SYNTHETIC BIOLOGY

Synthetic Biology, 2024, **9(1)**, ysae018

DOI: https://doi.org/10.1093/synbio/ysae018
Advance Access Publication 2 December 2024
Software

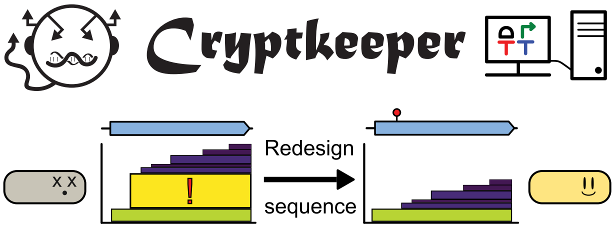
CryptKeeper: a negative design tool for reducing unintentional gene expression in bacteria

Cameron T. Roots and Jeffrey E. Barrick*

Department of Molecular Biosciences, Center for Systems and Synthetic Biology, The University of Texas at Austin, Austin, TX 78712, USA
*Corresponding author. Department of Molecular Biosciences, Center for Systems and Synthetic Biology, The University of Texas at Austin, 2500 Speedway A5000, Austin, TX 78712, USA. E-mail: jbarrick@cm.utexas.edu

Abstract

Foundational techniques in molecular biology—such as cloning genes, tagging biomolecules for purification or identification, and overexpressing recombinant proteins—rely on introducing non-native or synthetic DNA sequences into organisms. These sequences may be recognized by the transcription and translation machinery in their new context in unintended ways. The cryptic gene expression that sometimes results has been shown to produce genetic instability and mask experimental signals. Computational tools have been developed to predict individual types of gene expression elements, but it can be difficult for researchers to contextualize their collective output. Here, we introduce CryptKeeper, a software pipeline that visualizes predictions of *Escherichia coli* gene expression signals and estimates the translational burden possible from a DNA sequence. We investigate several published examples where cryptic gene expression in *E. coli* interfered with experiments. CryptKeeper accurately postdicts unwanted gene expression from both eukaryotic virus infectious clones and individual proteins that led to genetic instability. It also identifies off-target gene expression elements that resulted in truncations that confounded protein purification. Incorporating negative design using CryptKeeper into reverse genetics and synthetic biology workflows can help to mitigate cloning challenges and avoid unexplained failures and complications that arise from unintentional gene expression.



Graphical Abstract

Keywords: plasmid instability; reliability and reproducibility; computational DNA sequence design; design-build-test cycle; recombinant protein overexpression

1. Introduction

RNAs and proteins may be unexpectedly transcribed and translated from a DNA sequence. This type of cryptic gene expression can complicate studying and engineering biological systems. Cryptic gene expression can occur when natural DNA sequences contain promoters and ribosome-binding sites that are not annotated because they are redundant, antisense, or internal to genes. It can also emerge when sequences are moved into a new cellular context (e.g. cloning eukaryotic sequences in *Escherichia coli*) [1–4] or as a consequence of engineered changes to sequences (e.g. combining genetic parts, optimizing codon usage, or introducing artificial watermarks) [5, 6]. Cryptic gene expression products that interfere with the intended function of an engineered DNA construct may cause a design to be deemed a failure. Worse yet, cryptic gene expression may occur unbeknownst to researchers, causing them to misinterpret experimental results [7, 8]. Unintentional expression of genes, truncated pieces of genes, or out-of-frame products is often burdensome or even toxic to a host organism, creating a strong selection pressure favoring cells with mutations in the engineered DNA sequence [1–4, 9]. Rapid evolution of escape mutants that eliminate these or other sources of burden can be one reason that certain sequences are unreliable or even unconstructable [10, 11].

Both the process and products of gene expression can be burdensome to a cell. Studies of recombinant protein overexpression have shown that growth rates of bacterial cells decrease in proportion to how much of their translational capacity, usually determined by the number of ribosomes, is redirected to

Submitted: 29 November 2024; **Received (in revised form)**: 13 November 2024; **Accepted**: 29 November 2024

expressing exogenous proteins [12, 13]. Expression of some proteins is also directly deleterious due to their activities, whether they are enzymes that rewire metabolism in ways that redirect limiting resources away from cellular replication or disrupt cellular homeostasis in other ways [11, 12]. It is rarer for transcription of RNA alone to cause an appreciable burden on a bacterial cell, but it has been documented in yeast protein overexpression systems [14, 15].

Negative design is the process of eliminating undesirable qualities to engineer a safer or more effective system [16]. The relative rates at which ribosomes initiate translation from different start codons in *E. coli* and other bacteria can be accurately predicted from characteristics of their ribosome-binding sites (RBSs) and surrounding sequences [17–19]. Therefore, one negative design strategy for solving problems stemming from cryptic protein expression is to concentrate on redesigning a DNA sequence to eliminate the potential for unwanted translation, whether or not there is any evidence that a relevant mRNA is transcribed. Because translation and transcription are coupled in bacteria, disrupting translation is also expected to reduce RNA levels by short-circuiting transcription and promoting mRNA degradation [20, 21].

Another negative design approach would be to eliminate cryptic transcription so mRNAs are not produced in the first place. Unfortunately, tools for predicting bacterial promoters and terminators currently have limited accuracy, only make qualitative predictions, and/or are not openly accessible [22–40]. Furthermore, many promoter prediction tools that incorporate machine learning classifiers use protein coding sequences as their non-promoter group during training, an assumption which could lead them to systematically underpredict true cases of cryptic promoters within ORFs [30, 37, 40]. Even so, predictions of promoters and terminators could provide additional context for interpreting predictions of translated reading frames and warn of other potential problems with a sequence design, such as the accidental production of inhibitory transcripts that are antisense to known genes.

Here we describe Crypt-Keeper, an open-source software tool that integrates and displays predictions of *E. coli* gene expression elements in engineered DNA constructs, such as plasmids. CryptKeeper is designed to allow users to evaluate the potential for cryptic gene expression that may interfere with the construction or function of a DNA sequence. We demonstrate the utility of CryptKeeper by using it to analyze the results of several prior studies in which researchers identified cryptic gene expression that was problematic and then redesigned their sequences to avoid it.

2. Materials and methods

2.1 Software overview

CryptKeeper integrates the output of several tools that predict bacterial gene expression elements from DNA sequences (Fig. 1a). It accepts input sequences in GenBank or FASTA format. It displays predictions of translation initiation sites, promoters, Rho-dependent terminators, and intrinsic (Rho-independent) terminators. These predictions are summarized with a translational burden score and displayed in an interactive visualization so that a user can evaluate whether there is the potential for cryptic gene expression that may interfere with the function or stability of their DNA sequence. CryptKeeper is a Python package. It and all of its dependencies can be installed as Bioconda packages.

