Chapter 3

RNAi experiments in *D. melanogaster*: solutions to the overlooked problem of off-targets shared by independent dsRNAs

Erwin Seinen, Johannes G.M. Burgerhof, Ritsert C. Jansen, Ody C.M. Sibon

*Published in PLoS ONE, October 2010*
CHAPTER 3

Abstract

**Background:** RNAi technology is widely used to downregulate specific gene products. Investigating the phenotype induced by downregulation of gene products provides essential information about the function of the specific gene of interest. When RNAi is applied in *Drosophila melanogaster* or *Caenorhabditis elegans*, often large dsRNAs are used. One of the drawbacks of RNAi technology is that unwanted gene products with sequence similarity to the gene of interest can be down regulated too. To verify the outcome of an RNAi experiment and to avoid these unwanted off-target effects, an additional non-overlapping dsRNA can be used to down-regulate the same gene. However it has never been tested whether this approach is sufficient to reduce the risk of off-targets.

**Methodology:** We created a novel tool to analyse the occurrence of off-target effects in *Drosophila* and we analyzed 99 randomly chosen genes.

**Principle findings:** Here we show that nearly all genes contain non-overlapping internal sequences that do show overlap in a common off-target gene.

**Conclusion:** Based on our *in silico* findings, off-target effects should not be ignored and our presented on-line tool enables the identification of two RNA interference constructs, free of overlapping off-targets, from any gene of interest.

Introduction

Genes can be silenced using RNA interference (RNAi). This powerful method is widely used to study biological consequences induced by the down-regulation of selected genes [1-4]. Since its discovery, a great amount of valuable information has been collected using this technology. However, RNAi technology also has some drawbacks such as off-target effects [5-12]. Off-target effects are caused by short stretches of sequence similarity between the RNAi molecule and one or more genes other than the target. Because of high success rates, the fly and worm (*D. melanogaster* and *C. elegans*) model systems generally use large double strand RNAs (dsRNAs) of 300-800 bp. From (large) dsRNAs, numerous siRNAs are generated by the action of DICER and each of these can provoke an RNAi response and exert their gene down-regulating action [13]. Although this results in a favourable synergistic RNAi response towards the target gene, it may in theory also increase the number of off-target possibilities.
A straightforward method to reduce off-target effects, is to use 2 independent and non-overlapping dsRNAs to down-regulate a specific target. Because these dsRNAs are different in sequence composition, their individual off-targets are also assumed to be unique while they both silence the same on-target gene. Consequently, it is reasonable to assume that any shared phenotype which is observed after the independent use of both dsRNAs is an effect of down-regulating the on-target gene (Figure 1A). Although, this line of reasoning is rational, hypothetically it is possible that different non-overlapping siRNAs may actually target different sequences within one identical off-target gene (illustrated in Figure 1B). In such an unfortunate case, a shared off-target effect induced by 2 independent dsRNAs may be misinterpreted as an on-target effect. It has never been investigated what the occurrences are of shared off-target effects when dsRNA are randomly chosen. Here, we present a detailed analysis, based on sequence similarity and a randomized trial which suggest that most genes have independent dsRNA-spanning sequences showing sequence similarity with the same off-target gene. In addition, we present an on-line tool that allows to scan \textit{Drosophila} gene sequences for the occurrence of off-target overlapping regions and to design dsRNAs that have a reduced likelihood to induce identical off-target effects.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{A: Schematic presentation of the event in which identical phenotypes are induced because of shared on-target effects and at the same time different phenotypes are induced because of off-target effects. Phenotype 2 is due to down-regulation of the on-target gene and is induced by dsRNA1 and dsRNA2. Phenotype 1 and 3 are due to down regulation of the off-target gene X and Y respectively and are specific for the individual distinct dsRNAs. In this fortunate event, the individual off-target effects are not identical and are classified as off-target-effects; bona fide conclusions will be drawn from the outcome of this experiment. B: Schematic presentation of the event in which identical phenotypes are induced because of shared on-target effects but at the same time different phenotypes are induced because of off-target effects的发生。}  
\end{figure}
time an additional identical phenotype is induced by the use of the two independent dsRNAs caused by off-target effects. Phenotype 2 is due to down-regulating the on-target gene and is shared by dsRNA1 and dsRNA3. Phenotype 1 is due to down regulation of a shared off-target gene of the distinct dsRNAs. In this unfortunate event, the off-target effects are identical and will be classified as on-target effects; false conclusions will be drawn from the outcome of this experiment.

