FlyRNAi: the *Drosophila* RNAi screening center database

Ian Flockhart[1,*], Matthew Booker[1], Amy Kiger[3], Michael Boutros[4], Susan Armknecht[1], Nadire Ramadan[1], Kris Richardson[1], Andrew Xu[1], Norbert Perrimon[1,2] and Bernard Mathey-Prevot[1]

[1]Department of Genetics and [2]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA, [3]Department of Biological Sciences, University of California-San Diego, La Jolla, CA 92093-0346, USA and [4]German Cancer Research Center, D-69120 Heidelberg, Germany

ABSTRACT

RNA interference (RNAi) has become a powerful tool for genetic screening in *Drosophila*. At the *Drosophila* RNAi Screening Center (DRSC), we are using a library of over 21 000 double-stranded RNAs targeting known and predicted genes in *Drosophila*. This library is available for the use of visiting scientists wishing to perform full-genome RNAi screens. The data generated from these screens are collected in the DRSC database (http://flyRNAi.org/cgi-bin/RNAi_screens.pl) in a flexible format for the convenience of the scientist and for archiving data. The long-term goal of this database is to provide annotations for as many of the uncharacterized genes in *Drosophila* as possible. Data from published screens are available to the public through a highly configurable interface that allows detailed examination of the data and provides access to a number of other databases and bioinformatics tools.

INTRODUCTION

In recent years, we have witnessed the wide application of high-throughput screening (HTS) technologies to approach biological questions. Arguably, the most promising HTS approach for discovering gene function is based on RNA interference (RNAi). RNAi results in the silencing of a gene through the specific degradation of its mRNA, which is triggered by double-stranded RNA (dsRNA) fragments complementary to that transcript. In *Drosophila*, RNAi can be achieved in cell lines and primary cells simply by adding long dsRNA to the medium (1,2). The long dsRNA are taken up by the cells, and are rapidly processed to 21–24 nt short-interfering RNAs (siRNAs) that guide the specific degradation of target mRNAs. The biochemical steps involved in the processing of dsRNAs and loading of the siRNAs into the RNA-induced silencing complex (RISC) are carried out by a number of proteins, including members of the Dicer protein family (3). In the last 2 years that technique has been adapted to HTS, allowing genome-wide screens to be performed efficiently in 384-well assay plates (1). With the support of a grant from NIGMS, we created in May 2003 the *Drosophila* RNAi Screening Center (DRSC) (http://flyRNAi.org) with the following goals:

(i) Make genome-wide RNAi screening technology available to the scientific community by providing a facility with the required infrastructure and expertise.

(ii) Provide a common platform to diminish variables between screens, allowing for functional comparisons across studies.

(iii) Create a database, in a standardized format, for the repository of results from all screens, which, upon publication, are made available to the public. The public database is divided into sections that offer researchers several basic data viewing options as well as a number of bioinformatic tools and links to other databases. The long-term purpose of the database is to provide experimental information to functionally annotate uncharacterized genes in *Drosophila*, to give a comprehensive list of genes involved in distinct cell biological and cell signaling processes and to be a resource for data mining by the scientific community.

SCREENING AT THE DRSC

Visiting scientists typically perform their screens in duplicate—screening against two full-genome sets. Raw data from duplicate genome sets are collected along with phenotype and 'hit' information. Data are primarily stored by plate and well, rather than by dsRNA, which allows for comparison

between genome sets and facilitates plate-wide analysis. The scientists perform their screens blind—only learning the identity of the dsRNA in a particular well after entering the data from that well into the system. Once the screen is completed, the data are held privately until the results are published or a 2 year period passes after completion of the screen.

The scientists have password-protected accounts which give them access to data entry interfaces and direct links to their personal data, both published and unpublished. The scientist's data are broken down by assay. Any changes the scientist may make to the experimental data display (as described below) are saved between sessions. The logged in user also has access to some tools for viewing data a plate at a time, direct links to the bioinformatic tools (listed below) and functions for directly querying the quality control (QC) information for the source plates.

