Small regulatory RNAs

Seinen, Erwin

Chapter 2

A Genome-Wide analysis of the specificity of RNAi in Drosophila melanogaster using the novel tool RNAiSelect

Erwin Seinen, Ritsert C. Jansen, Ody C.M. Sibon

An adapted version is currently under revision for publication in Briefings in Functional Genomics (2011)

Author Summary

Genes can be silenced with short-interfering-RNA molecules (siRNA). siRNAs are widely used to identify gene functions and have high potential for therapeutic treatments. It is critical that the siRNA specifically targets the expression of the gene of interest but has no off-target effects on other genes. Although siRNAs were initially considered to be exclusively active on mature mRNAs in the cytoplasm, additional studies have shown that siRNAs are present in the nucleus as well. In this study we investigated whether nuclear intron-containing premature mRNAs should be considered in off-target profiling. By using *Drosophila melanogaster* micro-array data we indeed show a significant off-target occurrence on sequences with homology not only to exonic but also to intronic sequences. Therefore, accurately predicting off-targets at a genome-wide level is important, but validated algorithms that seek beyond the mature mRNA sequences were not available. We designed the novel tool *RNAiSelect* to make comprehensive off-target profiling, based on sequence homology throughout the genome available to the public. With this tool we profiled 1.5 million RNAi sequences derived from all *D. melanogaster* genes.

Introduction

Off-target effects are caused by unintended cross-hybridization between siRNAs and endogenous RNA sequences other than the targeted sequences (1-5). They obscure the functional interpretation of gene silencing experiments (1) and should therefore be avoided as much as possible. Potential hybridizations between siRNAs and mature mRNAs are generally *in silico* analyzed with the user friendly and popular tool BLAST (basic local alignment search tool (6). However, BLAST has insufficient sensitivity for short sequence alignments with partial homology and will likely result in many false negatives. Other more sensitive tools which are also available through a web-interface and specifically designed to find RNAi off-targets (7-9) only apply to mature mRNA sequences, but not to promoter, intron and intergenic sequences. However, studies have shown that siRNAs also act in the nucleus (10-15) where they target promoters (15) and introns (10), while intergenic sequences could also account for (as yet unknown) regulatory functions. Therefore, an accurate whole genome alignment tool to identify RNAi sequences that have the smallest chance of inducing off-targets is imperative, and such a tool was not available to the general public.

In order to fulfill this need, we designed a novel algorithm and tool called RNAiSelect. RNAiSelect rapidly scans the genome for potential partial homology with siRNAs of 21 nucleotides in length. Our tool searches for exact, or nearly exact, homology. It will report homology within a short stretch of contiguous nucleotides with up to 3 mismatches, even if there are evenly distributed mismatches which remain undetected in BLAST. It will identify near exact homologous sequences containing G:U wobble mismatches which exhibit a high binding energy (16,17). It will also allow for single nucleotide polymorphisms, increasing the general applicability of pre-analyzed siRNA sequences in studying multiple genotypes of the same species.

Our tool permits the user to freely and rapidly select the best RNAi constructs possible based on sequence homology and thereby keeping off-targets in general to an absolute minimum.

Results

Experimental validation of RNAiSelect

Currently no validated alignment tools exist that predict off-targets at the genome-wide level. In order to obtain such an instrument we first developed a tool to search the entire *Drosophila melanogaster* genome (including introns) for exact, or nearly exact, sequence homology for a short stretch of contiguous nucleotides with up to 3 mismatches. This tool enables an accurate prediction of off-target effects (see Methods) for any RNAi sequence of interest. In *D. melanogaster*, dsRNA molecules of 300-800 base pairs are commonly used to induce down-regulation of genes. From a specific dsRNA, several siRNAs will be formed through the endogenous RNA interference (RNAi) machinery and each siRNA in theory has its own set of potential off-targets. We have tested whether these individual siRNAs could have any off-target effects and if so whether introns are also being targeted. First, our tool was used to predict all off-targets of siRNAs derived from specific dsRNA constructs and subsequently biological data were used to validate the predictive capacity of our tool. We analyzed 6 independent dsRNA experiments for which microarray data are publicly available (see Methods). Using previously described criteria (18-20) (see also supplementary Table 1), we extracted all the potent siRNAs from the dsRNA sequences used and searched for homology against the genome for 21 nucleotides (nt) with up to three mismatches. With the use of our tool we found an average of 83 potential off-targets per dsRNA with 0% of them containing zero mismatches, 4-10% containing one or two mismatches, and 90-96% of them containing 3 mismatches (see below for homology searching with 19 nucleotides).

To estimate the number of true off-targets, we first identified a group of potential off-targets using RNAiSelect. Within this group, we calculated the percentage of transcripts that were actually downregulated on the microarrays (further referred to as “the predicted set”) and compared them to the percentage of downregulated transcripts of the total array which represents a random set collected from the same microarray (see Figure 1).
Together these data show that the percentages of downregulation in the predicted sets are 10.48%, 13.38%, 13.73%, 13.71%, 25.10%, and 17.91% enriched compared to a representative random set of the total array

(P<0.009, P<0.015, P<0.023, P<0.005, P<0.008, P<0.002) in studies 1-6 respectively (Figure 1, see Methods). This provides strong evidence that at least 8 (on average 13) out of the predicted 83 off-target genes were true (and unwanted) off-targets of the siRNAs. Currently web-based tools that allow a comparable type of analysis predicted none, or far fewer, of these validated off-targets, even when the most sensitive parameters for these tools were selected (see Supplementary Table 2).