Results and Discussion

Statistical analysis on a randomized genome shows that it is likely that 2 distinct 21 nt sequences from the same gene can map closely elsewhere on the genome (see Supplementary file). This hypothetical event (illustrated in Figure 1B) may cause distinct dsRNAs to have common off-targets and that particular combinations of dsRNAs should therefore be avoided. These calculations are based on a non-organized genome containing random sequences, while the Drosophila genome is highly functional and far from 'randomized'. To evaluate the risks of our hypothetical event more pragmatically, we used the following approach. First, we picked a dataset of 99 random chosen genes (see supplementary table 1) from the D. melanogaster genome. We investigated the occurrence of independent dsRNAs derived from one gene to have shared off-targets. dsRNAs are often derived from cDNA so for our analysis only the cDNA of the 99 genes were considered. Because the complete cDNA can be used to design dsRNAs from, and the dsRNAs are split into siRNAs of approximately 21nt by the RNAi machinery, we first created a list of all possible siRNA sequences that can be obtained from the cDNA sequences of each of these 99 genes. This complete list was subsequently reduced using established scoring rules to exclude 21-bp siRNAs that are most likely non-active (see supplementary table 3). We like to stress that this assumption will only underestimate our findings. Next we calculated the occurrence of all siRNA derived from one cDNA to have a shared off-target with another siRNA derived from the same gene. For this analysis we included pre-mRNA sequences of the complete D. melanogaster genome because of the following published results: 1) It has been demonstrated that the RNAi machinery can target pre-mRNAs in C. elegans [14]. 2) RNA silencing components have been shown to localize in the nucleus in other organisms (including human) [15-22], further suggesting that pre-mRNAs can be targeted by the RNAi machinery. 3) RNAi constructs can be complementary to miRNAs which are often derived from introns [23] and might act like antagonirs [24]. We therefore analysed the filtered list of
siRNA sequences against both mature and pre-mature RNA sequences to map all possible off-targets with up to 3 mismatches in their sequence alignments with the use of a new tool (see Methods and http://www.RNAiSelect.info/dsrna). By doing so, a list of potential off-targets for each of the individual genes was generated. Next we analyzed whether there was overlap between the potential off-targets of siRNAs derived from the same gene (see Methods). We used the term **cot-group** (Common Off-Target group); a cot-group consists of 2 or more siRNAs, derived from a single gene, that map to the same off-target gene (also illustrated in Table 1; the lines represent members of cot-groups). The generated siRNA lists of all 99 genes were scanned individually for the presence of cot-groups. The occurrence of cot-groups appeared to be present in all genes, with sometimes excessive high frequencies (Figure 2). As expected, the number of cot-groups are highly correlated with the length of the cDNA of the gene; the larger the sequence, the more cot-groups are formed (Figure 3).

**Figure 2 - Cot-group frequency distribution for all 99 analyzed genes.** A cot-group consists of 2 or more siRNAs, derived from a single gene, that map to the same off-target gene. The x-axis shows the number of cot-groups, the y-axis shows the frequency of genes that contain that number of cot-groups. A: On average, 42 cot-groups were found per gene with a large spread (SD=48.579; N=99) because of the high correlation with gene length (see figure 3). B: Filtering for introns shows that the COT group frequencies are much lower, mostly concentrating around 1-3. Table 1 and Supplementary table 1 illustrate the cot-groups per gene. Table 2 and Supplementary table 2 illustrate the cot-groups per gene after filtering for intron targets.
Figure 3 - As expected, the length of the gene correlates with the number of cot-groups that can be formed (Spearman's correlation coefficient of 0.44; \( P < 0.001 \)). This is because the number of potential siRNAs that may originate from one cDNA sequence increases proportionately with the length of the gene. Each potential siRNA adds up on the possibility to form a cot-group. Figure 2A plots the number of cot-groups against the length of the gene. Figure 2B shows the percentage of the sequence within the analyzed genes that map to common off-targets. These values were acquired by multiplying the members of the different cot-groups with the length of a single siRNA (21-bp) within every gene and this is divided by the whole gene length. These values become less reliable with large genes, because there is an increased chance of a single siRNA being part of multiple cot-groups. In that case the area may reach beyond the 100%. Figure 2C and 2D are the results after filtering out intron targets.
Table 1 - Non-overlapping sequences have a high prevalence of sharing off-targets. Example of genes for which –based on sequence similarities- overlapping off-targets exist. The number of items per off-target group is given both in numbers and percentages. As an example: the cDNA of gene CG11372-RA contains 30 events of duplicate sequences with a shared off-target and 1 event of triplicate sequences that share the same off-target and 1 event of quadruple sequences that share the same off-target. Green lines represent sites that share an identical off-target with one other site, red lines represent sites that share an identical off-target with 2 other sites, and blue lines with 3 other sites. Purple lines represent sites that share identical off-targets with 5 or more other sites. The complete report from the 99 randomly selected genes are presented in Supplementary table 1. Note that for some genes the lines representing the off-target events are in close proximity and cannot be distinguished as separate lines in this illustrative figure. Overall, there appears to be a tendency for the occurrence of overlapping off-targets at the boundary (UTR’s) of the genes, as is evident in for example CG5834 (last gene in the list). The insert shows a more detailed illustration for the analysis of the cDNA of gene CG11620. The green vertical lines represent sites that share an identical off-target with one other site. Sites that share the same off-target are connected with dotted lines and the shared off-target (as CG number) is indicated for each pair. To avoid off-target effects, dsRNA constructs should be chosen in such a way that the dsRNA constructs do not include both members of one pair. The green boxes represent areas which do not include both members of one pair. In order to reduce the likelihood of shared-off target effects, dsRNAs should be designed using sequences from the green regions. In contrast the red areas do include both members of one pair. When 2
independent dsRNA constructs will be designed from these areas, these dsRNA constructs do share sequence similarities with the same off-target gene. Our tool provides for all the genes present in the Drosophila genome the green areas.