2.2 Translational burden prediction

Thermodynamic models can accurately predict the relative rates at which ribosomes initiate translation from different start codons [17–19]. CryptKeeper calculates the translation initiation rates at all start codons in the input sequence using OSTIR version 1.1.2 [41]. The translational burden of a DNA construct on an *E. coli* host cell is expected to be proportional to the number of ribosomes bound to new mRNAs transcribed from it [12, 13]. If rates of translation elongation and termination are fast relative to initiation and uniform, then ribosome occupancy of a given open-reading frame (ORF) will be directly proportional to the rate of initiation at its start codon and its length. Therefore, we summarize these results as a translational burden score for each ORF that is the product of its predicted translation initiation rate and its length in base pairs. Very short ORFs (<45 bases) are unlikely to contribute much to the overall burden of a construct. They are predicted by Cryptkeeper but are not shown in its graphical output (to avoid unnecessary visual clutter) when using the default settings.

2.3 Promoter and terminator prediction

CryptKeeper displays predictions of *E. coli* σ^{70} promoters from a fork of the Promoter Calculator version 1.2.2 [29] that we created to add multithreading, reduce memory usage, and make it installable as a Bioconda package. Intrinsic (Rho-independent) terminators are predicted using TransTermHP version 2.09 [28]. Rho-dependent terminators are predicted using a fork of RhoTermPredict version 3.4.0 [23] that we created to make it installable as a Bioconda package. CryptKeeper does not attempt to integrate transcription predictions into an overall score because these are less accurate and complete than translation predictions. For example, the Promoter Calculator only predicts σ^{70} promoter initiation strength with a coefficient of determination of 0.45 for a test set of plasmid-encoded promoters in *E. coli* [29]. While some tools exist that predict promoters that use alternative sigma factors, they are classifiers that do not quantitatively predict strength or are not open-source tools that can be run at the command line [30, 37, 40]. By contrast, prediction of intrinsic terminators reaches >90% accuracy and specificity [28] and should generalize to many other bacterial species. However, there is almost always some read-through of these terminators [42, 43], and this characteristic is not predicted by current tools. Predictions of Rho-dependent terminators may also generalize across different bacterial groups, but they tend to have indistinct boundaries, and even less is known about how well their presence and efficiencies are predicted by current algorithms [23]. Despite these current shortcomings, predictions of transcription initiation and termination elements may provide additional context to the user and may be sufficient for spotting problems in certain cases.

2.4 Output

For an input DNA construct (Fig. 1b), CryptKeeper uses the Bokeh Python library [44] to output an HTML document that includes an interactive plot (Fig. 1c). This plot displays stacked boxes associated with different ORFs. The height of each box is proportional to the predicted rate of translation initiation at the start codon of the ORF, which makes its area proportional to the translational burden score. Boxes are also colored according to their burden scores on a linear scale. Promoter and terminator predictions are shown on two inner tracks, one for each DNA strand. The number of these predictions shown can be adjusted by the user. The default is to show the three strongest promoters, Rho-dependent terminators, and intrinsic terminators per kilobase of the input sequence. If a GenBank file was used as the input, features annotated in this

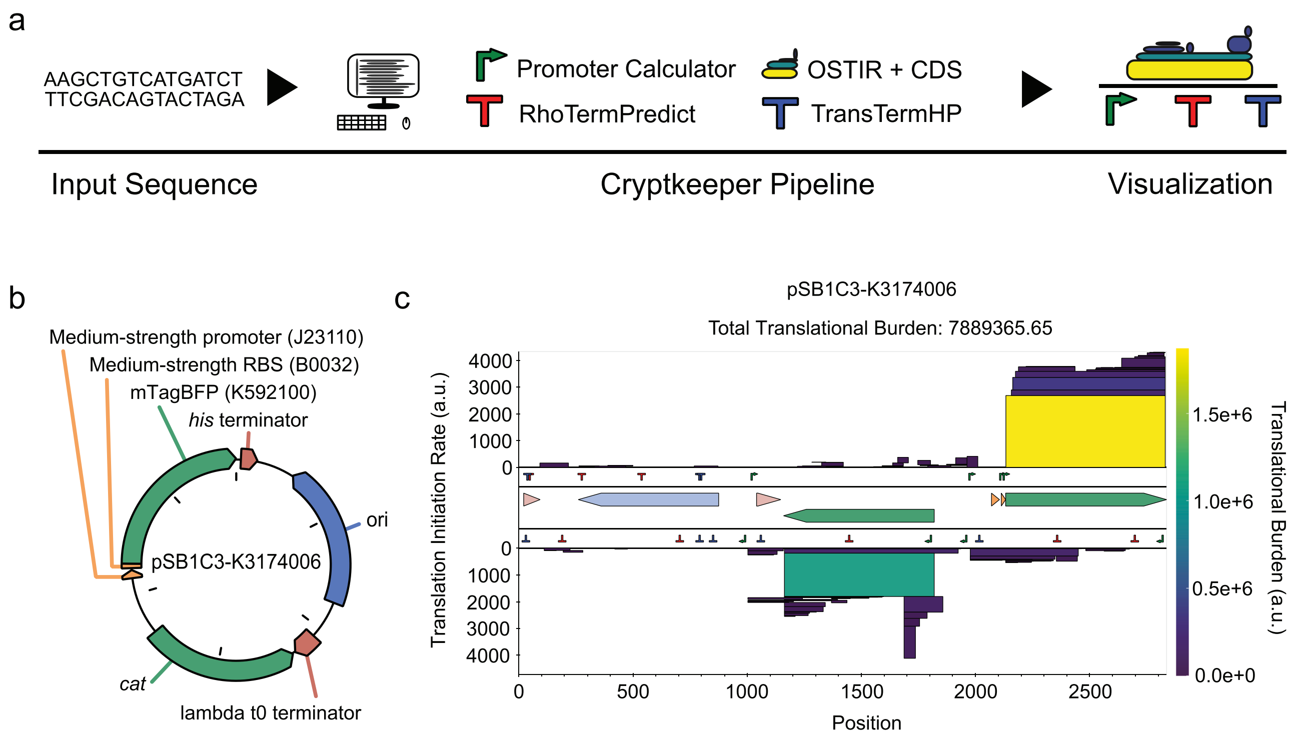


Figure 1. (a) CryptKeeper overview. Predictions of gene expression signals in an input DNA sequence are integrated into an interactive plot and used to calculate an overall translational burden score. (b) pSB1C3-K3174006, an example of a plasmid engineered to express a transgene. It contains a pUC origin of replication, chloramphenicol resistance cassette, and BioBrick K3174006, which expresses mTagBFP (K592100) under control of a medium-strength constitutive promoter (J23110) and a medium-strength ribosome binding site (B0032) [10]. Figure adapted from pLannotate output [46]. (c) CryptKeeper output for pSB1C3-K3174006. The outermost tracks display protein-coding sequences on the forward and reverse strands as stacked boxes with heights proportional to their predicted translation initiation rates and colors and areas proportional to their individual translational burden scores. The next inner two tracks display predictions of RNA expression signals on each strand: promoters, Rho-dependent terminators, and intrinsic terminators. The central track displays annotations from the input GenBank file (matching panel b).

