Genetics: Early Online, published on June 21, 2013 as 10.1534/genetics.113.151340

UP-TORR: online tool for accurate and up-to-date annotation of RNAi reagents

Yanhui Hu[*,§], Charles Roesel[§,†], Ian Flockhart[*,§], Lizabeth Perkins[*,§], Norbert Perrimon[*,‡], Stephanie Mohr[*,§]

[*] Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

[§] *Drosophila* RNAi Screening Center, Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

[†] Bioinformatics program, Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA

[‡] Howard Hughes Medical Institute, 77 Avenue Louis Pasteur, Boston, MA 02115, USA

Running title: UP-TORR, RNAi reagent annotation tool

Key words: Drosophila, RNAi, annotation, genome

Corresponding author:

Stephanie Mohr

New Research Building Room 336

77 Avenue Louis Pasteur

Boston, MA 02115

Email: smohr@hms.harvard.edu

Phone: (617)432-5626

Abstract

RNA interference (RNAi) is a widely adopted tool for loss-of-function studies but RNAi results only have biological relevance if the reagents are appropriately mapped to genes. Several groups have designed and generated RNAi reagent libraries for studies in cells or *in vivo* for *Drosophila* and other species. At first glance, matching RNAi reagents to genes appears to be a simple problem, as each reagent is typically designed to target a single gene. In practice, however, the reagent-gene relationship is complex. Although the sequences of oligonucleotides used to generate most types of RNAi reagents are static, the reference genome and gene annotations are regularly updated. Thus, at the time a researcher chooses an RNAi reagent or analyzes RNAi data, the most current interpretation of the RNAi reagent-gene relationship, as well as related information regarding specificity (*e.g.* predicted off target effects), can be different from the original interpretation. Here, we describe a set of strategies and an accompanying online tool, UP-TORR (for Updated Targets of RNAi Reagents; <www.flyrnai.org/up-torr>), useful for accurate and up-to-date annotation of cell-based and *in vivo* RNAi reagents. Importantly, UP-TORR automatically synchronizes with gene annotations daily, retrieving the most current information available, and for *Drosophila*, also synchronizes with the major reagent collections. Thus, UP-TORR allows users to choose the most appropriate RNAi reagents at the onset of a study, as well as to perform the most appropriate analyses of results of RNAi-based studies.

Introduction

RNA interference (RNAi) is an effective tool to study gene function. In particular, genome-scale RNAi screens in mammalian and *Drosophila* cultured cells, as well as *in vivo* in *Drosophila* and *C. elegans,* have made contributions to a number of areas of study (BOUTROS and AHRINGER 2008; DIETZL *et al.* 2007; KAMATH *et al.* 2003; MOHR *et al.* 2010; MOHR and PERRIMON 2012; PERRIMON *et al.* 2010; QU *et al.* 2011). RNAi screening is dependent not only on the availability of RNAi reagents but also on accurate information regarding the predicted gene targets of the reagents. Large-scale RNAi libraries are available for a number of model systems. Although different types of RNAi reagents are used in different systems, there is a common and significant need to keep RNAi reagent annotations up-to-date with new genome assemblies and gene annotations.

A large number of cell-based RNAi screens have been performed using various genome-scale RNAi reagent libraries (MOHR *et al.* 2010). RNAi reagents for *Drosophila* cells are usually long (~100-500 bp) double-stranded RNAs (dsRNAs) made by PCR using a genomic or cDNA template, followed by *in vitro* transcription. In the cell, dsRNAs are processed by the endogenous RNAi machinery, generating active RNAi reagents, *i.e.* small dsRNA segments typically 20-22 bp in length with a 2 bp 3' overhang (CLEMENS *et al.* 2000; HAMMOND *et al.* 2000). In *Drosophila,* dsRNAs can be easily introduced into cultured cells (CLEMENS *et al.* 2000; HAMMOND *et al.* 2000). Several large-scale facilities, including the *Drosophila* RNAi Screening Center (DRSC) at Harvard Medical School, Boutros lab at German Cancer Research Center (DKFZ), RNAi Core at New York University, and Sheffield RNAi Screening Facility (SRSF), support *Drosophila* cell-based RNAi screening and offer genome-wide libraries with multiple dsRNAs-per-gene coverage. For mammalian cells, RNAi screens are done using synthesized short interfering RNAs (siRNAs), endoribonuclease-prepared short interfering RNAs (esiRNAs), or plasmid- or viral-encoded short hairpin RNAs (shRNAs) (KITTLER *et al.* 2007; MICKLEM and LORENS 2007; ROOT *et al.* 2006). Similar to *Drosophila* cell screens, mammalian screens are typically performed in individual labs or in conjunction with one of several academic screening facilities that provide automation and database support for screens.

RNAi reagents have also been developed for *in vivo* screens in various systems. In *C. elegans*, RNAi is systemic, and gene expression can be knocked down efficiently by feeding worms with bacteria expressing a long dsRNA (FRASER *et al.* 2000). A genome-scale RNAi feeding library is available (KAMATH *et al.* 2003) and widely used for functional studies. For *Drosophila*, *in vivo* RNAi relies on transgenic flies carrying RNAi transgenes that can be combined with the Gal4/UAS system for developmental, stage- and/or tissue-specific knockdown (G. Dietzl, D. Chen et al, Nature 2007). *Drosophila in vivo* RNAi reagents are either long dsRNA hairpins, for which gene fragments are cloned as an inverted repeat, or short hairpins synthesized as oligonucleotides and then cloned into an expression vector (PERRIMON *et al.* 2010). Altogether, about 90% of annotated *Drosophila* genes are targeted by fly RNAi collections from Vienna *Drosophila* RNAi Center (VDRC), NIG RNAi Resources in Japan and Transgenic RNAi Project (TRiP) at Harvard Medical School (DIETZL *et al.* 2007; NI *et al.* 2009; NI *et al.* 2008; NI *et al.*; YAMAMOTO 2010). Several large-scale transgenic RNAi screens have been successfully performed (reviewed in (PERRIMON *et al.* 2010)) and numerous *in vivo Drosophila* RNAi projects are ongoing.

Obtaining meaningful results from RNAi-based studies is entirely reliant upon appropriate identification of the sequence-specific gene target(s) of the reagent. Target identification might appear to be a simple problem but this is not necessarily the case. Even though sequences associated with RNAi reagents are static (*e.g.* the sequences of oligonucleotides used to make a library do not change), the reference sequences and gene annotations, including gene boundaries, exon-intron boundaries and nomenclature, are constantly being updated. Re-evaluations of existing RNAi libraries have shown that by the time of re-analysis, a percentage of reagents do not target any gene or are no longer predicted to be specific (HORN *et al.* 2010; QU *et al.* 2011). For a genome-wide *Caenorhabditis elegans* RNAi feeding library made available in 2003, for example, re-analysis in 2011 revealed that 18% of reagents needed to be re-annotated (QU *et al.* 2011).

