The Wild Worm Codon Adapter: a web tool for automated codon adaptation of transgenes for expression in non-*Caenorhabditis* nematodes

Astra S. Bryant [ID] [1] and Elissa A. Hallem [ID] [1,2,*]

[1]Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA
[2]Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA

*Corresponding author: University of California, Los Angeles, MIMG, 237 BSRB, 615 Charles E. Young Dr. S., Los Angeles, CA 90095, USA. Email: ehallem@ucla.edu

Abstract

Advances in genomics techniques are expanding the range of nematode species that are amenable to transgenesis. Due to divergent codon usage biases across species, codon optimization is often a critical step for the successful expression of exogenous transgenes in nematodes. Platforms for generating DNA sequences codon-optimized for the free-living model nematode *Caenorhabditis elegans* are broadly available. However, until now such tools did not exist for non-*Caenorhabditis* nematodes. We therefore developed the Wild Worm Codon Adapter, a tool for rapid transgene codon optimization for expression in non-*Caenorhabditis* nematodes. The app includes built-in optimization for parasitic nematodes in the *Strongyloides*, *Nippostrongylus*, and *Brugia* genera as well as the predatory nematode *Pristionchus pacificus*. The app also supports custom optimization for any species using user-provided optimization rules. In addition, the app supports automated insertion of synthetic or native introns, as well as the analysis of codon bias in transgene and native sequences. Here, we describe this web-based tool and demonstrate how it may be used to analyze genome-wide codon bias in *Strongyloides* species.

Keywords: *Strongyloides*; *Brugia*; *Pristionchus*; *Nippostrongylus*; nematodes; codon optimization; introns; transgenesis

Introduction

Parasitic nematodes, including soil-transmitted gastrointestinal parasites in the genus *Strongyloides* and filarial nematodes such as *Brugia malayi*, are a major source of disease and economic burden (Lustigman *et al.* 2012). While *Caenorhabditis elegans* is often used as a model system for the study of parasitic nematodes, parasitic nematodes are behaviorally and genetically divergent from *C. elegans*; for example, they engage in a number of parasite-specific behaviors such as host seeking, host invasion, and intra-host migration (Haas 2003; Gang and Hallem 2016). Establishing methods that allow researchers to genetically manipulate parasitic nematodes directly is critical for understanding the genetic and cellular basis of parasitism in these species.

Historically, the application of functional genomics techniques to parasitic nematodes has lagged behind their use in the free-living model nematode *C. elegans*, due in part to the limited availability of parasite genomic information (Castelletto *et al.* 2020). High-quality reference genomes for many parasitic nematode species are now available (Hunt *et al.* 2016; Howe *et al.* 2017; International Helminth Genomes Consortium 2019), and are a critical resource for parasitic nematode functional genomics techniques such as transgenesis and CRISPR/Cas9-mediated mutagenesis (Lok *et al.* 2017; Castelletto *et al.* 2020; Liu *et al.* 2020). Transgenesis in parasitic nematodes is an essential tool for mechanistic studies of parasite development and behavior

(Bryant *et al.* 2018; Gang *et al.* 2020). As our technical understanding of nematode genomics continues to develop beyond *C. elegans*, establishing accessible tools that automate the transgene design process for a broad selection of nematode species will greatly facilitate the application of genomics techniques in these species.

Transgenesis protocols are increasingly well-established in non-*Caenorhabditis* nematode species, including three soil-transmitted gastrointestinal parasites in the *Strongyloididae* family—the human parasite *Strongyloides stercoralis*, the rodent parasite *Strongyloides ratti*, and the Australian brushtail possum parasite *Parastrongyloides trichosuri*—as well as the rodent gastrointestinal parasite *Nippostrongylus brasiliensis*, the human-parasitic filarial nematode *B. malayi*, the predatory nematode *Pristionchus pacificus*, and the free-living nematodes *Auanema rhodensis* and *Auanema freiburgensis* (Grant *et al.* 2006; Lok *et al.* 2017; Adams *et al.* 2019; Castelletto *et al.* 2021; Han *et al.* 2020). Intragonadal microinjection, in which exogenous DNA is injected directly into the gonad, has been used to generate progeny expressing a range of transgenes (Schlager *et al.* 2009; Lok *et al.* 2017; Adams *et al.* 2019; Hong *et al.* 2019; Carstensen *et al.* 2021; Castelletto *et al.* 2020; Han *et al.* 2020). Intragonadal microinjection has also been used to achieve CRISPR/Cas9-mediated mutagenesis in *S. stercoralis*, *S. ratti*, *P. pacificus*, and *Auanema* species (Witte *et al.* 2015; Gang *et al.* 2017; Lok *et al.* 2017; Bryant *et al.* 2018; Adams *et al.* 2019; Han *et al.*

2020). In *B. malayi*, transfection of infective larvae has been used to deliver reporter plasmids and CRISPR constructs (Liu *et al.* 2018, 2020). Most recently, lentiviral transduction of infective larvae was used to deliver RNA interference molecules and drive expression of fluorescent reporters in *N. brasiliensis* (Hagen *et al.* 2021).

In non-*Caenorhabditis* nematodes and other species, successful transgene expression often requires the use of species-specific preferred codons. Codon usage bias is pervasive in species from all taxa, including nematodes: most amino acids may be encoded by multiple synonymous codons, and individual species tend to favor a specific set of codons, particularly for encoding highly expressed genes (Sharp and Li 1987; Cutter *et al.* 2006; Mitreva *et al.* 2006). The use of preferred codons is thought to promote efficient translation, although the exact mechanisms are not clear (Plotkin and Kudla 2011). Codon usage bias is believed to regulate the expression of exogenous transgenes as well as endogenous genes (Redemann *et al.* 2011). In non-*Caenorhabditis* nematodes, expression of exogeneous genes from transgenes is promoted by the use of species-specific codon usage patterns (Han *et al.* 2020; Hagen *et al.* 2021).

Although the process of codon-adapting transgenes for *C. elegans* is simplified by web-based platforms (Grote *et al.* 2005; Redemann *et al.* 2011), transgenes codon-adapted for other nematode species were previously designed by hand. Therefore, we created a web-based application, the Wild Worm Codon Adapter, that automates the process of codon optimization for transgene expression in non-*Caenorhabditis* nematode species. Furthermore, the application permits automated insertion of synthetic or native introns into codon-optimized cDNA sequences, inclusion of which can significantly increase gene expression (Junio *et al.* 2008; Li *et al.* 2011; Crane *et al.* 2019; Han *et al.* 2020). Finally, the app enables users to rapidly assess relative codon bias of transgene sequences and native genes using genus-specific codon adaptation indices.