file are shown in the central track. The CryptKeeper plot can be zoomed and rescaled, and it displays information about each predicted feature and annotation on mouseover. To facilitate further analysis by users, a table describing each predicted element is provided below the plot and in a separate comma-separated values (CSV) output file.

2.5 Test datasets

We tested CryptKeeper on DNA sequences from six published studies that encountered and characterized cryptic gene expression from plasmids in *E. coli* (Table 1). In each case, the relevant sequences were recreated *in silico*. When sufficient information was available, the entire plasmid sequences were reconstructed and analyzed. All of these studies report how researchers introduced mutations that resolved their issues, which allowed us to further examine how well CryptKeeper output tracks with the experimentally validated outcomes of redesigning DNA sequences. Plasmid annotations were based on GenBank records [45], pLannotate predictions [46], and descriptions in the relevant studies.

3. Results

3.1 Virus infectious clone case studies

Cloning a plant or animal virus into an *E. coli* vector makes it possible to use standard molecular biology workflows to modify its sequence. These infectious clone plasmids are a widespread reverse genetics tool for studying viruses and their applications in biotechnology. Because the virus DNA in an infectious clone is replicating in a cell that is evolutionarily distant from its eukaryotic host, it should not produce toxic virus proteins or infectious particles. However, sometimes there are cryptic gene expression elements in virus sequences that direct transcription and translation in *E. coli* cells. These products can cause significant translational burden or toxicity, leading to plasmid instability.

Table 1. Test datasets

Cloned sequence	Complication	Solution	Citation
Zika virus	Plasmid instability	Eliminate two promoters	[1]
Dengue virus	Plasmid instability	Insert artificial intron	[3]
Mouse *mdr1a* cDNA	Plasmid instability	Eliminate ribosome binding site	[4]
Human *SCN1A* cDNA	Plasmid instability	Eliminate promoter and insert artificial intron	[2]
Yeast GCN5	Truncated protein	Eliminate internal ribosome binding site and/or start codon	[7]
Human *NISTmAb*	Truncated protein	Eliminate internal start codon	[8]

In a study that cloned Zika virus into a bacterial plasmid to create an infectious clone, researchers identified two putative *E. coli* promoters they designated ECP1 and ECP2 within the nucleotide sequence of the E envelope protein that they suspected were

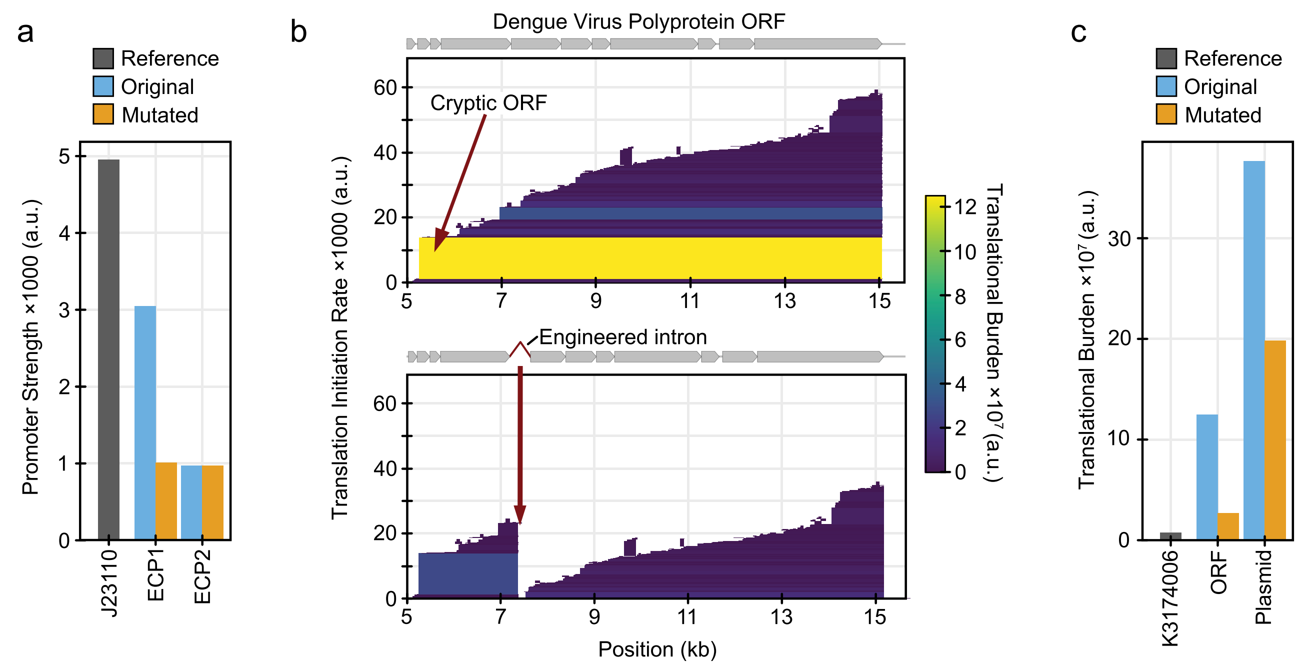


Figure 2. Case studies of virus infectious clone redesign. (a) Predicted strengths of promoters in a Zika virus infectious clone plasmid before and after redesign, compared to the promoter in BioBrick K3174006. (b) CryptKeeper burden plots for the Dengue virus sequence in an infectious clone plasmid before and after adding an intron that disrupts the ORF that makes the largest contribution to burden. (c) Predicted translational burden of a Dengue virus infectious clone before and after redesign compared to the predicted burden of the mTagBFP ORF from BioBrick K3174006.

responsible for the expression of toxic products [1]. CryptKeeper predicts ECP1 as the strongest promoter within the Zika genome. ECP2 is also identified by CryptKeeper, but it is among the weaker promoter predictions in the construct, so it is not shown in the plot with the default settings. The researchers found that introducing point mutations in both ECP1 and ECP2, as well as ECP1 alone, allowed them to stabilize the infectious clone. Their sequence changes reduced the transcription initiation rate predicted by CryptKeeper for ECP1 by 67% and did not change the predicted rate of ECP2 (Fig. 2a). Both promoters are located approximately 1800 bases upstream of an ORF with a predicted translational burden that is similar to what we found explained instability in other case studies. However, these researchers did not test for cryptic translation, so we cannot determine whether the instability they observed was due to translational burden or a toxic effect of a protein product.