For *Drosophila,* FlyBase is the primary resource of integrated genetic and genomic information, and FlyBase makes regular corrections and additions to gene models (FLYBASE-CONSORTIUM 2003) Since January 2008, FlyBase has released updated gene annotations about 10 times per year. Because several years can pass between the design of RNAi reagents and their use or data analysis, many new FlyBase annotations are released between reagent design and experimental design, and even more between reagent design and data analysis. Off-target effects (OTEs) are also relevant to the annotation of RNAi reagents. OTEs are induced by unintended cross-hybridization between RNAi reagents and endogenous sequences other than the target (KULKARNI *et al.* 2006; MOFFAT *et al.* 2007). As the sequences of gene and transcript change at each gene annotation release, annotation of potential OTEs can also change over time. Correcting for changes is not simply a matter of keeping up with new gene names and synonyms. Updates can change predictions as to the target gene, the number of predicted off-targets, isoform specificity, *etc.* As a result, it is critically important to regularly update the annotation of RNAi reagents and make this information readily accessible to the researchers who plan, execute and analyze RNAi-based experiments.

Several tools are available for the design of RNAi reagents, including SnapDragon for long dsRNAs (FLOCKHART *et al.* 2006; FLOCKHART *et al.* 2012), DSIR for siRNAs (FILHOL *et al.* 2012; VERT *et al.* 2006), and E-RNAi and NEXT-RNAi (ARZIMAN *et al.* 2005; HORN and BOUTROS 2010; HORN *et al.* 2010) for long dsRNAs and siRNAs. Nevertheless, a web-based tool that addresses the dynamic nature of gene annotation has not previously been available. Although E-RNAi can be used to evaluate long dsRNAs and siRNAs, the reference gene information for *Drosophila* in E-RNAi is currently out of date (FlyBase release5.19 from July 2009). NEXT-RNAi was designed to be integrated into a backend design/annotation pipeline and there is not currently an openly accessible web-based user-interface for the approach. In addition, NEXT-RNAi does not distinguish between RNAi reagents generated from genome DNA versus cDNA templates, a feature that is relevant to accurate annotation.

To best support community needs, the ideal tool would be based on regular, automated retrieval of new genome assemblies and gene annotation releases. The ideal tool would also handle the dynamic nature of reagent collections via regular, automatic retrieval of new reagent information from major public resources. To meet these needs, we developed a tool that allows users to query existing RNAi reagents from various sources based on the current gene annotation. The tool also allows researchers to query up-to-date information regarding gene target using user-provided RNAi reagent sequences.

Materials and Methods

Data Sources

Reference gene information is downloaded from the following sources: FlyBase for *Drosophila melanogaster* gene annotation (ftp://ftp.flybase.net/releases/current/); WormBase for *C. elegans* gene annotation (ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/sequence/); RefSeq for human and mouse gene annotation (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/).

RNAi reagent information is queried and downloaded from the following sources: FlyRNAi database for information regarding DRSC and TRiP reagents (http://www.flyrnai.org/); GenomeRNAi ftp site for DKFZ library (http://b110-wiki.dkfz.de/signaling/wiki/download/attachments/917513/Annotation_1stPCR_fulllibrary_HD2.xls); NIG catalog for NIG RNAi transgenic lines (http://www.shigen.nig.ac.jp/fly/nigfly/); VDRC catalog for VDRC RNAi transgenic lines (http://stockcenter.vdrc.at/control/fullCatalogueExcel).

Data Annotation Pipeline

The data annotation pipeline (Fig. 1) includes the following. (1) A module for automatic retrieval of reference genes and reagent information. This module downloads information from corresponding locations daily. The annotation pipeline is triggered whenever there is a new release from FlyBase or NCBI RefSeq, and/or when any new reagents become available. (2) A module that processes reference gene information for each species respectively to assemble a gene lookup table, the BLASTable database of genomic sequence for virtual PCR as well as the BLASTable database of transcript sequences for virtual PCR of the reagents made from cDNA library and on-target/off-target gene search. (3) A module that processes RNAi reagent information from each source in order to assemble the RNAi reagent lookup table and a BLASTable database of RNAi reagent sequences used when the end user queries UP-TORR by gene sequence. (4) A module that assembles the sequences of dsRNA reagents by *in silico* PCR. With the exception of the majority of NIG reagents, which have sequences assembled based on sequence validation data, long dsRNA reagents have not been sequenced and thus, the most accurate information is the sequences of PCR primers. Virtual PCR is performed with each new FlyBase release for all relevant reagents using either genomic sequence or transcript sequence, depending on how the reagents were initially generated, by BLASTing the primer sequences against the corresponding BLASTable database. (5) A module that matches the sequences of RNAi reagents to transcript sequences by BLAST. (6) A module that summarizes the on-target/off-target search results based on user defined parameters and presents the summary table to the end user. (7) A module that matches the gene sequences submitted by the end user to reagent sequences by BLAST. (8) A module that aligns reagents to genomic sequences of reference genes and reformats the information about the reference gene and RNAi reagents into the GFF3 format for upload to JBrowse, facilitating visual display of gene/reference alignment to the end user.

Software

The BLAST program from NCBI (ALTSCHUL *et al.* 1990) is among the research applications already installed on the Orchestra platform at Harvard Medical School. The BLAST parameters for virtual PCR: -W 10 -e 1 -G 5 -E 2; cutoff for virtual PCR: 100% identity; BLAST parameters for on-target/off-target searches: -W 14 -e 10 -G 5 -E 2 -F F; cutoff for on-target search: 27bp or longer with >=98% identify; cutoff for off-target search: 15bp alignment or longer. JBrowse was downloaded from jbrowse.org/install (SKINNER *et al.* 2009). More detailed information can be found at jbrowse.org/developer. Programs for reagent annotation were written in Perl and the user interface was developed using HTML, JavaScript, Java servlets and Lucene. A Perl program provided as part of the JBrowse download converts annotations from the GFF3 format to the JBrowse format.

Results and Discussion

Reference genes are 'moving targets' that change over time

For *Drosophila melanogaster,* FlyBase is the primary resource of integrated genetic and genomic information, including up-to-date genome assemblies and gene annotations (FLYBASE-CONSORTIUM 2003). Since the first assembly of the *Drosophila melanogaster* genome published in 2000, four subsequent genome assemblies, with the most recent one in February 2007, have occurred (CELNIKER *et al.* 2002; HOSKINS *et al.* 2007; MYERS *et al.* 2000). In addition to updates to the genome assembly, there have been numerous updates since 2000 to gene annotations. Particularly given the new availability of next-generation sequencing approaches, gene annotations continue to change, for example due to the addition of newly identified genes and newly identified isoforms of previously identified genes. Thus, despite the fact that *Drosophila* is arguably the best annotated genome among multi-cellular species, our knowledge of the fly genome and proteome continues to improve. Indeed, since the availability of the 5th genome assembly (*i.e.* over that last six years or so), the FlyBase consortium has released 49 updates to *Drosophila* gene annotations.

