ACS SyntheticBiology

Research Article

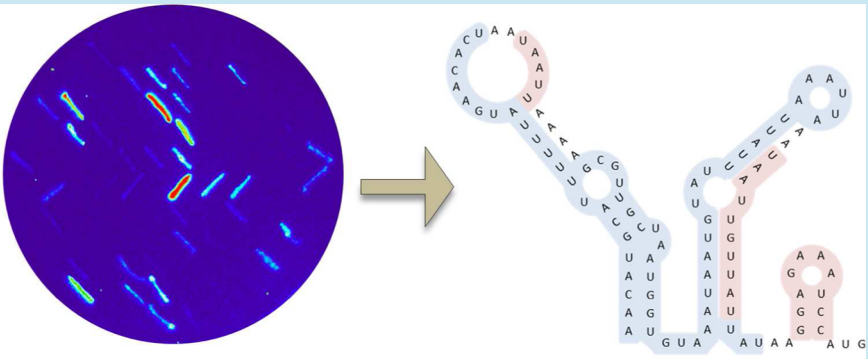
pubs.acs.org/synthbio

Synthetic Oligonucleotide Libraries Reveal Novel Regulatory Elements in *Chlamydomonas* Chloroplast mRNAs

Elizabeth A. Specht and Stephen P. Mayfield*

The San Diego Center for Algae Biotechnology, Division of Biological Sciences, University of California San Diego, La Jolla, California 92093, United States

Supporting Information



ABSTRACT: Gene expression in chloroplasts is highly regulated during translation by sequence and secondary-structure elements in the 5′ untranslated region (UTR) of mRNAs. These chloroplast mRNA 5′ UTRs interact with nuclear-encoded factors to regulate mRNA processing, stability, and translation initiation. Although several UTR elements in chloroplast mRNAs have been identified by site-directed mutagenesis, the complete set of elements required for expression of plastid mRNAs remains undefined. Here we present a synthetic biology approach using an arrayed oligonucleotide library to examine *in vivo* hundreds of designed variants of endogenous UTRs from *Chlamydomonas reinhardtii* and quantitatively identify essential regions through next-generation sequencing of thousands of mutants. We validate this strategy by characterizing the relatively well-studied 5′ UTR of the *psbD* mRNA encoding the D2 protein in photosystem II and find that our analysis generally agrees with previous work identifying regions of importance but significantly expands and clarifies the boundaries of these regulatory regions. We then use this strategy to characterize the previously unstudied *psaA* 5′ UTR and obtain a detailed map of regions essential for both positive and negative regulation. This analysis can be performed in a high-throughput manner relative to previous site-directed mutagenesis methods, enabling compilation of a large unbiased data set of regulatory elements of chloroplast gene expression. Finally, we create a novel synthetic UTR based on aggregate sequence analysis from the libraries and demonstrate that it significantly increases accumulation of an exogenous protein, attesting to the utility of this strategy for enhancing protein production in algal chloroplasts.

KEYWORDS: *oligonucleotide synthesis, translational regulation, chloroplast gene regulation, synthetic biology*

Microalgae have significant potential for mass production of industrial molecules such as biofuels, therapeutic proteins, or industrial enzymes, due to their ability to produce biomass at large scale in a rapid and cost-effective manner. However, current levels of recombinant protein accumulation in algae are well below that achieved in other production hosts,[1] making only the most valuable products economically suitable for algal production. The highest yields achieved to date in the model alga *Chlamydomonas reinhardtii* are around 10% of total soluble protein,[1] but most recombinant proteins accumulate to less than 1%.[2] To fully realize the potential of algae as a biotechnology platform, we need robust and controlled gene expression in both the nuclear and chloroplast genomes.

Although progress has been made understanding gene expression in algae, the regulatory processes that govern gene expression and protein accumulation in the chloroplast remain an area of active investigation. A better understanding of the regulatory processes that govern protein accumulation may enable us to design custom regulatory regions that overcome the current limitations in producing recombinant proteins.

Protein expression in *Chlamydomonas* chloroplasts is regulated primarily during translation, governed by regulatory sequences in the untranslated regions (UTRs) of mRNAs and by *trans*-acting factors that interact with elements in these UTRs. Chloroplast 3′ UTRs are processed at the distal end of a stem-loop structure that appears to protect the mature transcript from 3′ to 5′ exonucleases,[3] but these 3′ sequences appear to have little impact on protein expression[4] although

Received: July 30, 2012

there is some evidence that UG repeats in 3′ UTRs are involved in circadian expression.[5] For chloroplast gene expression, most research has focused on the 5′ UTRs of plastid mRNAs, as these have been shown to exert significant control over protein accumulation, impacting both translation rates and mRNA stability.

One of the most well-studied UTRs, the *psbA* 5′ UTR has been shown to interact with a four-member protein complex,[6] and an RNA stem-loop and adjacent ribosome binding site have been identified as key RNA elements required for translation.[7] Several nuclear mutants have been found that affect *psbA* translation or stability, revealing the complex regulatory interactions between the nuclear genome and the plastid genome.[8] *Cis*-acting elements have also been identified in the 5′ UTRs of *petD*,[9] *rbcL*, *atpB*,[10] *psbD*,[11] and a few other genes, but the mechanisms by which these elements exert control over translation are not well understood. The details of plastid RNA elements and translational regulation have been obtained piecemeal through individual biochemical or mutant analyses, making it difficult to know how many elements are present in any mRNA, and how important any one element may be in controlling gene expression. Using a more systematic approach to characterize many chloroplast UTRs may elucidate recurring structures or sequences, leading to a more comprehensive understanding of plastid gene regulation.

In addition to gaining a better understanding of native mRNA elements involved in gene regulation, the thorough systematic analysis of UTR regulatory elements will allow us to design truly synthetic UTRs for driving expression of exogenous genes in transgenic algae. A notable characteristic of some of the most commonly used endogenous chloroplast regulatory regions is that they exhibit autoattenuation and therefore can only be used effectively in a strain in which the native gene has been deleted. For example, the *psbA* regulatory regions are capable of high levels of recombinant protein production, but the *psbA* gene product inhibits expression of additional *psbA* transcript. Therefore these regulatory regions are only useful for driving exogenous gene expression in a non-photosynthetic *psbA* knockout strain, eliminating the energetic benefit of using a photosynthetic organism to produce recombinant proteins at large scale. By identifying regions useful for strong positive regulation across many UTRs, we may be able to mix and match sequence elements from many UTRs to create novel synthetic elements that evade the negative feedback mechanisms involved in autoattenuation while maintaining high expression of the transgene of interest.

Here we present a systematic synthetic biology approach for identifying important 5′ UTR sequence elements involved in modulating chloroplast gene expression. We used large-scale oligonucleotide synthesis to create libraries of variant UTRs and cloned these libraries into vectors driving the expression of a codon-optimized luciferase reporter gene. By selecting pools of transformants that have high, medium, or low expression of the reporter, followed by next-generation sequencing of these different pools, we were able to quantitatively identify effector elements from each of the pooled groups. We validate this method by confirming previous partially characterized elements within the *psbD* 5′ UTR, though we find that these elements extend far beyond the region initially identified through site-directed mutagenesis, highlighting the superiority of a comprehensive, unbiased approach. We then extend the analysis to *psaA* to identify novel regulatory regions within its 5′ UTR. Finally, we demonstrate the predictive power of our method by creating a synthetic UTR, based on the aggregate data from the *psaA* 5′ UTR, that outperforms the wild type version.

RESULTS AND DISCUSSION

Design and Representation of Libraries. Two highly expressed chloroplast mRNA 5′ UTRs, from the *psaA* and *psbD* genes, were selected for variant analysis to identify potential regulatory elements. These genes encode subunits of photosystem I (*psaA*) and photosystem II (*psbD*). The prevalence of these photosystem proteins in the thylakoid membranes, as well as their high turnover as a result of photodamage, contributes to their inherent high expression. These UTRs were also selected because of their short length, ensuring that they could be synthesized with high fidelity to make a variant library with a manageable number of members.

The *psbD* mRNA is transcribed with a 74 nucleotide 5′ UTR that is cleaved to 47 nucleotides through a processing step that appears to be coupled with translation.[11,12] Previous analyses of the regulatory elements in the *psbD* 5′ UTR have been somewhat complicated by the introduction of multiple mutations in a single strain, insertion of restriction sites into the UTR, or use of the *psbD* protein itself as a reporter for expression.[11,13] In this last instance there is the additional complication of autoregulation of synthesis in which unassembled subunits of the photosystem complexes can significantly impact their own synthesis.[14] Because D2 is the limiting assembly partner in PSII, a mutant UTR that raises or lowers the amount of D2 produced will have significant downstream effects on the regulation of all other subunits of the PSII complex.[14] As a result of this impact on chloroplast gene regulation, the readout may reflect the downstream consequences of the mutant UTR instead of providing an unbiased analysis of the UTR's performance alone.

In our system we have reduced these epistatic effects as much as possible to allow an unbiased analysis that can be compared to the results of previous studies. By constructing libraries using the USER cloning system[15] to clone the synthesized UTRs directly downstream of their respective promoter, there are no areas containing engineered restriction site sequences that could potentially alter UTR activity. All of our UTRs were specifically designed and synthesized, so the entire UTR was scanned and all possible variants generated to cover the entire UTR. All nucleotide replacements were made with homopolymer adenosine and not with random combinations of nucleotides, to ensure that properties like local GC content were not unpredictably affected and so that new RNA secondary structures were not inadvertently introduced.

We used the chloroplast codon-optimized luciferase reporter *luxAB* for expression analysis[16] instead of the D2 protein as a reporter, and all constructs were examined in a photosynthetic strain to ensure that there are no translational regulation artifacts from defective photosystem assembly. The selection for transformation was restoration of photosynthesis, as constructs were transformed into a strain with a deletion in the *psbH* gene which was subsequently repaired to wild type by the transformation vector as previously described.[17] In all constructs, the 3′ UTR remained unaltered; it is the native 3′ UTR from the *psbA* transcript.