In another study, researchers found that a Dengue virus infectious clone plasmid was unstable in *E. coli*, which they attributed to cryptic bacterial promoters within its 5′ untranslated region (UTR) [47]. Subsequent studies produced stabilized Dengue virus infectious clones through a variety of approaches. Several copies of the TetR binding site tetO were sufficient for its binding to the 5′ UTR to prevent transcription initiation from the upstream promoters [48]. Using a low-copy bacterial artificial chromosome instead of a high-copy plasmid also stabilized an infectious Dengue virus clone, presumably because this reduced all cryptic expression [48, 49]. Recently, researchers constructed a stable infectious clone in a high-copy plasmid by introducing a synthetic intron within the NS1 coding region [3]. This addition interrupts translation of the long Dengue polyprotein in *E. coli* cells where it is not spliced, but the intron is removed and the virus RNA becomes infectious when transfected into mammalian cells. CryptKeeper predicts several promoters in the Dengue 5′ UTR, including a weak promoter overlapping with previously predicted ones. It also predicts a strong *E. coli* ribosome-binding site that initiates translation beginning at M126 of the Dengue polyprotein (Fig. 2b). Adding the synthetic intron introduces a stop codon that interrupts this very long reading frame (Fig. 2b), which reduces the predicted translational burden from this ORF by 88.4% and reduces the total translational burden of the complete infectious clone plasmid by 48.6% (Fig. 2c).

3.2 Eukaryotic transgene case studies

Gene expression costs are not restricted to plasmids that encode viruses. Individual proteins or protein complexes are often cloned into bacteria to study their functions or for biomanufacturing. Long or toxic ORFs in these constructs can present cloning issues similar to those of the infectious clone plasmids.

When cloned in *E. coli*, the mouse *mdr1a* cDNA was found to contain a bacterial promoter and ribosome binding site near the 5′ end of its ORF [4]. These elements contributed to instability. Mutating the start codon associated with the RBS resulted in a stable plasmid. CryptKeeper predicts both the cryptic promoter and the cryptic RBS reported in the study. The researchers eliminated cryptic translation in *E. coli* by changing the M107 ATG start codon of *mdr1a* to CTG. CryptKeeper predicts that this edit should completely abolish expression of the highly burdensome ORF (Fig. 3a, b), thereby reducing the burden of the cDNA portion of the plasmid by 60.9% (Fig. 3d).

Another similar study found that the *SCN1A* cDNA encoding the human sodium channel $Na_v1.1$ contains a cryptic promoter and translation initiation site that results in strong expression of a truncated product in *E. coli* [2]. Introduction of a β-globin/IgG chimeric intron containing an in-frame stop codon was used to disrupt *E. coli* translation and establish plasmid stability in this case. CryptKeeper detects both the promoter and translation initiation site suspected of causing cryptic gene expression (Fig. 3c). Interruption of the cryptic ORF by the introduced intron reduces