Table 2 – Non-overlapping sequences have a significant prevalence of sharing exon targeting off-targets. All 99 cDNA sequences were re-analyzed, now ignoring any predicted off-targets that occur within intron sequences. Although this lowers the predicted off-targets, it still shows a significant occurrence of overlapping off-targets. Our analysis shows that after applying this filter, 74% of the genes show off-targets that occur more than once within the same derived cDNA sequence.
We then looked at the general profile of the cot-groups for each gene separately and tried to deduce the required number of dsRNAs to strongly reduce the event of common off-targets. If for example the number of members within the cot-groups does not exceed 2, than this implies that at most 2 siRNAs within the same gene map to a common off-target. For that particular situation, the use of 3 or more non-overlapping dsRNA will always generate bona fide data as there is no possibility for all 3 of them to share a common off-target (bases on sequence similarity). Unfortunately, most genes have cot-groups with at least 3 members (Figure 4, also depicted by red lines in Table 1 and in Supplementary table 1) or even 4 members (Figure 4, also depicted by blue lines in Table 1 and in Supplementary table 1). This finding demonstrates that just using multiple non-overlapping dsRNAs is not sufficient to exclude off-target events (see also insert Table 1), even if the number of independent dsRNAs is 3 or more. We therefore developed a bioinformatics approach to design dsRNAs that avoids all predicted off-targets. Our freely available website presents such a tool at http://www.RNAiSelect.info/dsrna. This web based tool accepts a gene name as input and presents a number of choices each containing a combination of 2 unique dsRNAs that lack overlapping –based on sequence similarity- off-targets (Supplementary Figure 1).

Figure 4 - Distribution of the maximum cot-group size for the 99 analyzed genes. For each cot-group with the indicated number of members on the x-axis, the number of genes where counted that have cot-groups with a corresponding maximum cot-group size. This shows for example that in figure 4A, there are 19 genes that have cot-groups with no more than 2 members. Most genes (38) in our analysis appear to have at least one cot-group present with up to 3 members. Due to the gene length with cot-group correlation, some large genes in our analysis also show very large cot-groups which account for the unexpected large 7+ count as plotted in the last bar (see also Table 1 and Supplementary Table 1). Figure 4B presents the results after filtering out intron targets, showing that most genes have at least one cot-group present with 2 members.
Next, we repeated the above analysis, but now only considering off-targets targeting mature RNA sequences, because these are maybe be more active in RNAi [25]. Filtering out the intron off-targets, causes much less off-targets to be found in general per cDNA (Table 2). Overall, both the sizes and occurrences of the COT groups are smaller (Figure 2B, Figure 4B). Nevertheless, there is still a significant number of overlapping potential off-targets to be expected in >74% of the genes. In 24% of the analysed genes there is at least 1 COT groups present of size 3 (Figure 4B), meaning that there are at least 3 areas within the cDNA that target the same off-target. Therefore, even if only mature RNA sequences are considered and 2 randomly chosen non-overlapping dsRNA’s are used, the experimental outcome can be obscured by off-target effects. This further underscores the utility of our tool.

Although not exclusively, we observed a strong tendency for overlapping off-targets to occur at the end of genes (see graphical illustrations of the common off-targets in Table 1 and in Supplementary Table 1), corresponding to the untranslated regions (UTR). The UTR sequences are less unique in the genome as compared to the coding region and therefore preferably should be avoided when dsRNA constructs are designed. Our tool includes an option to avoid UTR sequences to minimize off-targets when designing dsRNAs of interest.

**Conclusion**

Our analysis demonstrated that most genes in the *D. melanogaster* genome contain 2 or more (distinct) sequences that show sequence similarity (containing 3 or less mismatches) to the same off-target gene. The potential consequence of these overlapping occurrences is that 2 dsRNAs which are generated to down-regulate a specific target gene, may possess a common off-target gene as well. In case these 2 distinct dsRNAs are used, their common phenotype induced by down-regulation of their shared off-target gene may lead to misinterpretation of the experiment. We present a method to identify 2 distinct dsRNAs from a gene of choice that do not show any off-target overlap, -based on sequence similarity- by performing a thorough off-target overlap analysis. This tool is freely available at http://www.rnaiselect.info/dsrna and may be used the *Drosophila* community where dsRNAs are generally used for gene down-regulation.
Methods

Genomic data (build 45-43b for Drosophila) were downloaded from the Ensembl website (www.ensembl.org). The data from Ensembl and its derived seed tables were processed, stored and indexed in a MySQL database, version 5.0, running on top of Ubuntu 6.06. The on-line available RNAiSelect program (http://www.RNAiSelect.info/) was written in C#.NET and performs a comprehensive sequence alignment against the input genome for up to 3 mismatches.

The RNAiSelect algorithm was specifically designed for finding relationships between short nucleotide sequences. It has a high performance and usability for short-sequence studies, including siRNA (off-)targets. The complete source code and documentation for a standalone version of this algorithm may be downloaded from http://rnaiselect.sourceforge.net/. The algorithm is based on the following assumption:

“An example sequence TTTAATTTGGGCGGG consists of 18 nucleotides and may be split into two 9-nt child sequences; TTTAATTT and GGGCGGG. By plain observation, we know that the sequence GGGCGGG is exactly 9-nt separated from TTTAATTT in the original sequence.”

