

UP-TORR: Online Tool for Accurate and Up-to-Date Annotation of RNAi Reagents

Yanhui Hu,^{*,†} Charles Roesel,^{†,*} Ian Flockhart,^{*,†} Lizabeth Perkins,^{*,†}
Norbert Perrimon,^{*,§} and Stephanie E. Mohr^{*,†,1}

^{*}Department of Genetics and [†]*Drosophila* RNAi Screening Center, Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, [‡]Bioinformatics Program, Northeastern University, Boston, Massachusetts 02115, and [§]Howard Hughes Medical Institute, Boston, Massachusetts 02115

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.

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 *Caenorhabditis elegans*, have made contributions to a number of areas of study (Kamath *et al.* 2003; Dietzl *et al.* 2007; Boutros and Ahringer 2008; Mohr *et al.* 2010; Perrimon *et al.* 2010; Qu *et al.* 2011; Mohr and Perrimon 2012). 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 signif-

icant 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

Copyright © 2013 by the Genetics Society of America
doi: 10.1534/genetics.113.151340

Manuscript received March 21, 2013; accepted for publication May 24, 2013

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151340/-/DC1>.

¹Corresponding author: New Research Building Room 336, 77 Avenue Louis Pasteur, Boston, MA 02115. E-mail: smohr@hms.harvard.edu

(siRNAs), endoribonuclease-prepared short interfering RNAs (esiRNAs), or plasmid- or viral-encoded short hairpin RNAs (shRNAs) (Root *et al.* 2006; Kittler *et al.* 2007; Micklem and Lorenz 2007). 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 (Dietzl *et al.* 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, ~90% of annotated *Drosophila* genes are targeted by fly RNAi collections from the Vienna *Drosophila* RNAi Center (VDRC), National Institute of Genetics (NIG) RNAi Resources in Japan, and the Transgenic RNAi Project (TRiP) at Harvard Medical School (Dietzl *et al.* 2007; Ni *et al.* 2008, 2009, 2011; 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. Reevaluations of existing RNAi libraries have shown that by the time of reanalysis, 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 *C. elegans* RNAi feeding library made available in 2003, for example, reanalysis in 2011 revealed that 18% of reagents needed to be reannotated (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 ~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 genes and transcripts 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, 2012), DSIR for siRNAs (Vert *et al.* 2006; Filhol *et al.* 2012), 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 release 5.19 from July 2009). NEXT-RNAi was designed to be integrated into a back-end 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 vs. 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/releases/current-production-release/species/c_elegans/PRJNA13758/); 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.flyrna.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/>);

VDRRC catalog for VDRRC RNAi transgenic lines (<http://stockcenter.vdrc.at/control/fullCatalogueExcel>).

Data annotation pipeline

The data annotation pipeline (Figure 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 National Center for Biotechnology Information (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 the cDNA library and on-target/off-target gene search. (3) A module that processes RNAi reagent information from each source 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 Generic Feature Format 3 (GFF3) 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: 27 bp or longer with $\geq 98\%$ identity; cutoff for off-target search: 15-bp 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

Reference genes are “moving targets” that change over time

For *D. 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 *D. melanogaster* genome published in 2000, four subsequent genome assemblies, with the most recent one in February 2007, have occurred (Myers *et al.* 2000; Celniker *et al.* 2002; Hoskins *et al.* 2007). 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 multicellular species, our knowledge of the fly genome and proteome continues to improve. Indeed, since the availability of the fifth genome assembly (i.e., over the last 6 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 vary 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 1 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 $\sim 30\%$ 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 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 5 years, see [Supporting Information, Table S1](#).

Dynamic annotation of RNAi reagents

When a large amount of information is involved (in this case, information surrounding the sequence and targets of