Exemplifying the extent of changes, for the gene annotation release issued on September 7, 2012, 123 genes and 578 protein-coding transcripts were changed relative to the previous release. Moreover, the number and type of changes to gene annotations varies with each release. To obtain a more comprehensive picture of gene annotation changes, we looked at changes to the gene annotation over the period of one year (FlyBase version r5.34 vs 5.44). On the gene level, 412 new genes were added, 12 genes were retired, and the genome location of 2287 genes was changed. On the transcript level, 3407 new transcripts were added, 833 transcripts were retired, and the specific sequences of 2902 transcripts were changed. Thus, for a *Drosophila* RNAi reagent designed at the beginning of this period, there is an approximately

30% chance that the sequence of the gene target had changed a year later. Given that the time from RNAi reagent design to availability of the reagent for experiments can be months, and the practical reality that many RNAi reagents are put to use several years after they were designed, these changes have a significant impact on RNAi reagent annotation. Notably, gene annotation changes can affect not just the on-target predictions for a given RNAi reagent but also the number of predicted off-target effects (OTEs) associated with a given reagent and/or whether or not it is predicted to target all isoforms of the target gene. For a summary of annotation changes in FlyBase and WormBase over the past five years, see supplemental table 1.

Dynamic annotation of RNAi reagents

When a large amount of information is involved (in this case, information surrounding the sequence and targets of RNAi reagents), the typical approach is to use a back-end database to store the information. At the DRSC, the backend storage is a relational MySQL database (Flockhart *et al.* 2006; Flockhart *et al.*) in which a couple dozen tables are used to store information regarding gene annotations associated with DRSC and TRiP RNAi reagents. Updating gene annotations as frequently as FlyBase releases updates is not trivial and as a result, such databases are usually out of sync with the most current release, a situation that is acceptable for most RNAi reagents but potentially misleading for a sub-set of reagents for which the corresponding gene annotations have changed significantly. Moreover, forever associating the RNAi reagent with its originally intended target might bias interpretation of RNAi results, even when information about alternative targets is also presented.

To address this issue, we developed a new strategy and developed a dynamic annotation tool that is 'blind' to the original target gene annotations, basing the final reports presented online solely on updated information. The tool, which we named UP-TORR for updated targets of RNAi reagents, daily and automatically accesses the ftp sites available at FlyBase, WormBase as well as RefSeq database at NCBI and whenever a new release is available, retrieves all of the new sequence and gene annotation information. Thus, at any given time, a query of UP-TORR will generate the most updated results available. For cell-based RNAi reagents from the DRSC and DKFZ as well as *in vivo* long hairpin reagents generated by VDRC and Ahringer lab, PCR primer sequences are aligned to the up-to-date genome assembly sequence, generating virtual PCR products. The sequences of these PCR products are then BLASTed against transcript sequences in order to identify the current on-target and off-target predictions. The process is similar for *in vivo* long hairpin reagents generated by TRiP, except that for these, transcript sequences are used to generate the virtual PCR product, as the template used to generate these was cDNA rather than genomic DNA. For the *in vivo* long hairpin reagents generated by NIG, because most reagent sequences were assembled by end-to-end sequencing, for these reagents we skip the virtual PCR step and go directly to BLASTing RNAi sequences against transcript sequences. When a user enters a pair of primers for analysis, the user can specify if genomic DNA or transcript sequences should be used in the virtual PCR step. For shRNA reagents, both the 21 bp sense-strand and anti-strand sequences, which originated as synthetic oligonucleotides, can be directly BLASTed against transcript sequences (Fig. 1).

During the reagent 'live re-annotation' process, UP-TORR is designed to answer the following questions. (1) What are all of the possible gene targets? (2) Does the reagent target all isoforms or only some isoforms of the gene? (3) What region of the transcript(s) does the reagent target, *i.e.* the 5'UTR, CDS or 3'UTR? And (4) are there potential off-target genes that share a certain level of sequence similarity? The on-target matches are relative to the full reagent sequence with at least 17 bp matches for shRNA and 27 bp perfect match for long dsRNA, whereas off-target matches can be as short as 15 bp matches. The user can specify the cutoffs at the user interface.

Using this tool, we re-annotated all the RNAi reagents generated at DRSC, DKFZ, VDRC, NIG and TRiP based on FlyBase release 5.49 (Table 1). We found that a percentage of the reagents no longer met the original design goal. For example, within the TRiP shRNA collection, 3% of reagents were predicted at the time of our re-annotation with UP-TORR to target multiple genes. Some of these are due to high sequence similarity of the paralogous genes such as His1, His2A, His2B and His3 families respectively, making it impossible to design gene-specific RNAi reagents. Additionally, the *Drosophila* genome is more compact than the mammalian genome, and some genes are located close to each other or fully overlap on the genome as well as at the transcript level. For example, the genes *cup* and *CG34310* are both located at 6663968-6674780 on the + strand of chromosome 2L. Their transcripts are also identical and the only difference is the protein-coding regions (Fig. 2A). In cases like this, it is impossible to design any RNAi reagent targeting one gene but not the other. Another example is *eIF-2gamma* and *Su(var)3-9*. These genes partially overlap on both the genome and transcript levels. TRiP reagent HMS00279 happened to target exons shared by the two genes; therefore, the library could be improved by targeting the regions specific to each gene (Fig. 2B). In addition, 0.8% of reagents do not target any genes in the release we were testing. They aligned to introns (Fig. 2C), inter-gene regions (Fig. 2D) or pseudogenes (Fig. 2E) due to the changes in the intron-exon boundary, gene boundary or gene retirement.

Our comparison of FlyBase releases (r5.34 and r5.44) shows that 3407 new transcripts were added and 833 transcripts were removed. Thus, it is more likely that a new isoform will be added than that an existing isoform will be retired. An RNAi reagent may fail to target all isoforms even though it was initially designed to be isoform unspecific. According to FlyBase release 5.49, 38% of fly genes have more than one isoform. We found that 90% of TRiP shRNA reagents still target all isoforms whereas 6% target one or a subset of isoforms based on current isoform annotation. Some of these reagents are limited by the genes themselves, which lack exons common among all isoforms (Fig. 2F), whereas others could be improved (Fig. 2G) by targeting regions shared by all isoforms. Because isoforms can be expressed specifically in certain tissues or under certain pathological conditions, and/or might have divergent functions, providing annotation at the isoform level is important for the appropriate identification of RNAi reagents and interpretation of RNAi results.