For the RNAiSelect algorithm to work, we first wrote a program that generates a seed table which holds the exact genomic location(s) for every possible 9-nt sequence (4^9, or 262,144 sequences). Generating such an index is a general strategy used by many algorithms to rapidly look-up any sequence of fixed length for its positions in the genome. The used algorithm however uses a novel method to calculate the positional relationship between indexed seeds, instead of performing string-to-string comparisons for every nucleotide after a hit has been found. In other words, by searching 9-nt subsequences of the whole query sequence for consecutive matches of locations, it will find hits larger than 9nt without performing actual DNA comparisons. This following example, in layman code, shows how to find an 18-nt sequence in the genome by first splitting the sequence into its two 9-nt subsequences and comparing these sequences with the available index table with a word size of 9.
This example merely demonstrates how to find an exact 18-nt hit not allowing any mismatches. However, mismatches may be added by expanding the seed searches with variations so that all possible combinations will be found. We thus included variations of the 9-nt subsequences and then compared the distance relationship between the original locations of the seed hits, which has to be exactly 9. This may considerably increase the number of seed searches, but because these are relatively cheap in terms of processing time, the overall performance is very high while it guarantees that every possible alignment is evaluated.

The cDNA sequences from each 99 gene (Supplementary Table 1) were first analyzed for potentially active sequences as might be produced by endogenous DICER. A scoring schema was used (Supplementary Table 2) during this analysis to estimate and extract the most potential sequences. Each extracted sequence was analyzed for potential off-targets. The combined output from all these potential off-targets was cross-referenced with each other to map areas on the original cDNA sequence that are predicted to have overlapping off-targets. At the same time, regions can be identified that lack these areas and dsRNA sequences can be extracted that are completely devoid of off-targets. The results are presented and 2 or more dsRNA are indicated that originate from the same gene and that are predicted to lack overlapping off-targets. From the indicated areas, dsRNA can be designed. An identical analysis can be done for every Drosophila gene of interest through a web-interface at http://www.rnaisel ect.info/dsRNA which presents the output in a user friendly interface.

```
1  SPLIT QUERY SEQUENCE(18 nt) INTO dnacode_left(9 nt) AND dnacode_right(9 nt)
2  EXTRACT LOCATIONS FROM index_table FOR dnacode_left AND STORE IN seedtable_left
3  EXTRACT LOCATIONS FROM index_table FOR dnacode_right AND STORE IN seedtable_right
4  SELECT ALL HITS WHERE (LOCATIONS seedtable_left + 9) EQUALS (LOCATIONS seedtable_right)
```
References

1. Moazed D. Small RNAs in transcriptional gene silencing and genome defence, Nature 2009;457:413-420.
2. Castanotto D, Rossi JJ. The promises and pitfalls of RNA-interference-based therapeutics, Nature 2009;457:426-433.
3. Dietzl G, Chen D, Schnorrer F et al. A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila, Nature 2007;448:151-156.
4. Morris K, Simon WLC, Jacobsen S et al. Small Interfering RNA-Induced Transcriptional Gene Silencing in Human Cells, Science 2004;305:1289-1292.
5. Qiu S, Adema C, Lane T. A computational study of off-target effects of RNA interference, Nucleic Acids Res 2005;33:1834-1847.
6. Kulkarni M, Booker M, Silver S et al. Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays, Nat Methods 2006;3:833-838.
7. Moffat J, Reiling JH, Sabatini DM. Off-target effects associated with long dsRNAs in Drosophila RNAi screens, Trends in pharmacological sciences 2007;28:149-151.
8. Fedorov Y, Anderson EM, Birmingham A et al. Off-target effects by siRNA can induce toxic phenotype, RNA 2006;12:1188-1196.
9. Ma Y, Creanga A, Lum L et al. Prevalence of off-target effects in Drosophila RNA interference screens, Nature 2006;443:359-363.
10. Doench JG, Petersen CP, Sharp PA. siRNAs can function as miRNAs, Genes Dev 2003;17:438-442.
11. Saxena S, Jonsson Z, Dutta A. Small RNAs with imperfect match to endogenous mRNA repress translation. Implications for off-target activity of small inhibitory RNA in mammalian cells, The Journal of biological chemistry 2003;278:44312-44319.
12. Jackson A, Burchard J, Schelter J et al. Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity, RNA 2006;12:1179-1187.
13. Hannon GJ. RNA interference, Nature 2002;418:244-251.
14. Bosher J, Dufourcq P, Sookhareea S et al. RNA Interference Can Target Pre-mRNA: Consequences for Gene Expression in a Caenorhabditis elegans Operon, Genetics 1999;153:1245-1256.
15. Pal-Bhadra M, Bhadra U, Birchler JA. RNAi Related Mechanisms Affect Both Transcriptional and Posttranscriptional Transgene Silencing in Drosophila, Molecular cell 2002;9:315-327.
16. Verdel A, Jia S, Gerber S et al. RNAi-Mediated Targeting of Heterochromatin by the RITS Complex, Science 2004;303:672-676.
17. Langlois M-A, Boniface C, Wang G et al. Cytoplasmic and Nuclear Retained DMPK mRNAs Are Targets for RNA Interference in Myotonic Dystrophy Cells, J Biol Chem 2005;280:16949-16954.
18. Matzke MA, Birchler JA. RNAi-mediated pathways in the nucleus, Nat Rev Genet 2005;6:24-35.
19. Robb GB, Brown KM, Khurana J et al. Specific and potent RNAi in the nucleus of human cells, Nat Struct Mol Biol 2005;12:133-137.
20. Weinberg MS, Barichievy S, Schaffer L et al. An RNA targeted to the HIV-1 LTR promoter modulates indiscriminate off-target gene activation, Nucleic Acids Res 2007;35:7303-7312.
21. Lin S-L, Kim H, Ying S-Y. Intron-mediated RNA interference and microRNA (miRNA), Frontiers in bioscience : a journal and virtual library 2008;13:2216-2230.
22. Politz JC, Hogan EM, Pederson T. MicroRNAs with a nucleolar location, RNA 2009;15:1705-1715.
23. Lin SL, Miller JD, Ying SY. Intronic MicroRNA (miRNA), J Biomed Biotechnol 2006;2006:26818.
24. Krutzfeldt J, Rajewsky N, Braich R et al. Silencing of microRNAs in vivo with 'antagomirs', Nature 2005;438:685-689.
25. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs, Genes & development. 2001;15:188-200.
26. Adams J. Gene duplication and the birthday problem, Nature 1982;296:176-176.
27. Holen T, Moe S, Sorbo J et al. Tolerated wobble mutations in siRNAs decrease specificity, but can enhance activity in vivo, Nucleic Acids Res 2005;33:4704-4710.
28. Wakiyama M, Matsumoto T, Yokoyama S. Drosophila U6 promoter-driven short hairpin RNAs effectively induce RNA interference in Schneider 2 cells, Biochem Biophys Res Commun 2005;331:1163-1170.
29. Aleman L, Doench J, Sharp P. Comparison of siRNA-induced off-target RNA and protein effects, RNA 2007;13:385-395.
30. Haley B, Zamore PD. Kinetic analysis of the RNAi enzyme complex, Nature Structural & Molecular Biology 2004;11:599-606.
Supporting Information Chapter 3