The *psaA* 5′ UTR is far less well-characterized than the *psbD* 5′ UTR. The *psaA* 5′ UTR appears to extend up to 238 nucleotides upstream of the initiation codon, based on reverse transcription data from the gene's original identification in

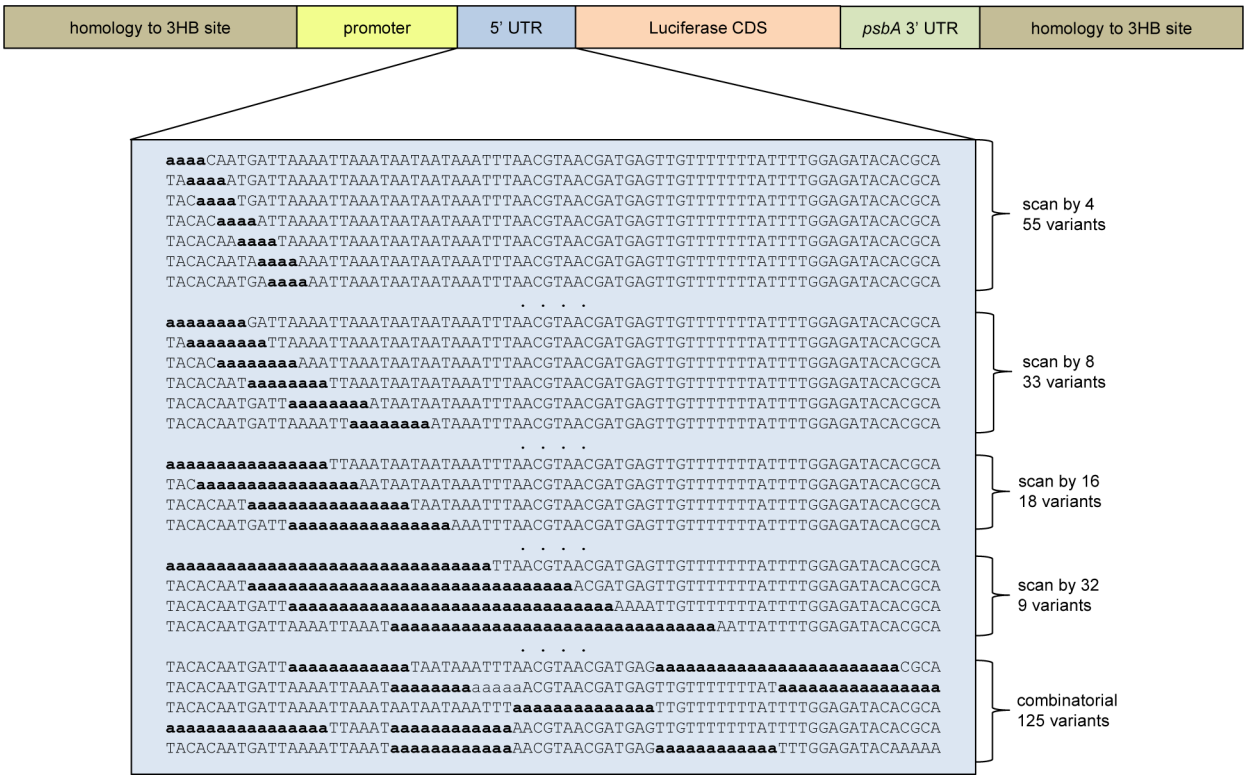


Figure 1. Schematic of the transformation vectors containing variant 5′ UTRs. The 3HB site is a silent site in the chloroplast genome near the *psbH* locus. The luciferase coding sequence has been codon-optimized for higher expression in the algal chloroplast. The 3′ UTR is kept as that from *psbA* regardless of the origin of the 5′ UTR, while the promoter correlates with the 5′ UTR. Representations of the *psbD* poly(A) scanning pools and combinatorial pools are illustrated below. Not all sequences are shown, but the total number of variants of each type is indicated; scans continue across the full length of the UTR, and the combinatorial pool contains all combinations of wild type sequence and poly(A) in 12-mer segments. The full list of sequences is available in the Supporting Information. Substitutions from wild type are shown in bold and lowercase.

Chlamydomonas.[18] However, subsequent chimeric constructs have relied on as little as 245 nucleotides upstream of the start codon to serve the function of both promoter *and* 5′ UTR,[19] suggesting that the promoter is only the seven preceding bases, which seems unlikely. For this study we manipulated only the 100 nucleotides immediately upstream of the start codon due to oligonucleotide synthesis length constraints. The full 245 nucleotides were included to serve as the promoter and UTR, but the first 145 nucleotides were kept as wild-type sequence while only the latter 100 were mutated.

Each UTR was manipulated in two ways. In the first permutation, the native sequence was scanned across and replaced with one of four different length stretches, either 4, 8, 16, or 32 nucleotides, of adenosine homopolymer. Each UTR in this set, 162 variants for *psaA* and 113 for *psbD*, had a single segment of sequence replaced by these poly(A) stretches. In the second set, the UTR was divided into segments of 12 nucleotides, and these segments were either left as native sequence or were replaced by 12 bases of poly(A) in all possible permutations. This set comprises 512 variants for *psaA* and 128 variants for *psbD* and captures combinatorial effects of removing sequence elements from parts of the UTR that are not immediately adjacent. In all of these UTRs, the total length of the UTR and the spacing remains unaltered to eliminate confounding effects of changing the spacing between important elements. This set also allows us to determine the minimal set of elements required for translation. Figure 1 provides representations of the library members, using mutated *psbD* sequences as examples, as well as a schematic for the algal transformation vector.

Poly(A) was chosen as the replacement sequence in both of these manipulations in keeping with the A-T rich nature of *Chlamydomonas* chloroplast UTRs and also to decrease the potential for forming structured elements by GC base pairing. There is some potential that addition of poly(A) stretches could enhance translation as there is evidence of a role for poly(A)-binding proteins in light-induced translational activation of chloroplast mRNAs.[20] In addition, the chloroplast homologue of the bacterial ribosomal protein S1, which binds stretches of poly(U) to assist in association of the ribosome, may in fact bind stretches of poly(A) instead, as demonstrated in higher plant chloroplasts.[21]

The synthetic oligonucleotides were converted into double-stranded fragments by second-strand synthesis and amplified by PCR using primers containing specific uracil residues such that the PCR fragments could be seamlessly cloned into transformation vectors by USER (uracil-specific excision reagent) cloning.[15] Illumina sequencing of bacterial plasmid minipreps from cells harboring the complete pool of algal transformation vectors revealed that all 241 designed *psbD* UTRs were unambiguously present in the DNA transformed into algae. One *psaA* sequence was potentially lost through the cloning steps from synthesized oligonucleotide to full transformation vector, but the other 673 out of 674 designed sequences were unambiguously identified in the final transformation DNA pool.

***psbD* 5′ UTR Element Identification.** Algal transformants containing a single variant from the library were characterized

dx.doi.org/10.1021/sb300069k | *ACS Synth. Biol.* XXXX, XXX, XXX−XXX

as individual clones by assaying plates containing arrays of individual colonies by luminescence following addition of a luciferase substrate (see Figure 2). Each clone was qualitatively

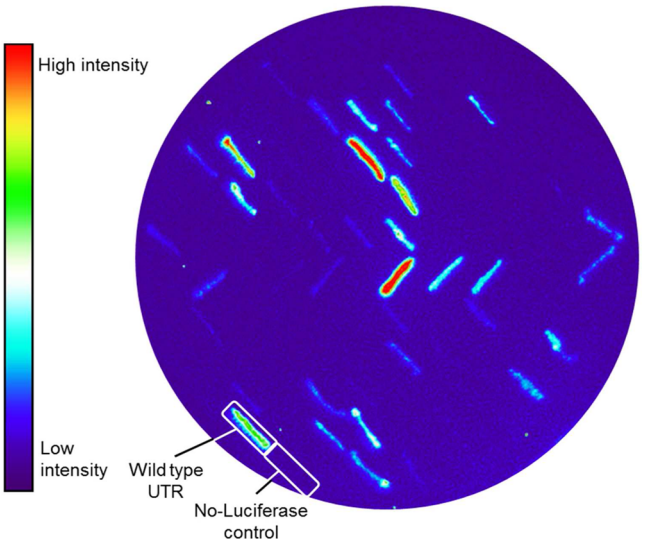


Figure 2. False-color heat-map image of one plate of 100 unique clones expressing luciferase. Indicated in the lower left is the patch of algae expressing luciferase driven by the wild-type *psaA* UTR. Immediately below is a patch of wild type *Chlamydomonas reinhardtii* that has not been transformed with luciferase; as expected, no luminescence is detected in this sample.

classified as high, intermediate, or low expression based on their luciferase signal. Clones with no visible expression were included in the low-expression pool. DNA from high-, low-, or medium-expression pools was sequenced using Illumina 150 bp paired-end reads (for further details see Methods). Reads were parsed by barcode and then mapped to the designed UTRs. Only those with a 100% match to the UTR reference sequences were retained for analysis. After removal of low quality and ambiguous reads, over 1.3 million reads from the *psbD* pools were analyzed.

We tallied whether each reference sequence had an A or the wild type nucleotide at each position along the UTR and then multiplied by the number of sequence reads that mapped to each reference. After summing across all sequence reads, we generated a likelihood map for the probability that an A (mutation) or the wild type nucleotide was present at any position of the UTR, for each expression-level pool. For instance, there was a 96% chance that a clone in the high-expression pool had a T at the −43 position relative to the AUG start codon in *psbD*, whereas a clone in the low-expression pool had only a 79% chance of having a T at that position. All clones that did not have a T at this position have an A instead, as this was the only substitution made (and all mutations to another nucleotide were filtered out in the read mapping). Figure 3 graphically illustrates the results for the *psbD* 5′ UTR analysis.