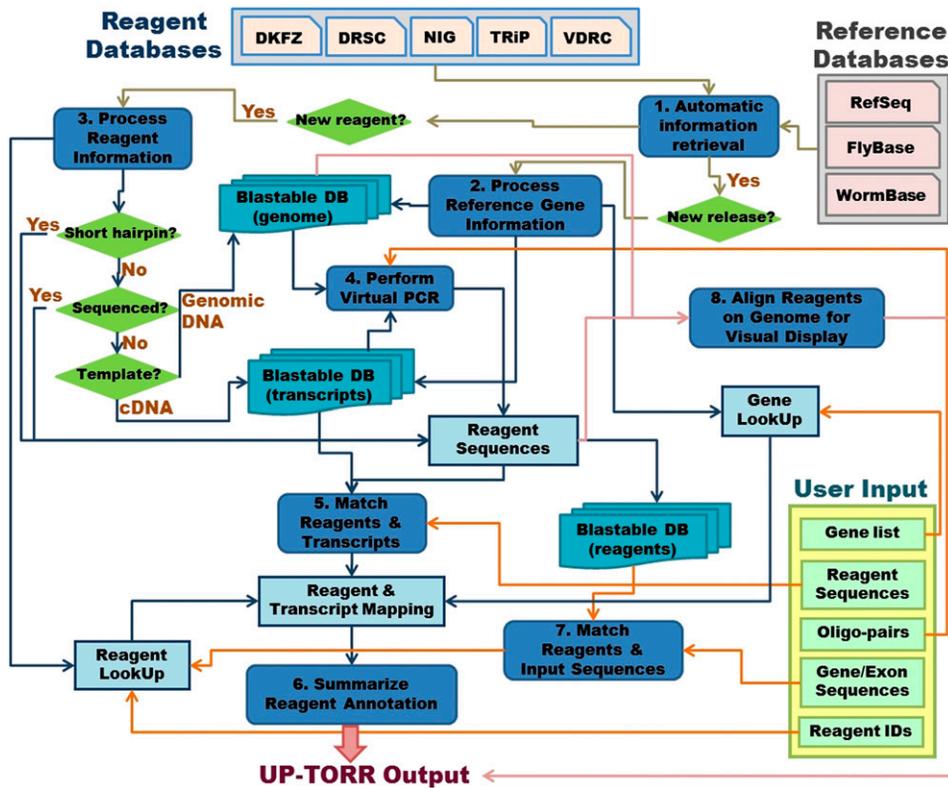


Figure 1 UP-TORR annotation pipeline. Features of the pipeline include automated daily checks for new reagents and gene annotations from the relevant public sources. When updates are available, the information is downloaded and used to rebuild the underlying databases and lookup tables, allowing UP-TORR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

RNAi reagents), the typical approach is to use a back-end database to store the information. At the DRSC, the back-end storage is a relational MySQL database (Flockhart *et al.* 2006, 2012) 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 subset 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 **u**pdated **t**argets 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 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 antistrand sequences, which originated as synthetic oligonucleotides, can be directly BLASTed against transcript sequences (Figure 1).

During the reagent “live reannotation” 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? (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 reannotated 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 reannotation 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 (Figure 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 (Figure 2B). In addition, 0.8% of reagents do not target any genes in the release we were testing. They aligned to introns (Figure 2C), intergene regions (Figure 2D), or pseudogenes (Figure 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 (Figure 2F), whereas others could be improved (Figure 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 reannotation, 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/antisense strand corresponding to a siRNA or short hairpin, or a DNA sequence corresponding to a dsRNA); (2) enter PCR primers for dsRNA,

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

RNAi collection	Reagent type	All reagents	Target 1 gene, all isoforms	Target 1 gene, not all isoforms	Target multiple genes	Reagents with > 5 OTEs (19 bp)	Target pseudogene	No gene target
DKFZ	dsRNA, cell based	20,016	15,948 (80%)	816 (4%)	1466 (7%)	394 (2%)	78 (0.4%)	1708 (9%)
DRSC-Genome Library	dsRNA, cell based	24,037	19,615 (82%)	1664 (7%)	979 (4%)	552 (2%)	84 (0.4%)	1695 (7%)
DRSC-Followup Library	dsRNA, cell based	9448	8519 (90%)	470 (5%)	296 (3%)	15 (0.2%)	14 (0.2%)	149 (2%)
NIG	dsRNA, transgenic fly	11,725	10,328 (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%)
VDRG-GD Library	dsRNA, transgenic fly	21,808	18,607 (85%)	962 (4%)	1357 (6%)	745 (3%)	60 (0.3%)	822 (4%)
VDRG-KK Library	dsRNA, transgenic fly	10,748	9135 (85%)	431 (4%)	378 (4%)	67 (1%)	27 (0.3%)	777 (7%)
Ahringer Library	dsRNA, worm feeding	16,256	11,002 (68%)	678 (4%)	1733 (11%)	1074 (7%)	283 (2%)	2843 (16%)

DKFZ, German Cancer Research Center; DRSC, *Drosophila* RNAi Screening Center; NIG, National Institute of Genetics (Japan); TRiP, Transgenic RNAi Project; VDRG, Vienna *Drosophila* RNAi Center.