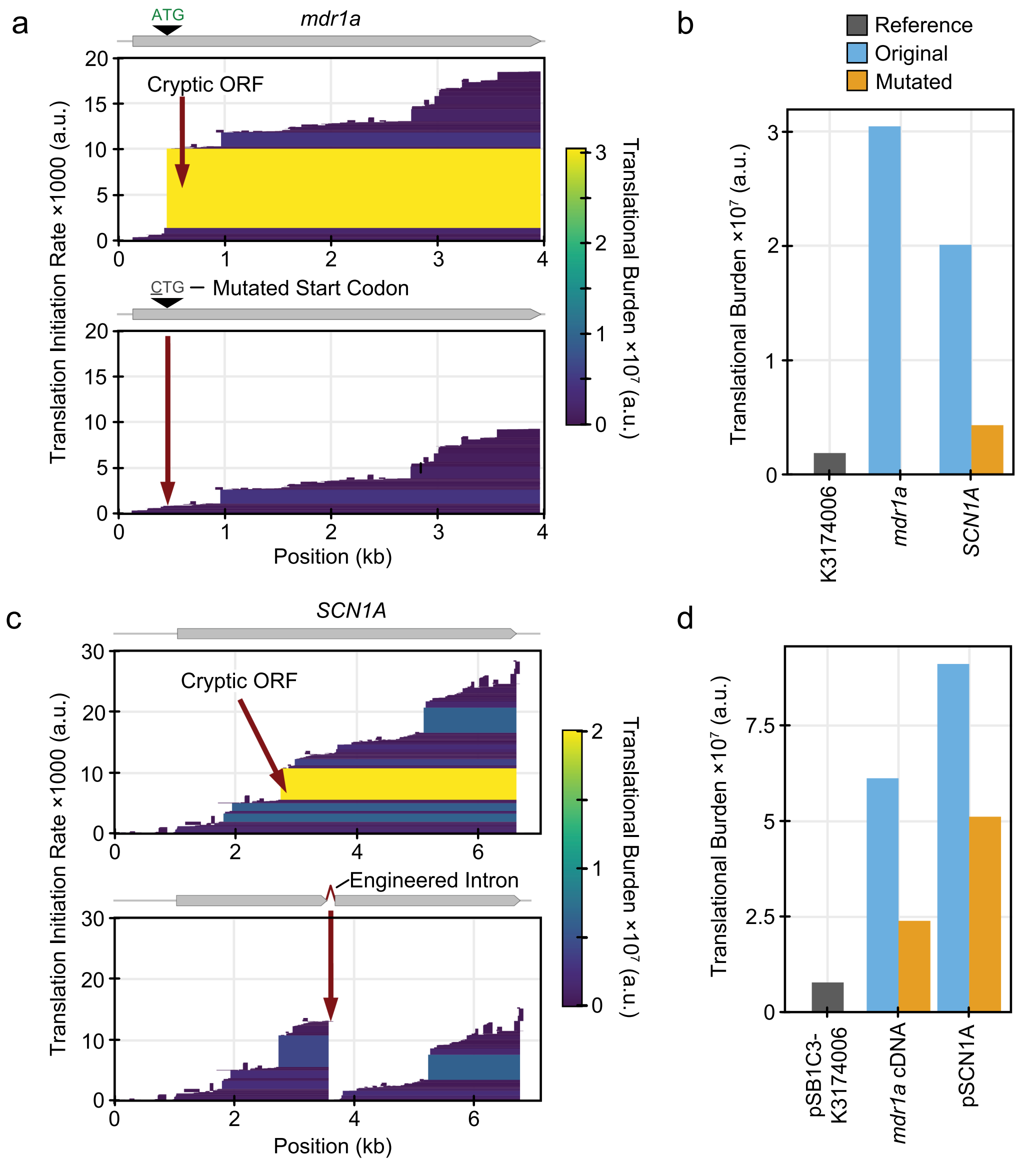


Figure 3. Case studies of eliminating cryptic translation from eukaryotic transgenes. (a) CryptKeeper translation predictions for a cloned mouse *mdr1a* cDNA sequence before and after mutating the start codon associated with a cryptic ORF. (b) CryptKeeper burden predictions for the mTagBFP ORF from BioBrick K3174006, the cryptic ORF of *mdr1a* before and after mutating its start codon, and the cryptic ORF of *SCN1A* before and after introducing an engineered intron. (c) CryptKeeper translation predictions for a human *SCN1A* cDNA sequence before and after redesigning it to include an engineered intron. (d) Predicted total burden of plasmid pSB1C-K3174006, the complete *mdr1a* cDNA before and after mutating the cryptic ORF start codon, and the full plasmid encoding *SCN1A* before and after introducing the engineered intron.

the predicted translational burden associated with its initiation site by 78.4% (Fig. 3b) and the score for the complete plasmid by 44.3% (Fig. 3d).

3.3 Protein truncation case studies

Experimental complications from cryptic translation are not limited to instability from burden. Truncated proteins produced by internal translation initiation sites can disrupt fusions to purification tags, antibody epitopes, or fluorescent reporters, decoupling these sequences from the protein of interest. It has been demonstrated that mutating predicted translation initiation sites can be used eliminate the unintentional production of truncated proteins [50]. CryptKeeper is able to detect and visualize these internal ribosome-binding sites, which can help researchers diagnose and redesign their sequences to prevent the unintentional translation of truncated proteins.

Researchers who cloned the yeast *GCN5* gene in *E. coli* to express and purify the yeast SAGA histone acetyltransferase observed a smaller product that was suspected to be a proteolytic degradation product of full-length SAGA on SDS-PAGE gels [7]. Editing the construct to disrupt an internal RBS or start codon reduced or eliminated this band, revealing that the smaller product was a truncated protein resulting from cryptic translation. CryptKeeper predicts that the rate of translation initiation at this internal start codon is ∼350% than that of the upstream start codon for the full-length protein (Fig. 4a). The mutated RBS sequence used in the study reduced the truncated protein's predicted expression to just 8% of the unmodified truncation and 29% of the full-length protein. As shown in the study, mutating this

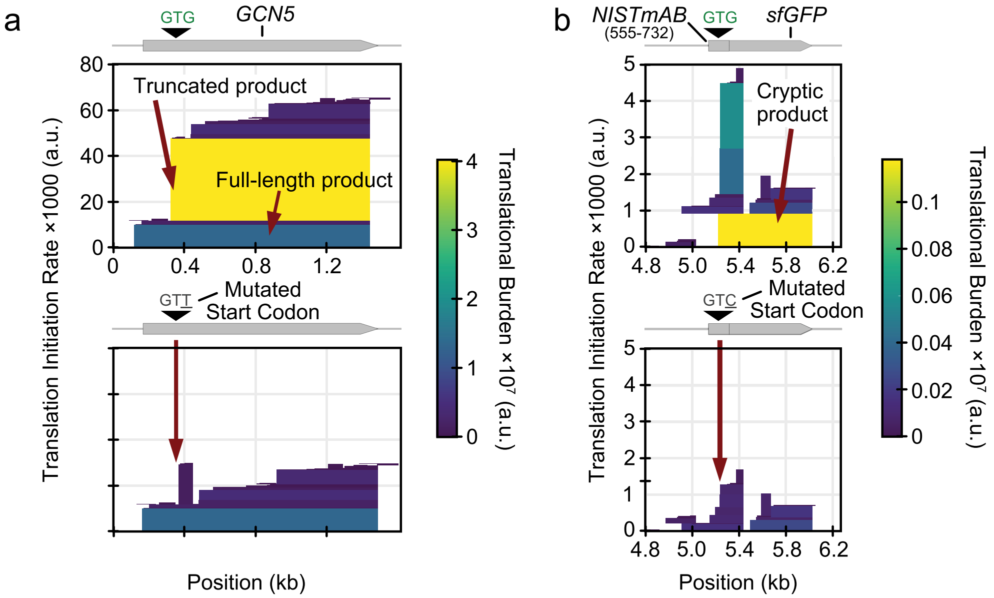


Figure 4. Case studies of redesign to eliminate truncated protein expression. (a) CryptKeeper translation predictions for a *GCN5* expression cassette for producing yeast SAGA histone acetyltransferase. Predictions are shown before and after mutating an internal start codon. (b) CryptKeeper translation predictions for a construct consisting of a fragment of the codon-optimized antibody NISTmAB ORF (nucleotides 555 to 732) placed upstream of a sfGFP reporter. Predictions are shown before and after mutating an internal start codon in the NISTmAB ORF.

start codon from GTG to GTT entirely eliminated the truncated product.

A NISTmAb human antibody gene sequence for *E. coli* expression produced a shorter protein that copurified with the complete antibody [6]. Later, it was demonstrated that this product was a truncated heavy chain produced by an RBS and GTG start codon that were unintentionally introduced during the codon optimization process [8]. In a construct in which the researchers used this putative RBS and GTG codon to drive sfGFP expression to investigate the source of this unwanted product, mutating the start codon to GTC fully eliminated fluorescence. Crypt-Keeper identifies the unintended RBS in the sfGFP expression construct and correctly predicts that no full-length protein will be produced after the GTG to GTC start codon mutation (Fig. 4b).