Online features of UP-TORR

To provide researchers with the most current and accurate annotation of RNAi reagents, we developed a freely accessible web-based application. To accommodate the full spectrum of community needs regarding reagent identification and live re-annotation, we have provided users with five different ways to query UP-TORR. After selecting the species (*Drosophila, C. elegans*, mouse or human) from the appropriate menu tab, users can (1) enter the gene-specific region of an RNAi reagent sequence (*i.e.* a 19-21 bp sense/anti-sense strands corresponding to a siRNA or short hairpin, or a DNA sequence corresponding to a dsRNA); (2) enter PCR primers for dsRNA, then choose the proper PCR template (genomic DNA or cDNA); (3) enter a list of RNAi reagent IDs (*e.g.* DRSC amplicon ID, GenomeRNAi amplicon ID, TRiP stock ID, NIG stock ID, VDRC transformant ID or Ahringer primer pair ID); (4) enter a list of gene identifiers for which all relevant reagents will be retrieved (*e.g.* FlyBase FBgn IDs, CG numbers and/or Gene Symbols); or (5) enter the sequence to be targeted (*e.g.* a full-length transcript or exon sequence). For query types (1), (2) and (3), in which an RNAi reagent is the input, UP-TORR returns a summary of all of the potentially targeted genes, including gene identifiers such as FlyBase FBgn number for fly and NCBI Entrez GeneID for other species, gene symbol, and gene isoform information, as well as the region and location of each isoform that is targeted. UP-TORR also reports the number of possible off-target genes, which is hyperlinked to detailed information about the genes (Fig. 3). For query types (4) and (5), in which a target gene is the input, all of the RNAi reagents deemed relevant by the live re-annotation are reported, along with a similar summary of information about isoform specificity and predicted OTEs. These search options allow users to retrieve all the available RNAi reagents quickly without searching individual resources. In addition, users can easily compare all RNAi reagents available for a given gene and select the best one(s). There has been ongoing effort evaluating the efficiency of TRiP RNAi transgenic lines by phenotyping and/or qPCR analysis. To help UP-TORR users select the most efficient reagent(s), TRiP stock IDs are hyper-linked to a page that includes validation results. With query type (5), in addition to full gene or transcript sequences, users can also enter specific exon or domain sequences to identify reagents specifically targeting the transcript region of interest. For all query types, results are hyperlinked to an instance of JBrowse, where alignment of the RNAi reagents with genes and transcripts is displayed visually. Users also have the option to download a summary table of results and supporting information.

Finally, we note that when the output species is *Drosophila* or *C. elegans*, the output page from a DIOPT ortholog search (flyrnai.org/diopt) or DIOPT-DIST disease-gene ortholog search (flyrnai.org/diopt-dist) (HU *et al.* 2011) has been modified to include a button that carries the gene list forward from DIOPT or DIOPT-DIST to UP-TORR. We expect this should help facilitate identification of RNAi reagents relevant to conserved and disease-related genes.

Discussion

There is a necessary passage of time between the design of RNAi reagents and their use, as well as between design and analysis of results (and later re-interpretation of RNAi data, such as in meta-analyses) (HORN *et al.* 2010; QU *et al.* 2011). As we have presented, gene annotations change over time (Supplemental table 1), leading to changes in what the latest evidence suggests is the appropriate interpretation of RNAi on-target and off-target potential. The UP-TORR approach and accompanying freely accessible user interface make it possible for researchers to identify RNAi reagents and/or interpret the results of RNAi studies based on the most current annotation available from FlyBase. Our analysis of RNAi reagents from all the public RNAi collections show that a small percentage of RNAi reagents did not meet initial design goals upon re-annotation (*i.e.* they were no longer predicted to be gene specific and isoform non-specific with regards to the intended target gene). By comparing the different FlyBase releases, we further found that the coding sequences (CDS) are less likely to change as compared with un-translated regions (5' or 3' UTRs). This likely reflects the fact that it has historically been easier both computationally and experimentally to identify coding sequences than to identify full-length transcripts.

Because UP-TORR checks for updates at FlyBase, WormBase as well as RefSeq database daily and incorporates these new data, facilitating what we refer to as a 'live re-annotation' of RNAi reagent information, the tool will be valuable to anyone interested in designing, analyzing or re-analyzing RNAi results, including results from high-throughput screens. We recognize, however, that results from UP-TORR or any another up-to-date comparison with the current annotation of genomes and/or transcriptomes does not necessarily provide the 'final word' on RNAi on-target and off-target effects. For example, RNAi treatments can have generalized, gene non-specific effects (Muller *et al.* 2008). In addition, SNPs (Chen *et al.* 2009), RNA editing (Rodriguez *et al.* 2012) and chimeric transcripts (Frenkel-Morgenstern *et al.* 2013) can complicate the prediction of the on-target as well as off-target genes of RNAi reagents. Nevertheless, UP-TORR is the first tool available to address the issue of genome annotation and RNAi sequences. Importantly, the tool provides up-to-date annotation for RNAi reagents targeting human (Fig. S1), mouse genes as well as for *Drosophila* and *C. elegans*, and could easily be expanded to include more species. In the future, this tool might be applied to other methods (*e.g*. TALEs (Christian *et al.* 2010) and CRISPRs (Cong *et al.* 2013)) for which gene annotation impacts interpretation of the reagents.

Acknowledgements

We would like to thank the members of the *Drosophila* RNAi Screening Center, Transgenic RNAi Project and Perrimon lab, in particular Laura Holderbaum and Dong Yan for helpful comments and support. We thank Thomas Micheler from the VDRC for help with data access. We also thank all of the past and current users of the DRSC and TRiP for providing valuable feedback. This work was supported in large part by NIGMS R01 GM067761 and GM084947, with additional support from NCRR/ORIP R24 RR032668. S.E.M. is also supported in part by the Dana Farber/Harvard Cancer Center and N.P. is an investigator of the Howard Hughes Medical Institute.

Supplemental files

Supplemental table 1: Summary of changes in gene and transcript annotations at FlyBase and WormBase over the past six years.

Supplemental figure 1: UP-TORR user interface for human RNAi reagents.

Table 1. Summary of major public *Drosophila* and *C. elegans* RNAi reagent collections.

RNAi Collection	Reagent Type	All Reagents	Target 1 gene, all isoform(s)	Target 1 gene, not all isoform(s)	Target multiple genes	Reagents with >5 OTEs (19bp)	Target pseudogene	No gene target
DKFZ	dsRNA, cell-based	20016	15948 (80%)	816 (4%)	1466 (7%)	394 (2%)	78 (0.4%)	1708 (9%)
DRSC - GenomeLibrary	dsRNA, cell-based	24037	19615 (82%)	1664 (7%)	979 (4%)	552 (2%)	84 (0.4%)	1695 (7%)
DRSC-FollowupLibrary	dsRNA, cell-based	9448	8519 (90%)	470 (5%)	296 (3%)	15 (0.2%)	14 (0.2%)	149 (2%)
NIG	dsRNA, transgenic fly	11725	10328 (88%)	532 (5%)	436 (4%)	416 (4%)	2 (0.02%)	427 (4%)
TRiP-LongHairPin	dsRNA, transgenic fly	2483	2255 (91%)	114 (5%)	72 (3%)	8 (0.3%)	2 (0.1%)	40 (2%)
TRiP-ShortHairPin	shRNA, transgenic fly	4132	3738 (90%)	242 (6%)	120 (3%)	0 (0%)	2 (0.1%)	30 (1%)
VDRC-GD Library	dsRNA, transgenic fly	21808	18607 (85%)	962 (4%)	1357 (6%)	745 (3%)	60 (0.3%)	822 (4%)
VDRC-KK Library	dsRNA, transgenic fly	10748	9135 (85%)	431 (4%)	378 (4%)	67 (1%)	27 (0.3%)	777 (7%)
Ahringer Library	dsRNA, worm feeding	16256	11002 (68%)	678 (4%)	1733 (11%)	1074 (7%)	283 (2%)	2843 (16%)

Figure 1. UP-TORR annotation pipeline. Features of the pipeline include automated daily checks for new reagents and gene annotations from the relevant public sources. When updates are available, the information is downloaded and used to rebuild the underlying databases and lookup tables, allowing UP-TORR to provide the most updated interpretation of the relationship between RNAi reagents and gene annotations.