We mapped the potential elements identified using this analysis onto a secondary structure map of the *psbD* 5′ UTR as predicted by Quikfold[22] and confirmed by RNaseH mapping.[13] Previously identified RNA stability elements[11] are shown in orange, while regions previously implicated in translational activation[11] are shown in green. In Figure 4b, we highlight in blue the regions where our data indicate the high-expression pool has greater than 90% conservation of the wild type sequence; in red are regions where this conservation is less than 90% in the high expression lines. These cut-offs were determined empirically by comparison with the previously identified elements, as χ-square tests using 3 × 2 contingency tables to determine the statistical significance of the differences between each pool at each position indicate that all the points are significant ($p \ll 10^{-6}$) due to the large read counts. Adenosine nucleotides that lie on the boundaries of these regions cannot be assigned either way, as all sequences have adenosine at these positions regardless of whether that region was mutated or left as wild type sequence.

Using this method, we confirmed two of the three previous characterized elements that were defined by biochemical and

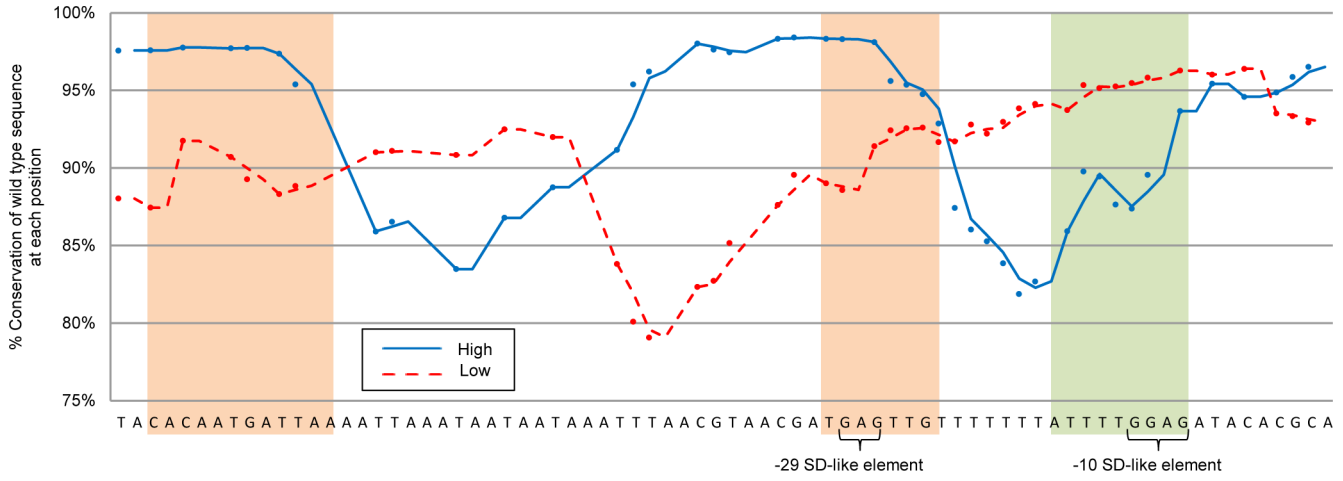


Figure 3. Conservation of wild-type sequence at each position along the *psbD* 5′ UTR. The ATG start codon is immediately downstream of the sequence shown. Blue represents aggregate data from the high-expression pool and red represents the low-expression pool. Trend lines are two-point moving averages of positions where the wild type nucleotide is not already an adenosine (i.e., only positions where mutation caused an actual nucleotide change are considered here). Sequences highlighted in orange have been previously identified as important for RNA stability. The sequence in green was previously implicated in translation, but later it was determined that it is merely the spacing it provides rather than the sequence itself that is important.

dx.doi.org/10.1021/sb300069k | *ACS Synth. Biol.* XXXX, XXX, XXX−XXX

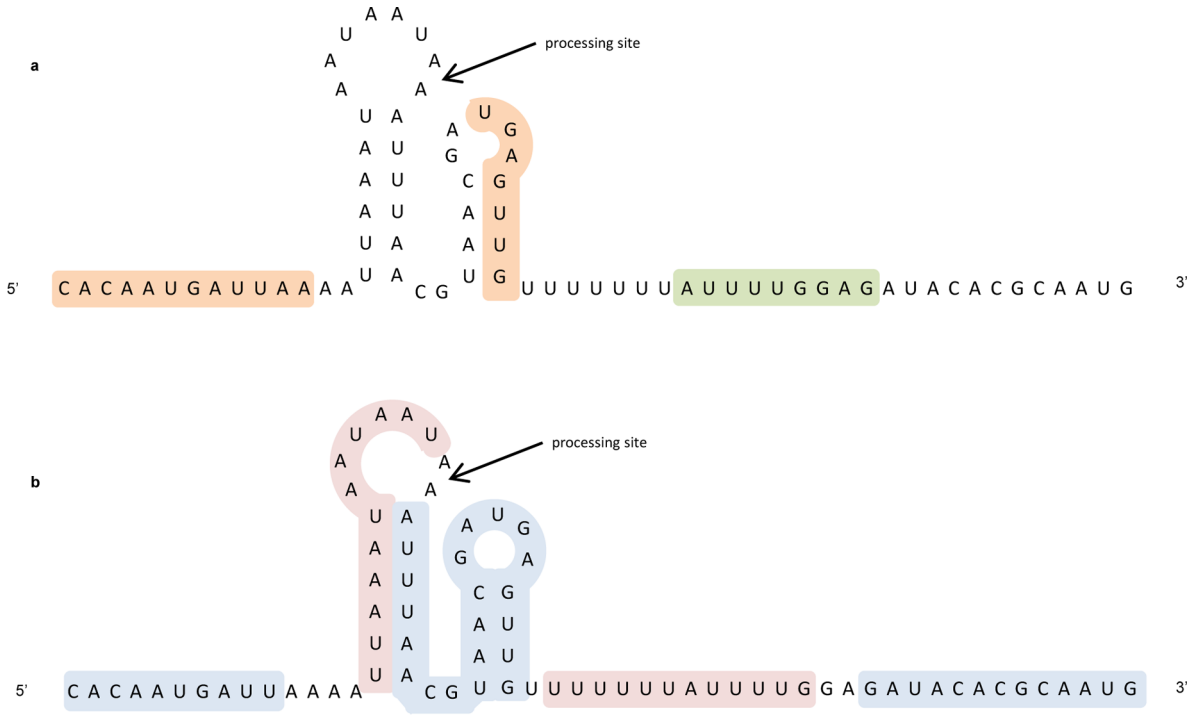


Figure 4. Depiction of the *psbD* 5′ UTR secondary structure. (a) The previously identified regulatory elements for message stability (orange) or translation (green) are highlighted. (b) Regions identified by our method as conserved in the high expression pool are highlighted in blue; regions that favor or tolerate mutation in the high expressers are highlighted in red. The previously described stem-loop containing the AUG is not depicted here because the stem requires additional nucleotides in the *psbD* coding sequence that are not present in the *lux* coding sequence, so this structure was not relevant to our studies. Furthermore, this secondary structure was shown to not exert much effect in a background in which the spacing between all other 5′ UTR elements remains unaltered.[13] Though these secondary structures have been verified experimentally *in vitro*,[13] it is possible that they do not exist *in vivo*, particularly upon protein binding. At the least, the stem-loop nearest the 5′ end cannot exist in the processed form of the transcript.

site-directed mutagenic studies of the *psbD* UTR. Our data significantly enlarge one of these elements and clarify the boundaries of these previously identified elements. These two regions, thought to be important in mRNA stability,[11] are both identified in our study as among regions highly conserved in our high-expression pool (see orange boxes in Figure 3).[11] The first element, at the 5′ end of the unprocessed mRNA, may form the binding site for a previously identified RNA-binding protein that interacts only with the unprocessed *psbD* 5′UTR and is critical for message stability,[23] and our data maps quite closely. The second stability element is implicated in processing and stable accumulation of the processed form of the transcript,[11] and our analysis shows that this element is likely twice as large as previously predicted. Our results indicate that this element extends all the way up to the processing site, which agrees with previous evidence that RNA-binding proteins act as protective caps at the ends of chloroplast transcripts to protect them from exonuclease degradation.[24] This processing site, therefore, may not be a single site of action by specific processing machinery; it may simply be the site where the cap protein no longer protects the 5′ end of the transcript from degradation.

The region indicated in green in Figure 3 had been identified, along with the adjacent uracil tract, to be important for translation.[11] Our results do not find this sequence to be important for expression, and this is supported by more recent studies that revealed that in fact neither the sequence nor the secondary structure of this U-rich region was essential for *psbD* translation; this region predominantly functions as a spacer for elements located on either side of it.[25] Since our mutational constructs do not alter spacing between unmutated regions, we do not observe the same importance of this region, in agreement with the work by Ossenbühl et al.[25] Again, this demonstrates the strength of our unbiased method for assessing the contribution of individual sequence elements without unintended secondary effects.