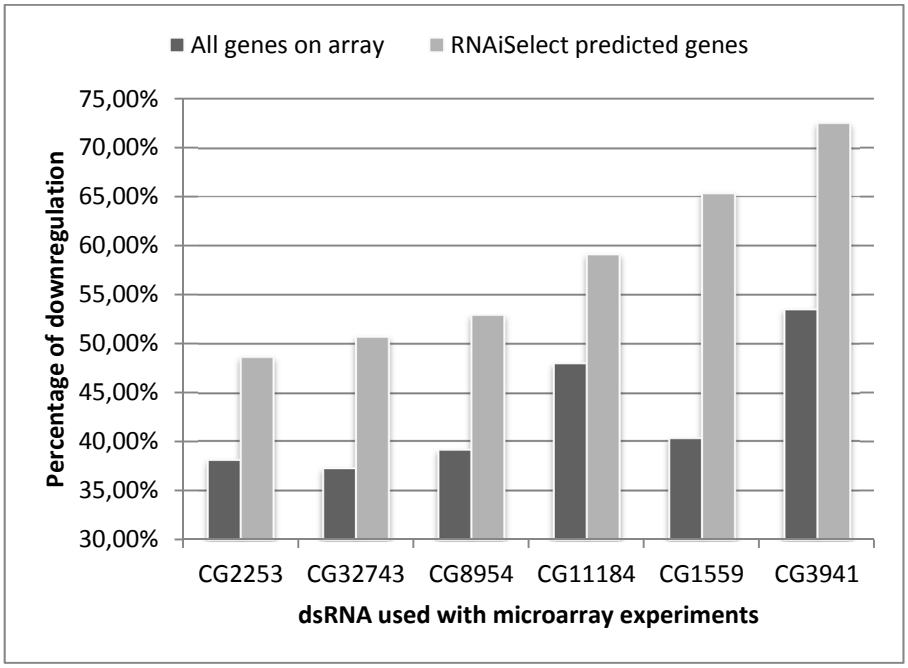


Figure 1. *To validate RNAiSelect, we compared the number of transcripts downregulated on the whole array with the number of predicted off-targets that were actually downregulated. This was performed for six independent experiments. The dark grey bars represent the percentage of downregulated genes of each experiment on the whole array. The light grey bars represent the percentage of predicted off-targets that are downregulated. This analysis shows that the off-target set predicted by RNAiSelect contains an increased fraction of downregulated genes compared to genes randomly selected from the total array (P<0.01 on average).*

Separating exonal from intronal off-targets

RNAiSelect can search for near exact homologies in exons, promoter regions, introns and untranslated regions (UTRs). It can therefore search for

homologies in intergenic sequences that may account for (as yet unknown) regulatory functions. From the potential off-targets predicted by RNAiSelect, we focus on the subset containing exon sequences (set A) and another subset containing intron sequences, including sequences overlapping intron/exon boundaries (set B). The dataset containing the intron sequences (set B) showed the fraction of downregulated genes to be significantly enriched compared to the whole data sets for 3 out of 6 dsRNAs ($P<0.184$, $P<0.026$, $P<0.233$, $P<0.013$, $P<0.1096$, $P< 0.001$ in studies 1-6, respectively in Table 1, see Methods). This finding is consistent with previous findings that there is indeed RNAi activity in the nucleus and specific pre-mRNAs might be exposed to silencing (10,13). This data supports our hypothesis that accurate *in silico* off-target screenings should include exons as well as introns. Examples of predicted off-targets that are strongly downregulated are listed in Supplementary Table 3, many of which are intronic and functionally unrelated to the target gene (Supplementary Table 4). So far, RNAiSelect is the only off-target search tool that allows the user to extend the search to intron-containing pre-mRNA sequences.

	CG2253 (n=142)	**CG32743 (n=73)**	**CG8954 (n=70)**	**CG11184 (n=93)**	**CG1559 (n=26)**	**CG3941 (n=70)**
intron+exon (A+B)	10%	13%	13%	14%	25%	18%
exon only (A)	13%	9%	14%	5%	22%	13%
intron only (B)	8%	19%	11%	20%	30%	32%

Table 1. *The set of predicted potential off-targets (predicted set A+B) was split into one set containing exon sequences (predicted set A) and another set containing intron sequences and sequences overlapping intron/exon boundaries (predicted set B). Enrichment of downregulated genes within the predicted off-targets compared to the microarray background is presented for each set. All the analyzed microarrays showed an increased downregulated fraction in the intronic set (including the boundaries) compared to the background, three of which were statistically significant (see Methods). The number of predicted off-targets per dsRNAi construct for the six independent experiments is indicated with n.*

Allowing for even weaker partial homology

Our analysis shows that homology for 21 nucleotides containing up to 3 mismatches can cause significant off-target effects. Since it has been shown that a homology for 17 contiguous nucleotides can cause off-target effects

(1), it could be important to search for partial homology for 19 contiguous nucleotides with up to 3 mismatches, despite the fact that the set of potential off-targets will most likely contain many more false positives. Performing the analysis using these parameters (including exons, introns and intergenic regions) revealed 5,669, 3,956, 4,255, 6,125, 1,435 and 2,456 potential off-targets for the total sum of the siRNAs that may derive from the dsRNAs of CG2253, CG32743, CG8954, CG11184, CG1559 and CG3941, respectively (Figure 2). It is likely that this set of potential off-targets will contain many false positives, so that there is no longer a significant enrichment of downregulation in the predicted off-target set. However, to reduce the risks in RNAi experiments as much as possible, one can use our tool to select those siRNA constructs that have the lowest number of predicted off-targets (see Figure 2). RNAiSelect can instantly identify these siRNA constructs as well as their potential predicted off-targets, allowing users to perform experiments with the best possible RNAi constructs available based on sequence homology.