To investigate the probability of shared off-targets, we first addressed the question whether a random sequence will likely match or nearly match to any other sequence present in the *D. melanogaster* genome. To estimate this, a formula previously designed by others for the same purpose [26], was used. With this formula we calculated what the occurrence is of any random sequence of 21 nt to find one other (k=2) identical sequence within the genome. For this calculation, the 21 nt sequence and the complete genome are for the sake of simplicity considered to consist of random sequences. In this formula, *n* is the number of possible different sequences of a specific sequence length, equal to $4^{21}$ for a sequence of 21 nt long (assuming 4 possible different nucleotides; A, T, C or G). 21 nt was chosen because siRNAs of this size may be generated when large dsRNAs are used in RNAi experiments for the *D. melanogaster* genome. The number of copies to be present is represented by k (=2).

$$\lambda = n \frac{e^{-\frac{r}{n}}}{k!} \left( \frac{r}{n} \right)^k$$

The size of the Drosophila genome is 168.7 Mb (r), therefore it can be expected that, within a randomized genome of this size, there are 3,235 occurrences ($\lambda$) of random 21 nt sequence that exactly match to another sequence within that same genome. However, in this estimation mismatches are neglected while sequences containing mismatches may also contribute to a measurable RNAi response [27-30]. When mismatches are taken into account, this considerably increases the number of different possibilities for an siRNA to find successful sequence hybridization partners with other mRNAs. The number of sequences that are possible with length *n* and up to *k* mismatches is calculated with the standard formula:

$$\frac{n!}{k!(n-k)!}$$
Note that in this standard formula the symbols \( n \) and \( k \) are also used, however they are different from \( n \) and \( k \) presented in formula 1. We used 3 mismatches in this and following calculations as this has been known to elicit a successful RNAi response [6]. Using \( k=3 \) results in 1,330 different combinations for a 21-bp sequence. Combining this with formula (1), \( n \) is divided by 1,330 and \( \lambda \) is recalculated to be 4,089,178 occurrences instead of the previous calculated 3,235 for exact matches. Each occurrence may span up to 42-nt (2x the size of a single occurring siRNA; 21-nt), and therefore all events together (42 x 4,089,178 ) may cover the complete genome.

**Supplementary Figure 1**

*Illustration of the on-line website (www.rnaiseselect.info/dsrna), enabling the identification of dsRNA constructs that do not have shared off-targets. The table lists different dsRNAs combinations (red and blue boxes) that do not have predicted overlapping off-targets (off-targets are indicated by vertical lines) on top of a schematic view of the complete cDNA sequence. By using these dsRNA sequences, the chances that the experimental outcome is influenced by shared off-target effects will most likely be reduced.*
Supplementary Table 1

| Description                              | Score            |
|------------------------------------------|------------------|
| 30%-52% GC Content                      | 1 point          |
| 3 or more A/Us at positions 15-19       | 1 point per A/U  |
| Tm>20°C                                  | 1 point          |
| A at position 19                         | 1 point          |
| A at position 3                          | 1 point          |
| U at position 10                         | 1 point          |
| G/C at position 19                       | -1 point         |
| G at position 13                         | -1 point         |
| >4 sequential nucleotide repeat          | -9 points        |
| > 4 diplet repeat                        | -9 points        |
### Supplementary Table 2