Materials and methods
Data source and preferred codon selection
Codon usage rules for *Strongyloides*, *Brugia*, *Pristionchus*, and *C. elegans* were calculated based on previously published codon count and frequency data (Supplementary File S1). For *Strongyloides*, codon count data are from 11,458 codons from the 50 most abundant *S. ratti* expressed sequence tag (EST) sequences (Mitreva *et al.* 2006). For *C. elegans*, codon count data are from 178 genes (73,164 codons) with the highest bias toward translationally optimal codons (Sharp and Bradnam 1997). For *Brugia* and *Pristionchus*, codon frequency data are from the ∼10% of genes (bin 7) with the highest expression (Han *et al.* 2020). For *Nippostrongylus*, RNA-seq expression data across three *N. brasiliensis* life stages [infective 3rd-stage larvae (iL3s), activated iL3s, red-blood-cell-feeding iL3s] was retrieved from WormBase ParaSite (Eccles *et al.* 2018). The 10% of genes with the highest median expression across all life stages were identified (2279 genes). Coding sequences for these highly expressed genes were retrieved from WormBase ParaSite with the biomaRt package v2.42.1 (Durinck *et al.* 2005, 2009), and total codon usage counts (Supplementary File S1) were calculated using the uco function in the seqinr package v3.6-1.

Count and frequency data were used to quantify the relative adaptiveness of individual codons: the frequency that codon "i" encodes amino acid "AA" ÷ the frequency of the codon most often used for encoding amino acid "AA" (Sharp and Li 1987; Jansen

et al. 2003). Preferred codons were defined as the codons with the highest relative adaptiveness value for each amino acid (Supplementary File S1).

Codon adaptation index
The codon biases of individual sequences are quantified by calculating a Codon Adaptation Index (CAI), defined as the geometric average of relative adaptiveness of all codons in the sequence (Sharp and Li 1987; Jansen *et al.* 2003). CAI values relative to species-specific relative adaptiveness values are calculated using the seqinr package v3.6-1.

Fractional GC content
The fraction of $G + C$ bases in a sequence is calculated using the seqinr package v3.6-1.

Intron insertion
Intron sequences are either: the three canonical artificial intron sequences used for *C. elegans* (Fire *et al.* 1995), *P. pacificus* native introns (Han *et al.* 2020), Periodic A_n/T_n Cluster (PATC)-rich introns from the *C. elegans* gene *smu-2* (introns 3–5) (Aljohani *et al.* 2020), or custom introns provided by users via FASTA file upload. Built-in intron sequences are flanked by canonical 5′-GT...AG-3′ splice recognition sequences (Shapiro and Senapathy 1987; Blumenthal and Steward, 1997; Wheeler *et al.* 2020). For intron placement, the optimized cDNA sequence is divided at three predicted intron insertion sites spaced approximately equidistantly. Users may choose to insert introns at the equidistant sites, or may further refine insertion site locations by identifying the closest conserved invertebrate exon splice sites (5′-AG^G-3′, 5′-AG^A-3′; the "^" symbol indicates the exact insertion site) (Shapiro and Senapathy 1987; Blumenthal and Steward, 1997). The user-specified number of introns (up to a maximum of three) are inserted into the sequence using the 5′ insertion site first and continuing in the 3′ direction (Crane *et al.* 2019; Han *et al.* 2020). When inserting introns using conserved invertebrate exon splice sites, if there are fewer insertion sites than the user-requested number of introns, the program will insert as many introns as there are available insertion sites.

Coding sequence lookup
In "Analyze Sequences" mode, users may submit search terms including stable gene IDs, *C. elegans* gene names, or matched gene IDs and cDNA sequences as either a two-column CSV file or a FASTA file. If users supply only gene IDs, either via a text box or file upload, the app first fetches the coding sequences from WormBase ParaSite via the biomaRt package v2.42.1 (Durinck *et al.* 2005, 2009). The following types of gene IDs may be used: stable gene or transcript IDs with prefixes "SSTP," "SRAE," "SPAL," "SVE," "Ppa," "Bma," "NBR," or "WB"; *C. elegans* stable transcript IDs; or *C. elegans* gene names prefaced with the string "Ce-" (*e.g.*, *Ce-tax-4*).

Genome-wide codon bias and gene ontology (GO) analysis
FASTA files containing all coding sequences (CDS) for *S. stercoralis*, *S. ratti*, *S. papillosus*, *S. venezuelensis*, *N. brasiliensis*, *B. malayi*, *P. pacificus*, and *C. elegans* were downloaded from WormBase ParaSite (WBPS15) and analyzed using the app. For *C. elegans* and *Strongyloides* species, results were filtered to identify six functional subsets: the 2% of genes with highest and lowest *S. ratti* CAI values, the 2% of genes with the highest and lowest *C. elegans* CAI values, and the 2% of genes with highest and lowest RNA-seq

expression in free-living females. Log$_2$ counts per million (CPM) expression in free-living adult females was downloaded from the *Strongyloides* RNA-seq Browser (Bryant *et al.* 2021). For statistical comparisons of the expression of the highest and lowest *Strongyloides*-codon-adapted genes, relative to all genes, a 2-way ANOVA (type III) with Tukey *post-hoc* tests was performed in R using the car package v3.0-8. GO analyses of functional subsets were performed using the gprofiler2 package v0.1.9, with a false discovery rate (FDR)-corrected *p*-value of <0.05. Commonly enriched GO terms in each subset were defined as GO terms that were enriched in all four of the *Strongyloides* species (or at least three *Strongyloides* species in the case of gene-expression-based subsets) with an FDR-corrected *p*-value of ≤0.001. For statistical comparisons of genome-wide CAI values and fractional GC content between species, Kruskal-Wallis tests with *post-hoc* Dunn's tests were performed in R using the dunn.test package v1.3.5. For *post-hoc* Dunn's tests, *p*-values were corrected using the Bonferroni method.

Data availability

Preprocessing and analysis source code, plus codon frequency data, intron sequences, and supplementary files are available at: https://github.com/HallemLab/Bryant-and-Hallem-2021.

The following supplementary files have been uploaded to fig-share: https://doi.org/10.25387/g3.14462481. Supplementary File S1 contains codon usage frequencies and optimal codons for: highly abundant *S. ratti* EST transcripts (Mitreva *et al* 2006); highly expressed *C. elegans, P. pacificus, B. malayi*, and *N. brasiliensis* genes (Sharp and Bradnam 1997; Eccles *et al.* 2018; Han *et al* 2020); and all *S. ratti* ESTs (Mitreva *et al* 2006). Supplementary File S2 contains a code freeze for the Wild Worm Codon Adapter. Supplementary File S3 contains gene IDs, CAI values, GC ratios, and GO term accession numbers of the 2% of genes with the highest and lowest CAI values for each *Strongyloides* species and *C. elegans*. Supplementary File S4 contains GO analysis results for the 2% of genes with the highest and lowest CAI values for each *Strongyloides* species and *C. elegans*. Supplementary File S5 contains GO terms significantly enriched in all four *Strongyloides* species for the highest (top 2%) and lowest (bottom 2%) *Strongyloides* codon-adapted sequences, as well as GO terms significantly enriched in at least three *Strongyloides* species for genes with the highest (top 2%) and lowest (bottom 2%) expression in free-living females.