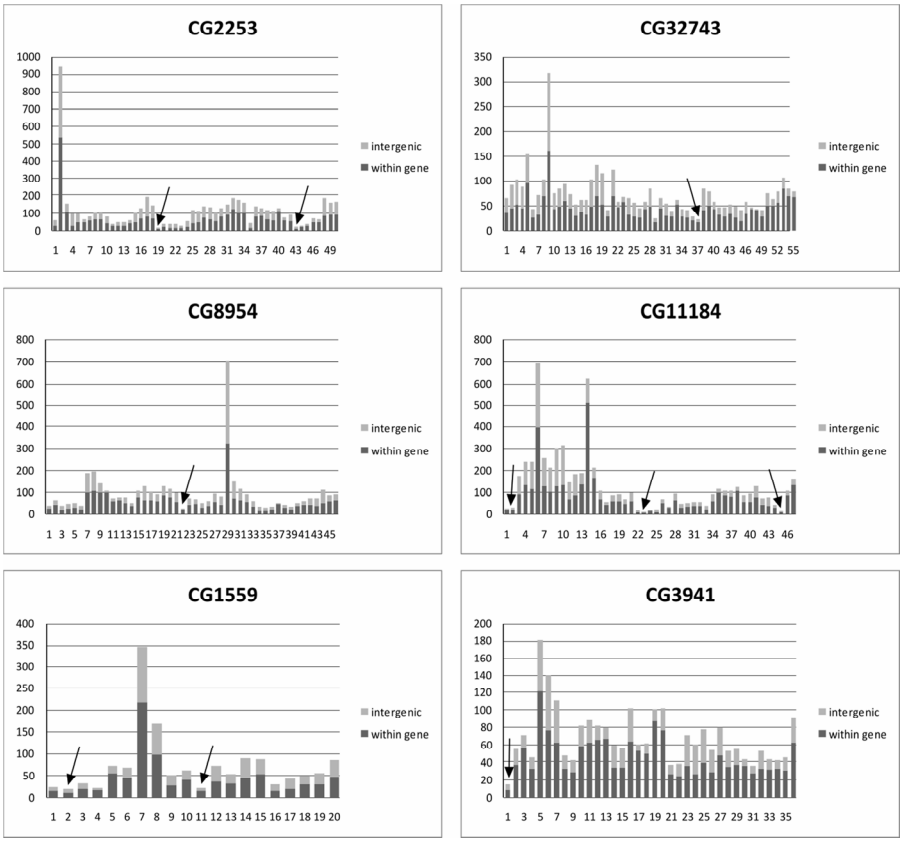


Figure 2. *Predicted off-targets identified by RNAiSelect for the 6 analyzed genes. On the x-axis, all the siRNAs derived from the original dsRNA strand are represented as numbers. On the y-axis, the number of identified off-targets is given. Arrows indicate siRNAs with a low number of predicted off-targets, classifying these rare specific siRNAs as highly specific.*

Analysis of candidate off-targets for all *D. melanogaster* genes

Our analysis of the 6 genes (CG2253, CG32743, CG8954, CG11184, CG1559 and CG3941, shown in Figure 2), reveals a large number of possible off-targets per gene. Next we investigated whether these findings were representative for the complete genome. We used RNAiSelect to scan for all possible siRNAs derived from all annotated *D. melanogaster* genes. Using the same criteria as before (see supplementary Table 1), we extracted 1.5 million candidate siRNA sequences for *D. melanogaster*. On average, we found 121 potential off-targets per siRNA when searching for homology

with a minimum of 19 consecutive nucleotides and 3 mismatches, respectively (median 76). These data show that the dsRNA sequences as analyzed in Figure 2 are not an exception with regard to their large number of off-targets.

MicroRNA effects

It is known that siRNAs can act as microRNAs (miRNAs) based on homology with short seeds of no more than 6 nucleotides on the 3' UTR sequences of the mRNA (21). To find these possible interactions, RNAiSelect offers the possibility to perform a 3' UTR seed pairing analysis using known seed pairing rules (22). Due to the short sequence length (6-nt), there is only a limited number of sequences possible ($4^{\wedge 6}$=4.096). Therefore, if a random genome with the size of D. melanogaster is considered (140 Mb; Flybase R5.20), than any particular seed sequence of 6-nt would theoretically be present a little over 34.000 times throughout the genome (not counting overlapping sequences). Considering 14.419 3' UTR sequences with an average size of 372 nt (Flybase, R5.20), almost 1 out of 10 genes will have any random 6-nt sequence present within its 3' UTR. To show more relevant data that surpasses the background, we only considered the less frequent multiple seed hits within the same 3' UTR as possible miRNA targets. This approach has been demonstrated to decrease the false positive rates (21). In addition, we have recently described a method to *in silico* validate miRNA targets that were identified by *in vivo* experiments (23). Both strategies (21, 23) are employed in the miRNA seed scanning tool within RNAiSelect. This results in the identification of genes that are potentially regulated by siRNAs acting in a miRNA-like manner. From these lists the user can decide either not to use the siRNA or to keep the candidate off-targets in mind while interpreting the results of the experiment.

Generating a resource to select highly specific siRNAs

When selecting the most specific siRNA construct it is not only important to select a sequence with a low number of predicted off-targets (as shown in Figure 2), but also the type of off-target sequences may be relevant. In order to analyze the bulk of data for the type of off-targets at the genome-wide

level we classified the off-targets of an siRNA using the criteria described in Table 2. For this qualification the lowest score (1) was given to an off-target that contains 0 mismatches; score 2 represented an off-target containing 1 G:U wobble mismatch; score 3 represented an off-target containing 1 true mismatch etc. and the highest score was given to an off-target containing 3 true (no G:U wobble) mismatches. The average classification score of all the off-targets of the 1.5 million siRNAs aligned against the genome was calculated to be 7.8. This indicates that, on average, any off-target has 3 mismatches (consisting of 3 G:U wobble mismatches or consisting of 2 G:U wobble mismatches and 1 true mismatch). To further analyze the nature of off-targets, the 1.5 million siRNAs were therefore given the score of their lowest class. With these criteria, the 1.5 million siRNAs have a calculated average of 4.2. This implies that, on average, any siRNA has at least one off-target with 2 G:U wobble mismatches. Together, our genome-wide analysis shows that a randomly selected siRNA construct could have a large number of predicted off-targets with 2 G:U wobble mismatches. The results of these analyses are available as open source (https://sourceforge.net/projects/rnaiselect/files/) and can be used to design future RNAi studies by filtering out these cross-reacting RNAi molecules and instead actively selecting specific criteria for the siRNAs present to downregulate the gene of interest. In case future research reveals specific criteria to allow accurate prediction of the activity of siRNA sequences, the classification table can be adjusted accordingly. Due to the fast speed and internal design of RNAiSelect, it is possible to accommodate a webserver to provide real-time online accessibility.