4. Discussion

DNA synthesis, assembly, and cloning workflows are critical for a wide variety of bioengineering tasks, including vector construction, protein purification, enzyme engineering, genetic circuit design, and more. Cryptic gene expression can disrupt these workflows and obfuscate experimental results, leading to abandoning constructs, time-consuming troubleshooting, or incorrect conclusions. Since many cloning failures go unexplained and unpublished, problems with cryptic gene expression are undoubtedly underreported. Currently, there is no freely accessible, open-source solution for integrating output from the ecosystem of tools for predicting gene expression elements into a visual dashboard that makes potential design issues immediately evident to a researcher. We show that CryptKeeper can effectively diagnose these issues, as described in the troubleshooting case studies.

Ultimately, the utility of CryptKeeper is limited by the computational tools that are available for predicting gene expression. For *E. coli*, these challenges currently include high rates of false-positives/false-negatives and poor quantitative accuracy when predicting promoters, even with state-of-the-art algorithms trained on large sets of experimental data [22, 29, 30, 33, 37, 40]. Tools for predicting transcription initiation rates driven by alternative sigma factors, transcription driven by T7 RNA polymerase, and quantitative predictions of terminator read-through are also needed to complete the picture of transcription in this host. Some CryptKeeper predictions are expected to become less accurate for other bacteria as their evolutionary distance from *E. coli* increases. OSTIR and the Promoter Calculator, two main CryptKeeper dependencies, are calibrated specifically for *E. coli* [29, 41]. OSTIR can be configured to use alternative anti-Shine–Dalgarno sequences, but its effectiveness in other bacteria has not been evaluated. On the other hand, CryptKeeper predictions of terminators by TransTermHP and RhoTermPredict are likely to be accurate for diverse species, including Gram-positive bacteria such as *Bacillus subtilis* [23, 28]. Ideally, it would be possible to tailor all of these tools for different bacterial hosts used as alternative chassis for cloning (e.g. *Vibrio natriegens*) [51] or for specific bioengineering applications (e.g. *Pseudomonas putida*) [52]. Another goal for the field should be to extend these approaches to widely used eukaryotic chassis, such as yeast (*Saccharomyces cerevisiae*), where cryptic gene expression also poses challenges [53]. We expect tools for predicting gene expression to improve as laboratory automation makes more extensive training sets available and as new machine learning approaches (e.g. large language models) are adopted [54, 55].

The main summary output from CryptKeeper is a translational burden score for each ORF in the input sequence. This score reflects, in relative terms, how much of a cell's capacity for translation is expected to be redirected to this ORF, as this has been shown to be the major cause of burden for many constructs [10, 12, 13]. The translational burden score is currently calculated simply as the translation initiation rate multiplied by the ORF length. More detailed models could account for how rare codons that slow translation or mRNA structures that act as pause sites exacerbate this burden by leading to more ribosomes than expected from the simple model becoming sequestered on certain mRNAs. This effect has been experimentally demonstrated by comparing constructs with rare codons early versus late in a reading frame [12]. Incorporating these refinements into CryptKeeper's score could be especially important for evaluating burden from cloning eukaryotic sequences with very different codon usage into *E. coli* plasmids, as is the case when constructing virus infectious clones.

CryptKeeper is most useful as a tool for negative design. In this paradigm, one takes care to avoid issues that could arise from off-target interactions when engineering a system. Other examples of negative design in synthetic biology include adding genetic insulators between modules [56], avoiding crosstalk between metabolic pathways [57], and editing DNA sequences to remove mutational hotspots [58, 59]. In the context of negative design, it is fine to overpredict problems, as long as alternatives without these potential issues exist in the space of possible designs. Biological sequences have so many degrees of freedom that this is often the case. For example, one could eliminate an internal start codon with a single amino acid substitution that does not compromise folding of the encoded protein.

Researchers can—and we argue, should—take precautionary steps to avoid off-target translation and unintentional translational burden even if it is unclear whether any RNA containing a problematic ORF will be transcribed. CryptKeeper can inform this redesign process by highlighting gene expression elements and allowing users to evaluate the effects of editing a sequence in different ways to eliminate potential issues. To address unwanted translation, we recommend making synonymous substitutions in an ORF to eliminate antisense or alternative (non-ATG) start codons and weaken RBS sequences upstream of internal ATG start codons. Future versions of CryptKeeper could automate recoding sequences for this design objective. Natural gene sequences have experienced selection against off-target gene expression elements in their original biological contexts [60, 61]. CryptKeeper can help researchers follow suit and apply negative design to sequences they create *de novo* or transplant into new contexts to improve the reliability and reproducibility of synthetic biology.

Acknowledgments

We thank Sean Leonard and Peng Geng for helpful discussions and software testing. We also thank Davy Wybiral for helping to resolve a Python Package Index naming conflict. We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high-performance computing resources.

Author contributions

Conceptualization: Cameron T. Roots and Jeffrey E. Barrick;
Funding Acquisition: Jeffrey E. Barrick;
Investigation: Cameron T. Roots;
Methodology: Cameron T. Roots and Jeffrey E. Barrick;
Software: Cameron T. Roots;
Visualization: Cameron T. Roots;
Writing—Original draft preparation: Cameron T. Roots and Jeffrey E. Barrick; and
Writing—review & editing: Cameron T. Roots and Jeffrey E. Barrick.

Conflict of interest: None declared.

Funding

This work was supported by the Defense Advanced Research Projects Agency (HR0011-17-2-0052), National Institutes of Health (R01GM088344), Army Research Office (W911NF-20-1-0195), and National Science Foundation (IOS-2103208 and MCB-2123996).

Data availability

CryptKeeper is open-source software released under a GPL-3.0 license. Source code, instructions, and data used for testing are available at https://github.com/barricklab/CryptKeeper. The current version of the repository has been archived on Zenodo (DOI:10.5281/zenodo.13308762). Additionally, CryptKeeper can be installed as a Bioconda package.