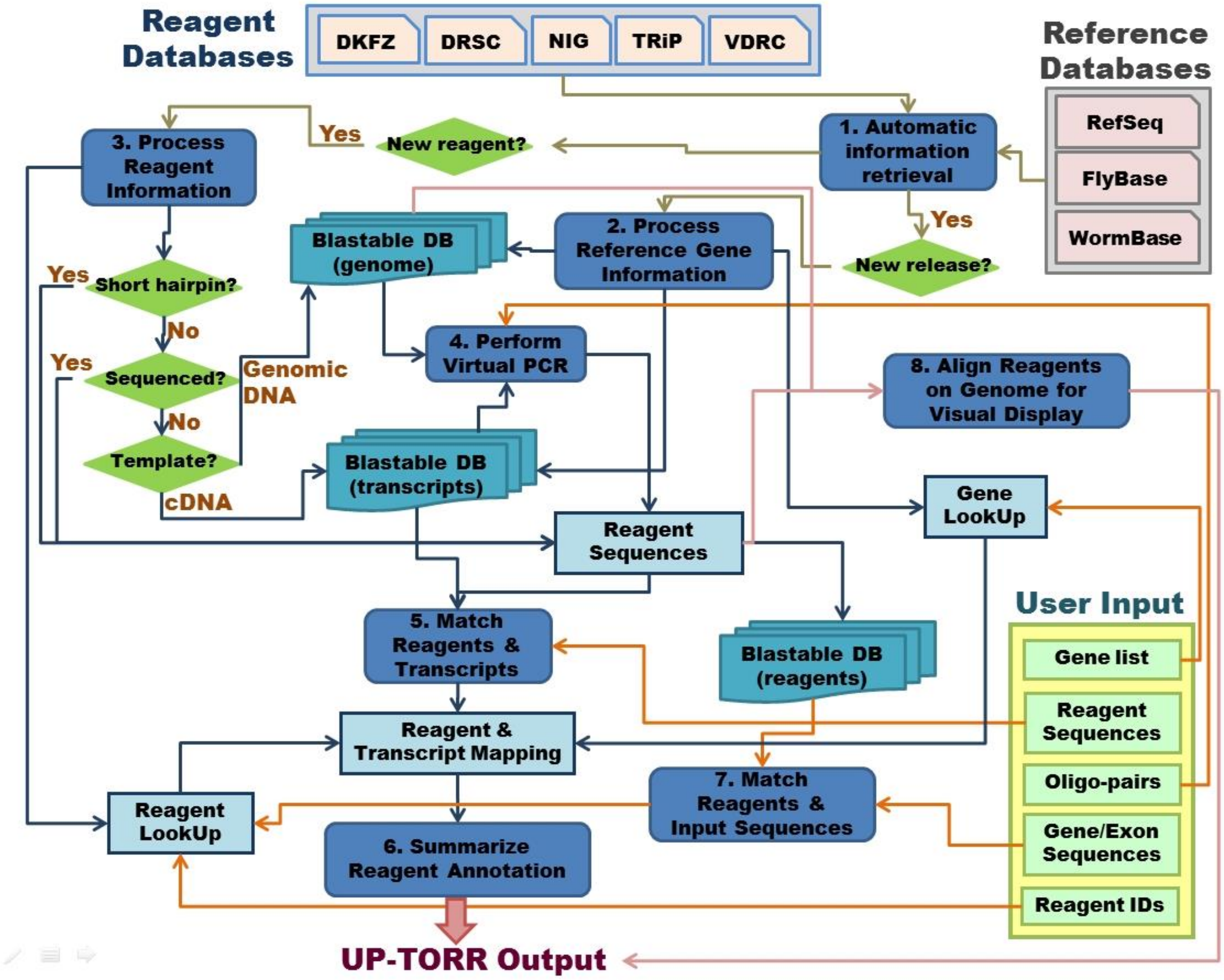


Figure 2. Issues associated with RNAi reagents and annotations. (A) and (B), examples of reagents that target multiple genes. (A) TRiP line GL00327 targets both *cup* and *CG34310* because both the genome and transcript sequences of these two genes fully overlap. (B) TRiP line HMS00279 targets the common exon shared by both the *eIF-2gamma* and *Su(var)3-9* genes. Since the transcript sequence of both of these genes only partially overlap, it is possible to design specific RNAi reagents targeting either *eIF-2gamma* or *Su(var)3-9*. (C) to (E), examples of reagents that do not target any gene. (C) TRiP reagent HMS00286 aligns to the intron of gene *loqs* due to a change in the intron-exon junction(s) of *loqs* gene annotation. (D) TRiP reagent HMS01233 aligns to an inter-gene region due to a change in the *Parp* gene boundary. (E) TRiP line HMS00620 aligns to the newly annotated pseudogene *CR43361.* (F) and (G), examples of reagents that do not target all isoforms. (F) TRiP reagent HMS00621 targets five of the eight isoforms of gene *CG42724*. *CG42724* lacks any common exon shared by all isoforms. (G) TRiP reagent HMS01241 targets two of the four isoforms of gene *qkr54B*. An improved reagent can be designed against the exons shared by all four isoforms.