Class	Description
1	0 mismatches
2	1 G:U wobble mismatch
3	1 true mismatch
4	2 G:U mismatches
5	1 G:U and 1 true mismatch
6	2 true mismatches
7	3 G:U wobble mismatches
8	2 G:U mismatch + 1 true mismatches
9	1 G:U mismatch + 2 true mismatches
10	3 true mismatches

Table 2. *Classification of off-targets, with class 1 having a perfect off-target match and higher numbers (class 2, class 3, etc.) representing classes with increasing numbers of mismatches. G:U wobble mismatches are considered to be more stable than other mismatches and are therefore more represented in the lower classes.*

Computational challenges

The genome-wide alignment of 1.5 million potential siRNA sequences, as performed in this study, is a major computational task. Standard algorithms like BLAST and Smith-Waterman (SW) (23) for local sequence alignment are either not sensitive enough (BLAST) or computationally too intensive (SW). We have therefore developed the RNAiSelect algorithm based on using simple look-up tables. The look-up table contains the exact *D. melanogaster* genomic location for every short sequence of 9 nucleotides. Multiple look-ups of consecutive 9 nucleotide subsequences of a siRNA can find off-target sequences without actual alignment of the siRNA to the genome but by using integer calculations (see Methods). Such look-up tables can easily be constructed, so that the same exercise we have done for *D. melanogaster* can be repeated for other organisms: we anticipate that a similar level of improvement of siRNA specificity can be achieved. RNAiSelect is up to 10 times quicker than SW (24) and equally sensitive.

Statistical analysis

Table 3 and Table 4 show a detailed chi-square analysis of the 6 individual dsRNA experiments from the off-target data predicted by RNAiSelect, with the exception of CG1559 in Table 4 where a two-tailed binomial test was used due to a low transcript number. Table 3 confirms that when considering both introns and exons, the six different dsRNAs show a significant number of off-targets by comparing the observed number of downregulated transcripts with the microarray background ($\alpha = 0.05$). Table 4 shows that when only introns are considered, 3 out of 6 analyses still show a significant number of downregulated transcripts due to off-targets ($\alpha = 0.05$). From this data we concluded that intron-based off-targets were indeed occurring.

Gene	**CG2253**		**CG32743**		**CG8954**		**CG11184**		**CG1559**		**CG3941**	
H0: π =	0.38		0.37		0.39		0.48		0.4		0.53	
Expression	+	-	+	-	+	-	+	-	+	-	+	-
Expected	88.0	54.0	46.0	27.0	43.3	27.7	55.6	51.4	15.6	10.4	32.9	37.1
Observed	73	69	36	37	34	37	41	66	9	17	20	50
Chi-Square	6.761		5.865		5.132		8.025		6.981		9.543	
P-value	0.009		0.015		0.023		0.005		0.008		0.002	

Table 3. *Statistical analysis for both intron and exon data in the microarray experiments For each dsRNA, we analyzed the off-targets predicted by RNAiSelect. The second row presents the percentage of genes that are downregulated on the complete microarray. The number of off-targets as predicted by RNAiSelect are divided in upregulated and downregulated genes (expected values; fourth row) based on the microarray background. The fifth row presents the actual number of upregulated (+) and downregulated (-) genes within the set of off-targets predicted by RNAiSelect (observed values). Chi-square and binomial analysis of these data shows that for every experiment the predicted set of off-targets contains a significant larger fraction of downregulated genes as compared to the complete microarray (H0; second row).*

Gene	**CG2253**		**CG32743**		**CG8954**		**CG11184**		**CG1559**		**CG3941**	
H0: π =	0.38		0.37		0.39		0.48		0.4		0.53	
Regulation	+	-	+	-	+	-	+	-	+	-	+	-
Expected	43.3	26.7	18.8	11.2	17	11	32.3	29.7	6	4	9.4	10.6
Observed	38	32	13	17	14	14	20	42	3	7	3	17
Chi- Square	1.768		4.978		1.424		6.156		[binomial]		11.594	
P-value	0.184		0.026		0.233		0.013		0.1096		0.001	

Table 4*. Statistical analysis for intron data in the microarray experiments As in Table 3, we have analyzed the off-targets predicted by RNAiSelect for each dsRNA, except we now filtered for intron*

targeted regions. For each dsRNA, we further analyzed the by RNAiSelect predicted off-targets. The second row shows the percentage of genes that are downregulated on the complete microarray. The third row presents the number of off-targets as predicted by RNAiSelect. This number is divided in upregulated and downregulated genes (expected values; fourth row) based on the microarray background. The fifth row presents the actual number of upregulated (+) and downregulated (-) genes within set of off-targets predicted by RNAiSelect (observed values). Chi-square and binomial analysis of these data shows that for every experiment the predicted set of off-targets contains a significant larger fraction of downregulated genes as compared to the complete microarray (H0; second row).