Some *Chlamydomonas* chloroplast genes contain a sequence resembling a Shine-Dalgarno consensus sequence at around −10 from the start of translation, but these elements are often dispensable. For example, in *atpB*, *atpE*, *rps4*, and *rps7*, eliminating the −10 SD-like sequence or replacing it with a canonical SD appears not to affect expression of any of these mRNAs.[26] In Figure 3 we see that indeed for *psbD* this −10 SD sequence is not important for expression, but that a SD-like GAG further upstream at position −29 is indeed strongly conserved in the high expression lines, indicating that this may be the true ribosome binding site as previously suggested.[11]

Our scanning mutagenesis analysis can also elucidate potential elements that function as negative regulatory elements. In order to determine whether a minimally conserved region is merely unimportant for high expression or whether mutations in that region are indeed favored in the high expression lines, we compared the high expression pool to the sum of all sequence reads obtained in all pools (high, medium, and low expression). Areas where the wild-type sequence conservation is *lower* for the high expression lines than for the sum of all reads indicate that mutations in these sequence

dx.doi.org/10.1021/sb300069k | *ACS Synth. Biol.* XXXX, XXX, XXX−XXX

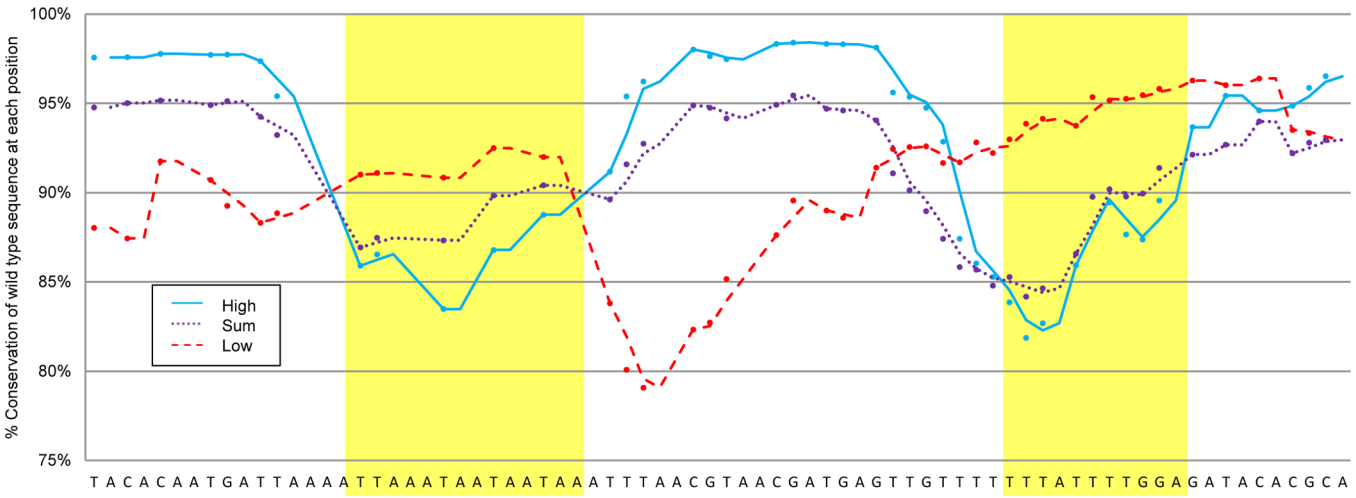


Figure 5. Overlaid plots of the high-expression pool (blue), low-expression pool (red), and the sum of all *psbD* UTR reads (purple). Highlighted in yellow are the only two regions where the following criteria are met: the high expressers show lower conservation than both the low-expresser pool and the sum of all reads, and the low expressers show higher conservation than both the high-expresser pool and the sum of all reads. These regions may be elements involved in negative regulation.

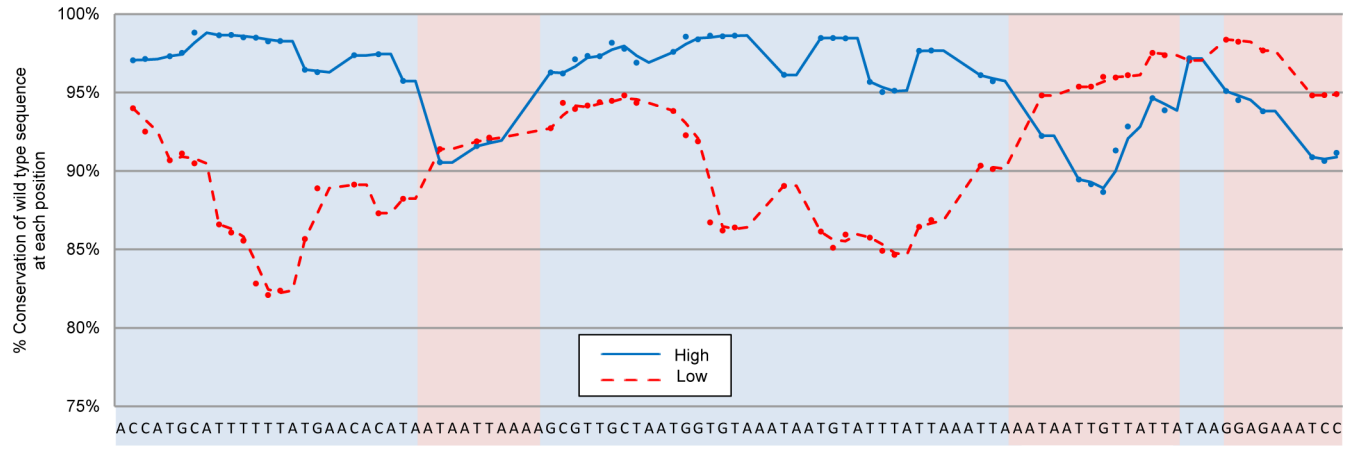


Figure 6. Conservation of wild type sequence at each position along the last 100 nucleotides of the *psaA* 5′ UTR. Blue represents the high-expression pool, and red represents the low-expression pool. Regions that exhibit greater than 95% wild-type sequence conservation in the aggregate data of the high expressers are designated with a blue background; regions with less than 95% conservation have a red background, indicating that they are either unimportant for high expression or indeed may be sites of negative regulation.

elements are in fact favored for high expression, suggesting that they may be regions where negative regulatory factors interact.

The plots of the mutation frequency for the high and low expression lines and the sum of all reads are shown in Figure 5, with the percent conservation of the wild type sequence at each position graphed along the entire *psbD* UTR. There are two regions where the high expression lines contain mutations at a rate much higher than observed in all the algal reads from the high, medium, and low pools combined. Lending additional evidence for the negative regulatory impact of these elements, the low-expression pool actually shows increased conservation of these regions above that observed in the sum of all reads.

This indicates that while some mutants certainly lose their ability to express the reporter due to mutation in an essential element, there is overall a significant bias toward retaining these two putative negative regulatory elements in the low-expression pool. Interestingly, one of these regions is upstream of the processing site, suggesting that the negative control it may exert must occur prior to processing. Note that the plot for the sum of all reads does not always fall exactly between the high- and low-expresser pools, as the majority of our clones were classified as intermediate expressers. In certain regions, these intermediate expressers conserved wild-type sequence below the levels observed in either the high or the low expressers. Further work will be needed to determine whether these regions may have dual functionality or are involved in more complex regulatory mechanisms.

Identification of Regulatory Elements within the *psaA* 5′ UTR. Phenotyping, sequencing, mapping, and data analysis for the *psaA* clones were performed exactly as described above for the *psbD* clones. We have no prior knowledge of important sequence elements in the *psaA* 5′ UTR, so all findings from the sequence analysis are novel. These results are based on greater than 12.8 million reads from *psaA* clones that were retained after all the quality and mapping constraints were applied.

As with the *psbD* results, overall we observe that increased sequence conservation in the high expression pool directly corresponds with poor conservation in the low expression pool, as we would expect. In other words, elements that are critical for expression are retained in the high-expression lines and have

dx.doi.org/10.1021/sb300069k | *ACS Synth. Biol.* XXXX, XXX, XXX–XXX

been lost in the low-expression lines. In Figure 6, we designate with a blue background the regions that are highly conserved in the high expression lines and with a red background the regions that tolerate or favor alteration in the high expression lines. In this analysis, our empirical cutoff for defining these regions was raised to 95% conservation of the wild type sequence in the high expresser pool, since nearly all of the positions were conserved at greater than 90%. In Figure 7 we map these regions onto a predicted secondary structure of the *psaA* 5′ UTR.

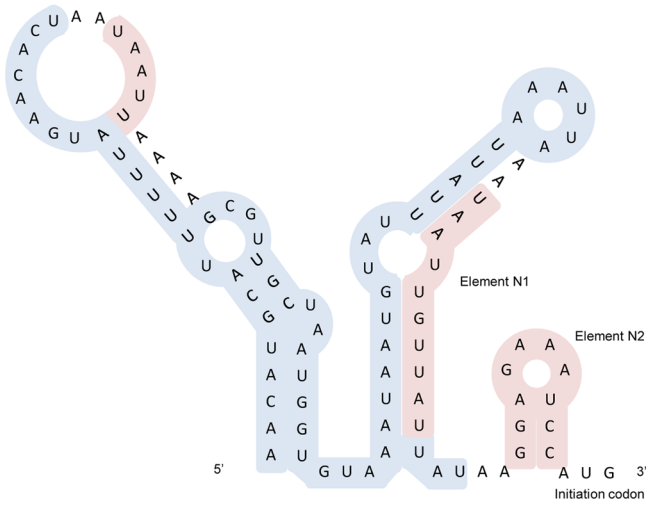


Figure 7. Depiction of the secondary structure of the mutagenized region of the *psaA* 5′ UTR. This structure is predicted by QuikFold version 2.3 at 25 °C but has not been confirmed experimentally. Regions identified as conserved in the high expression pool are highlighted in blue; regions that favor or tolerate mutation in the high expressers are highlighted in red.

In our analysis of the *psaA* 5′ UTR, we find a lack of conservation at the RBS-like GGAG immediately upstream of the start codon, similar to what we observed for the −10 GGAG in the *psbD* UTR. In fact, we observe a strong propensity to eliminate this sequence in the high-expressing clones. Again similar to the *psbD* UTR, in the *psaA* UTR we observe very strong conservation of a similar RBS-like sequence further upstream in the high expression lines. The GGUG at position −51 is among the most highly conserved regions across the entire analyzed region and forms a perfect complementary match with four nucleotides in the anti-Shine-Dalgarno sequence at the 3′ end of the 16s rRNA.[27] Though this site is significantly further upstream from the start codon than in the *psbD* mRNA, it has been noted that the spacing between ribosome binding sites and initiation codons is much more variable in chloroplasts than in bacteria,[28] and there are other *Chlamydomonas* chloroplast genes with similar spacing such as the −51 SD-like element in the *rps12* transcript.[29]

Aside from *cis*-acting secondary structure, many UTR elements act as recognition sites for nuclear-encoded *trans*-acting factors that promote RNA stability or translation initiation. There is evidence that the *psaA* 5′ UTR participates in a feedback loop, as exogenous genes under the control of the *psaA* 5′ UTR accumulate to higher levels in a strain with a *psaA* splicing defect than when the *psaA* gene product is being synthesized normally.[30] The sequence elements that participate in this feedback may have already been identified in this analysis, and we can test candidate mutated UTRs in this splicing defect background to determine which UTR elements are responsible.