| Gene         | Length (nt) | # of overlapping off-targets |   |   |   |   |   |   | Total |
|--------------|-------------|------------------------------|---|---|---|---|---|---|-------|
| CG11455-RA   | 471         | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 1     |
| CG11454-RA   | 852         | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 2     |
| CG11488-RA   | 898         | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 6     |
| CG31975-RA   | 1215        | 80%                          | 1 | 20%| 0 | 0 | 0 | 0 | 5     |
| CG17691-RA   | 1238        | 85%                          | 2 | 15%| 0 | 0 | 0 | 0 | 13    |
| CG10465-RA   | 1241        | 90%                          | 2 | 10%| 0 | 0 | 0 | 0 | 21    |
| CG11617-RA   | 1267        | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 7     |
| CG1712-RA    | 1284        | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 2     |
| CG8245-RA    | 1306        | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 10    |
| CG11374-RA   | 1338        | 88%                          | 2 | 12%| 0 | 0 | 0 | 0 | 17    |
| CG3436-RA    | 1381        | 89%                          | 1 | 11%| 0 | 0 | 0 | 0 | 9     |
| CG3388-RA    | 1452        | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 2     |
| CG5680-RB    | 1530        | 88%                          | 2 | 13%| 0 | 0 | 0 | 0 | 16    |
| CG17684-RA   | 1546        | 85%                          | 4 | 15%| 0 | 0 | 0 | 0 | 26    |
| CG11377-RA   | 1585        | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 13    |
| CG11125-RA   | 1591        | 83%                          | 1 | 8% | 0 | 0 | 0 | 0 | 1     |
| CG11127-RA   | 1617        | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 5     |
| CG2863-RA    | 1625        | 100%                         | 0 | 0 | 0 | 0 | 0 | 0 | 3     |
| Gene ID     | Start | Length | % CG | % Tp | % Ap | % Gp | % CG | % Tp | % Ap | % Gp | % CG | % Tp | % Ap | % Gp |
|------------|-------|--------|------|------|------|------|------|------|------|------|------|------|------|------|
| CG1903-RA  | 1666  | 34     | 94%  | 2%   | 6%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 36   |
| CG2657-RA  | 1674  | 12     | 92%  | 1%   | 8%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 13   |
| CG40449-RA | 1707  | 10     | 100% | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 10   |
| CG3709-RA  | 1719  | 35     | 95%  | 5%   | 1%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 37   |
| CG31974-RA | 1736  | 24     | 86%  | 4%   | 14%  | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 28   |
| CG14489-RA | 1744  | 21     | 91%  | 1%   | 4%   | 0%   | 0%   | 0%   | 0%   | 0%   | 1%   | 4%   | 23   |     |     |
| CG11023-RA | 1773  | 23     | 96%  | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 1%   | 4%   | 24   |     |     |
| CG2248-RA  | 1805  | 30     | 86%  | 5%   | 14%  | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 35   |     |
| CG18292-RA | 1812  | 8      | 89%  | 1%   | 11%  | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 9    |     |     |
| CG11372-RA | 1832  | 30     | 94%  | 1%   | 3%   | 1%   | 3%   | 0%   | 0%   | 0%   | 0%   | 0%   | 32   |     |
| CG5725-RA  | 1866  | 4      | 100% | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 4    |     |
| CG30110-RA | 1908  | 27     | 79%  | 1%   | 18%  | 0%   | 0%   | 0%   | 0%   | 0%   | 1%   | 3%   | 34   |     |
| CG32669-RA | 1977  | 1      | 50%  | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 1%   | 50%  | 2    |     |
| CG1605-RA  | 1992  | 4      | 100% | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 4    |     |
| CG2061-RA  | 1993  | 5      | 100% | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 5    |     |
| CG2720-RA  | 2000  | 15     | 100% | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 15   |
| CG11620-RA | 2004  | 4      | 100% | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 4    |
| CG11430-RA | 2051  | 27     | 93%  | 2%   | 7%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 29   |
| CG2522-RA  | 2074  | 9      | 100% | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 9    |
| CG11123-RA | 2135  | 23     | 88%  | 2%   | 8%   | 1%   | 4%   | 0%   | 0%   | 0%   | 0%   | 0%   | 0%   | 26   |
| CG15862-RA | 2166  | 30     | 86%  | 2%   | 6%   | 3%   | 9%   | 0%   | 0%   | 0%   | 0%   | 0%   | 35   |     |
| CG11450-RA | 2185  | 31     | 89%  | 3%   | 9%   | 1%   | 3%   | 0%   | 0%   | 0%   | 0%   | 0%   | 35   |     |
| Code          | Value | Percent | Count | Percent | Count | Percent | Count | Percent | Count | Percent | Count | Percent | Count |
|--------------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|---------|-------|
| CG1916-RA    | 2239  | 7       | 88%   | 0       | 0%    | 1       | 13%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 8      |
| CG30325-RA   | 2253  | 74      | 73%   | 22      | 22%   | 5       | 5%    | 0       | 0%    | 0       | 0%    | 1       | 1%    | 102    |
| CG2674-RA    | 2294  | 22      | 76%   | 6       | 21%   | 1       | 3%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 29     |
| CG11186-RA   | 2299  | 26      | 90%   | 2       | 7%    | 0       | 0%    | 1       | 3%    | 0       | 0%    | 0       | 0%    | 29     |
| CG11140-RA   | 2317  | 21      | 95%   | 1       | 5%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 22     |
| CG18455-RA   | 2325  | 13      | 93%   | 1       | 7%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 14     |
| CG5779-RA    | 2344  | 7       | 78%   | 1       | 11%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 1       | 11%   | 9      |
| CG3048-RA    | 2370  | 28      | 88%   | 4       | 13%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 32     |
| CG4822-RA    | 2453  | 38      | 79%   | 8       | 17%   | 2       | 4%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 48     |
| CG12017-RA   | 2466  | 32      | 89%   | 2       | 6%    | 2       | 6%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 36     |
| CG8411-RA    | 2467  | 11      | 85%   | 2       | 15%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 13     |
| CG12178-RA   | 2531  | 23      | 88%   | 3       | 12%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 26     |
| CG8390-RA    | 2539  | 32      | 86%   | 5       | 14%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 37     |
| CG5834-RA    | 2591  | 28      | 74%   | 3       | 8%    | 0       | 0%    | 1       | 3%    | 1       | 3%    | 5       | 13%   | 38     |
| CG11665-RA   | 2621  | 34      | 76%   | 9       | 20%   | 2       | 4%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 45     |
| CG3164-RA    | 2768  | 32      | 84%   | 6       | 16%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 38     |
| CG5748-RA    | 2777  | 24      | 86%   | 4       | 14%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 28     |
| CG1616-RA    | 2801  | 16      | 89%   | 1       | 6%    | 1       | 6%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 18     |
| CG1464-RA    | 2844  | 16      | 80%   | 3       | 15%   | 1       | 5%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 20     |
| CG11486-RA   | 2862  | 24      | 100%  | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 24     |
| CG5036-RA    | 2869  | 32      | 89%   | 4       | 11%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 36     |
| CG4889-RA    | 2907  | 21      | 84%   | 4       | 16%   | 0       | 0%    | 0       | 0%    | 0       | 0%    | 0       | 0%    | 25     |
| Gene ID     | CHR  | Length | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 |
|------------|------|--------|----|----|----|----|----|----|----|----|----|-----|
| CG11490-RA | 2909 | 24     | 89%| 11%| 0% | 0% | 0% | 0% | 0% | 0% | 0% | 27  |
| CG2144-RA  | 2914 | 18     | 95%| 5% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 19  |
| CG32465-RB | 2948 | 18     | 82%| 14%| 1% | 5% | 0% | 0% | 0% | 0% | 0% | 22  |
| CG10283-RA | 3001 | 55     | 90%| 10%| 0% | 0% | 0% | 0% | 0% | 0% | 0% | 61  |
| CG4648-RA  | 3003 | 22     | 85%| 15%| 0% | 0% | 0% | 0% | 0% | 0% | 0% | 26  |
| CG2331-RA  | 3018 | 8      | 100%| 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 8   |
| CG4637-RA  | 3089 | 41     | 87%| 11%| 0% | 0% | 1% | 2% | 0% | 0% | 0% | 47  |
| CG11166-RA | 3103 | 30     | 88%| 12%| 0% | 0% | 0% | 0% | 0% | 0% | 0% | 34  |
| CG11121-RA | 3104 | 33     | 80%| 7% | 1% | 2% | 0% | 0% | 0% | 0% | 0% | 41  |
| CG15207-RA | 3114 | 48     | 89%| 6% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 54  |
| CG11579-RA | 3136 | 13     | 93%| 1% | 7% | 0% | 0% | 0% | 0% | 0% | 0% | 14  |
| CG9211-RA  | 3252 | 25     | 89%| 11%| 0% | 0% | 0% | 0% | 0% | 0% | 0% | 28  |
| CG4698-RA  | 3323 | 49     | 84%| 14%| 1% | 2% | 0% | 0% | 0% | 0% | 0% | 58  |
| CG11066-RA | 3391 | 36     | 88%| 4% | 10%| 1% | 2% | 0% | 0% | 0% | 0% | 41  |
| CG18492-RA | 3392 | 25     | 96%| 4% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 26  |
| CG8426-RA  | 3405 | 76     | 83%| 12%| 13%| 2% | 2% | 2% | 0% | 0% | 0% | 92  |
| CG3836-RA  | 3637 | 49     | 94%| 3% | 6% | 0% | 0% | 0% | 0% | 0% | 0% | 52  |
| CG3938-RA  | 3658 | 56     | 86%| 9% | 14%| 0% | 0% | 0% | 0% | 0% | 0% | 65  |
| CG5076-RA  | 4228 | 22     | 85%| 3% | 12%| 1% | 4% | 0% | 0% | 0% | 0% | 26  |
| CG31973-RA | 4324 | 62     | 74%| 14%| 17%| 5% | 6% | 3% | 4% | 0% | 0% | 84  |
| CG2186-RA  | 4449 | 28     | 93%| 2% | 7% | 0% | 0% | 0% | 0% | 0% | 0% | 30  |
| CG1624-RA  | 4516 | 75     | 83%| 11%| 12%| 4% | 4% | 0% | 0% | 0% | 0% | 90  |
Detailed report of all 99 analyzed genes showing the overlap between sequence similarities based off-targets. The number of items per off-target group is given both in numbers and percentages. As an example: the cDNA of gene CG11372-RA contains 30 events of duplicate sequences with a shared off-target and 1 event of triplicate sequences that share the same off-target and 1 event of quadruple sequences that share the same of target. Green lines represent sites that share an identical off-target with one other site, red lines represent sites that share an identical off-target with 2 other sites, and blue lines with 3 other sites. Purple lines represent sites that share identical off-
targets with 5 or more other sites. Note that for some genes the lines representing the off-target events are in close proximity and cannot be distinguished as separate lines in this illustrative figure.
## Supplementary Table 3