Discussion

There are 2 explanations why a non-targeted transcript is downregulated as a results of an RNAi experiment: (i) The transcript is a true off-target of the used RNAi constructs, (ii) The downregulated on-target gene triggers a cascade of regulatory effects which result in down regulation of seemingly unrelated gene products. Explanation (ii) complicates the validation of off-target prediction algorithms and because of this possibility a significantly down regulated transcript does not necessarily represent an off-target effect. Consequently, to enable a proper validation, (a) the 'background noise' has to be corrected for and (b) the dataset to work with must be large enough to enable testing whether there is a significant correlation between downregulated transcripts and identified off-targets. For an optimal correction of the background noise, we compared the number of down regulated genes in the by RNAiSelect predicted set with a randomized group from the same microarray data. Any background noise due to technical limitations or due to specific on-target effects are present in both sets and is corrected for by this approach.
For our analysis we did not use a specific threshold or cutoff value but we divided the transcripts in 2 groups: downregulated and not downregulated. We used this approach for the following reasons. (i) Our aim was not to identify individual transcripts to be downregulated but instead we were interested in a general trend. (ii) Previously, it has been demonstrated that RNAi can induce off-target effects resulting in less than 2-fold reduced expression (25), while still inducing strong protein reduction and subsequently biological effects. (iii) By using no threshold, the analysis contains many more transcripts which enhanced the sensitivity to a great extent. Moreover this allows a proper statistical analysis, which would not be possible when small groups were used. In addition, by maximizing the sensitivity, small significant expression changes that might have real

biological effects are not overlooked. With these considerations, we validated our methods on six experimental data sets from *D. melanogaster*. *In silico* we predicted the potential off-targets of specific double-stranded RNAs (dsRNAs) and empirically show that predicted off-target genes were significantly more frequently silenced than other genes. Moreover, we show that intron containing off-target effects and homologies up to 3 mismatches should not be ignored.
It is to be expected that the off-target RNAi activity is reversely proportional with the number of mismatches. Unfortunately we were unable to make any statistical distinction between the number of mismatches in our dataset, for the reason that off-targets with < 3 mismatches are relatively rare in comparison to 3 or more mismatches. However, because 90-96% of the predicted off-targets do contain 3 mismatches and therefore account for the majority in the significant enrichment during our validation, our data demonstrate that this type of off-target should not be ignored and that partial homology searches are indeed necessary.

Due to the less stringent homology requirements, RNAiSelect will most likely over-predict the number of off-targets. However, this approach is valid, because RNAiSelect is used to find RNAi molecules that have a low number of predicted off-targets to minimize potential side-effects. These sequences with the least amount of off-targets can then be used for knock-down experiments. From that view, over-predicting off-targets is far better than being unaware of possible off-targets, while at least suspicious candidates (like the examples listed in Supplementary Table 3) can now be identified.
Fortunately, in our genome-wide analysis of all potential siRNA sequences, we have found that, when selecting for the most specific siRNAs, the majority of genes have potent siRNAs with 24 or fewer predicted off-targets (arrows in Figure 2). These selected siRNAs have 80% fewer predicted off-targets than the average 121 off-targets generally found. In addition to select for specific siRNAs, the user can make an inventory of possible off-targets when using a particular dsRNAs. This will allow to evaluate existing expression profiles derived from experiments using RNAi technology, e.g. to identify false-positive results caused by homologous induced off-target effects of the used dsRNA sequences. When performing a genome-wide screen using dsRNA molecules, RNAiSelect might also be useful to assess the positive hits within this screen and assist in the decision which of them are most promising to proceed with.

In addition to a real-time analysis tool, we supply a comprehensive database containing 1.5 million pre-analyzed siRNAs covering the whole genome of *D. melanogaster*. RNAiSelect uses this database to allow the generation of a detailed report containing the number and type of mismatches which assists in rapidly selecting specific siRNAs while keeping potential off-targets to a minimum. These siRNAs can then be used instead of the less specific and more generally used dsRNAs (26). Selecting for the most specific siRNAs will be even more important when cocktails of siRNAs are used to downregulate multiple gene products as will be of value for complex traits studies (27).
In conclusion, our tool identifies many validated off-targets, which results in simple and rapid identification of those rare siRNAs with few potential off-target effects. Information on the most selective siRNAs for any individual gene is generated for the users of RNAiSelect, allowing them to choose those siRNAs with the smallest likelihood for off-target effects. Although our approach does not give detailed insights in why specific RNAi constructs are effective, our tool permits the user to work with the best RNAi constructs possible based on sequence homology and thereby keeping off-targets in general to an absolute minimum. Considering the conservation of RNAi mechanisms across species, our findings in *D. melanogaster* will also be of interest for research based on other model systems in which RNAi technology is applied.

Methods

Computer hardware and software

Genomic data (build 45-43b) were downloaded from the Ensembl website (www.ensembl.org). The data from Ensemble and its derived seed tables were processed, stored and indexed in a MySQL database, version 5.0, running on top of Ubuntu 6.06. RNAiSelect was written in C#.NET and can run as a standalone command line executable or within a Microsoft Internet Information Server environment connected to the MySQL database server. Both database and application were hosted on a single system with dual XEON 5140 2.33 GHz processors and 16 GB of 667 ECC memory.

RNAiSelect algorithm

The RNAiSelect algorithm was specifically designed for finding relationships between short nucleotide sequences. It has a high performance and usability for any short-sequence study, including siRNA (off-)targets or miRNA docking sites. The algorithm is based on the following assumption:
"An example sequence TTTTAATTTGGGCCCGGG consists of 18 nucleotides and may be split into two 9-nt child sequences; TTTTAATTT and GGGCCCGGG. By plain observation, we know that the sequence GGGCCCGGG is exactly 9-nt separated from TTTTAATTT in the original sequence."
For the RNAiSelect algorithm to work, we first wrote a program that generates a seed table which holds the exact *D. melanogaster* genomic location(s) for every possible 9-nt sequence ($4^{\wedge 9}$, 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. Our 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.

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)*

This example merely demonstrates how to find an exact 18-nt hit not allowing any mismatches. However, users can allow mismatches by expanding the seed searches by variations in such a way that all possible

combinations will be found. Supplementary Figure 1 shows how mismatches may be distributed on a single 18-nt sequence. We thus included variations of the 9-nt sub-sequences 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.