Regions where negative regulators may interact, indicated by particularly low conservation in the high-expression pool and high conservation in the low-expression lines, are highlighted in Figure 8. The involvement of these two elements in negative regulation is strongly supported by preliminary analysis of several mutant-*psaA* 5′ UTR clones that express the reporter luciferase at levels higher than the wild-type *psaA* UTR. Twenty-four of the brightest individuals from the initial luciferase assays were grown in liquid culture and then assayed for *lux* expression using equal numbers of cells. Of these, 15 were found to express *lux* at higher levels than the wild-type *psaA* UTR; these individuals were sequenced with traditional Sanger sequencing. The results shown in Table 1 indicate that the putative negative regulatory elements identified by the

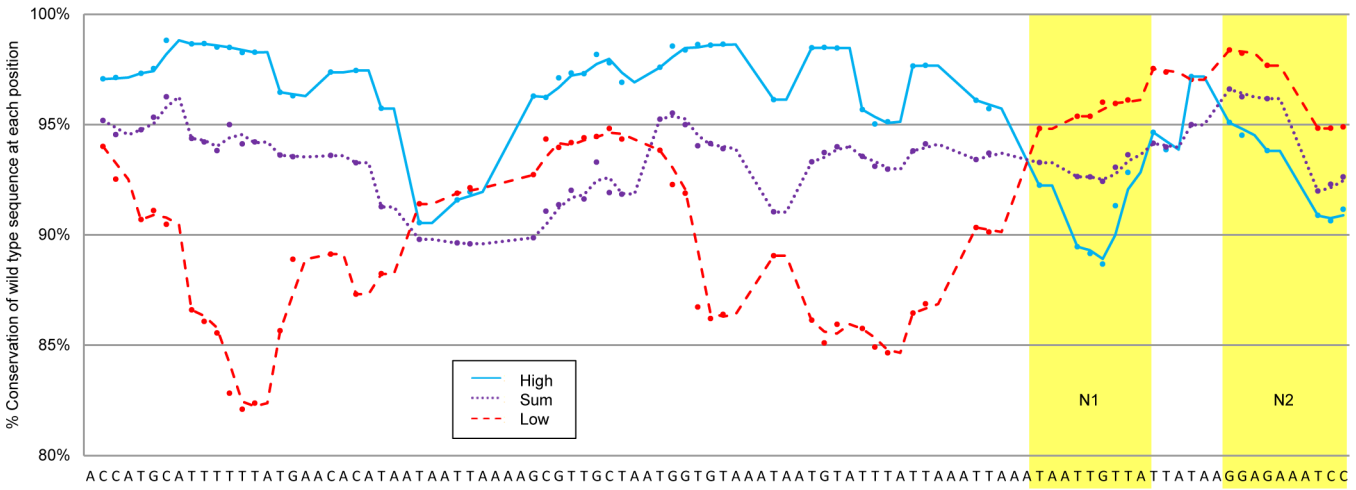


Figure 8. Overlaid plots of the high-expresser pool (blue), low-expresser pool (red), and the sum of all *psaA* UTR reads (purple). As in Figure 5 above, highlighted in yellow are the only two regions where high expressers show lower conservation than the sum of all reads *and* low expressers show higher conservation than the sum of all reads. These elements are candidates for recognition sites for negative regulatory factors, and preliminary analysis of brighter-than-wild-type *psaA* mutant UTR individuals supports this notion.

dx.doi.org/10.1021/sb300069k | *ACS Synth. Biol.* XXXX, XXX, XXX−XXX

Table 1. Sequence Analysis of 15 Individual Clones Expressing Luciferase at Levels Higher than the *psaA* UTR

clone number	% of *psaA* UTR-driven luciferase expression	mutated in region N1	mutated in region N2
A1-2	115	no	yes
A1-8	126	yes	no
A1-15	106	no	yes
A1-25	159	no	yes
A1-26	238	yes	yes
A1-98	112	yes	yes
A2-20	104	no	yes
A2-56	122	yes	no
A2-58	137	no	yes
A2-81	107	yes	yes
A2-85	115	yes	yes
A4-37	116	yes	yes
A4-51	126	no	yes
A4-60	120	yes	yes
A5-32[a]	295	no	yes

[a]Clone A5-32 has a two-nucleotide deletion at the 3′ end in addition to the poly(A) substitution disrupting region N2.

aggregate sequence analysis are indeed highly mutated in all 15 of these individuals. This suggests that it is necessary to release inhibitory regulation conferred by these regions in order to achieve expression levels beyond those observed with the wild-type *psaA* UTR.

Ongoing work to more thoroughly characterize the highest-expressing candidates from the libraries will lend insight into the mechanisms underlying the varying expression levels we observe.

Constructing a Synthetic UTR. The initial aim of this work was to elucidate positive and negative regulatory elements in chloroplast 5′ UTRs using an unbiased scanning approach to better characterize the regulatory mechanisms of gene expression in algal chloroplasts. Identifying these key regulators of chloroplast gene expression allows us to develop tools for increased recombinant protein accumulation in algal plastids. In particular, this information also allows us to develop modified or wholly synthetic UTRs designed to overcome the limitations currently encountered using algae as a protein production platform for complex therapeutic molecules or industrial enzymes, or in metabolic engineering to produce high-value small molecules.

Designing a heterologous but highly effective UTR may avoid the negative feedback issues currently encountered when using a native UTR from a photosystem gene to drive recombinant protein production. Furthermore, homologous recombination in the chloroplast is so efficient that we have observed crossover events between endogenous UTRs and the introduced corresponding regulatory elements driving exogenous transgenes, resulting in loss of transgene expression over time (unpublished data). With multiple unique synthetic UTRs at our disposal, we can introduce multiple exogenous genes at once without recombination between repeated native UTRs. This will be especially important for introducing multiple genes, for example, novel metabolic pathways for producing biofuels or high-value secondary metabolites.

Taking the aggregate sequencing results from the *psaA* UTR variant library, we designed a simple synthetic UTR to drive expression of a useful industrial enzyme. The putative negative regulatory regions N1 and N2 indicated in Figure 8 were substituted with poly(A) stretches, and the remainder of the

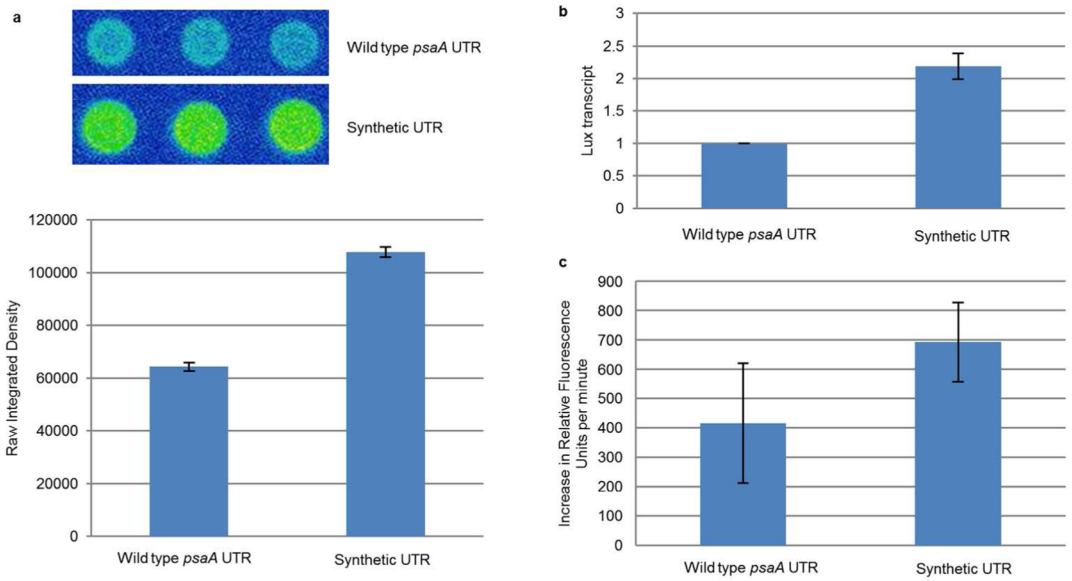


Figure 9. A designed synthetic UTR significantly exceeds activity from the wild-type UTR. (a) Results of a quantitative luciferase spotting assay are shown for the synthetic UTR compared to the wild-type *psaA* UTR, using the same intensity scale depicted in Figure 2. Below, this difference is quantitated by integrated density over the spots; the synthetic UTR produces 67% higher expression. (b) Fold change in transcript level relative to the *lux* transcript in the wild-type *psaA* UTR strain, as determined by quantitative RT-PCR. The error bar indicates standard deviation among four independent synthetic UTR clones. The synthetic UTR produces approximately 2.1 times as much transcript as the endogenous UTR. (c) Xylanase activity of four independent homoplasmic lines of each construct (synthetic UTR and wild-type *psaA* UTR) was analyzed in triplicate using a fluorogenic substrate. Activity is expressed as the increase in relative fluorescent units (RFU) per minute in the linear range between 2 and 6 min after substrate addition. Error bars indicate one standard deviation from the mean. In very close agreement with the luciferase assays, when driving xylanase the synthetic UTR on average results in 66% greater expression than the wild-type *psaA* UTR.

dx.doi.org/10.1021/sb300069k | *ACS Synth. Biol.* XXXX, XXX, XXX–XXX

UTR was left as wild-type sequence. This synthetic UTR and the wild-type *psaA* UTR as a control were cloned into identical chloroplast transformation vectors driving luciferase expression. Quantitative luciferase spotting assays were used to determine the difference in expression, and quantitative RT-PCR was used to examine transcript level differences between the two projects. Indeed, the transcript level for luciferase mRNA is approximately 2-fold higher with the synthetic UTR compared to the endogenous *psaA* UTR, which is close to the observed increase in expression as determined by luciferase assay. This suggests that elements N1 and N2 may negatively affect mRNA stability, and that their replacement can enable the transcript to persist and therefore result in increased translation.