Results and discussion
Software functionality

The Wild Worm Codon Adapter app (https://hallemlab.shinyapps.io/Wild_Worm_Codon_Adapter/) features two usage modes: "Optimize Sequences" mode and "Analyze Sequences" mode (Figures 1 and 2, Supplementary File S2). "Optimize Sequences" mode automates the process of transgene codon adaptation and intron insertion. For codon optimization, the app features built-in preferred codons for four non-*Caenorhabditis* nematode genera for which transgenesis protocols are increasingly well-established: *Strongyloides, Nippostrongylus, Pristionchus*, and *Brugia*. The app also includes the option to codon-optimize transgenes for expression in *C. elegans* (*e.g.*, for researchers wishing to express parasite sequences heterologously in *C. elegans*), similar to established platforms (Grote *et al.* 2005; Redemann *et al.* 2011). This mode also supports custom codon optimization based on a user-provided list of preferred codons.

To generate codon-optimized sequences suitable for expression in *Strongyloides, Nippostrongylus, Brugia, Pristionchus*, or *C. elegans*, users select the appropriate codon usage rule and then submit cDNA or amino acid sequences via a text window or file upload. To codon-optimize transgenes for expression in any additional organism of interest, users may prefer to supply a custom list of preferred codons. For example, users wishing to codon-optimize parasite sequences for heterologous expression in mammalian cell culture would upload a list of mammalian optimal codons. The uploaded custom list of preferred codons must use the following format: a 2-column CSV file listing single-letter amino acid symbols and corresponding 3-letter optimal codon sequence; only one optimal codon should be provided per amino acid and stop codons should be designated using the "*" symbol. An example custom preferred codon table is available to download from the website.

For *Strongyloides* species, codon optimization is based on codon usage patterns in *S. ratti* (Mitreva *et al.* 2006). Codon optimizations for *Brugia, Pristionchus*, and *Nippostrongylus* species are based on codon bias in *B. malayi, P. pacificus*, and *N. brasiliensis*, respectively (Eccles *et al.* 2018; Han *et al.* 2020). Codon usage is often well-conserved between closely related nematode species (Mitreva *et al.* 2006; Han *et al.* 2020). Thus, codon usage rules generated from individual species are likely effective across closely related species (*e.g.*, across members of a genus).

We calculated built-in codon usage rules from the codon usage patterns observed in highly expressed genes (Sharp and Bradnam 1997; Mitreva *et al.* 2006; Eccles *et al.* 2018; Han *et al.* 2020), since the codon usage patterns of highly expressed genes are thought to correlate with higher protein expression (Sharp and Li 1987; Plotkin and Kudla 2011). However, the codon usage rules for highly expressed *S. ratti* genes are extremely similar to those observed across all *S. ratti* coding sequences (Supplementary File S1). In contrast, the preferred codon usage rules we implement for *Strongyloides, Nippostrongylus, Pristionchus*, and *Brugia* are distinct from the *C. elegans* codon usage rules (Supplementary File S1), consistent with observations that individual transgenes show limited expression across nematode species with divergent genomes (Hunt *et al.* 2016; Lok *et al.* 2017; Castelletto *et al.* 2020; Han *et al.* 2020).

Previous studies have suggested that highly divergent codon usage patterns between nematodes are driven in part by the extreme AT-richness of some nematode genomes (Cutter *et al.* 2006; Mitreva *et al.* 2006; Hunt *et al.* 2016; Han *et al.* 2020). As expected, in *S. ratti* and *B. malayi*, which have AT-rich genomes, the preferred codons that comprise the built-in usage rules are AT-biased (preferred codon fractional GC content: *S. ratti* = 0.33, *B. malayi* = 0.32, *N. brasiliensis* = 0.54, *P. pacificus* = 0.55, *C. elegans* = 0.51). Thus, codon optimization for expression in *Strongyloides* and *Brugia* species will likely yield AT-rich optimized sequences. Given the use of a single codon sequence per amino acid and the AT-bias in preferred codons for some species, users may want to eliminate repeated nucleotide patterns, hairpin loops, or unwanted restriction sites prior to final gene synthesis.

To insert introns into optimized cDNA sequences, users select the type and number of introns for insertion into the optimized sequence (up to three). Users may choose between three sets of built-in unique intron sequences: canonical *C. elegans* artificial introns (Fire *et al.* 1995), *P. pacificus* native introns (Han *et al.* 2020), or Periodic A$_n$/T$_n$ Cluster (PATC)-rich introns that enhance germline expression of transgenes in *C. elegans* (Aljohani *et al.* 2020). Alternatively, users may upload a FASTA file containing a custom set of introns. The app identifies three putative intron insertion

User selects operation mode	
Optimize Sequences	Analyze Sequences

User inputs sequence and optimization parameters	
User Inputs	1. Use text window or file upload control to load a cDNA or amino acid sequence 2. Use dropdown menu to select a built-in codon usage rule *-or-* Use file upload control to load a custom list of preferred codons 3. Use dropdown menu to select an intron sequence source *-or-* Use file upload control to load a custom set of introns 4. Use dropdown menu to select the desired number of introns (0-3)
Content Displayed in Browser	1. Sequence information for original and optimized sequences: • fractional GC ratio • Codon Adaptation Index (CAI) value relative to selected codon usage rules 2. Original sequence 3. Optimized sequence • without introns • with desired number of introns
Downloadable Files	Optimized sequence (.txt file) • without introns • with desired number of introns

User inputs gene(s) of interest	
User Inputs	Use textbox to input sequence *-or-* Use second textbox to input gene/ transcript IDs, using the formats: • Stable IDs with the prefix: SSTP|SRAE|SVE|SPAL|NBR| Bma|Ppa|WB • *C. elegans* gene names with 'Ce-' prefix • *C. elegans* transcript IDs *-or-* Use the file upload control to load: • One-column .csv file listing geneIDs • Two-column .csv file listing names and coding sequences • .fasta file containing named coding sequences
Content Displayed in Browser	Sequence information, including: • GC content • Codon Adaptation Index (CAI) value relative to codon usage rules for the following species: - *C. elegans* - *S. ratti* - *N. brasiliensis* - *B. malayi* - *P. pacificus*
Downloadable Files	Sequence information (.xlsx file)