| Gene            | Total length | 2 | %  | 3 | %  | 4 | %  | 5 | %  | 6 | %  | >6 | %  | Total |
|-----------------|--------------|---|----|---|----|---|----|---|----|---|----|----|----|-------|
| CG11455-RA      | 471          | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG11454-RA      | 852          | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG11488-RA      | 898          | 1 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 1     |
| CG31975-RA      | 1215         | 1 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 1     |
| CG17691-RA      | 1238         | 2 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 2     |
| CG10465-RA      | 1241         | 3 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 3     |
| CG11617-RA      | 1267         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG1712-RA       | 1284         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG8245-RA       | 1306         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG11374-RA      | 1338         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG3436-RA       | 1381         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG3388-RA       | 1452         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG5680-RB       | 1530         | 1 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 1     |
| CG17684-RA      | 1546         | 2 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 2     |
| CG11377-RA      | 1585         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0     |
| CG11125-RA      | 1591         | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 1 | 100% | 1    |
| CG11127-RA      | 1617         | 1 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 1     |
| CG2863-RA       | 1625         | 1 | 100% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 0 | 0% | 1     |
| CG11490-RA   | 2909 | 2  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 2  |
|--------------|------|----|-------|-----|-----|-----|-----|-----|-----|-----|----|
| CG2144-RA    | 2914 | 0  | 0%   | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0  |
| CG32465-RB   | 2948 | 0  | 0%   | 100%| 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 1  |
| CG10283-RA   | 3001 | 5  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 5  |
| CG4648-RA    | 3003 | 0  | 0%   | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0  |
| CG2331-RA    | 3018 | 3  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 3  |
| CG4637-RA    | 3089 | 5  | 83%  | 0%  | 0%  | 0%  | 17% | 0%  | 0%  | 0%  | 6  |
| CG11166-RA   | 3103 | 4  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 4  |
| CG11121-RA   | 3104 | 2  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 2  |
| CG15207-RA   | 3114 | 9  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 9  |
| CG11579-RA   | 3136 | 0  | 0%   | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0  |
| CG9211-RA    | 3252 | 2  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 2  |
| CG4698-RA    | 3323 | 5  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 5  |
| CG11066-RA   | 3391 | 1  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 1  |
| CG18492-RA   | 3392 | 1  | 50%  | 1  | 50% | 0%  | 0%  | 0%  | 0%  | 0%  | 2  |
| CG8426-RA    | 3405 | 9  | 90%  | 1  | 10% | 0%  | 0%  | 0%  | 0%  | 0%  | 10 |
| CG3836-RA    | 3637 | 4  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 4  |
| CG3938-RA    | 3658 | 5  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 5  |
| CG5076-RA    | 4228 | 1  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 1  |
| CG31973-RA   | 4324 | 8  | 100% | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 8  |
| CG2186-RA    | 4449 | 0  | 0%   | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0%  | 0  |
| CG1624-RA    | 4516 | 10 | 91%  | 1  | 9%  | 0%  | 0%  | 0%  | 0%  | 0%  | 11 |