The synthetic UTR and endogenous control were also cloned into a chloroplast transformation vector driving a chloroplast codon-optimized xylanase II (*xyn2*) gene from *Trichoderma reesei*[31] but otherwise identical to the luciferase vector previously described. Xylanase is an industrially important enzyme for the pulp and paper industries and will become increasingly valuable for converting agricultural wastes and hemicellulosic feedstocks into animal feed and biofuels.[32] Based on xylanase activity in cell lysates, our designed synthetic UTR produces nearly 70% more enzyme than the wild-type *psaA* UTR. This correlates well with the effect observed in the luciferase strains, demonstrating that synthetic UTRs can be effective for increasing transgene expression from a number of genes of interest.

We have shown the utility of using a synthetic oligonucleotide library to generate a large variant library coupled with phenotype screening and next-generation sequencing to rapidly identify critical regulatory elements in plastid 5′ UTRs. There are many cases where this synthetic high-throughput technique could allow rapid identification of RNA elements required for gene expression. RNA binding assays have identified many binding partners for specific mRNAs or leader sequences, but in many instances the follow-up work to determine exactly where and how these factors interact with the UTR has not been done. Our approach can significantly accelerate this process and allow us to gather enough information from distinct UTRs to begin to view chloroplast gene regulation from a systems level, rather than on a gene-by-gene basis. There is already evidence that some RNA binding proteins (RBPs) such as RBP46 may interact with multiple mRNAs,[33] but the current information is not comprehensive enough to allow us to formulate consensus binding sites or estimate tolerance of variations on these elements.

New advances in computational prediction of RBP recognition sites may provide avenues for synergistic discovery of RBP−RNA interactions, further accelerating our understanding of gene regulation in the chloroplast. Future work to map *trans*-acting RPBs to each *cis*-element can contribute to predictability of RBP sites for engineering rationally designed RBPs. Recently, the maize chloroplast RNA-binding protein PPR10 has been engineered to bind novel RNA sequences,[34] akin to the combinatorial engineering of TAL Effectors that have garnered much attention as tools for genetic engineering. A better understanding of RNA-binding proteins and their recognition sequences may soon facilitate engineering both RNAs and their binding sites to enable highly optimized gene regulation.

This method is far more comprehensive, high-throughput, and unbiased than previous methods involving discrete mutated versions of UTRs. We can examine the entire length of UTRs from a single synthesized library and identify regions of these UTRs that may not be recognizable *a priori* as important for chloroplast gene expression. Furthermore, we demonstrate the practical utility of this method by introducing a novel synthetic UTR based on the aggregate sequencing data from the synthetic library of the previously uncharacterized *psaA* UTR. We show that this designed UTR exhibits increased protein accumulation compared to the wild-type *psaA* UTR from which it was derived, and that this enhanced expression is maintained when the synthetic UTR is used to drive expression of an industrial enzyme, demonstrating that the elements identified by our method will function with a variety of exogenous recombinant genes.

METHODS

Design and Synthesis of UTR Libraries. We designed 915 sequences as synthetic variants of the 5′ UTRs of two *Chlamydomonas reinhardtii* genes, *psbD* and *psaA*. The *psbD* and *psaA* genes were selected because of their high expression levels and relatively short 5′ UTRs. Library sequences were designed with two distinct goals, manifested in two sets of mutant UTRs for each gene. In one set, we scanned across the entire UTR, altering a single sliding window of nucleotides changing them from the wild type sequence to poly adenosine. The length of this window was either 4, 8, 16, or 32 nucleotides, and the window moved along the UTR by 1, 2, 4, or 8 nucleotides at a time, respectively to the length of the window. In the second set, we divide the UTRs into windows of 12 nucleotides each and then create all combinations of these windows with the background sequence being poly adenosine. This set was created to capture potential interactions between spatially disparate regions of the UTR and to determine the minimal set of elements required for translation in each of these UTRs. These strategies are represented in Figure 1, along with examples of each set of variant UTRs.

UTR variants were synthesized as single-stranded DNA oligonucleotides using the Agilent Oligo Library Synthesis (OLS) platform[35,36] and consolidated into a single pool. The 915 unique sequences for *psaA* and *psbD* were synthesized with 28-fold redundancy. In addition to the UTR sequence, the oligonucleotides include a portion of the *lux* gene downstream, as well as regions unique for each gene corresponding to the promoter region immediately upstream of the UTR, enabling PCR amplification of gene-specific subpools from the total OLS pool.

Cloning and Vector Creation. Using gene-specific primer sets, two subpools were amplified, one for the *psaA* UTRs and one for the *psbD* UTRs, using the original oligonucleotide library as template. The reverse primer was common to all oligos, annealing to the 3′ end of the oligos corresponding to the first 20 nucleotides of the *lux* gene. The forward primer defined the subpool, as it annealed to the endogenous promoter upstream of the UTR. All oligos had to be synthesized with equal length, so all were precisely 140 nucleotides, which included a minimum of 20 nucleotides on either side of the UTR for the amplification primers to anneal. These PCR products were then cloned using USER fusion technology[15] into a truncated version of the transformation vector, consisting of just the UTRs and *Chlamydomonas* codon-optimized luciferase reporter coding sequence[16] inserted in a pGEM backbone. Two fragments, one consisting of the UTR and one for the backbone, were amplified by PCR using the uracil-containing primers provided in Supplementary Table S3,

mixed in equimolar quantities along with the USER enzyme (New England BioLabs, Ipswich, MA, USA), and then incubated for 30 min at 37 °C followed by 20 min at 25 °C prior to transformation into DH5α chemically competent *E. coli*.

This step was necessary because the full chloroplast transformation vectors proved too large to amplify by PCR for the USER fusion technique.[15] Classical restriction ligation cloning techniques were not considered for this project because they would have left a restriction site scar between the promoter and UTR, which could have affected sequence or spatial elements important for translation.

After USER fusion cloning into the pGEM-based vectors, the entire luciferase expression cassette (Figure 1) was cloned into a chloroplast transformation vector[17] by restriction ligation using a *BamH*I site upstream of the promoter and a *Spe*I site downstream of the 3′ UTR. As with the first cloning step, this procedure was performed on gene-specific pools, not on individual clones.

To determine if there was bias in the original oligonucleotide library or if bias was introduced into the library at any of these cloning steps, we analyzed a pool of plasmids at each step with next-generation sequencing. In each cloning step we generated greater than 1,000-fold more colonies than unique library sequences, as determined by plating a small aliquot of the transformation before pooling the rest into liquid LB culture. Analysis of loss of library diversity is addressed in the results.

Algal Transformation, Strain Isolation, and Expression Analysis. Chloroplast transformation vectors were purified using the GeneJet bacterial miniprep kit (Fermentas Inc., Glen Burnie, MD, USA) from the pooled liquid bacterial culture described above, containing all variants of the library for either *psaA* (675 variants) or *psbD* (240 variants). As previously described, particle bombardment with gold particles (S550d, Seashell Technologies, San Diego, CA, USA) was used to transform the algal chloroplast of a *psbH*- strain[37] using a PDS-1000/HE biolistic transformation chamber (Bio-Rad, Hercules, CA, USA). Transformants were recovered on minimal media to select for restoration of the *psbH* locus following homologous recombination with the transformation vector.[17]

For the *psaA* library, 6979 individual algal clones were analyzed, giving 97.9% certainty that all 674 designed sequences were present in the algal library assuming unbiased selection with replacement. For the *psbD* library, 2704 clones were analyzed, giving 99.7% certainty that all 241 designed sequences were present in the library. Individual algal clones were replated three times on minimal media to drive the cells toward homoplasmy, reducing the effect of variable heteroplasmy on expression phenotype. Before assaying expression levels of the luciferase reporter, stronger expression was induced with a light shift of the plated cells. Plates were kept in the dark for 36 h and then shifted to bright light for 8 h prior to assaying. After adding 150 μL of a mixture of 3% decanal (Sigma, St. Louis, MO, USA) in heavy mineral oil (Sigma) to the plate lid and spreading it evenly with a sterile plastic inoculating loop, the plate was covered with the lid and the clones were incubated in the dark with the substrate for 5 min. Colonies were then assayed for luciferase expression using a 5-min exposure on an EG&G Berthold NightOwl Imager (Berthold Technologies, Bad Wildbad, Germany). Clones were pooled into three categories by visual estimation: high expression, low expression, and intermediate expression. Figure 2 shows a false-color example of the output of this assay for one sample plate. Expression was not normalized to cell count because the patches were assayed directly, but no notable differences in growth were observed among the clones, allowing the intensities to be directly compared between patches.

For a more quantitative phenotypic analysis of 24 of the brightest *psaA* library members and of the synthetic UTR, clones were grown in 6 mL of minimal media for 3 days, cell density was determined by hemocytometer, and the cultures were spun down and resuspended in HSM. Five microliters of each culture, containing a total of 3×10^6 cells, was spotted in triplicate onto minimal media plates alongside a strain with the wild-type *psaA* UTR driving luciferase prepared in the same manner. These equal-cell-count spots were allowed to grow for 2 days and then were subjected to a light-shift and assayed exactly as described above. Luciferase expression was quantitated using integrated density measurements on ImageJ software (available at http://rsbweb.nih.gov/ij/) across equal pixel areas; the average of the three spots for each clone was divided by the average for the three *psaA* UTR-driven spots on the same plate to obtain percentage of *psaA* UTR-driven expression. Those expressing luciferase at a level higher than the wild-type *psaA* control were sequenced with traditional Sanger sequencing by PCR-amplifying the UTR of interest using a forward primer in the promoter and a reverse primer in the *lux* gene (see Supplementary Table S3 for primer sequences). The PCR product was treated with Exonuclease I (New England BioLabs) and Shrimp Alkaline Phosphatase (SAP, Fermentas Inc.) (10 μL PCR product, 0.2 μL Exonuclease I, 0.2 μL SAP; 37 °C for 30 min followed by 85 °C for 15 min) and sent directly for sequencing using the reverse primer.