Validation by microarray analysis

Microarray data were obtained via the EMBL-EBI online repository (http://www.ebi.ac.uk/) (28). We have used the microarray data with the IDs MEXP-202 and E-GEOD-2623. The downloaded raw CEL-data were imported into ArrayAssist 5.5.1 and PLIER normalized. Because our analysis requires all information available at the micro-arrays, we used an approach that allows the evaluation of transcript levels within a large group. First, all transcripts of the whole micro array derived from the dsRNA experiments (the primers used to construct the dsRNA sequences are listed in Supplementary Table 5) were divided in two groups: one group representing all upregulated transcripts and another group representing all downregulated transcripts as compared to the control array. N.B for this specification of groups no cut-off values were used. Although this analysis is not appropriate for single probe analysis, this approach does make it possible to gather sufficient information to identify a general trend within the chosen groups as compared to the background. All ratio comparisons were subjected to chi-square or binomial analysis (see Table 3 and 4) to find significant trends. For all 6 experiments the ratio of down- versus upregulated transcripts was defined and referred to as the background ratio (presented in Figure 1). RNAiSelect was used to define the predicted set of off-targets and within these sets, the ratio of down- versus upregulated transcripts was determined. This ratio was compared with a randomized group representing the background ratio (presented in Figure 1).

Acknowledgements

We thank Gerald de Haan and Lenoid Bystrykh for critical reading of the manuscript. We thank Norbert Perrimon, Matthew Brooker and Bernard Mathey-Prevot for essential advice and stimulating discussions. We also thank Hans Burgerhof for his detailed statistical analysis.

Financial Disclosure

This work was supported by a VIDI grant from the Netherlands Organization for Scientific Research (NWO; 971-36-400) to O.C.M.S. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Kulkarni, M., Booker, M., Silver, S., Friedman, A., Hong, P., Perrimon, N. and Mathey-Prevot, B. (2006) Evidence of off-target effects associated with long dsRNAs in Drosophila melanogaster cell-based assays. *Nat Methods*, **3**, 833-838.
2. Moffat, J., Reiling, J.H. and Sabatini, D.M. (2007) Off-target effects associated with long dsRNAs in Drosophila RNAi screens. *Trends in pharmacological sciences*, **28**, 149-151.
3. Fedorov, Y., Anderson, E.M., Birmingham, A., Reynolds, A., Karpilow, J., Robinson, K., Leake, D., Marshall, W.S. and Khvorova, A. (2006) Off-target effects by siRNA can induce toxic phenotype. *RNA*, **12**, 1188-1196.
4. Ma, Y., Creanga, A., Lum, L. and Beachy, P. (2006) Prevalence of off-target effects in Drosophila RNA interference screens. *Nature*, **443**, 359-363.
5. Jackson, A., Burchard, J., Schelter, J., Chau, B., Cleary, M., Lim, L. and Linsley, P. (2006) Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA*, **12**, 1179-1187.
6. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.
7. Iyer, S., Deutsch, K., Yan, X. and Lin, B. (2007) Batch RNAi selector: a standalone program to predict specific siRNA candidates in batches with enhanced sensitivity. *Comput Methods Programs Biomed*, **85**, 203-209.
8. 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**, W589-591.
9. Naito, Y., Yamada, T., Ui-Tei, K., Morishita, S. and Saigo, K. (2004) siDirect: highly effective, target-specific siRNA design software for mammalian RNA interference. *Nucleic Acids Res*, **32**, W124-129.
10. Bosher, J., Dufourcq, P., Sookhareea, S. and Labouesse, M. (1999) RNA Interference Can Target Pre-mRNA: Consequences for Gene Expression in a Caenorhabditis elegans Operon. *Genetics*, **153**, 1245-1256.
11. Langlois, M.-A., Boniface, C., Wang, G., Alluin, J., Salvaterra, P., Puymirat, J., Rossi, J. and Lee, N. (2005) Cytoplasmic and Nuclear Retained DMPK mRNAs Are Targets for RNA Interference in Myotonic Dystrophy Cells. *J Biol Chem*, **280**, 16949-16954.

12. Matzke, M.A. and Birchler, J.A. (2005) RNAi-mediated pathways in the nucleus. *Nat Rev Genet*, **6**, 24-35.
13. Robb, G.B., Brown, K.M., Khurana, J. and Rana, T.M. (2005) Specific and potent RNAi in the nucleus of human cells. *Nat Struct Mol Biol*, **12**, 133-137.
14. Weinberg, M.S., Barichievy, S., Schaffer, L., Han, J. and Morris, K.V. (2007) An RNA targeted to the HIV-1 LTR promoter modulates indiscriminate off-target gene activation. *Nucleic Acids Res*, **35**, 7303-7312.
15. Morris, K., Simon, W.L.C., Jacobsen, S. and Looney, D. (2004) Small Interfering RNA-Induced Transcriptional Gene Silencing in Human Cells. *Science*, **305**, 1289-1292.
16. Xu, D., Landon, T., Greenbaum, N.L. and Fenley, M.O. (2007) The electrostatic characteristics of G.U wobble base pairs. *Nucleic Acids Res*, **35**, 3836-3847.
17. Holen, T., Moe, S., Sorbo, J., Meza, T., Ottersen, O. and Klungland, A. (2005) Tolerated wobble mutations in siRNAs decrease specificity, but can enhance activity in vivo. *Nucleic Acids Res*, **33**, 4704-4710.
18. Jagla, B., Aulner, N., Kelly, P.D., Song, D., Volchuk, A., Zatorski, A., Shum, D., Mayer, T., De Angelis, D.A., Ouerfelli, O. *et al.* (2005) Sequence characteristics of functional siRNAs. *RNA*, **11**, 864-872.
19. Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. and Khvorova, A. (2004) Rational siRNA design for RNA interference. *Nature biotechnology.*, **22**, 326-330.
20. Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W. and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J*, **20**, 6877-6888.
21. Birmingham, A., Anderson, E., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J. *et al.* (2006) 3' UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature methods*, **3**, 199-204.
22. Brennecke, J., Stark, A., Russell, R.B. and Cohen, S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol*, **3**, e85.
23. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147**, 195-197.
24. Yamada, T. and Morishita, S. (2005) Accelerated off-target search algorithm for siRNA. *Bioinformatics*, **21**, 1316.