- CG11490-RA: 2909, 100%; CG2144-RA: 2914, 0%; CG32465-RB: 2948, 100%; CG10283-RA: 3001, 5%; CG4648-RA: 3003, 0%; CG2331-RA: 3018, 3%; CG4637-RA: 3089, 5%; CG11166-RA: 3103, 4%; CG11121-RA: 3104, 2%; CG15207-RA: 3114, 9%; CG11579-RA: 3136, 0%; CG9211-RA: 3252, 2%; CG4698-RA: 3323, 5%; CG11066-RA: 3391, 1%; CG18492-RA: 3392, 1%; CG8426-RA: 3405, 9%; CG3836-RA: 3637, 4%; CG3938-RA: 3658, 5%; CG5076-RA: 4228, 1%; CG31973-RA: 4324, 8%; CG2186-RA: 4449, 0%; CG1624-RA: 4516, 10%.
All 99 cDNA sequences were re-analyzed, now ignoring any predicted off-targets that occur within intron sequences. Although this lowers the predicted off-targets, it still shows a significant occurrence of overlapping off-targets. Our analysis shows that after applying this filter, 74% of the genes show off-targets that occur more than once within the same derived cDNA sequence. The number of items per off-target group is given both in numbers and percentages. Green lines represent sites that share an identical off-target with one other site, red lines represent sites that share an identical off-target with 2 other sites, and blue lines with 3 other sites. Purple lines
represent sites that share identical off-targets with 5 or more other sites. Note that for some genes the lines representing the off-target events are in close proximity and cannot be distinguished as separate lines in this illustrative figure.