Next-Generation Sequencing and Data Analysis. Up to 250 colonies from a single expression pool were inoculated into each of 60 flasks of tris acetate phosphate (TAP) media such that each flask contained as many as 250 unique clones of roughly equivalent expression. These cultures were allowed to grow for 48 h, and then genomic DNA was extracted with a protocol modified from that of Newman et al.[38] Ten milliliters of log-phase culture were concentrated by centrifugation at 3000 rpm and resuspended in 0.5 mL of TEN buffer (10 mM Tris-HCl, 10 mM EDTA, 150 mM NaCl). This solution was transferred to a 1.5 mL tube and spun at 10,000 rpm for 10 s, and the cell pellet was resuspended in 150 μL of H_2O and 300 μL of SDS-EB (2% SDS, 400 mM NaCl, 40 mM EDTA, 100 mM Tris-HCl, pH 8.0) and vortexed. Then 350 μL of 1:1 phenol/chloroform was added, and the mixture was vortexed again. Phases were separated by a 5 min 14,000 rpm spin, and the aqueous phase was transferred to a new tube, where was it again extracted by adding 300 μL of pure chloroform, vortexing, and spinning as above to separate phases. The aqueous phase was again transferred to a new tube, added to 2 vol of ice-cold 100% ethanol, and kept on ice for 30 min. Then the tubes were spun for 10 min at 14,000 rpm, and the pellet was washed with 200 μL of 70% ethanol before drying and resuspending in 100 μL of EB.

From these DNA preps, we amplified subpools for sequencing by PCR using Kapa Biosystems Library Amplification Kit (Kapa Biosystems Inc., Woburn, MA, USA), which reduces PCR-based bias especially with A/T-rich templates.[39,40] Primers for this step had internal barcodes where the forward primer barcode designates the gene origin of the UTR, and the reverse barcode designates the pool (low, medium, high expression) from which the sample originated (see barcoding

primers in Supplementary Table S3). Standard Illumina adapters were ligated onto the PCR fragments according to Illumina product literature on Paired-End Sample Preparation, and all samples were sequenced on a single lane of an Illumina GAIIx (Illumina Inc., San Diego, CA, USA) at The Scripps Research Institute DNA Array Core using paired-end 150 bp reads. Because the nucleotide chain growth is blocked after the addition of each nucleotide, the Illumina platform is particularly well-suited for sequences with long homopolymer regions, compared to other next-generation sequencing technologies. Though some slippage in sequencing may still have occurred, we mitigated this risk with the paired-end reads: these allowed us to read a single sequence from both ends and determine the consensus sequence by aligning both reads.

Reads were analyzed in CLC Genomics Workbench (CLC bio, Aarhus, Denmark). First, subpools were isolated on the basis of the pairs of forward and reverse internal barcodes, and then reads were filtered and trimmed by quality. Reads were removed if any one of the following criteria were met: if P_{error} exceeded 0.05, if there were more than two ambiguous nucleotides, or if read length was less than 64 nucleotides. In cases where one read of a pair did not meet these stringency cutoffs, the remaining read was mapped to the reference sequences as a single read.

Trimmed reads were mapped to a set of reference sequences of the designed oligonucleotides synthesized on the Agilent OLS platform. Mapping tolerated zero mismatch along the full length of the UTR, and nonspecific mappings were ignored, such that all mapped reads unambiguously map to a single unique reference sequence. All subsequent positional mutation likelihood analysis was done in Microsoft Excel after importing read counts from each pool for each reference UTR sequence. Each reference was parsed into individual residues, and then references where the wild-type nucleotide was replaced by an adenosine were assigned a value of 1 at that position and a value of 0 at all positions identical to wild type. The resulting matrices of zeros and ones for each reference UTR were multiplied by the number of times a read mapped to that reference from each expression pool, producing a matrix of aggregate data on how frequently a given position was mutated to adenosine for the high, low, or intermediate expression lines. The data presented in Figures 3, 5, 6, and 8 are essentially inverse of these aggregate matrices: the graphs depict the likelihood that the wild-type sequence was conserved at a given position, rather than the likelihood that the position was mutated to adenosine. To determine statistical significance of these positional likelihoods among each of the three expression level pools, a 3 × 2 contingency table was constructed for each position, delineating the aggregate wild type and "A" read counts for each pool. Pearson's χ-square tests indicate that even at positions exhibiting the most similar % conservation of wild type sequence for all expression pools (i.e., the points where the plots most nearly overlap in Figures 3 and 6), the p-value approaches zero ($p \ll 10^{-6}$), indicating that the differences observed between the pools are highly significant at every position along both of these UTRs.

Synthesizing and Cloning the Synthetic UTR. The synthetic UTR was made by ordering one oligonucleotide primer (see primer LS191 in Supplementary Table S3) containing all the desired nucleotide substitutions (all nucleotides in the negative regulatory regions N1 and N2 substituted with adenosines) to PCR-amplify using the wild-type *psaA* UTR as template. This fragment was cloned into the luciferase expression vector previously described and simultaneously into an identical chloroplast transformation vector except with a chloroplast codon-optimized β-xylanase gene from *Trichoderma reesei*. Another fragment containing the wild-type *psaA* UTR was cloned into these same vectors, for comparison. The vectors were amplified in three fragments with uracil-containing primers for USER fusion cloning with the synthetic UTR fragment (see Supplementary Table S3 for primer sequences). All fragments were gel-purified and mixed in equimolar amounts with 1 μL of the USER enzyme in a 10 μL reaction and incubated as described above. These plasmids were transformed into the *psbH-* strain of *C. reinhardtii* and restreaked to homoplasmy on minimal media. Four homoplasmic lines of each, as verified by a homoplasmic PCR screen as described[41] (see primers in Supplementary Table S3), were assayed for luciferase expression using the same equal-cell spotting assay described above and for xylanase activity.

Xylanase Assays. The four homoplasmic lines of each construct were inoculated into 50 mL flasks of TAP media and allowed to grow to mid-log phase. Cells were harvested at 3,000 rpm, resuspended in 1 mL of lysis buffer (TBS plus 0.5% Tween-20), and sonicated for two intervals of 10 s each at 20% amplitude (S450D digital sonifier, Branson) on ice. Cellular debris was separated from the soluble fraction by centrifugation at 14,000 rpm at 4 °C for 20 min. Next 400 μL of the soluble lysate was transferred to a new tube, and a Bio-Rad DC Protein Assay was performed as described by the manufacturer (Bio-Rad). This quantitation of total soluble protein was used to normalize the volume of lysate used in the subsequent xylanase assays, such that each sample contained equal amounts of total protein.

The xylanase assays were performed according to the manufacturer using the EnzChek Ultra Xylanase assay kit (Invitrogen) in 96-well black flat-bottom plates and read on an Infinite 200 Pro platereader (Tecan, Männedorf, Switzerland) at 355 nm excitation and 455 nm emission. The temperature was held at 42 °C, and readings were taken every 2 min until the increase in fluorescence was no longer linear. In the linear range, the strength of xylanase expression was calculated as the increase in relative fluorescence units (RFU) per minute. Each homoplasmic line was analyzed in triplicate; the average of all four homoplasmic lines is represented in Figure 9 for each construct.

Quantitative PCR. From cultures grown to mid-log phase, RNA was extracted using the standard protocol for the Plant RNA Reagent (Invitrogen) and eluted at a concentration of approximately 500 ng/μL in nuclease-free water. Complementary DNA (cDNA) was synthesized using the Verso cDNA Kit (Thermo Scientific) as described by the manufacturer. This template was diluted 1:10 in nuclease-free water before adding to the qPCR reactions. qPCRs were all performed in triplicate, using a 126 bp amplicon in the luciferase gene and a 139 bp amplicon in the chlorophyll B gene as a control (see primers in Supplementary Table S3). SsoFast EvaGreen Supermix (Bio-Rad) was used to perform the qPCR, according to the manufacturer's recommendations for cDNA templates. The efficiency of the luciferase primer set was calculated to be >97% at an annealing temperature of 54.8 °C across serial dilutions from 1:5 to 1:625. Under the same conditions, the chlorophyll B primer set also gave an efficiency >97%, and melt curves reveal single products devoid of off-target amplification for both primer sets, so calculations for fold change in transcript level have assumed perfect exponential amplification.

■ ASSOCIATED CONTENT

Supporting Information

Sequences of all synthesized variants of the *psbD* and *psaA* UTRs and sequences of all primers used in this work. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: smayfield@ucsd.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Agilent Technologies for their support in constructing the UTR libraries, and David Masica for his assistance writing code to generate and filter the mutagenized reference sequences. We gratefully acknowledge Kevin Hoang for his help patching, screening, and maintaining the transformant libraries and for his assistance with the genomic DNA preparations for sequencing. We also express many thanks to Steve Head and Lana Schaffer at The Scripps Research Institute DNA Array Core facility for their assistance with the Illumina sequencing and data analysis. Ryan Georgianna and Hussam Auis graciously reviewed the manuscript and provided valuable feedback, and we thank Vineet Bafna for his input on the data analysis and statistical methods. This work was supported by the U.S. Department of Energy (DE-EE0003373) and California Energy Commission (CILMSF #500-10-039). E.S. is supported by a National Science Foundation Graduate Research Fellowship.

■ ABBREVIATIONS

HSM, high salt minimal media; OLS, oligonucleotide library synthesis; RBS, ribosome binding site; SD, Shine-Dalgarno; TAP, tris acetate phosphate media; USER, uracil-specific excision reagent; UTR, untranslated region

■ REFERENCES

(1) Surzycki, R., Greenham, K., Kitayama, K., Dibal, F., Wagner, R., Rochaix, J. D., Ajam, T., and Surzycki, S. (2009) Factors effecting expression of vaccines in microalgae. *Biologicals 37*, 133−138.

(2) Specht, E., Miyake-Stoner, S., and Mayfield, S. (2010) Micro-algae come of age as a platform for recombinant protein production. *Biotechnol. Lett. 32*, 1373−1383.