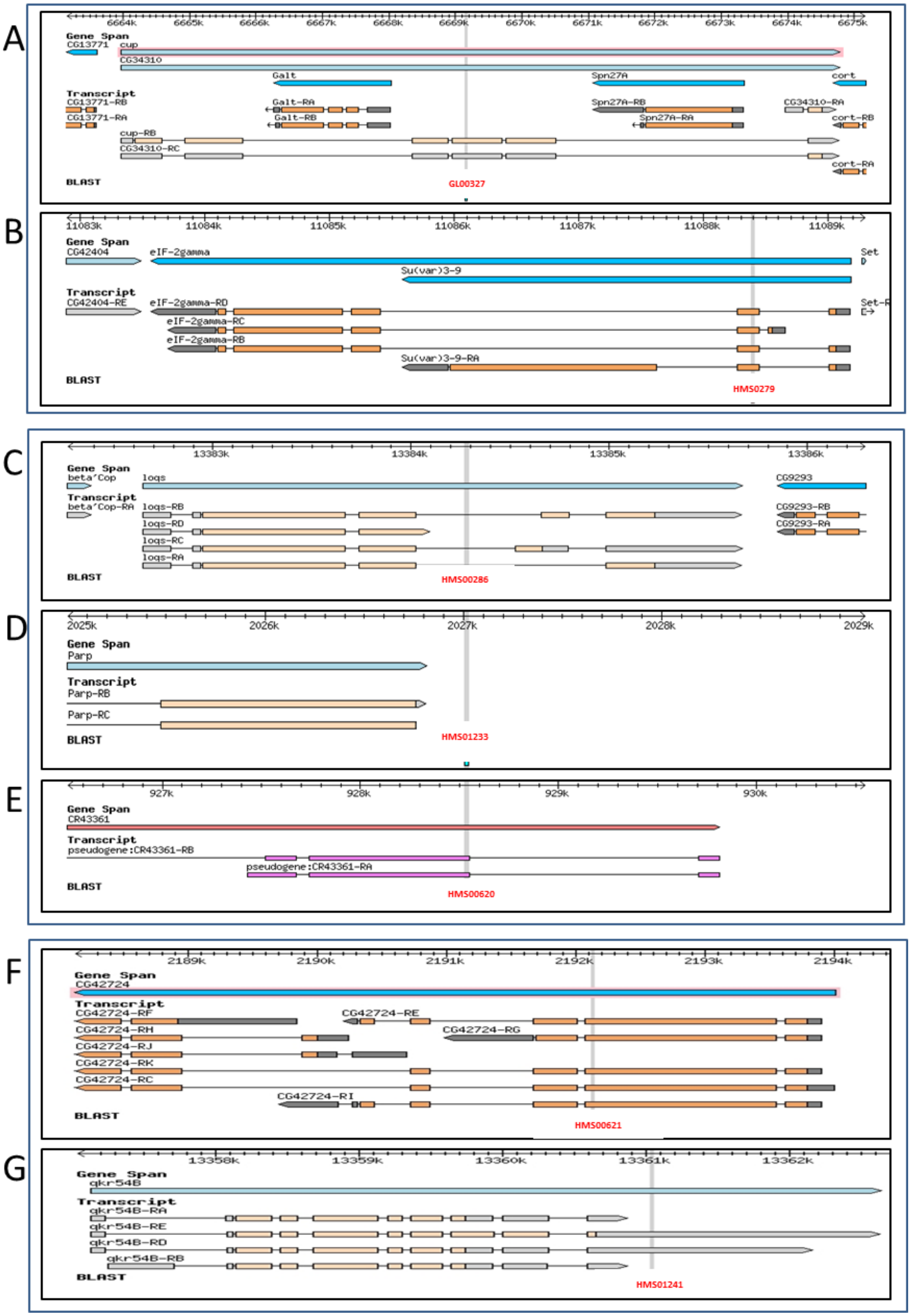
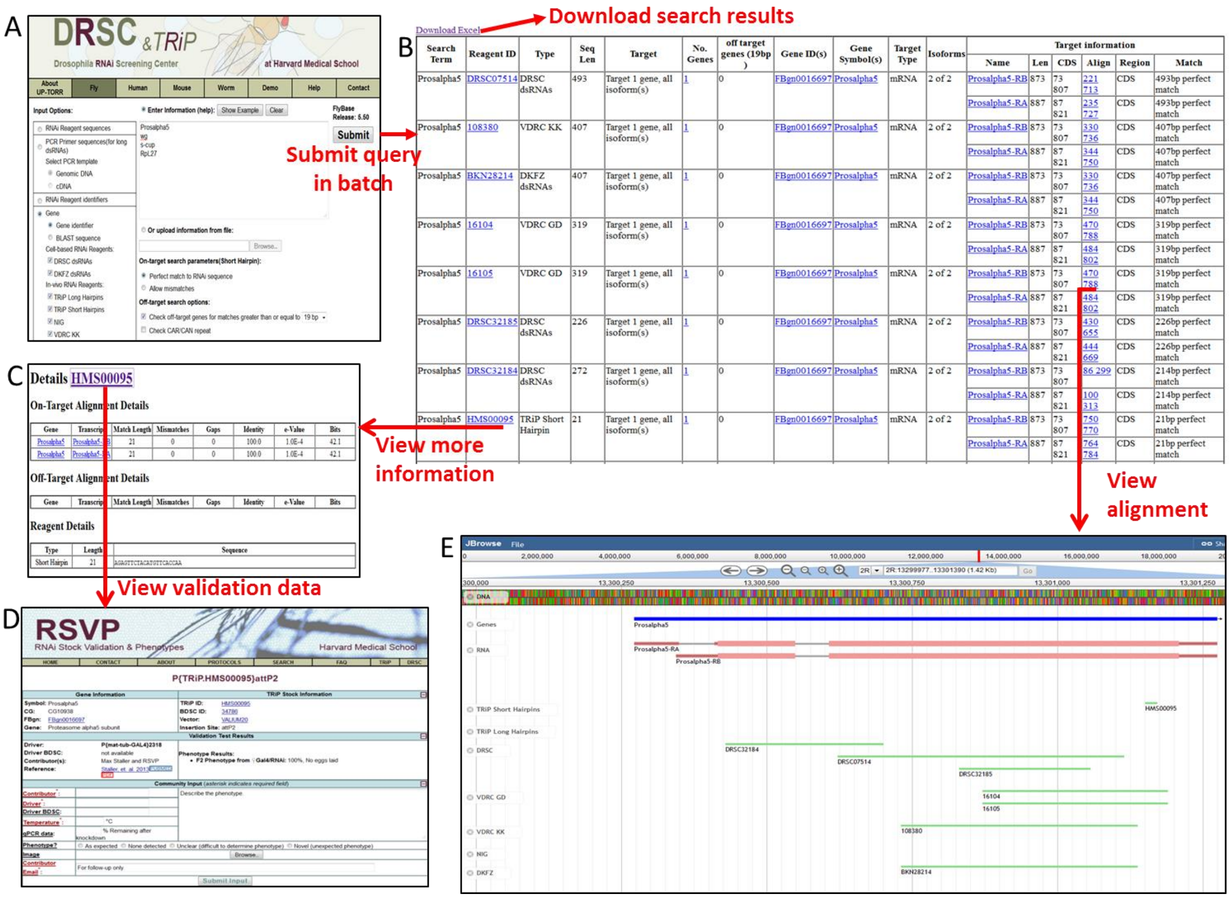

A
Gene Span
CG13771
cup
CG34310
Galt
Spn27A
cort
Transcript
CG13771-RB
CG13771-RA
Galt-RA
Galt-RB
Spn27A-RB
Spn27A-RA
CG34310-RA
cort-RB
cup-RB
CG34310-RC
cort-RA
BLAST
GL00327
B
Gene Span
CG42404
eIF-2gamma
Su(var)3-9
Set D
Transcript
CG42404-RE
eIF-2gamma-RD
eIF-2gamma-RC
eIF-2gamma-RB
Su(var)3-9-RA
BLAST
HMS0279
C
Gene Span
beta'Cop
loqs
CG9293
Transcript
beta'Cop-RA
loqs-RB
loqs-RD
loqs-RC
loqs-RA
CG9293-RB
CG9293-RA
BLAST
HMS00286
D
Gene Span
Parp
Transcript
Parp-RB
Parp-RC
BLAST
HMS01233
E
Gene Span
CR43361
Transcript
pseudogene:CR43361-RB
pseudogene:CR43361-RA
BLAST
HMS00620
F
Gene Span
CG42724
Transcript
CG42724-RF
CG42724-RE
CG42724-RH
CG42724-RG
CG42724-RJ
CG42724-RK
CG42724-RC
CG42724-RI
BLAST
HMS00621
G
Gene Span
qkr54B
Transcript
qkr54B-RA
qkr54B-RE
qkr54B-RD
qkr54B-RB
BLAST
HMS01241