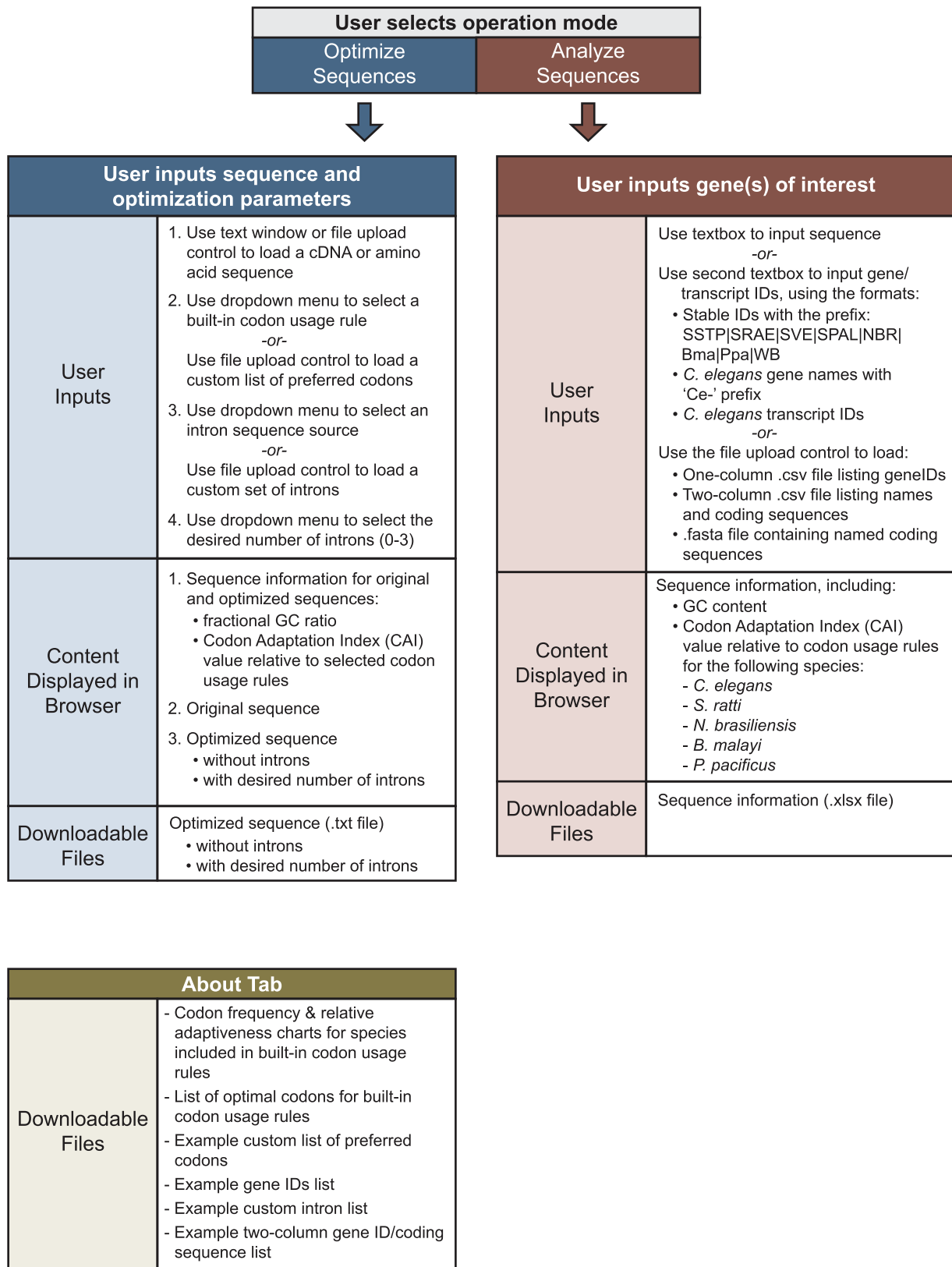| About Tab	
Downloadable Files	- Codon frequency & relative adaptiveness charts for species included in built-in codon usage rules - List of optimal codons for built-in codon usage rules - Example custom list of preferred codons - Example gene IDs list - Example custom intron list - Example two-column gene ID/coding sequence list

Figure 1 A UI overview of the Wild Worm Codon Adapter app. The overview shows user inputs, content displayed in the browser, and downloadable files. The application features two modes: Optimize Sequences (blue panels) and Analyze Sequences (red panels) that are accessed via separate tabs in the browser window. The app also includes an About tab that presents methods information and a menu for downloading files. Available files include example input files as well as tables of species-specific optimal codons, species-specific codon frequencies, and relative adaptiveness values used to calculate codon adaptation indices.
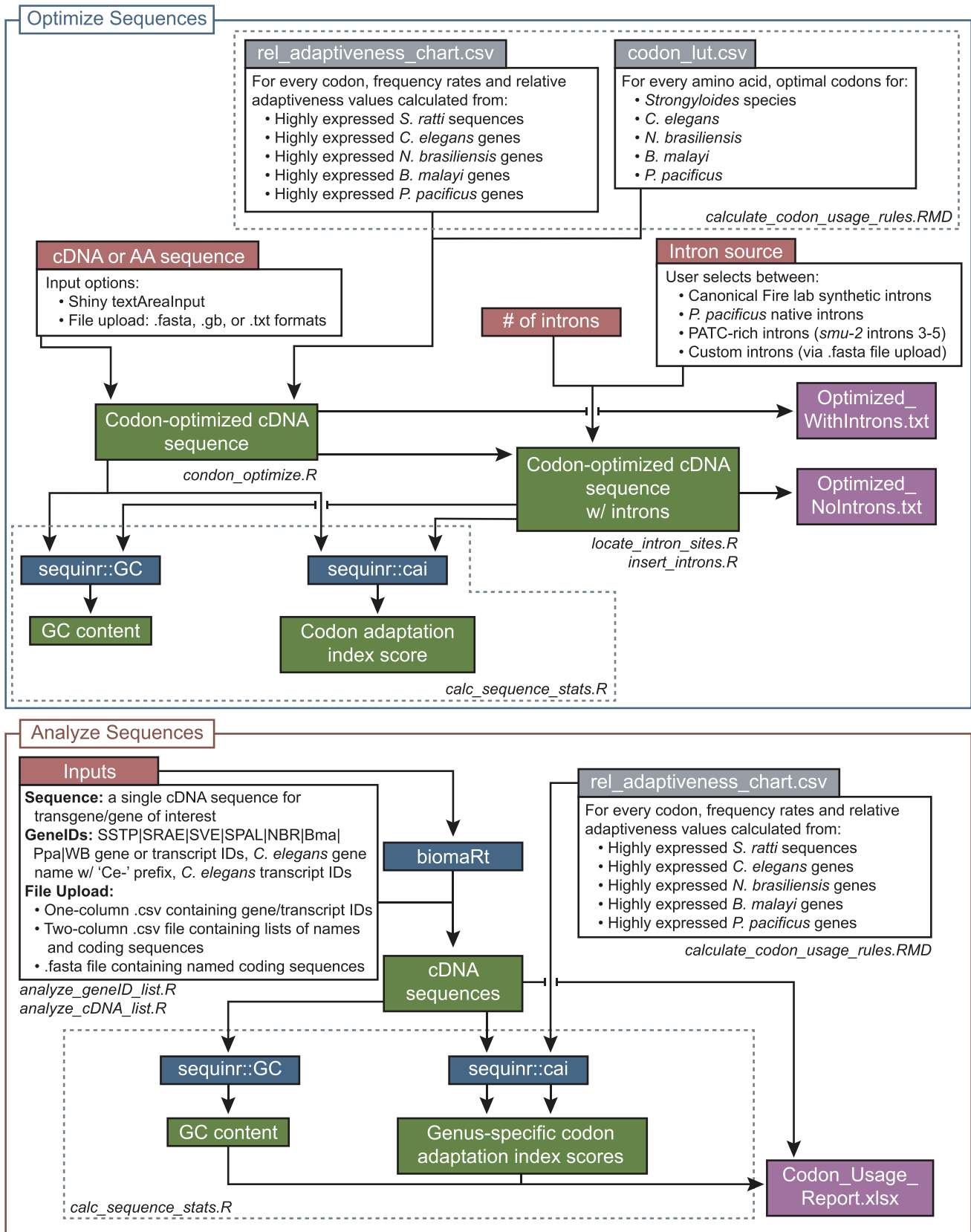
Optimize Sequences

rel_adaptiveness_chart.csv

For every codon, frequency rates and relative adaptiveness values calculated from:
• Highly expressed *S. ratti* sequences
• Highly expressed *C. elegans* genes
• Highly expressed *N. brasiliensis* genes
• Highly expressed *B. malayi* genes
• Highly expressed *P. pacificus* genes

codon_lut.csv

For every amino acid, optimal codons for:
• *Strongyloides* species
• *C. elegans*
• *N. brasiliensis*
• *B. malayi*
• *P. pacificus*

calculate_codon_usage_rules.RMD

cDNA or AA sequence

Input options:
• Shiny textAreaInput
• File upload: .fasta, .gb, or .txt formats

Intron source

User selects between:
• Canonical Fire lab synthetic introns
• *P. pacificus* native introns
• PATC-rich introns (*smu-2* introns 3-5)
• Custom introns (via .fasta file upload)