25. Aleman, L., Doench, J. and Sharp, P. (2007) Comparison of siRNA-induced off-target RNA and protein effects. *RNA*, **13**, 385-395.
26. Wakiyama, M., Matsumoto, T. and Yokoyama, S. (2005) Drosophila U6 promoter-driven short hairpin RNAs effectively induce RNA interference in Schneider 2 cells. *Biochem Biophys Res Commun*, **331**, 1163-1170.
27. Jansen, R.C. (2003) Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet*, **4**, 145-151.
28. Brazma, A. and Parkinson, H. (2006) ArrayExpress service for reviewers/editors of DNA microarray papers. *Nat Biotech*, **24**, 1321-1322.

Supporting Information Chapter 2

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

Scoring scheme used to define most potent siRNAs, based on a summary from several publications. Sequences scoring at least 6 point were considered by RNAiSelect.

References:

- http://www.protocol-online.org/prot/Protocols/Rules-of-siRNA-design-for-RNA-interference--RNAi--3210.html
- Jagla, B., Aulner, N., Kelly, P.D., Song, D., Volchuk, A., Zatorski, A., Shum, D., Mayer, T., De Angelis, D.A., Ouerfelli, O. et al. (2005) Sequence characteristics of functional siRNAs. RNA, 11, 864-872.
- Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. and Khvorova, A. (2004) Rational siRNA design for RNA interference. Nature biotechnology., 22, 326-330.
- Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W. and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. EMBO J, 20, 6877-6888.

Supplementary Table 2

Comparison of RNAiSelect with BLAST (most widely accepted web-based tool) and dsCheck/siDirect (most accurate web-based tool).

	RNAiSelect	**BLAST**	**dsCheck / siDirect**
Online available web-application	Yes	Yes	Yes
Speed	Fast	Fast	Fast
Accurate short sequence alignment	Yes	**No**	Yes
Finds 1 (G:U or other) mismatches	Yes	Yes	Yes
Finds 2 (G:U or other) mismatches	Yes	**No**	Yes
Finds 3+ (G:U or other) mismatches	Yes	**No**	**No**
Finds exon or UTR based siRNA off-targets	Yes	Some	Yes
Finds whole genome siRNA off-targets, including introns	Yes	Some	**No**
Finds seed based miRNA off-targets	Yes	**No**	**No**
Designs specific shRNAs to knockdown *D. melanogaster* genes[1]	Yes	**No**	**No**
Average validated off-targets found per dsRNA[2]	13	0	2
Identification of non-overlapping dsRNAs with no shared off-targets	Yes	**No**	**No**

[1]In general, long dsRNAs are used for D. melanogaster RNAi experiments. RNAiSelect is the only tool that allows the design of short, specific 21 bp shRNAs. [2]Average number of verified off-targets (by 6 independent microarray data) with 700 bp dsRNAs against 6 different genes.

Supplementary Table 3

Examples of predicted and downregulated off-targets containing various types of mismatches

21-nt sequence from dsRNA	Targeted gene	Off-targeted gene	Regular mismatches	G:U mismatches	exon/intron	fold downregulation
Q: AGCCGAAGGUGCUGAACAAGU ||||||| |||| ||||| R: GGCCGAAGCUGCUGUACAAGU	CG3941	CG3629	3	0	intron	>50
Q: ACAACGACAACGACAUCGAUA ||||| |||| |||||| | R: ACAACAACAACAACAUCGACA	CG2253	CG4128	3	0	intron	>50
Q: CUUUUCGGCUUUGUUUUGAUU ||||| ||||||:|||| || R: CUUUUUGGCUUUGGUUUGUUU	CG11184	CG4128	2	1	intron	>50
Q: AGCACGAAAUCGAAGAGAAAC |||| ||| ||||||||||| R: AGCAGAAAACCGAAGAGAAAC	CG8954	CG2507	3	0	exon	>50
Q: ACAACGACAACGACAUCGAUA ||||| ||||||||:||| | R: ACAACAACAACGACAGCGACA	CG2253	CG2507	2	1	exon	33
Q: ACAACGACAACGACAUCGAUA ||||||||||| ||| || || R: ACAACGACAACAACAACGUUA	CG2253	CG3315	3	0	exon	33
Q: UCGAGGCCAAACUGAAAAUGA |||| |||||||||| ||| R: CCGAGCCCAAACUGAAACUGA	CG2253	CG9656	3	0	intron	20
Q: ACAUCAUGUUUGCAUUUGUUG ||| ||||||||||| |||: R: ACACCAUGUUUGCAUUCGUUU	CG11184	CG1133	2	1	intron	16
Q: AGAACGCGAUCCACCCAGAAA || |||| |||| ||||||| R: AGGACGCAAUCCUCCCAGAAA	CG32743	CG13185	3	0	exon	12

Q: UUAUCAACCGCAAGUCGUAUC |||| |||| |||||||||| R: UUAUGAACCACAAGUCGUAUA	CG32743	CG7978	3	0	intron	11
Q: CUUUUCGGCUUUGUUUUGAUU |||||||| ||||||||| | R: CUUUUCGGUUUUGUUUUGGCU	CG11184	CG12290	3	0	exon	7
Q: UCGGCCUGAUUGGCUUUAUCA ||||| || |||:|||||||| R: UCGGCUUGUUUGUCUUUAUCA	CG32743	CG12819	2	1	exon	7
Q: UGCAACAACUGCCGCAAAUGG |||||||| |||||| |||| R: UGCAACAAAUGCCGCGAAUGC	CG1559	CG15295	3	0	exon	6
Q: UGCAACAACUGCCGCAAAUGG |||||||| ||| |||||||| R: UGCAACAAGUGCAGCAAAUGG	CG1559	CG32046	2	0	intron	6
Q: ACAUCAAGGCCACCGAGAAGA ||||| |:||| ||||||| R: ACAUCACGUCCACGGAGAAGA	CG2253	CG3234	2	1	exon	6
Q: ACAACGACAACGACAUCGAUA ||| |||| |||||||||:| R: ACAUCGACAUCGACAUCGAGA	CG2253	CG32130	2	1	Intron	6
Q: CUGCGUCUGUCCAAGAUCAUC ||| ||||||||||||| R: GUGCCUCUGUCCAAGAUCAUA	CG32743	CG4678	3	0	exon	6
Q: AUCUGCGUCUGUCCAAGAUCA ||||| || |||||||| | R: AUCUGCCUCCGUCCAAGAUGA	CG32743	CG3359	3	0	intron	5

