maximizing mRNA folding energy introduced 122 less frequent codons into the final construct. The ChimeraMap algorithm, in this work, yielded a less biased codon distribution, whereas construct 3 was the least biased of these three. This could be quantified by different measures for codon bias, including the Codon Adaptation Index (CAI; Sharp and Li, 1987) (Figure S5a) and the effective number of codons (ENC; Wright, 1990) (Figure S5b). Interestingly, the manner in which the mRNA hierarchy follows the codon bias hierarchy for these three constructs could point towards a causal connection between the codon ‘optimality’ of a gene and its corresponding mRNA levels, given that all other features are kept constant.

The position effect

During nuclear transformation in C. reinhardtii, integration occurs by non-homologous end joining, causing a ‘position effect’ in which the genomic region of integration affects the transcription rates of the heterologous gene. This effect caused clones transformed with the same sequence to yield different levels of protein. Indeed, the ranking between clones derived from measuring transcript abundance was in accordance with the protein level hierarchy. To overcome these problems, we pooled all Rhodobacter-positive clones and statistically compared the groups, thus averaging out the noise introduced by the ‘position effect’. The only scenario in which this would not have solved the issue of the ‘position effect’ is if an individual sequence were somehow biased in its tendency to be incorporated into a certain genomic locus. Little is known about such biases; in general, it is thought that integration can occur all along the genome and PCR-based methods have shown that insertions in specific areas could be screened for (Gonzalez-Ballester et al., 2011). To conclude, the genomic integration mechanism is not completely understood; however, if sequence-based position biases do exist, we do not expect there to be any here because all sequences we used were highly similar (Figure S6).
reinhardtii sequence optimization services. By synonymously mutating eight codons in the translation initiation area we were able to increase the mean protein abundance received from transformants by 2.4-fold (Figure 5b). These modifications were engineered to increase the folding energy of mRNA secondary structures around the START codon, matching the conserved feature we found while studying highly expressed C. reinhardtii genes (Figure 1a).

While there were also seven base substitutions in the linker region (introduced into construct 8 due to cloning restrictions of the synthesis company) which could potentially affect translation elongation, the discrepancy in protein abundance is most likely to be caused by higher translation initiation rates (Figure 5c,d). This successful increment in heterologous expression shows that single base substitutions can cause dramatic effects on overall expression. A similar explanation could contribute to understanding the superiority of construct 5 over construct 4 (Figure 3d). Here, too, there is a discrepancy in folding energy favoring the more highly expressed group 5, although in this case the differences are relatively subtle.

**Recommendations for transgene design**

Based on this work, we were able to provide several general recommendations for optimizing a DNA sequence towards high heterologous expression in C. reinhardtii (Figure 6). In this work the best results were achieved in constructs 5 and 8 (the difference between these two was found to be insignificant) in which a combination of sequence optimization methods was used, thus we recommend following the whole pipeline. Additionally, our original data set alongside a code for C. reinhardtii sequence optimization can be found at https://www.mathworks.com/matlabcentral/fileexchange/65416-coding-sequence-optimization-for-chlamydomonas-reinhardtii

**EXPERIMENTAL PROCEDURES**

**Detection and deletion of splicing signals**

To detect splicing signals, the highly expressed set of 100 nuclear C. reinhardtii genes was used as reference for calculation of the CAI (Sharp and Li, 1987) in all intron-containing genes. Of the genes with the top 10% CAI values, half were randomly selected (resulting in 281 genes) and the sequence of their first intron was examined. A window containing both exon and intron NTs was manually fine-tuned to include the most informative subsequence. The received thresholds were: 5\text{PSSM} = 0.42, 3\text{PSSM} = 0.33.

While traversing the synthetic genes, windows exceeding one of the thresholds were synonymously mutated to achieve window scores within the approved limits. This was implemented by examining all alternative codon options for the problematic window, considering only codons with more than 15% usage. From all options, the one with the lowest PSSM score was selected. This typically "fixed" the window’s score and brought it below the threshold. If this objective could not be achieved, we chose the lowest-score option.

**Optimizing mRNA folding energy (FE) in synthetic sequences**

The target function was to find the highest mean FE over all 40 NT windows, ranging from -39 to +39 (relative to the START codon). The degrees of freedom derived from synonymous mutations in the CDS region and base substitutions in the 7-NT area that was added ('CGCAAAA') immediately upstream of the START codon, which resulted in about $3 \times 10^9$ different combinations to test.

---

**Figure 6. Recommended sequence optimization pipeline.**

AA, amino acid; UTR, untranslated region; CDS, coding sequence; FE, folding energy [Colour figure can be viewed at wileyonlinelibrary.com].

The FE per window was calculated using the Vienna package (Hofacker et al., 1999) V2.19.

**Methyl viologen protein quantification**

Carried out precisely as described in Eilenberg et al. (2016). Following 2 h of dark anaerobiosis, cells were transferred into a buffer containing reduced methyl viologen and Triton-X for lysing the cells. A 500-μl sample was drawn from the headspace and the H₂ concentration was determined by GC. The amount of enzyme was calculated based on the constant fd-hyd specific activity (see Table S1 for all protein levels measured).

**Quantitative PCR**

One hundred milligrams of the cell pellet was taken for total RNA extraction using an RNeasy Plant Mini Kit (Qiagen 74903, http://www.qiagen.com/). Two micrograms of purified RNA from each sample was used for synthesis of cDNA. Serial template dilutions provided a standard curve to which the samples were fitted. The qPCR reaction was performed with Applied Biosystems StepOnePlus™ real time PCR system (http://www.thermofisher.com/il/en/home/life-science/pcr/real-time-pcr). A sample from all cDNA pools (each isolated from a different clone) was put through a subsequent qPCR analysis, amplifying a sequence of the reference gene (each isolated from a different clone) was put through a subsequent qPCR analysis, amplifying a sequence of the reference gene.

**ACKNOWLEDGEMENTS**

This study was supported by: The Israelni Ministry of Science, Technology and Space; a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University; a fellowship from the Rieger foundation for environmental studies; and The Gordon Center for Energy Studies. We would like to thank Professor Matthew Posewitz for the hydA1,2 double hydrogenase knockout mutant and Professors Matt Wecker and Maria Ghirardi for the H₂-sensing *R. capsulatus*.

**AUTHOR CONTRIBUTIONS**

IW, IY and TT designed the research; SA, TT and IW performed the computational analyses; IW, SS, HE, IY, YF, MA, MB and AD performed the experimental procedures; IW, IY and TT wrote the paper.

**CONFLICT OF INTEREST**

The authors declare no conflict of interest.

**SUPPORTING INFORMATION**

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

- **Figure S1.** codon bias in the reference set.
- **Figure S2.** *Rhodobacter* high-throughput screening. Algal plates overlaid with *R. capsulatus*, used to detect positive fd-hyd clones.
- **Figure S3.** Verification of the methyl viologen quantification assay by immunoblot quantification.
- **Figure S4.** mRNA levels for clones from groups 2, 3, 4 and 5.
- **Figure S5.** Codon bias measurements for different fd-hyd versions.
- **Figure S6.** Number of mismatches between sequences.
- **Table S1.** Summary of all protein levels measured.

**Methods S1.** Supplemental experimental procedures.

**REFERENCES**


Chlamydomonas reinhardtii. BMC Genom., 11, 114.