# of introns

Codon-optimized cDNA sequence

condon_optimize.R

Codon-optimized cDNA sequence w/ introns

locate_intron_sites.R
insert_introns.R

Optimized_WithIntrons.txt

Optimized_NoIntrons.txt

sequinr::GC

sequinr::cai

GC content

Codon adaptation index score

calc_sequence_stats.R

Analyze Sequences

Inputs

Sequence: a single cDNA sequence for transgene/gene of interest
GeneIDs: SSTP|SRAE|SVE|SPAL|NBR|Bma| Ppa|WB gene or transcript IDs, *C. elegans* gene name w/ 'Ce-' prefix, *C. elegans* transcript IDs
File Upload:
• One-column .csv containing gene/transcript IDs
• Two-column .csv file containing lists of names and coding sequences
• .fasta file containing named coding sequences

analyze_geneID_list.R
analyze_cDNA_list.R

biomaRt

cDNA sequences

rel_adaptiveness_chart.csv

For every codon, frequency rates and relative adaptiveness values calculated from:
• Highly expressed *S. ratti* sequences
• Highly expressed *C. elegans* genes
• Highly expressed *N. brasiliensis* genes
• Highly expressed *B. malayi* genes
• Highly expressed *P. pacificus* genes

calculate_codon_usage_rules.RMD

sequinr::GC

sequinr::cai

GC content

Genus-specific codon adaptation index scores

calc_sequence_stats.R

Codon_Usage_Report.xlsx

Figure 2 Detailed graphic view of the codebase of the Wild Worm Codon Adapter app. The application features two usage modes: Optimize Sequences (upper, blue panel) and Analyze Sequences (lower, red panel). Gray boxes are preprocessed data inputs generated from published data. Red boxes are user inputs. Green boxes are output elements displayed within the browser; output values directly depend on inputs provided via red elements. Blue boxes are commands run in R. White boxes are code details or input options. Purple boxes are downloadable output file options. Dashed lines show division of code elements into the named files.

sites spaced approximately equidistantly within the optimized cDNA sequence, and inserts the user-specified type and number of introns (Fire *et al.* 1995; Blumenthal and Steward, 1997; Redemann *et al.* 2011). In *C. elegans* and *P. pacificus*, a single 5′ intron is sufficient for intron-mediated enhancement of gene expression, whereas a single 3′ intron is not (Crane *et al.* 2019; Han *et al.* 2020). Thus, when the user chooses to insert fewer than three introns, the three hypothetical insertion sites are filled as needed, starting from the 5′ site.

The app displays the original (non-optimized) sequence and the optimized sequence with and without introns; users may download optimized sequences as plain text (.txt) files. "Optimize Sequences" mode reports the fractional GC content for the original and optimized sequences. When performing optimization using built-in codon usage rules, the app also provides a measure of the codon usage bias for both the original and optimized cDNA sequences by reporting CAI values relative to the selected usage rule.

Finally, the "Analyze Sequences" mode (Figure 1) was designed to support descriptive analyses of transgene sequences as well as native *Strongyloides*, *B. malayi*, *P. pacificus*, *N. brasiliensis*, or *C. elegans* coding sequences. To measure how well codon-adapted a transgene is for expression in *Strongyloides*, *Brugia*, *Nippostrongylus*, *Pristionchus*, or *C. elegans*, users may submit the relevant coding sequence(s). For individual sequences, the following values are calculated and displayed as a downloadable table: fractional GC content; and CAI values relative to *Strongyloides*, *Brugia*, *Nippostrongylus*, *Pristionchus*, and *C. elegans* codon usage rules (Sr-CAI, Bm-CAI, Nb-CAI, Pp-CAI, and Ce-CAI, respectively). When presented with a list of gene or transcript IDs, the app first retrieves associated coding sequences from WormBase ParaSite before calculating the quantifications listed above.

Benchmarking and example usage

To benchmark the use of the CAI to quantify codon adaptiveness of native genes, we used "Analyze Sequences" mode to assess and compare genome-wide codon bias patterns in *Strongyloides* species, *B. malayi*, *N. brasiliensis*, *P. pacificus*, and *C. elegans*. Consistent with previous findings (Mitreva *et al.* 2006), we found that the distribution of genome-wide codon bias varied by genus, such that the genus-specific CAI values of *C. elegans* coding sequences were lower than those for parasitic species (Figure 3A; $p < 0.0001$ for all comparisons to *C. elegans*, Kruskal-Wallis test with Dunn's *post-hoc* tests). Also consistent with previous observations that *Strongyloides* genomes are highly AT-rich (Mitreva *et al.* 2006; Cutter *et al.* 2006; Hunt *et al.* 2016), we observed that *Strongyloides* coding sequences displayed lower fractional GC content than *C. elegans*, *N. brasiliensis*, *P. pacificus*, and *B. malayi* coding sequences (Figure 3B; $p < 0.0001$ for *Strongyloides* species
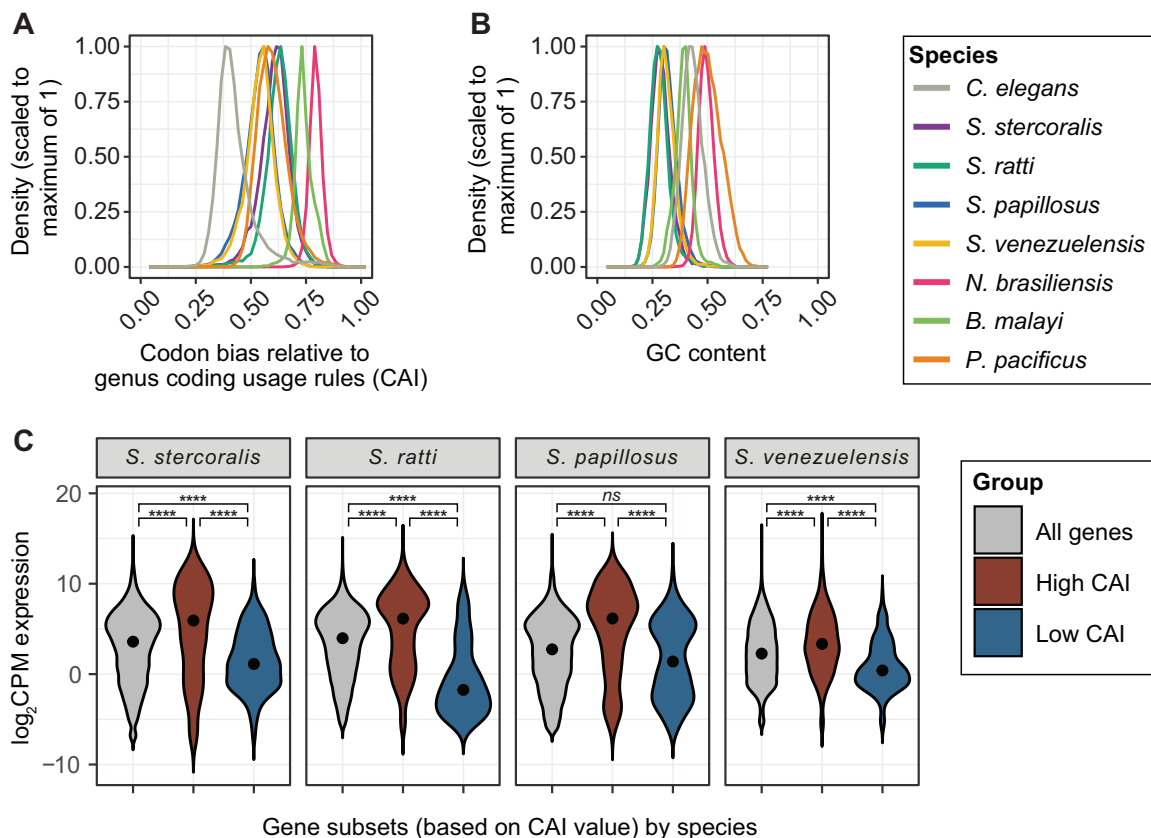