Collection of 18 identified potential off-targets from the six available datasets which appeared to be 5-fold or more downregulated. The first column shows the alignments as found by RNAiSelect which would possibly not have been identified by online available alignment tool. The targeted genes as well as the off-targeted genes predicted by RNAiSelect are listed. The number of regular mismatches and the number of G:U mismatches is given for each off-target. It is listed whether the off-target sequence is present within intron or exon containing sequences of the gene. The fold downregulation compared to the control group (as derived from the available dataset) is presented for each predicted off-target. Functional comparison (using the UniProt Protein knowledgebase; http://www.uniprot.org) did not indicate any functional relation between the targeted gene and these 18 off-targeted genes.

Supplementary Table 4

Targeted gene	Targeted gene description (Uniprot)	Off-targeted gene	Off-target gene description (Uniprot)
CG3941	mitosis; DNA endoreduplication; DNA replication	CG3629	Transcription factor that plays a role in larval and adult appendage development.
CG11184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG4128	Ionic channel
CG11184		CG1133	Transcription factor essential for parasegmental subdivision of the embryo.
CG11184		CG12290	G-protein coupled receptor protein signaling pathway
CG8954	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG2507	Putative epidermal cell surface receptor
CG2253	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG4128	Ionic channel
CG2253		CG2507	Putative epidermal cell surface receptor
CG2253		CG3315	Belongs to the thioredoxin family.
CG2253		CG9656	Transcription factor that is vital to the development of multiple organ systems.
CG2253		CG3234	Forms a heterodimer with period (PER); the complex then translocates into the nucleus. Required for the production of circadian rhythms.
CG2253		CG32130	Apoptosis
CG32743	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG13185	Hydrolase
CG32743		CG7978	This is a membrane-bound, calmodulin-insensitive adenylyl cyclase
CG32743		CG12819	nucleolus organization and biogenesis

CG32743		CG4678	Carboxypeptidase
CG32743		CG3359	Unknown
CG1559	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	CG15295	protein binding
CG1559		CG32046	Unknown

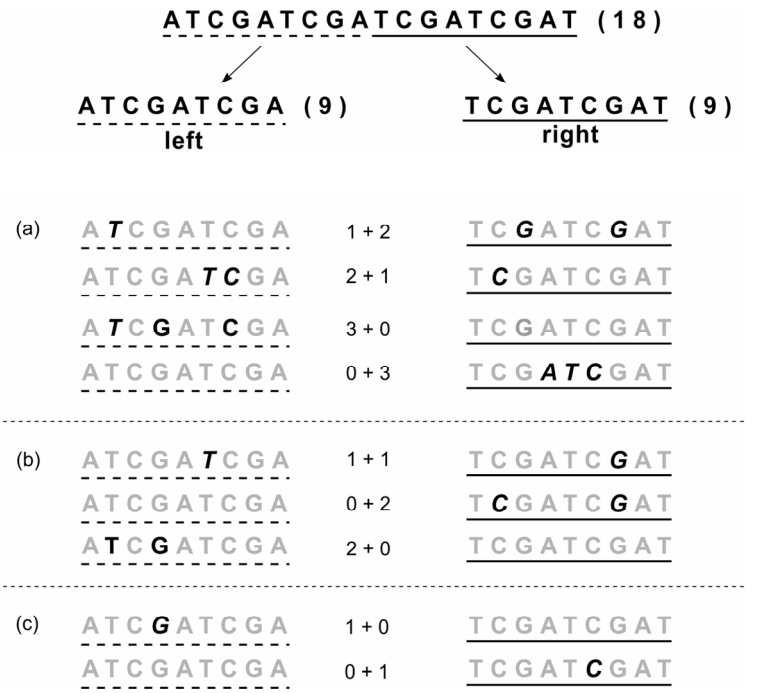
List of functions (as defined by UniProt) of the on-targeted genes from the 6 analyzed dsRNAs and the 18 potential off-target genes listed in Supplementary Table 3.

Supplementary Table 5

Gene	Forward primer	Reverse primer
CG2253	ATGCTAGCCAACGATTCT	CCAGGGCAATCAAATGCA
CG32743	ATGAAGAACGCGATCCAC	GGAGGGCCATGATCATGT
CG8954	ATGGAGGTGACATTCAGC	TGCTTAGTTTGCTGTCGA
CG11184	GTGCCATCTCTATCGGTT	GCTTCCGCTTCTCCTCGT
CG1559	ATGAGCGTGGACACGTACG	TTTGGCGAGCTCGCAGCT
CG3941	GCAGATGTGCAAGCGGGC	TCTCGCACAGGAGACACT

Primers to construct the dsRNAs used in the micro-array experiments.

Supplementary Figure 1



Schematic overview of the distribution of mutations (in bold) along a split 18-nt sequence. (a) The distribution of 3 mismatches (mm) is described by having either 0 mm left and 3 mm right, 1 mm left and 2 mm right, 2 mm left and 1 mm right, or 3 mm left and 0 mm right. (b) The distribution of 2 mismatches is described by having either 0 mm left and 2 mm right, 1 mm left and 1 mm right, or 2 mm left and 0 mm right. (c) The distribution of 1 mismatch is described by having either 0 mm left and 1 mm right, or 1 mm left and 0 mm right.