An important aspect of how the database is organized is in the distinction among genes, amplicons and dsRNAs. The dsRNA library was designed in collaboration with R. Paro's group and collaborators (ZMBH Heidelberg) (4). The approach taken was to generate gene-specific primers to 21 396 putative open reading frames (ORFs) covering the entire *Drosophila* genome. Choice of the primers was based on the combined genome annotations available from BDGP/Celera (5) and the Sanger Center. A pair of a specific forward and reverse primer was used to amplify a genomic region (henceforth called amplicon) corresponding to each predicted ORF. Each amplicon, flanked by RNA polymerase T7 promoters, was in turn used as a template in an *in vitro* transcription reaction to generate dsRNA. As new releases of the *Drosophila* genome result in slightly revised annotations, the predicted gene target of each amplicon may change accordingly. Because of this unavoidable issue, result and 'hit' information from screens is focused more on amplicons and the corresponding dsRNA than on genes. However, the key piece of information that remains invariant is the nucleotide sequence of the specific region encompassed by an amplicon and by extension of its related dsRNA. For the purpose of data tracking in the database, the term amplicon can refer to two distinct biological entities that are related through their sequences: the DNA fragment amplified with specific primers or its corresponding dsRNA.

WEBSITE OVERVIEW

The main URL for the public database is http://flyRNAi.org/cgi-bin/RNAi_screens.pl.

This page has four major parts; a menu bar to the left, a link to the Gene Lookup Page (Search for Genes in Public Screens), a list of public screens below it for which all data are accessible, and a list of ongoing screens for which the data are kept confidential until the time of publication or the 2 year limit after their completion, whichever comes first (Figure 1).

The menu bar provides links to a number of informational resources. The 'About Us' link provides general information about the DRSC, such as personnel, location, equipment, funding and DRSC news. The 'Screening' header opens up a 'how-to' section and summarizes the current protocols in use at the DRSC to conduct RNAi screens in the 384-well plate format (1). The 'Applications' link is for scientists interested in submitting an application to come and carry out a screen at the DRSC. The 'Literature' and 'Links' pages are lists of external resources that the DRSC wishes to highlight for RNAi screeners. The 'Tools' section consists of a number of small bioinformatics applications that we have developed to help screeners or interested scientists search for and manage information displayed in the Data section of the database. It also offers links to other databases with the purpose of providing additional information on particular genes or gene function.

The published screens page is organized by screen publication date, the most recent of which are listed first. Each screen field consists of the title and the authors of the screen. A pdf file of the publication, as well as the supplementary data (when available), are included whenever possible. One or more direct links are provided to access the raw data, with a listing of the dsRNAs (or their corresponding amplicons) found to have an effect in the assay under study. Immediately below the record, which is updated as is appropriate, of the published screens (6–13) follows a list of screens that have just been completed at the DRSC. These screens are yet to be fully analyzed and have not yet reached the stage of publication. For each unpublished screen listed, the title and brief summary of the screen are given along with details about the scientists involved in the screen (names, academic affiliation, etc.). A contact email address is provided for each screen.

EXPERIMENTAL DATA INTERFACE

Each screen has links to one or more listings of the amplicons for which phenotypic data was entered during that assay. Multiple listings appear in cases when several different assays were combined in the screen, such as when two different cell lines are screened for comparison. Each listing is organized in rows representing DRSC amplicons. The data being displayed are fully configurable by the user (Figure 2).

The user can use search criteria to display a subset of rows, or use check boxes to indicate the rows to display when a redisplay button is pressed. More significantly, the user can sort the data, based on any displayed column of information, by clicking the column header. The set of columns that is displayed is likewise configurable via a selection menu at the bottom of the page, which offers a choice of over 40 columns of information per row (Table 1). When the page displays data to the user's liking, the displayed contents of the page can be saved as a tab-delimited text file, to be imported into the user's spreadsheet program of choice.

For further details about any row, the user can click the triangle icon at the start of the row to go to the amplicon detail page for that amplicon.