(3) Blowers, A. D., Klein, U., Ellmore, G. S., and Bogorad, L. (1993) Functional in vivo analyses of the 3′ flanking sequences of the Chlamydomonas chloroplast Rbcl and Psab genes. *Mol. Gen. Genet. 238*, 339−349.

(4) Barnes, D., Franklin, S., Schultz, J., Henry, R., Brown, E., Coragliotti, A., and Mayfield, S. (2005) Contribution of 5′- and 3′-untranslated regions of plastid mRNAs to the expression of Chlamydomonas reinhardtii chloroplast genes. *Mol. Genet. Genomics 274*, 625−636.

(5) Kiaulehn, S., Voytsekh, O., Fuhrmann, M., and Mittag, M. (2007) The presence of UG-repeat sequences in the 3 ′-UTRs of reporter luciferase mRNAs mediates circadian expression and can determine acrophase in Chlamydomonas reinhardtii. *J. Biol. Rhythms 22*, 275−277.

(6) Yohn, C. B., Cohen, A., Danon, A., and Mayfield, S. P. (1996) Altered mRNA binding activity and decreased translational initiation in a nuclear mutant lacking translation of the chloroplast psbA mRNA. *Mol. Cell. Biol. 16*, 3560−3566.

(7) Mayfield, S. P., Cohen, A., Danon, A., and Yohn, C. B. (1994) Translation of the psbA mRNA of Chlamydomonas reinhardtii requires a structured RNA element contained within the 5′ untranslated region. *J. Cell Biol. 127*, 1537−1545.

(8) Boudreau, E., Nickelsen, J., Lemaire, S. D., Ossenbuhl, F., and Rochaix, J.-D. (2000) The Nac2 gene of Chlamydomonas encodes a chloroplast TPR-like protein involved in psbD mRNA stability. *EMBO J. 19*, 3366−3376.

(9) Higgs, D. C., Shapiro, R. S., Kindle, K. L., and Stern, D. B. (1999) Small cis-acting sequences that specify secondary structures in a chloroplast mRNA are essential for RNA stability and translation. *Mol. Cell. Biol. 19*, 8479−8491.

(10) Anthonisen, I. L., Salvador, M. L., and Klein, U. (2001) Specific sequence elements in the 5′ untranslated regions of rbcL and atpB gene mRNas stabilize transcripts in the chloroplast of Chlamydomonas reinhardtii. *RNA 7*, 1024−1033.

(11) Nickelsen, J., Fleischmann, M., Boudreau, E., Rahire, M., and Rochaix, J. D. (1999) Identification of cis-acting RNA leader elements required for chloroplast psbD gene expression in chlamydomonas. *Plant Cell 11*, 957−970.

(12) Schwarz, C., Elles, I., Kortmann, J., Piotrowski, M., and Nickelsena, J. (2007) Synthesis of the D2 protein of photosystem II in Chlamydomonas is controlled by a high molecular mass complex containing the RNA stabilization factor Nac2 and the translational activator RBP40. *Plant Cell 19*, 3627−3639.

(13) Klinkert, B., Elles, I., and Nickelsen, J. (2006) Translation of chloroplast psbD mRNA in Chlamydomonas is controlled by a secondary RNA structure blocking the AUG start codon. *Nucleic Acids Res. 34*, 386−394.

(14) Wobbe, L., Schwarz, C., Nickelsen, J., and Kruse, O. (2008) Translational control of photosynthetic gene expression in phototrophic eukaryotes. *Physiol. Plant 133*, 507−515.

(15) Nour-Eldin, H. H., Hansen, B. G., Norholm, M. H., Jensen, J. K., and Halkier, B. A. (2006) Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. *Nucleic Acids Res. 34*, e122.

(16) Mayfield, S. P., and Schultz, J. (2004) Development of a luciferase reporter gene, luxCt, for Chlamydomonas reinhardtii chloroplast. *Plant J. 37*, 449−458.

(17) Bateman, J. M., and Purton, S. (2000) Tools for chloroplast transformation in Chlamydomonas: expression vectors and a new dominant selectable marker. *Mol. Gen. Genet. 263*, 404−410.

(18) Kuck, U., Choquet, Y., Schneider, M., Dron, M., and Bennoun, P. (1987) Structural and transcription analysis of two homologous genes for the P700 chlorophyll a-apoproteins in Chlamydomonas reinhardii: evidence for in vivo trans-splicing. *EMBO J. 6*, 2185−2195.

(19) Wostrikoff, K., Girard-Bascou, J., Wollman, F. A., and Choquet, Y. (2004) Biogenesis of PSI involves a cascade of translational autoregulation in the chloroplast of Chlamydomonas. *EMBO J. 23*, 2696−2705.

(20) Yohn, C. B., Cohen, A., Danon, A., and Mayfield, S. P. (1998) A poly(A) binding protein functions in the chloroplast as a message-specific translation factor. *Proc. Natl. Acad. Sci. U.S.A. 95*, 2238−2243.

(21) Franzetti, B., Carol, P., and Mache, R. (1992) Characterization and RNA-binding properties of a chloroplast S1-like ribosomal protein. *J. Biol. Chem. 267*, 19075−19081.

(22) Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res. 31*, 3406−3415.

(23) Nickelsen, J., van Dillewijn, J., Rahire, M., and Rochaix, J. D. (1994) Determinants for stability of the chloroplast psbD RNA are located within its short leader region in Chlamydomonas reinhardtii. *EMBO J. 13*, 3182−3191.

(24) Pfalz, J., Bayraktar, O. A., Prikryl, J., and Barkan, A. (2009) Site-specific binding of a PPR protein defines and stabilizes 5′ and 3′ mRNA termini in chloroplasts. *EMBO J. 28*, 2042−2052.

(25) Ossenbuhl, F., and Nickelsen, J. (2000) cis- and trans-Acting determinants for translation of psbD mRNA in Chlamydomonas reinhardtii. *Mol. Cell. Biol. 20*, 8134−8142.

(26) Fargo, D. C., Zhang, M., Gillham, N. W., and Boynton, J. E. (1998) Shine-Dalgarno-like sequences are not required for translation of chloroplast mRNAs in Chlamydomonas reinhardtii chloroplasts or in Escherichia coli. *Mol. Gen. Genet. 257*, 271−282.
(27) Dron, M., Rahire, M., and Rochaix, J. D. (1982) Sequence of the chloroplast 16S rRNA gene and its surrounding regions of Chlamydomonas reinhardii. *Nucleic Acids Res. 10*, 7609−7620.
(28) Harris, E. H., Boynton, J. E., and Gillham, N. W. (1994) Chloroplast ribosomes and protein synthesis. *Microbiol. Rev. 58*, 700−754.
(29) Liu, X. Q., Gillham, N. W., and Boynton, J. E. (1989) Chloroplast ribosomal protein gene rps12 of Chlamydomonas reinhardtii. Wild-type sequence, mutation to streptomycin resistance and dependence, and function in Escherichia coli. *J. Biol. Chem. 264*, 16100−16108.
(30) Michelet, L., Lefebvre-Legendre, L., Burr, S. E., Rochaix, J. D., and Goldschmidt-Clermont, M. (2011) Enhanced chloroplast transgene expression in a nuclear mutant of Chlamydomonas. *Plant Biotechnol. J. 9*, 565−574.
(31) Torronen, A., Mach, R. L., Messner, R., Gonzalez, R., Kalkkinen, N., Harkki, A., and Kubicek, C. P. (1992) The two major xylanases from Trichoderma reesei: characterization of both enzymes and genes. *Biotechnology (N. Y.) 10*, 1461−1465.
(32) Subramaniyan, S., and Prema, P. (2002) Biotechnology of microbial xylanases: enzymology, molecular biology, and application. *Crit. Rev. Biotechnol. 22*, 33−64.
(33) Zerges, W., and Rochaix, J. D. (1994) The 5′ leader of a chloroplast mRNA mediates the translational requirements for two nucleus-encoded functions in Chlamydomonas reinhardtii. *Mol. Cell. Biol. 14*, 5268−5277.
(34) Barkan, A., Rojas, M., Fujii, S., Yap, A., Chong, Y. S., Bond, C. S., and Small, I. (2012) A combinatorial amino Acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet. 8*, e1002910.
(35) Kosuri, S., Eroshenko, N., LeProust, E. M., Super, M., Way, J., Li, J. B., and Church, G. M. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol. 28*, 1295−U1108.
(36) LeProust, E. M., Peck, B. J., Spirin, K., McCuen, H. B., Moore, B., Namsaraev, E., and Caruthers, M. H. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res. 38*, 2522−2540.
(37) Kindle, K. L., Richards, K. L., and Stern, D. B. (1991) Engineering the chloroplast genome − Techniques and capabilities for chloroplast transformation in Chlamydomonas reinhardtii. *Proc. Natl. Acad. Sci. U.S.A. 88*, 1721−1725.
(38) Newman, S. M., Boynton, J. E., Gillham, N. W., Randolph-Anderson, B. L., Johnson, A. M., and Harris, E. H. (1990) Transformation of chloroplast ribosomal RNA genes in Chlamydomonas: molecular and genetic characterization of integration events. *Genetics 126*, 875−888.
(39) Quail, M. A., Otto, T. D., Gu, Y., Harris, S. R., Skelly, T. F., McQuillan, J. A., Swerdlow, H. P., and Oyola, S. O. (2011) Optimal enzymes for amplifying sequencing libraries. *Nat. Methods 9*, 10−11.
(40) Oyola, S. O., Otto, T. D., Gu, Y., Maslen, G., Manske, M., Campino, S., Turner, D. J., Maclnnis, B., Kwiatkowski, D. P., Swerdlow, H. P., and Quail, M. A. (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. *BMC Genomics 13*, 1.
(41) Rasala, B. A., Muto, M., Lee, P. A., Jager, M., Cardoso, R. M., Behnke, C. A., Kirk, P., Hokanson, C. A., Crea, R., Mendez, M., and Mayfield, S. P. (2010) Production of therapeutic proteins in algae, analysis of expression of seven human proteins in the chloroplast of Chlamydomonas reinhardtii. *Plant Biotechnol. J. 8*, 719−733.