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 1–3, 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 (Figure 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 reannotation 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 hyperlinked 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 DRSC Integrative Ortholog Prediction Tool (DIOPT) (flyrnai.org/diopt) or DIOPT diseases and traits (DIOPT-DIST) 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 reinterpretation 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 (Table S1), 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 reannotation (i.e., they were no longer predicted to be gene specific and isoform nonspecific with regards to the intended target gene). By comparing the different FlyBase releases, we further found that the coding CDS are less likely to change as compared with untranslated 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 reannotation of RNAi reagent information, the tool will be valuable to anyone interested in designing, analyzing, or reanalyzing RNAi results, including results from high-throughput screens. We recognize, however, that results from UP-TORR or any other 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 nonspecific 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 (Figure S1) and 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., Transcription Activator-Like Effectors

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 *elf-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 *elf-2gamma* or *Su(var)3-9*. (C–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 intergene 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.

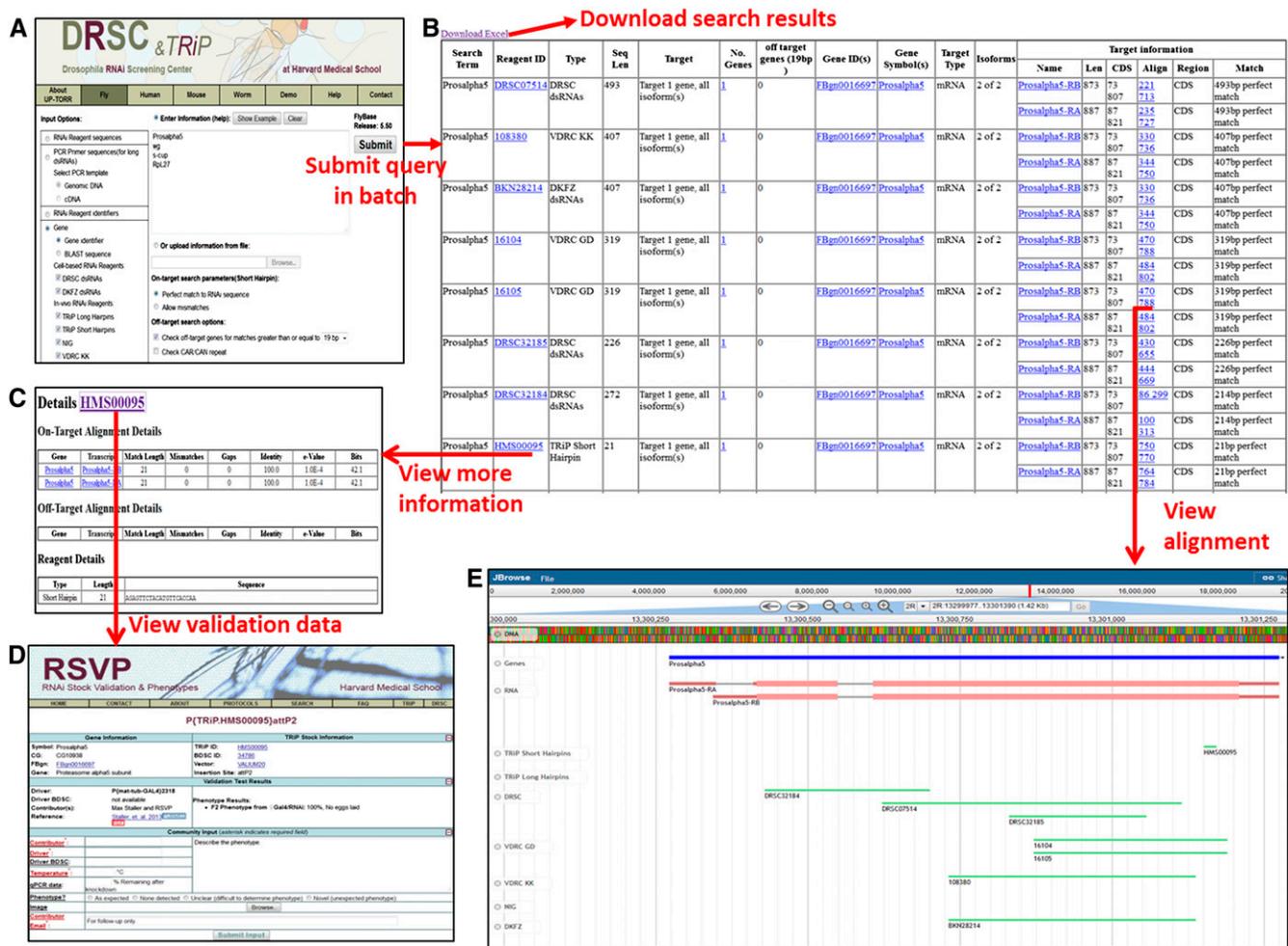


Figure 3 UP-TORR user interface. At UP-TORR, the user first specifies a species (*Drosophila*, *C. elegans*, mouse, or human) by selecting the appropriate menu tab. The example shown is for fly genes (see Figure S1 for an example using human RNAi reagent sequences). (A) At the appropriate starting page, 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 1); (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.