INFORMATIC TOOLS

This page provides a link to the Gene and Amplicon Lookup page (Figure 3). The Gene and Amplicon Lookup function provides a simple interface that allows visitors to get information on any amplicon in the DRSC library related to a particular gene. The search function is flexible and accepts queries by gene name, gene alias, FlyBase gene number (FBgn), CG accession number or by our internal DRSC ID number. When a gene is queried, the user is presented with a detailed page of information about the amplicon(s) found

Figure 1. The DRSC Data page. Published works are bordered in black. Screens awaiting publication are listed below, bordered in gray.

and any gene(s) targeted by it. A gene may have more than one amplicon as the original collection was based on early annotations of the *Drosophila* genome (5). Subsequent revision of the annotation resulted in several cases in the merging of two ORFs into a single gene. Currently, 1365 genes are targeted by more than one amplicon in our collection. If any amplicon targeting a queried gene has been associated with a phenotype in any public screens, then that information is presented at the top of the page.

The top section of the results page shows whether a particular gene and its related amplicon(s) have been identified in a public screen as 'hit'(s). If so, the targeting amplicon is listed with the name of the screen, the phenotype and the screener's evaluation of the strength of that hit (see also Table 1). Below that, there is a section for each amplicon in the DRSC library that targets that gene. It is broken into subsections detailing general information about the amplicon (predicted sequence, primer sequences, length, etc.), genomic position, detailed information about the gene, links to other databases and some historical information relating to the creation of the amplicon.