Figure 3 Codon adaptiveness and GC content across species. (A) Density plot of CAI values in a species' genome. Codon adaptiveness varies across species; *C. elegans* genome-wide codon adaptiveness is significantly lower than other species ($p < 0.0001$ comparing *C. elegans* to each remaining species, Kruskal-Wallis test with Dunn's *post-hoc* tests). For each species, codon bias is calculated relative to the genus-level codon usage rules (Supplementary File S1). (B) Density plot of fractional GC content across species. Consistent with previous reports (Cutter *et al.* 2006; Mitreva *et al.* 2006; Hunt *et al.* 2016), *Strongyloides* coding sequences tend to be more AT-rich than the coding sequences of other species ($p < 0.0001$ for the grouped *Strongyloides* species versus each remaining species, Kruskal-Wallis test with Dunn's *post-hoc* tests). For both panels, values were calculated by submitting a list of all predicted coding sequences to the Wild Worm Codon Adapter app in Analyze Sequences mode. (C) Violin plot of $\log_2$ counts per million (CPM) expression in free-living females for all genes (gray), genes in the top 2% of CAI values for each species (red), and genes in the bottom 2% of CAI values for each species (blue). Dots indicate median values. ****$p < 0.00005$, 2-way ANOVA with Tukey *post-hoc* tests. ns = not significant ($p > 0.05$).

A GO Analysis: 2% most *Strongyloides*-codon-adapted sequences
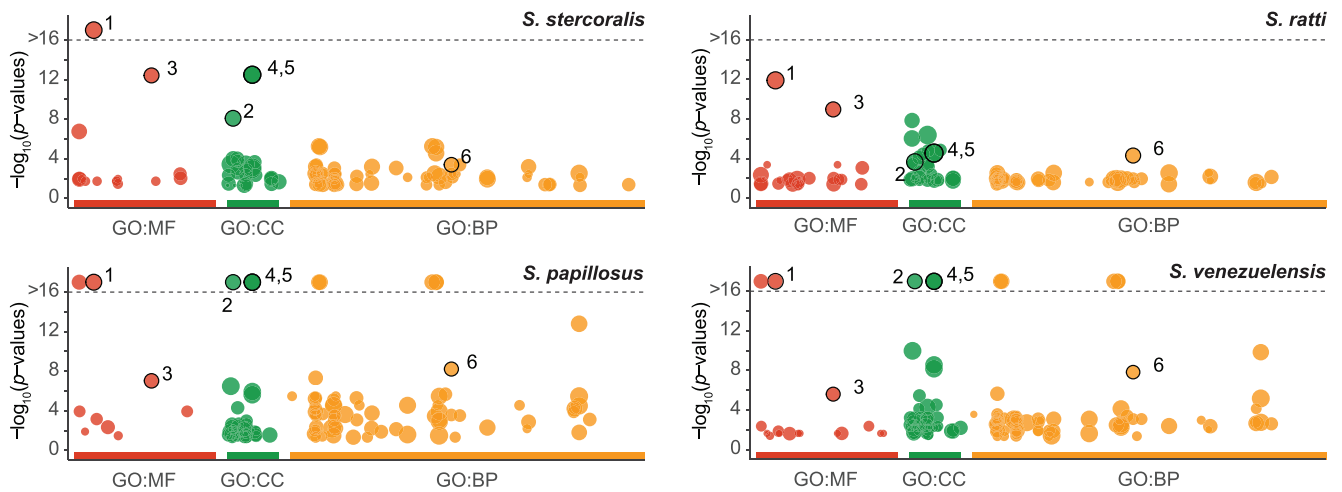
B GO terms significantly enriched in all *Strongyloides* species (*p*<0.001)

ID	Source	Term ID	Term Name
1	GO:MF	GO:0005198	structural molecule activity
2	GO:CC	GO:0005840	ribosome
3	GO:MF	GO:0042302	structural constituent of cuticle
4	GO:CC	GO:0043228	non−membrane−bounded organelle
5	GO:CC	GO:0043232	intracellular non−membrane−bounded organelle
6	GO:BP	GO:0046034	ATP metabolic process

C GO Analysis: 2% least *Strongyloides*-codon-adapted sequences
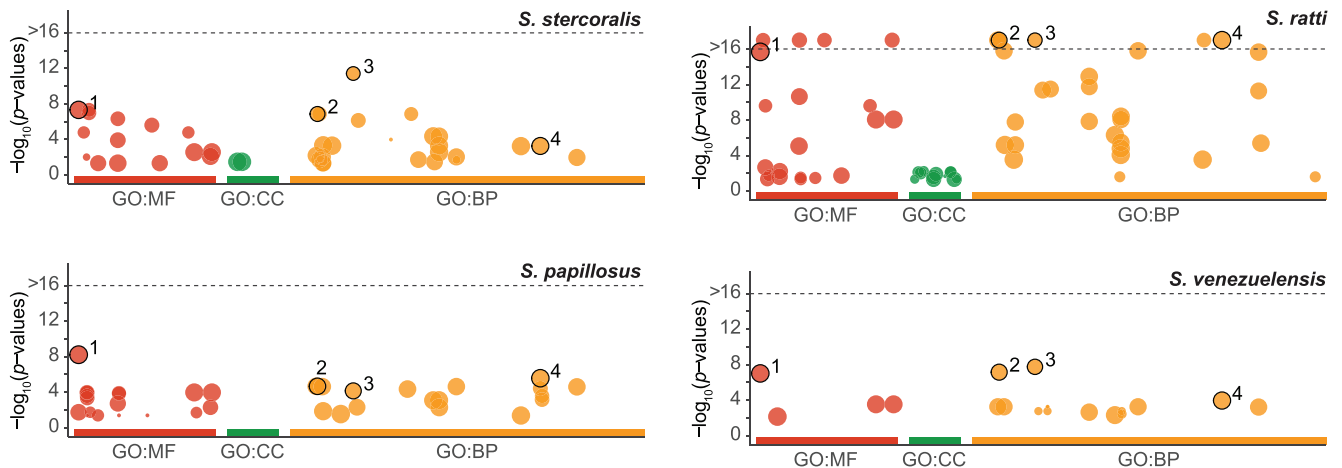
D GO terms significantly enriched in all *Strongyloides* species (*p*<0.001)