References

1. Chen Y, Liu T, Zhang Z *et al.* Novel genetically stable infectious clone for a Zika virus clinical isolate and identification of RNA elements essential for virus production. *Virus Res* 2018;**257**:14–24. https://doi.org/10.1016/j.virusres.2018.08.016
2. DeKeyser J-M, Thompson CH, George AL. Cryptic prokaryotic promoters explain instability of recombinant neuronal sodium channels in bacteria. *J Biol Chem* 2021;**296**:100298. https://doi.org/10.1016/j.jbc.2021.100298
3. Holliday M, Corliss L, Lennemann NJ. Construction and rescue of a DNA-launched DENV2 infectious clone. *Viruses* 2023;**15**:275. https://doi.org/10.3390/v15020275
4. Pluchino KM, Esposito D, Moen JK *et al.* Identification of a cryptic bacterial promoter in mouse (*mdr1a*) P-glycoprotein cDNA. *PLOS ONE* 2015;**10**:e0136396. https://doi.org/10.1371/journal.pone.0136396
5. Espah Borujeni A, Zhang J, Doosthosseini H *et al.* Genetic circuit characterization by inferring RNA polymerase movement and ribosome usage. *Nat Commun* 2020;**11**:5001. https://doi.org/10.1038/s41467-020-18630-2
6. Reddy PT, Brinson RG, Hoopes JT *et al.* Platform development for expression and purification of stable isotope labeled monoclonal antibodies in *Escherichia coli*. *mAbs* 2018;**10**:992–1002. https://doi.org/10.1080/19420862.2018.1496879
7. Jennings MJ, Barrios AF, Tan S. Elimination of truncated recombinant protein expressed in *Escherichia coli* by removing cryptic translation initiation site. *Protein Expr Purif* 2016;**121**:17–21. https://doi.org/10.1016/j.pep.2015.12.001
8. Leith EM, O'Dell WB, Ke N *et al.* Characterization of the internal translation initiation region in monoclonal antibodies expressed in *Escherichia coli*. *J Biol Chem* 2019;**294**:18046–56. https://doi.org/10.1074/jbc.RA119.011008
9. Umenhoffer K, Fehér T, Balikó G *et al.* Reduced evolvability of *Escherichia coli* MDS42, an IS-less cellular chassis for molecular and synthetic biology applications. *Microb Cell Factories* 2010;**9**:38. https://doi.org/10.1186/1475-2859-9-38
10. Radde N, Mortensen GA, Bhat D *et al.* Measuring the burden of hundreds of BioBricks defines an evolutionary limit on constructability in synthetic biology. *Nat Commun* 2024;**15**:6242. https://doi.org/10.1038/s41467-024-50639-9
11. Rugbjerg P, Myling-Petersen N, Porse A *et al.* Diverse genetic error modes constrain large-scale bio-based production. *Nat Commun* 2018;**9**:787. https://doi.org/10.1038/s41467-018-03232-w
12. Ceroni F, Algar R, Stan G-B *et al.* Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nat Methods* 2015;**12**:415–18. https://doi.org/10.1038/nmeth.3339

13. Scott M, Gunderson CW, Mateescu EM *et al.* Interdependence of cell growth and gene expression: origins and consequences. *Science* 2010;**330**:1099–102. https://doi.org/10.1126/science.1192588

14. Kafri M, Metzl-Raz E, Jona G *et al.* The cost of protein production. *Cell Rep* 2016;**14**:22–31. https://doi.org/10.1016/j.celrep.2015.12.015

15. Segall-Shapiro TH, Meyer AJ, Ellington AD *et al.* A 'resource allocator' for transcription based on a highly fragmented T7 RNA polymerase. *Mol Syst Biol* 2014;**10**:742. https://doi.org/10.15252/msb.20145299

16. Richardson JS, Richardson DC. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci* 2002;**99**:2754–59. https://doi.org/10.1073/pnas.052706099

17. Reis AC, Salis HM. An automated model test system for systematic development and improvement of gene expression models. *ACS Synth Biol* 2020;**9**:3145–56. https://doi.org/10.1021/acssynbio.0c00394

18. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* 2009;**27**:946–50. https://doi.org/10.1038/nbt.1568

19. Seo SW, Yang J-S, Kim I *et al.* Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab Eng* 2013;**15**:67–74. https://doi.org/10.1016/j.ymben.2012.10.006

20. Deana A, Belasco JG. Lost in translation: the influence of ribosomes on bacterial mRNA decay. *Genes Dev* 2005;**19**:2526–33. https://doi.org/10.1101/gad.1348805

21. Kim S, Wang Y-H, Hassan A *et al.* Re-defining how mRNA degradation is coordinated with transcription and translation in bacteria. *bioRxiv* 2024. https://doi.org/10.1101/2024.04.18.588412

22. de Avila e Silva S, Echeverrigaray S, Gerhardt GJL. BacPP: Bacterial Promoter Prediction—a tool for accurate sigma-factor specific assignment in enterobacteria. *J Theor Biol* 2011;**287**:92–99. https://doi.org/10.1016/j.jtbi.2011.07.017

23. Di Salvo M, Puccio S, Peano C *et al.* RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on *Escherichia coli, Bacillus subtilis* and *Salmonella enterica* databases. *BMC Bioinf* 2019;**20**: 117. https://doi.org/10.1186/s12859-019-2704-x

24. Feng C-Q, Zhang Z-Y, Zhu X-J *et al.* iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 2019;**35**:1469–77. https://doi.org/10.1093/bioinformatics/bty827

25. Gardner PP, Barquist L, Bateman A *et al.* RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res* 2011;**39**:5845–52. https://doi.org/10.1093/nar/gkr168

26. Huang Y-K, Yu C-H, Ng I-S. Precise strength prediction of endogenous promoters from *Escherichia coli* and J-series promoters by artificial intelligence. *J Taiwan Inst Chem Eng* 2024;**160**:105211. https://doi.org/10.1016/j.jtice.2023.105211

27. Jin Y, Ma H, Xu ZZ *et al.* BATTER: accurate prediction of Rho-dependent and Rho-independent transcription terminators in metagenomes. 2023.

28. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 2007;**8**: R22. https://doi.org/10.1186/gb-2007-8-2-r22

29. LaFleur TL, Hossain A, Salis HM. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat Commun* 2022;**13**:5159. https://doi.org/10.1038/s41467-022-32829-5

30. Lai H-Y, Zhang Z-Y, Su Z-D *et al.* iProEP: a computational predictor for predicting promoter. *Mol Ther Nucleic Acids* 2019;**17**:337–46. https://doi.org/10.1016/j.omtn.2019.05.028

31. Lesnik EA, Sampath R, Levene HB *et al.* Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res* 2001;**29**:3583–94. https://doi.org/10.1093/nar/29.17.3583

32. Lin H, Deng E-Z, Ding H *et al.* iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res* 2014;**42**:12961–72. https://doi.org/10.1093/nar/gku1019

33. Liu B, Li K. iPromoter-2L2.0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features. *Mol Ther Nucleic Acids* 2019;**18**:80–87. https://doi.org/10.1016/j.omtn.2019.08.008

34. Nadiras C, Eveno E, Schwartz A *et al.* A multivariate prediction model for Rho-dependent termination of transcription. *Nucleic Acids Res* 2018;**46**:8245–60. https://doi.org/10.1093/nar/gky563

35. Naville M, Ghuillot-Gaudeffroy A, Marchais A *et al.* ARNold: a web tool for the prediction of Rho-independent transcription terminators. *RNA Biol* 2011;**8**:11–13. https://doi.org/10.4161/rna.8.1.13346

36. Salamov VS, Solovyev V. Automatic annotation of microbial genomes and metagenomic sequences. In: Li RW (ed.), *Metagenomics and Its Applications in Agriculture, Biomedicine and Environmental Studies*. Hauppauge, NY: Nova Science Publishers, 2011, 61–78.