An additional important tool is the off-target sequence search tool. As in the mammalian field, the issue of off-target effects caused by siRNAs is emerging as a significant issue in *Drosophila* (14,15) and potential off-targets associated with dsRNAs for *Caenorhabditis elegans* have been annotated at RNAiDB (16). An initial review of the data at the DRSC confirms that off-target effects do happen in *Drosophila* and need to be taken into account when interpreting knock-down data by long dsRNAs (M. Booker, S. Silver, M. Kulkarni, A. Friedman, N. Perrimon and B. Mathey-Prevot, manuscript in preparation). As discussed earlier, long dsRNA (typically 400 nt) are processed to 21–23 nt siRNAs by the Dicer protein. Dicer does not appear to have a sequence preference for where processing will occur and as a result we do not know in advance which and how many siRNAs are produced from any particular dsRNA. However, we can check any possible 21mer sequence that is included in a given dsRNA (or within its corresponding amplicon) for a possible match with other mRNAs which are not the intended target. Ideally, only one match corresponding to the targeted mRNA should
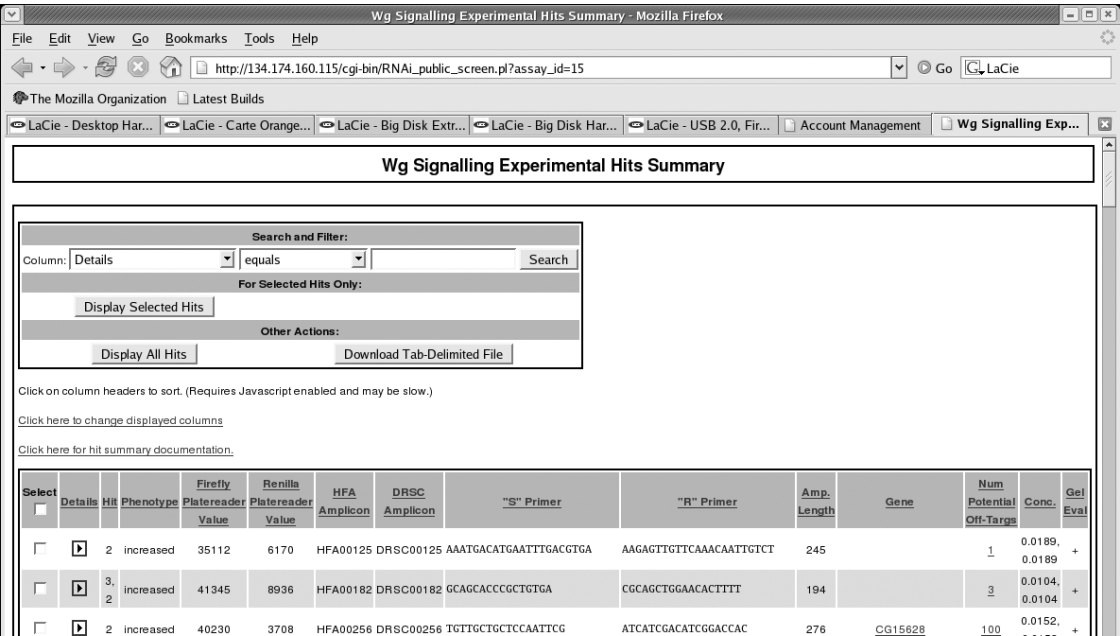
Figure 2. The Experimental Data Summary page. The columns shown are a typical example of the data that may be displayed. Data columns are chosen by the user.
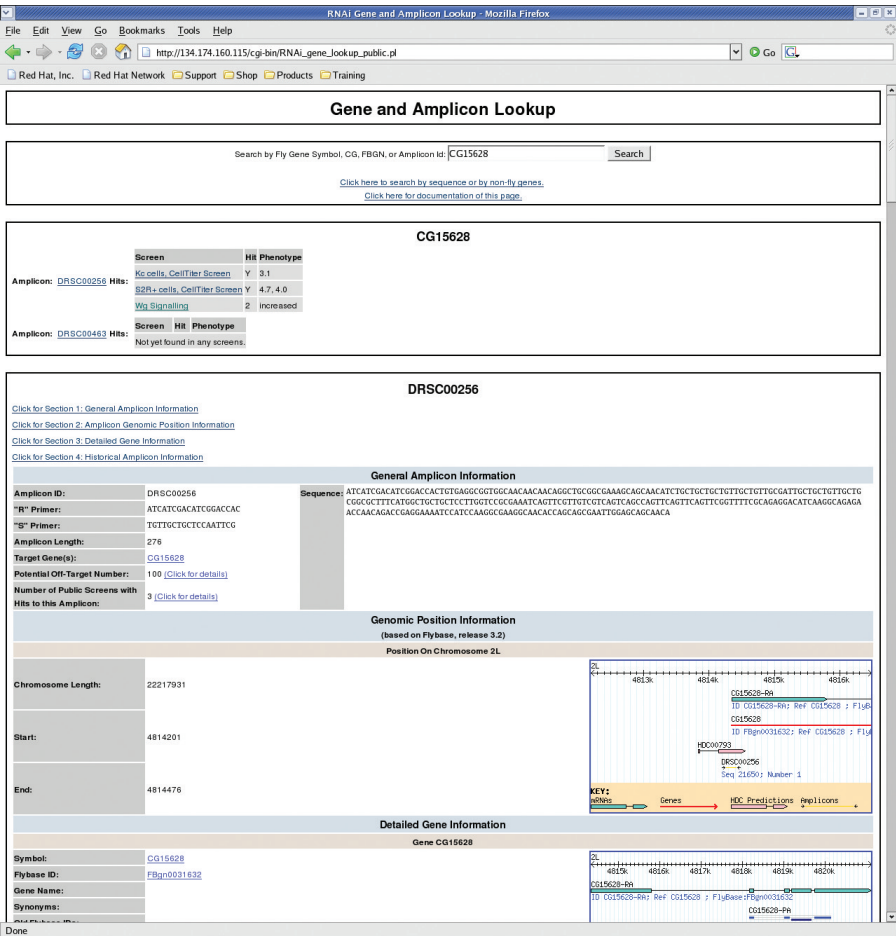