(TALE) (Christian *et al.* 2010) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) (Cong *et al.* 2013)) for which gene annotation impacts interpretation of the reagents.

Acknowledgements

We thank the members of the *Drosophila* RNAi Screening Center (DRSC), Transgenic RNAi Project (TRiP), and Perrimon lab, in particular Laura Holderbaum and Dong Yan, for helpful comments and support. We thank Thomas Micheler from the Vienna *Drosophila* RNAi Center 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 the National Institute of General Medical Sciences R01 GM067761 and GM084947, with additional

support from the National Center for Research Resources and Office of Research Infrastructure Programs (NCR/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.

Note added in proof: See Kulathinal, 2013 (pp. 7–8) in this issue, for a related work.

Literature Cited

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–W588.
 Boutros, M., and J. Ahninger, 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. Schnorrrer, 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–D494.
- 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–W339.
- 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., 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., 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., 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.

Communicating editor: L. M. McIntyre

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.151340/-/DC1>

UP-TORR: Online Tool for Accurate and Up-to-Date Annotation of RNAi Reagents

**Yanhui Hu, Charles Roesel, Ian Flockhart, Lizabeth Perkins,
Norbert Perrimon, and Stephanie E. Mohr**

DRSC & TRiP
Drosophila RNAi Screening Center at Harvard Medical School

About UP-TORR | Fly | Human | Mouse | Worm | Demo | Help | Contact

Input Options:

- RNAi Reagent sequences
- PCR Primer sequences (for long dsRNAs)
- Select PCR template
- Genomic DNA
- cDNA

Enter information (help):

Human RefSeq Release: 59

Sequences of human RNAi reagents

```

1 ACCTCTTCGACTGTGACTTIG
2 ATTGCAGTCATGAAGAACCA
3 CCTACATCCCTGCTACTTCA
4 GTCCGATGAATGTACTGCTT
5 CTTCAATGTTTCAGATGCTCT
6 AGGGCTGGAATCTGCTACAA
7 ACGAAGATGTGCTACTTATA
8 AGAAGAGCGAAGGACGCTACT
9 TGAGAAATACCGTGACTGCCT
  
```

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

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
1		21	Target 1 gene, all isoform(s)	1	0	1869	E2F1	mRNA	1 of 1	E2F1	2722	141 1454	1408 1428	CDS	21bp perfect match
2		21	Target 1 gene, not all isoform(s)	1	0	4094	MAF	mRNA	2 of 4	MAF	2656	824 2035	2084 2104	3UTR	21bp perfect match
3		21	Target 1 gene, all isoform(s)	1	0	27232	GNMT	mRNA	1 of 1	GNMT	1091	7 894	848 868	CDS	21bp perfect match
4		21	Target 1 gene, all isoform(s)	1	0	56937	PMEP1	mRNA	5 of 5	PMEP1	4580	190 903	4087 4107	3UTR	21bp perfect match
										PMEP1	4934	394 1257	4441 4461	3UTR	21bp perfect match
										PMEP1	4663	207 986	4170 4190	3UTR	21bp perfect match
										PMEP1	4645	255 968	4152 4172	3UTR	21bp perfect match
										PMEP1	4587	152 910	4094 4114	3UTR	21bp perfect match
5		21	Target 1 gene, not all isoform(s)	1	0	2260	FGFR1	mRNA	6 of 10	FGFR1	5895	943 3405	1287 1307	CDS	21bp perfect match
										FGFR1	5917	943 3411	1287 1307	CDS	21bp perfect match
										FGFR1	5375	324 2885	767 787	CDS	21bp perfect match
										FGFR1	5963	1035 3473	1355 1375	CDS	21bp perfect match
										FGFR1	5215	263 2725	607 627	CDS	21bp perfect match
										FGFR1	5911	943 3405	1287 1307	CDS	21bp perfect match

Link gene to NCBI EntrezGene

Link transcript to NCBI RefSeq

NCBI Entrez Gene page for E2F1

Gene: E2F1 [Homo sapiens (human)]