37. Xiao X, Hu Z, Luo Z *et al.* iPSI(2L)-EDL: a two-layer predictor for identifying promoters and their types based on ensemble deep learning. *Curr Bioinforma* 2023;**19**:327–40. https://doi.org/10.2174/0115748936264316230926073231

38. Zhai W, Duan Y, Zhang X *et al.* Sequence and thermodynamic characteristics of terminators revealed by FlowSeq and the discrimination of terminators strength. *Synth Syst Biotechnol* 2022;**7**:1046–55. https://doi.org/10.1016/j.synbio.2022.06.003

39. Zhang H, Li J, Hu F *et al.* AMter: An end-to-end model for transcriptional terminators prediction by extracting semantic feature automatically based on attention mechanism. *Concurr Comput Pract Exp* 2024;**36**:e8056. https://doi.org/10.1002/cpe.8056

40. Zhang M, Jia C, Li F *et al.* Critical assessment of computational tools for prokaryotic and eukaryotic promoter prediction. *Brief Bioinform* 2022;**23**:bbab551. https://doi.org/10.1093/bib/bbab551

41. Roots C, Lukasiewicz A, Barrick JE. OSTIR: open source translation initiation rate prediction. *J Open Source Softw* 2021;**6**:3362–3362. https://doi.org/10.21105/joss.03362

42. Chen Y-J, Liu P, Nielsen AAK *et al.* Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat Methods* 2013;**10**:659–64. https://doi.org/10.1038/nmeth.2515

43. Tarnowski MJ, Gorochowski TE. Massively parallel characterization of engineered transcript isoforms using direct RNA sequencing. *Nat Commun* 2022;**13**:434. https://doi.org/10.1038/s41467-022-28074-5

44. Bokeh Development Team. Bokeh: Python library for interactive visualization. 2018. http://www.bokeh.pydata.org (12 August 2024, date last accessed).

45. Sayers EW, Bolton EE, Brister JR *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2021;**50**:D20–D26. https://doi.org/10.1093/nar/gkab1112

46. McGuffie MJ, Barrick JE. pLannotate: engineered plasmid annotation. *Nucleic Acids Res* 2021;**49**:W516–W522. https://doi.org/10.1093/nar/gkab374

47. Li D, Aaskov J, Lott WB. Identification of a cryptic prokaryotic promoter within the cDNA encoding the 5′ end of dengue virus RNA genome. *PLOS ONE* 2011;**6**:e18197. https://doi.org/10.1371/journal.pone.0018197

48. Pu S-Y, Wu R-H, Tsai M-H *et al*. A novel approach to propagate flavivirus infectious cDNA clones in bacteria by introducing tandem repeat sequences upstream of virus genome. *J Gen Virol* 2014;**95**:1493–503. https://doi.org/10.1099/vir.0.064915-0

49. Usme-Ciro JA, Lopera JA, Enjuanes L *et al*. Development of a novel DNA-launched dengue virus type 2 infectious clone assembled in a bacterial artificial chromosome. *Virus Res* 2014;**180**:12–22. https://doi.org/10.1016/j.virusres.2013.12.001

50. Whitaker WR, Lee H, Arkin AP *et al*. Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. *ACS Synth Biol* 2015;**4**:249–57. https://doi.org/10.1021/sb500003x

51. Weinstock MT, Hesek ED, Wilson CM *et al*. *Vibrio natriegens* as a fast-growing host for molecular biology. *Nat Methods* 2016;**13**:849–51. https://doi.org/10.1038/nmeth.3970

52. Martínez-García E, de Lorenzo V. *Pseudomonas putida* as a synthetic biology chassis and a metabolic engineering platform. *Curr Opin Biotechnol* 2024;**85**:103025. https://doi.org/10.1016/j.copbio.2023.103025

53. Wei W, Hennig BP, Wang J *et al*. Chromatin-sensitive cryptic promoters putatively drive expression of alternative protein isoforms in yeast. *Genome Res* 2019;**29**:1974–84. https://doi.org/10.1101/gr.243378.118

54. Stephenson A, Lastra L, Nguyen B *et al*. Physical laboratory automation in synthetic biology. *ACS Synth Biol* 2023;**12**:3156–69. https://doi.org/10.1021/acssynbio.3c00345

55. Zhang S, Fan R, Liu Y *et al*. Applications of transformer-based language models in bioinformatics: a survey. *Bioinforma Adv* 2023;**3**:vbad001. https://doi.org/10.1093/bioadv/vbad001

56. Lou C, Stanton B, Chen Y-J *et al*. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat Biotechnol* 2012;**30**:1137–42. https://doi.org/10.1038/nbt.2401

57. Agapakis CM, Ducat DC, Boyle PM *et al*. Insulation of a synthetic hydrogen metabolism circuit in bacteria. *J Biol Eng* 2010;**4**:3. https://doi.org/10.1186/1754-1611-4-3

58. Jack BR, Leonard SP, Mishler DM *et al*. Predicting the genetic stability of engineered DNA sequences with the EFM Calculator. *ACS Synth Biol* 2015;**4**:939–43. https://doi.org/10.1021/acssynbio.5b00068

59. Menuhin-Gruman I, Arbel M, Amitay N *et al*. Evolutionary Stability Optimizer (ESO): a novel approach to identify and avoid mutational hotspots in DNA sequences while maintaining high expression levels. *ACS Synth Biol* 2022;**11**:1142–51. https://doi.org/10.1021/acssynbio.1c00426

60. Itzkovitz S, Hodis E, Segal E. Overlapping codes within protein-coding sequences. *Genome Res* 2010;**20**:1582–89. https://doi.org/10.1101/gr.105072.110

61. Yang C, Hockenberry AJ, Jewett MC *et al*. Depletion of Shine-Dalgarno sequences within bacterial coding regions is expression dependent. *G3* 2016;**6**:3467–74. https://doi.org/10.1534/g3.116.032227

Synthetic Biology, 2024, 9(1), ysae018 , DOI: https://doi.org/10.1093/synbio/ysae018, Advance Access Publication 2 December 2024, Software

Submitted: 29 November 2024; **Received (in revised form)**: 13 November 2024; **Accepted**: 29 November 2024