Figure 3. Gene Lookup Results page. This master page of amplicon information is available from the Gene and Amplicon Lookup and Data Display pages.

Table 1. Definitions of columns

Column	Information
Results information	
Hit	This indicates that this amplicon was a positive result in the screen
Phenotype	The result or phenotypic information that the scientist recorded
Amplicon information	
DRSC amplicon ID	An identifier assigned to the amplicon by the DRSC
'S' primer	The primer at the 5′ end of the ORF this amplicon was designed to target
'R' primer	The primer at the 3′ end of the ORF this amplicon was designed to target
Amplicon length(s)	The predicted length in base pairs of the amplicon
HDC target	The predict ORF the amplicon was designed to target
Hits in public screens	The number of screens which report this amplicon as a hit
Chromosomal location information	
Number of chromosome locations	The number of sequences the primer pair for the amplicon is predicted to amplify
Chromosome locations	Details regarding the genomic positions of the sequences the primer pair for the amplicon is predicted to amplify
Chromosome	The chromosome(s) which contains the sequence targeted by the amplicon
Chromosome start	The genomic position of the first nucleotide of the amplicon
Chromosome stop	The genomic position of the last nucleotide of the amplicon
Gene information	
Gene symbol	The symbol or name of the gene targeted by this amplicon
FlyBase gene ID	The FlyBase FBgn identifier of the gene targeted by this amplicon
Deprecated FlyBase gene IDs	Former FBgn identifiers of the gene targeted by this amplicon
Gene name	The gene name of the gene targeted by this amplicon according to FlyBase
Gene synonyms	Alternative names and symbols
NCBI nucleotide accessions	Accession numbers of nucleotide sequences in GenBank of the gene targeted by this amplicon
Entrez gene identifiers	An Entrez Gene identifier of the gene targeted by this amplicon
UniGene identifiers	A UniGene identifier of the gene targeted by this amplicon
RefSeq RNA accessions	A RefSeq identifier of the gene targeted by this amplicon
HomoloGene group identifiers	A HomoloGene group identifier of the gene targeted by this amplicon
Homo sapiens gene(s) via HomoloGene	The human orthologue of the gene targeted by this amplicon according to HomoloGene
Mus musculus gene(s) via HomoloGene	As above for mouse
Caenorhabditis elegans gene(s) via HomoloGene	As above for nematode
Saccharomyces cerevisiae gene(s) via HomoloGene	As above for yeast
Homo sapiens putative orthologues according to inParanoid	The human orthologue of the gene targeted by this amplicon according to InParanoid
Mus musculus putative orthologues according to inParanoid	As above for mouse
Caenorhabditis elegans putative orthologues according to inParanoid	As above for nematode
Saccharomyces cerevisiae putative orthologues according to inParanoid	As above for yeast
Miscellaneous organism putative orthologues according to inParanoid	As above for miscellaneous organisms
Homophila—possible human disease homologs	A link to Homophila which provides a fly to human disease gene cross-reference
Gene Ontology	GO identifiers for the gene targeted by this amplicon
Protein domains	Protein domains contained in the gene targeted by this amplicon
Amplicon 21 bp fragment information	
Number of potential off-targets based on 21 bp fragments	Self-explanatory
Maximum continuous sequence overlap of potential off-target(s)	Maximum sequence length of a match to an off-target
Quality control information	
RNA concentration	RNA concentration provided for quality control purposes
Gel evaluation	Gel-based quality control assay result
PCR evaluation	PCR-based quality control assay result
PCR T7 evaluation	PCR-based quality control assay result
Gel eval comment	Additional comments about the gel-based quality control assay result

be found. To facilitate this search, we have developed a bioinformatics tool based on our own faster algorithm, somewhat similar to that published by Arziman *et al.* (17) except that it does not have a built-in primer design component. Our off-target search tool allows a user to provide one or more DNA sequences in FASTA format and search those sequences for predicted off-targets among all fly gene transcripts. The user can specify an off-target length (16–50 bp) with a default value of 21 bp. A color-coded map of gene matches for a given sequence is returned to the user. The map shows regions of the submitted sequences that are devoid of predicted matches with genes other than the intended target (no off-target) as well as stretches which do have matches with off-target

genes. The intended or primary target is determined based on a match over all or most of the length of submitted sequence. The tool also reports the number of off-target genes.

The sequence extraction tool allows users to retrieve multiple FASTA-formatted DRSC amplicon sequences. It can also be used to retrieve fly gene sequences.

The FlyBase Identifier retrieval tool allows the user to do a batch query for FlyBase FBgn identifiers by giving a list of fly gene symbols, names, synonyms and CG accessions.

The genetic interactions tool allows the user to construct a graph of genes of interest for their reported genetic interactions based on data stored at FlyBase. The user may query for these

by submitting one or more gene symbols, synonyms or FBgn identifiers.

The Screen Analysis Tutorial is a series of web pages that provides screen analysis assistance to DRSC screeners as well as to public users interested in mining the data available in our database. This guide offers the user multiple resources and some graphical approaches to help integrate and explore relational associations between DRSC hits and other gene function, ontology or expression data sources.