Official Symbol: E2F1

Official Full Name: E2F transcription factor 1

Primary source: HGNC:3112

Gene type: protein coding

RefSeq status: REVIEWED

Language: English

Also known as: RBP28, E2F-1, RBP21, RBP22

Summary: The protein encoded by this gene is a member of the E2F family of transcription factors. The E2F family plays a crucial role in the control of cell cycle and action of tumor suppressor proteins and is also a target of the transforming proteins of small DNA tumor viruses. The E2F proteins contain several evolutionarily conserved domains found in most members of the family. These domains include a DNA-binding domain, a dimerization domain which determines interaction with the differentiation regulated transcription factor proteins (DRF), a transcription domain enriched in acidic amino acids, and a tumor suppressor protein association domain which is embedded within the transcription domain. This protein and another 2 members, E2F2

NCBI RefSeq page for E2F1 mRNA

Gene: Homo sapiens E2F transcription factor 1 (E2F1), mRNA

NCBI Reference Sequence: NM_005225.2

Class: mRNA

LOCUS: NM_005225 2722 bp mRNA linear PRE 20-APR-2013

DEFINITION: Homo sapiens E2F transcription factor 1 (E2F1), mRNA.

ACCESSION: NM_005225

VERSION: NM_005225.2 0110080109

KEYWORDS: Homo sapiens (human)

ORFID: 121222

REFERENCE: GenBank: U00001.1 (1981) The action-tumor suppressor protein p53 is frequently overexpressed in high-grade lipomas and its expression is positively correlated to that of c-Myc and E2F1.

REFERENCE: Wang, H. R., Shi, Y., Wang, P. R., Wang, W. M., Shi, J. Y., Wang, J. Q., Wang, J. P., Shi, Y. D., Shi, C. F., Shi, Y. R., Wang, J. J., Wang, J. S., Shi, P. C. (2013) E2F1 transcription factor 1-dependent apoptosis pathway regulates E2F1 transcription factor 1-dependent apoptosis pathway. *Biotechnology* 31 (1): 1008-1017 (2013)

REFERENCE: GenBank: U00001.1 (1981) The action-tumor suppressor protein p53 is frequently overexpressed in high-grade lipomas and its expression is positively correlated to that of c-Myc and E2F1.

Figure S1 UP-TORR user interface for human RNAi reagents. For human RNAi reagents, UP-TORR supports input of the gene-specific region from one or more specific RNAi reagent sequence (e.g. siRNA or shRNA 21-mer sequences). (A) The user first selects human tab, then enters the gene-specific sequence regions of RNAi reagents; (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) Genes are hyperlinked to NCBI EntrezGene page. (D) Transcripts are hyperlinked to NCBI RefSeq page.

Table S1 Summary of changes in gene and transcript annotations at FlyBase and WormBase over the past six years.

FlyBase release	r5.3	r5.6	r5.16	r5.26	r5.34	r5.44	r5.50
Date	2007- Aug	2008- Mar	2009- Mar	2010- Mar	2011- Mar	2012- Mar	2013- Mar
All genes	15096	15089	15073	14721	15183	15569	16093
Genes added (compared with r5.50)	2929	2833	2414	2148	1348	722	NA
Genes retired (compared with r5.50)	1932	1829	1394	776	588	371	NA
Genes with different genome location (compared with r5.50)	10487	10508	9985	9372	8811	7728	NA
All transcripts	21707	21984	22661	22938	24523	27055	30933
Transcripts added (compared with r5.50)	13674	13211	11572	10447	8395	5195	NA
Transcripts retired (compared with r5.50)	4498	4262	3300	2452	2124	1498	NA
Transcripts with different sequence (compared with r5.50)	13730	14040	14098	13665	13471	12889	NA
Transcripts with different CDS sequence (compared with r5.50)	47	4	2	1	0	1	NA
Genes with any transcript changes (compared with r5.50)	10095	10200	10016	9613	9247	8258	NA
WormBase release	WS170	WS190	WS200	WS212	WS225	WS230	WS237
Date	2007- Feb	2008- May	2009- Mar	2010- Mar	2011- Apr	2012- Mar	2013- Apr
All transcripts	27322	27417	27687	28278	30745	31249	28983
Transcripts added (compared with WS237)	8574	7128	6672	5538	4483	2928	NA
Transcripts retired (compared with WS237)	6913	5562	5376	4833	6245	5194	NA
Transcripts with different CDS sequence (compared with WS237)	3162	2440	2048	1603	1077	618	NA