ID	Source	Term ID	Term Name
1	GO:MF	GO:0003676	nucleic acid binding
2	GO:BP	GO:0006259	DNA metabolic process
3	GO:BP	GO:0015074	DNA integration
4	GO:BP	GO:0090304	nucleic acid metabolic process

Figure 4 Functional enrichment of highly and poorly codon-adapted genes. GO enrichment across *Strongyloides* species for sequences displaying high and low degrees of *Strongyloides* codon bias. (A, B) Manhattan plots and table showing GO enrichment for sequences in the top 2% of CAI values for each species. *X*-axis labels (A) and table Source values (B) indicate the three GO vocabulary categories: biological process (BP), molecular function (MP), and cellular component (CC). Numbered circles (A) and Term ID numbers (B) indicate GO terms that are significantly enriched ($p \leq 0.001$) in all four *Strongyloides* species. For Manhattan plots, $-\log_{10}(p\text{-values})$ along the y-axis are capped if ≥ 16; this threshold is indicated by a gray dashed line. All GO terms included in (B) are also significantly enriched in the top 2% of *C. elegans*-codon-adapted *C. elegans* sequences. Red terms are also significantly enriched among genes with the highest expression (top 2%) in free-living females, in at least 3 *Strongyloides* species. (C, D) Manhattan plots and table of GO enrichment for sequences that compose the bottom 2% of CAI values. None of the GO terms included in (D) are also significantly enriched ($p \leq 0.001$) in poorly *C. elegans*-codon-adapted *C. elegans* sequences or in genes with the lowest expression (bottom 2%) in free-living females, in at least three *Strongyloides* species. For all plots, *p*-values are FDR-corrected.

grouped together and compared to each remaining species, Kruskal-Wallis test with Dunn's *post-hoc* tests). Together these observations emphasize the likely benefit of transgene codon-optimization for promoting successful transgenesis in non-*Caenorhabditis* species. Finally, consistent with our use of *Strongyloides* codon usage rules based on highly abundant *S. ratti* ESTs, the highest and lowest *Strongyloides*-codon-adapted sequences generally displayed significantly higher and lower expression, respectively, in free-living female life stages relative to the expression of all genes in the genome (Figure 3C). However, the range of gene expression values between the highest and lowest codon-adapted sequences are largely overlapping; thus, the degree of codon adaptation exhibited by individual genes is not the sole determinant of mRNA expression level in these species.

Next, we used GO analyses to assess the putative functions of the most highly or poorly *Strongyloides*-codon-adapted coding sequences (defined as the top/bottom 2% of *Sr*-CAI values; Supplementary Files S3 and S4). For the highest *Strongyloides*-codon-adapted sequences, GO terms that are significantly enriched in all four *Strongyloides* species are primarily associated with structural integrity and ribosomal components (Figure 4A-B, Supplementary File S5). As expected, given that the *Strongyloides* codon usage rules are based on highly abundant sequences, there is significant overlap between GO terms enriched in the top 2% of *Strongyloides*-codon-adapted sequences and GO terms enriched in the genes that are most highly expressed in free-living females (Figure 4B, Supplementary File S5). Despite the differences between codon usage patterns in *Strongyloides* and *C. elegans*, GO terms enriched in the top 2% of *Strongyloides*-codon-adapted sequences are also significantly enriched in the top 2% of *C. elegans*-codon-adapted sequences (*i.e.*, *C. elegans* sequences that display high codon usage bias relative to *C. elegans* codon usage rules) (Figure 4B). In contrast, for parasite sequences that are the least *Strongyloides*-codon-adapted, commonly enriched GO terms are associated with DNA metabolism and interactions (Figure 4C-D, Supplementary File S5); these terms are not enriched in low-expressing *Strongyloides* genes or poorly *C. elegans*-codon-adapted *C. elegans* sequences (Figure 4D, Supplementary File S5). Although the causes and consequences of codon bias are not fully understood (Sharp *et al.* 2010; Plotkin and Kudla 2011), our observations of a common set of GO terms enriched in both *Strongyloides*- and *C. elegans*-codon-adapted genes suggest that in nematodes, heighted codon bias reflects a nonrandom process that systematically imposes divergent codon usage patterns on genes associated with a common set of biological functions.

Conclusions and future directions

Codon optimization of transgenes significantly improves the likelihood of successful transgenesis in non-*Caenorhabditis* nematodes. Here, we present the Wild Worm Codon Adapter, a web-based tool designed to replicate and extend the functionality of popular codon-optimization tools to include non-*Caenorhabditis* nematode genera such as *Strongyloides*, *Pristionchus, Nippostrongylus*, and *Brugia* (Grote *et al.* 2005; Redemann *et al.* 2011). The open-source code is extendable; users may perform codon-optimization for an unlimited selection of species by providing a custom list of optimal codons, and additions to the list of built-in optimization rules can be implemented as requested. Going forward, we hope that the Wild Worm Codon Adapter will simplify the process of transgene design for researchers using functional genomics to study non-*Caenorhabditis* nematodes.

Web resources

A web-hosted version of the app is available at: https://hallem lab.shinyapps.io/Wild_Worm_Codon_Adapter/

App source code and deployment instructions are available at: https://github.com/HallemLab/Wild_Worm_Codon_Adapter

Conflicts of interest

None declared.

Literature cited

Adams S, Pathak P, Shao H, Lok JB, Pires-daSilva A. 2019. Liposome-based transfection enhances RNAi and CRISPR-mediated mutagenesis in non-model nematode systems. Sci Rep. 9:483.

Aljohani MD, Mouridi SE, Priyadarshini M, Vargas-Velazquez AM, Frøkjær-Jensen C. 2020. Engineering rules that minimize germ-line silencing of transgenes in simple extrachromosomal arrays in *C. elegans*. Nat Commun. 11:6300.

Blumenthal T, Steward K. 1997. RNA Processing and Gene Structure, in *C. elegans* II, edited by DL Riddle, T Blumenthal, BJ Meyer, and JR Priess. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Bryant AS, DeMarco SF, Hallem EA. 2021. *Strongyloides* RNA-Seq Browser: a web-based software platform for on-demand bioinformatics analyses of *Strongyloides* species. G3 (Bethesda). 6:jkab104.