FUTURE DEVELOPMENTS

The DRSC database/interface is constantly evolving and new experimental data accrue at a pace of ∼20–30 genome-wide screens a year. The DRSC library of amplicons (and dsRNAs) is also evolving as new *Drosophila* genome annotations come on line, and new insights about the specificity of our dsRNAs come to our attention (e.g. the need to replace amplicons associated with off-target effects). We are committed to the idea that our database be an important resource for public data mining and will make available all screen data as soon as it is permitted by the general agreement signed between the screeners and the DRSC. We will work to provide additional bioinformatic tools and search capabilities to enhance our current database, and we welcome any suggestion or collaboration to improve the integration of our database with others. Finally, we encourage comments to make our database more useful to scientists and hope that similar databases will be created to collect information from RNAi screens in mammalian cells.

ACKNOWLEDGEMENTS

Conflict of interest statement. None declared.

REFERENCES

1. Armknecht,S., Boutros,M., Kiger,A., Nybakken,K., Mathey-Prevot,B. and Perrimon,N. (2005) High-throughput RNA interference screens in *Drosophila* tissue culture cells. *Methods Enzymol.*, **392**, 55–73.
2. Clemens,J.C., Worby,C.A., Simonson-Leff,N., Muda,M., Maehama,T., Hemmings,B.A. and Dixon,J.E. (2000) Use of double-stranded RNA interference in *Drosophila* cell lines to dissect signal transduction pathways. *Proc. Natl Acad. Sci. USA*, **97**, 6499–6503.
3. Meister,G. and Tuschl,T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature*, **431**, 343–349.
4. Hild,M., Beckmann,B., Haas,S.A., Koch,B., Solovyev,V., Busold,C., Fellenberg,K., Boutros,M., Vingron,M., Sauer,F. *et al*. (2003) An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.*, **5**, R3.
5. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
6. Agaisse,H., Burrack,L.S., Philips,J., Rubin,E.J., Perrimon,N. and Higgins,D.E. (2005) Genome-wide RNAi screen for host factors required for intracellular bacterial infection. *Science*, **14**, 14.
7. Baeg,G.H., Zhou,R. and Perrimon,N. (2005) Genome-wide RNAi analysis of JAK/STAT signaling components in *Drosophila*. *Genes Dev.*, **29**, 29.
8. Boutros,M., Kiger,A.A., Armknecht,S., Kerr,K., Hild,M., Koch,B., Haas,S.A., Consortium,H.F., Paro,R. and Perrimon,N. (2004) Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science*, **303**, 832–835.
9. Cherry,S., Doukas,T., Armknecht,S., Whelan,S., Wang,H., Sarnow,P. and Perrimon,N. (2005) Genome-wide RNAi screen reveals a specific sensitivity of IRES-containing RNA viruses to host translation inhibition. *Genes Dev.*, **19**, 445–452.
10. DasGupta,R., Kaykas,A., Moon,R.T. and Perrimon,N. (2005) Functional genomic analysis of the Wnt-wingless signaling pathway. *Science*, **308**, 826–833.
11. Eggert,U.S., Kiger,A.A., Richter,C., Perlman,Z.E., Perrimon,N., Mitchison,T.J. and Field,C.M. (2004) Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol.*, **2**, e379.
12. Kiger,A., Baum,B., Jones,S., Jones,M., Coulson,A., Echeverri,C. and Perrimon,N. (2003) A functional genomic analysis of cell morphology using RNA interference. *J. Biol.*, **2**, 27.
13. Philips,J.A., Rubin,E.J. and Perrimon,N. (2005) *Drosophila* RNAi screen reveals CD36 family member required for mycobacterial infection. *Science*, **14**, 14.
14. Qiu,S., Adema,C.M. and Lane,T. (2005) A computational study of off-target effects of RNA interference. *Nucleic Acids Res.*, **33**, 1834–1847.
15. Naito,Y., Yamada,T., Matsumiya,T., Ui-Tei,K., Saigo,K. and Morishita,S. (2005) dsCheck: highly sensitive off-target search software for double-stranded RNA-mediated RNA interference. *Nucleic Acids Res.*, **33**, W589–W591.
16. Gunsalus,K.C., Yueh,W.C., MacMenamin,P. and Piano,F. (2004) RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res.*, **32**, D406–D410.
17. Arziman,Z., Horn,T. and Boutros,M. (2005) E-RNAi: a web application to design optimized RNAi constructs. *Nucleic Acids Res.*, **33**, W582–W588.