Figure 3. UP-TORR user interface. At UP-TORR, the user first specifies a species (*Drosophila, C. elegans*, mouse or human) by selecting the appropriate menu tab. The example shown is for fly genes (see Supplemental Figures for an example using human RNAi reagent sequences). (A) At the appropriate starting page, the user 1) enters the gene-specific sequence region of an RNAi reagent; 2) enters PCR primers for dsRNA and selects the relevant PCR template (genomic DNA or cDNA); 3) enters a list of RNAi reagent IDs (*e.g.* DRSC or TRiP IDs, NIG IDs, GenomeRNAi IDs; see Table 2); 4) enters a list of gene identifiers for which all relevant reagents will be retrieved (*e.g.* FlyBase FBgn IDs); or 5) enters a specific sequence (*e.g.* a full-length transcript or exon sequence). (B) UP-TORR outputs a table that summarizes gene and isoform specificity of the reagents and provides information about the target region, location and alignment length. (C) Alignment results are hyperlinked to an instance of JBrowse that visually displays an alignment of the RNAi reagents with genes and transcripts. (D) Reagent identifiers are hyperlinked to detailed information pages with reagent sequence(s), on-target gene(s) and off-target gene(s) information. (E) Reagent identifiers on the detail information page are hyperlinked to verification and phenotype data at TRiP RSVP.



Download search results

Submit query in batch

View more information

View validation data

View alignment

Input Options:

○ RNAi Reagent sequences

○ PCR Primer sequences(for long dsRNAs)

Select PCR template

● Genomic DNA

○ cDNA

○ RNAi Reagent identifiers

● Gene

● Gene identifier

○ BLAST sequence

Cell-based RNAi Reagents:

☑ DRSC dsRNAs

☑ DKFZ dsRNAs

In-vivo RNAi Reagents:

☑ TRiP Long Hairpins

☑ TRiP Short Hairpins

☑ NIG

☑ VDRC KK

● Enter Information (help):

Prosalpha5
wg
s-cup
RpL27

FlyBase Release: 5.50

○ Or upload information from file:

On-target search parameters(Short Hairpin):

● Perfect match to RNAi sequence

○ Allow mismatches

Off-target search options:

☑ Check off-target genes for matches greater than or equal to 19 bp

□ Check CAR/CAN repeat

Download Excel

Search Term	Reagent ID	Type	Seq Len	Target	No. Genes	off target genes (19bp)	Gene ID(s)	Gene Symbol(s)	Target Type	Isoforms	Target information					
											Name	Len	CDS	Align	Region	Match
Prosalpha5	DRSC07514	DRSC dsRNAs	493	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	221 713	CDS	493bp perfect match
											Prosalpha5-RA	887	87 821	235 727	CDS	493bp perfect match
Prosalpha5	108380	VDRC KK	407	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	330 736	CDS	407bp perfect match
											Prosalpha5-RA	887	87 821	344 750	CDS	407bp perfect match
Prosalpha5	BKN28214	DKFZ dsRNAs	407	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	330 736	CDS	407bp perfect match
											Prosalpha5-RA	887	87 821	344 750	CDS	407bp perfect match
Prosalpha5	16104	VDRC GD	319	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	470 788	CDS	319bp perfect match
											Prosalpha5-RA	887	87 821	484 802	CDS	319bp perfect match
Prosalpha5	16105	VDRC GD	319	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	470 788	CDS	319bp perfect match
											Prosalpha5-RA	887	87 821	484 802	CDS	319bp perfect match
Prosalpha5	DRSC32185	DRSC dsRNAs	226	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	430 655	CDS	226bp perfect match
											Prosalpha5-RA	887	87 821	444 669	CDS	226bp perfect match
Prosalpha5	DRSC32184	DRSC dsRNAs	272	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	86 299	CDS	214bp perfect match
											Prosalpha5-RA	887	87 821	100 313	CDS	214bp perfect match
Prosalpha5	HMS00095	TRiP Short Hairpin	21	Target 1 gene, all isoform(s)	1	0	FBgn0016697	Prosalpha5	mRNA	2 of 2	Prosalpha5-RB	873	73 807	750 770	CDS	21bp perfect match
											Prosalpha5-RA	887	87 821	764 784	CDS	21bp perfect match

Details HMS00095

On-Target Alignment Details

Gene	Transcript	Match Length	Mismatches	Gaps	Identity	e-Value	Bits
Prosalpha5	Prosalpha5-RB	21	0	0	100.0	1.0E-4	42.1
Prosalpha5	Prosalpha5-RA	21	0	0	100.0	1.0E-4	42.1

Off-Target Alignment Details

Gene	Transcript	Match Length	Mismatches	Gaps	Identity	e-Value	Bits

Reagent Details

Type	Length	Sequence
Short Hairpin	21	AGAGTTCTACATGTTCACCAA

P{TRiP.HMS00095}attP2

Gene Information

Symbol: Prosalpha5
CG: CG10938
FBgn: FBgn0016697
Gene: Proteasome alpha5 subunit

TRiP Stock Information

TRiP ID: HMS00095
BDSC ID: 34786
Vector: VALIUM20
Insertion Site: attP2

Validation Test Results

Driver: P{mat-tub-GAL4}2318
Driver BDSC: not available
Contributor(s): Max Staller and RSVP
Reference: Staller, et. al. 2013

Phenotype Results:

- F2 Phenotype from Gal4/RNAi: 100%, No eggs laid

Community Input (asterisk indicates required field)

Contributor*:
Driver*:
Driver BDSC:
Temperature*: °C
qPCR data: % Remaining after knockdown
Describe the phenotype
Phenotype? ○ As expected ○ None detected ○ Unclear (difficult to determine phenotype) ○ Novel (unexpected phenotype)
Image
Contributor Email*: For follow-up only

References

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS and D. J. LIPMAN, 1990 Basic local alignment search tool. J Mol Biol **215:** 403-410.

ARZIMAN, Z., T. HORN and M. BOUTROS, 2005 E-RNAi: a web application to design optimized RNAi constructs. Nucleic Acids Res **33:** W582-588.

BOUTROS, M., and J. AHRINGER, 2008 The art and design of genetic screens: RNA interference. Nat Rev Genet **9:** 554-566.

CELNIKER, S. E., D. A. WHEELER, B. KRONMILLER, J. W. CARLSON, A. HALPERN *et al.*, 2002 Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol **3:** RESEARCH0079.