Bryant AS, Ruiz F, Gang SS, Castelletto ML, Lopez JB, *et al.* 2018. A critical role for thermosensation in host seeking by skin-penetrating nematodes. Curr Biol. 28:2338–2347.

Carstensen HR, Villalon RM, , Banerjee N, Hallem EA, Hong, R. 2021. Steroid hormone pathways coordinate developmental diapause and olfactory remodeling in *Pristionchus pacificus*. Genetics. *In press*.

Castelletto ML, Gang SS, Hallem EA. 2020. Recent advances in functional genomics for parasitic nematodes of mammals. J Exp Biol. 223:jeb206482.

Crane MM, Sands B, Battaglia C, Johnson B, Yun S, *et al.* 2019. *In vivo* measurements reveal a single 5′-intron is sufficient to increase protein expression level in *Caenorhabditis elegans*. Sci Rep. 9:9192.

Cutter AD, Wasmuth JD, Blaxter ML. 2006. The evolution of biased codon and amino acid usage in nematode genomes. Mol Biol Evol. 23:2303–2315.

Durinck S, Moreau Y, Kasprzyk A, Davis S, Moor BD, *et al.* 2005. BioMart and Bioconductor: a powerful link between biological

databases and microarray data analysis. Bioinformatics. 21: 3439–3440.

Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 4:1184–1191.

Eccles D, Chandler J, Camberis M, Henrissat B, Koren S, *et al.* 2018. *De novo* assembly of the complex genome of *Nippostrongylus brasiliensis* using MinION long reads. BMC Biol. 16:6.

Fire A, Ahnn J, Seydoux G, Xu S-Q. 1995. Fire Lab 1995 Vector Kit Documentation. www.ciwemb.edu.

Gang SS, Castelletto ML, Bryant AS, Yang E, Mancuso N, *et al.* 2017. Targeted mutagenesis in a human-parasitic nematode. PLoS Pathog. 13:e1006675.

Gang SS, Castelletto ML, Yang E, Ruiz F, Brown TM, *et al.* 2020. Chemosensory mechanisms of host seeking and infectivity in skin-penetrating nematodes. Proc Natl Acad Sci USA. 117: 17913–17923.

Gang SS, Hallem EA. 2016. Mechanisms of host seeking by parasitic nematodes. Mol Biochem Parasitol. 208:23–32.

Grant WN, Skinner SJM, Newton-Howes J, Grant K, Shuttleworth G, *et al.* 2006. Heritable transgenesis of *Parastrongyloides trichosuri*: a nematode parasite of mammals. Int J Parasitol. 36:475–483.

Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, *et al.* 2005. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Res. 33:W526–W531.

Haas W. 2003. Parasitic worms: strategies of host finding, recognition and invasion. Zoology. 106:349–364.

Hagen J, Sarkies P, Selkirk ME. 2021. Lentiviral transduction facilitates RNA interference in the nematode parasite *Nippostrongylus brasiliensis*. PLoS Pathog. 17:e1009286.

Han Z, Lo W-S, Lightfoot JW, Witte H, Sun S, *et al.* 2020. Improving transgenesis efficiency and CRISPR-associated tools through codon optimization and native intron addition in *Pristionchus* nematodes. Genetics. 216:947–956.

Hong RL, Riebesell M, Bumbarger DJ, Cook SJ, Carstensen HR, *et al.* 2019. Evolution of neuronal anatomy and circuitry in two highly divergent nematode species. eLife. 8:e47155.

Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. 2017. WormBase ParaSite—a comprehensive resource for helminth genomics. Mol Biochem Parasitol. 215:2–10.

Hunt VL, Tsai IJ, Coghlan A, Reid AJ, Holroyd N, *et al.* 2016. The genomic basis of parasitism in the *Strongyloides* clade of nematodes. Nat Genet. 48:299–307.

International Helminth Genomes Consortium 2019. Comparative genomics of the major parasitic worms. Nat Genet. 51:163–174.

Jansen R, Bussemaker HJ, Gerstein M. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res. 31:2242–2251.

Junio AB, Li X, Massey HC, Nolan TJ, Lamitina ST, *et al.* 2008. *Strongyloides stercoralis*: cell- and tissue-specific transgene expression and co-transformation with vector constructs incorporating a common multifunctional 3' UTR. Exp Parasitol. 118:253–265.

Li X, Shao H, Junio A, Nolan TJ, Massey HC, *et al.* 2011. Transgenesis in the parasitic nematode *Strongyloides ratti*. Mol Biochem Parasitol. 179:114–119.

Liu C, Grote A, Ghedin E, Unnasch TR. 2020. CRISPR-mediated transfection of *Brugia malayi*. PLoS Negl Trop Dis. 14:e0008627.

Liu C, Mhashilkar AS, Chabanon J, Xu S, Lustigman S, *et al.* 2018. Development of a toolkit for piggyBac-mediated integrative transfection of the human filarial parasite *Brugia malayi*. PLoS Negl Trop Dis. 12:e0006509.

Lok JB, Shao H, Massey HC, Li X. 2017. Transgenesis in *Strongyloides* and related parasitic nematodes: historical perspectives, current functional genomic applications and progress towards gene disruption and editing. Parasitology. 144:327–342.

Lustigman S, Prichard RK, Gazzinelli A, Grant WN, Boatin BA, *et al.* 2012. A research agenda for helminth diseases of humans: the problem of helminthiases. PLoS Negl Trop Dis. 6:e1582.

Mitreva M, Wendl MC, Martin J, Wylie T, Yin Y, *et al.* 2006. Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. Genome Biol. 7:R75.

Plotkin JB, Kudla G. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 12:32–42.

Redemann S, Schloissnig S, Ernst S, Pozniakowsky A, Ayloo S, *et al.* 2011. Codon adaptation-based control of protein expression in *C. elegans*. Nat Methods. 8:250–252.

Schlager B, Wang X, Braach G, Sommer RJ. 2009. Molecular cloning of a dominant roller mutant and establishment of DNA-mediated transformation in the nematode *Pristionchus pacificus*. Genesis. 47:300–304.

Shapiro MB, Senapathy P. 1987. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 15:7155–7174.

Sharp PM, Bradnam KR. 1997. Appendix 3: Codon Usage in *C. elegans*, in *C. elegans* II, edited by DL Riddle, T Blumenthal, BJ Meyer, and JR Priess. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Sharp PM, Emery LR, Zeng K. 2010. Forces that influence the evolution of codon bias. Philos Trans R Soc Lond B Biol Sci. 365:1203–1212.

Sharp PM, Li WH. 1987. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.

Wheeler NJ, Airs PM, Zamanian M. 2020. Long-read RNA sequencing of human and animal filarial parasites improves gene models and discovers operons. PLoS Negl Trop Dis. 14:e0008869.

Witte H, Moreno E, Rödelsperger C, Kim J, Kim J-S, *et al.* 2015. Gene inactivation using the CRISPR/Cas9 system in the nematode *Pristionchus pacificus*. Dev Genes Evol. 225:55–62.

Communicating editor: J. Ward