CHEN, D., J. BERGER, M. FELLNER and T. SUZUKI, 2009 FLYSNPdb: a high-density SNP database of Drosophila melanogaster. Nucleic Acids Res **37:** D567-570.

CHRISTIAN, M., T. CERMAK, E. L. DOYLE, C. SCHMIDT, F. ZHANG *et al.*, 2010 Targeting DNA double-strand breaks with TAL effector nucleases. Genetics **186:** 757-761.

CLEMENS, J. C., C. A. WORBY, N. SIMONSON-LEFF, M. MUDA, T. MAEHAMA *et al.*, 2000 Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. Proc Natl Acad Sci U S A **97:** 6499-6503.

CONG, L., F. A. RAN, D. COX, S. LIN, R. BARRETTO *et al.*, 2013 Multiplex genome engineering using CRISPR/Cas systems. Science **339:** 819-823.

DIETZL, G., D. CHEN, F. SCHNORRER, K. C. SU, Y. BARINOVA *et al.*, 2007 A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature **448:** 151-156.

FILHOL, O., D. CIAIS, C. LAJAUNIE, P. CHARBONNIER, N. FOVEAU *et al.*, 2012 DSIR: assessing the design of highly potent siRNA by testing a set of cancer-relevant target genes. PLoS One **7:** e48057.

FLOCKHART, I., M. BOOKER, A. KIGER, M. BOUTROS, S. ARMKNECHT *et al.*, 2006 FlyRNAi: the Drosophila RNAi screening center database. Nucleic Acids Res **34:** D489-494.

FLOCKHART, I. T., M. BOOKER, Y. HU, B. MCELVANY, Q. GILLY *et al.*, 2012 FlyRNAi.org--the database of the Drosophila RNAi screening center: 2012 update. Nucleic Acids Res **40:** D715-719.

FLYBASE-CONSORTIUM, 2003 The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res **31:** 172-175.

FRASER, A. G., R. S. KAMATH, P. ZIPPERLEN, M. MARTINEZ-CAMPOS, M. SOHRMANN *et al.*, 2000 Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. Nature **408:** 325-330.

FRENKEL-MORGENSTERN, M., A. GOROHOVSKI, V. LACROIX, M. ROGERS, K. IBANEZ *et al.*, 2013 ChiTaRS: a database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data. Nucleic Acids Res **41:** D142-151.

HAMMOND, S. M., E. BERNSTEIN, D. BEACH and G. J. HANNON, 2000 An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature **404:** 293-296.

HORN, T., and M. BOUTROS, 2010 E-RNAi: a web application for the multi-species design of RNAi reagents--2010 update. Nucleic Acids Res **38:** W332-339.

HORN, T., T. SANDMANN and M. BOUTROS, 2010 Design and evaluation of genome-wide libraries for RNA interference screens. Genome Biol **11:** R61.

HOSKINS, R. A., J. W. CARLSON, C. KENNEDY, D. ACEVEDO, M. EVANS-HOLM *et al.*, 2007 Sequence finishing and mapping of Drosophila melanogaster heterochromatin. Science **316:** 1625-1628.

HU, Y., I. FLOCKHART, A. VINAYAGAM, C. BERGWITZ, B. BERGER *et al.*, 2011 An integrative approach to ortholog prediction for disease-focused and other functional studies. BMC Bioinformatics **12:** 357.

KAMATH, R. S., A. G. FRASER, Y. DONG, G. POULIN, R. DURBIN *et al.*, 2003 Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature **421:** 231-237.

KITTLER, R., V. SURENDRANATH, A. K. HENINGER, M. SLABICKI, M. THEIS *et al.*, 2007 Genome-wide resources of endoribonuclease-prepared short interfering RNAs for specific loss-of-function studies. Nat Methods **4:** 337-344.

KULKARNI, M. M., M. BOOKER, S. J. SILVER, A. FRIEDMAN, P. HONG *et al.*, 2006 Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays. Nat Methods **3:** 833-838.

MICKLEM, D. R., and J. B. LORENS, 2007 RNAi screening for therapeutic targets in human malignancies. Curr Pharm Biotechnol **8:** 337-343.

MOFFAT, J., J. H. REILING and D. M. SABATINI, 2007 Off-target effects associated with long dsRNAs in Drosophila RNAi screens. Trends Pharmacol Sci **28:** 149-151.

MOHR, S., C. BAKAL and N. PERRIMON, 2010 Genomic screening with RNAi: results and challenges. Annu Rev Biochem **79:** 37-64.

MOHR, S. E., and N. PERRIMON, 2012 RNAi screening: new approaches, understandings, and organisms. Wiley Interdiscip Rev RNA **3:** 145-158.

MULLER, P., M. BOUTROS and M. P. ZEIDLER, 2008 Identification of JAK/STAT pathway regulators--insights from RNAi screens. Semin Cell Dev Biol **19:** 360-369.

MYERS, E. W., G. G. SUTTON, A. L. DELCHER, I. M. DEW, D. P. FASULO *et al.*, 2000 A whole-genome assembly of Drosophila. Science **287:** 2196-2204.

NI, J. Q., L. P. LIU, R. BINARI, R. HARDY, H. S. SHIM *et al.*, 2009 A Drosophila resource of transgenic RNAi lines for neurogenetics. Genetics **182:** 1089-1100.

NI, J. Q., M. MARKSTEIN, R. BINARI, B. PFEIFFER, L. P. LIU *et al.*, 2008 Vector and parameters for targeted transgenic RNA interference in Drosophila melanogaster. Nat Methods **5:** 49-51.

NI, J. Q., R. ZHOU, B. CZECH, L. P. LIU, L. HOLDERBAUM *et al.*, 2011 A genome-scale shRNA resource for transgenic RNAi in Drosophila. Nat Methods **8:** 405-407.

PERRIMON, N., J. Q. NI and L. PERKINS, 2010 In vivo RNAi: today and tomorrow. Cold Spring Harb Perspect Biol **2:** a003640.

QU, W., C. REN, Y. LI, J. SHI, J. ZHANG *et al.*, 2011 Reliability analysis of the Ahringer Caenorhabditis elegans RNAi feeding library: a guide for genome-wide screens. BMC Genomics **12:** 170.

RODRIGUEZ, J., J. S. MENET and M. ROSBASH, 2012 Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila. Mol Cell **47:** 27-37.

ROOT, D. E., N. HACOHEN, W. C. HAHN, E. S. LANDER and D. M. SABATINI, 2006 Genome-scale loss-of-function screening with a lentiviral RNAi library. Nat Methods **3:** 715-719.

SKINNER, M. E., A. V. UZILOV, L. D. STEIN, C. J. MUNGALL and I. H. HOLMES, 2009 JBrowse: a next-generation genome browser. Genome Res **19:** 1630-1638.

VERT, J. P., N. FOVEAU, C. LAJAUNIE and Y. VANDENBROUCK, 2006 An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinformatics **7:** 520.

YAMAMOTO, M. T., 2010 Drosophila Genetic Resource and Stock Center; The National BioResource Project. Exp Anim **59:** 125-138.