

Internship Report

Optimization of RNAi Targets on the Human Transcriptome

Ahmet Arslan Kurdoglu
Computational Biosciences Program
Arizona State University

Jeff Touchman
Internship Advisor
Translational Genomics Research Institute

Internship:
Fall 2006

Report Number:
06-13

Table of Contents

Table of Contents	2
Abstract	3
Introduction	4
Methods and Results	7
<i>Statistical Analysis of the Whole Human Genome siRNA Library</i>	9
<i>Identifying BLAST hits that belong to the same gene family</i>	13
<i>Guidelines for designing siRNAs to silence multiple genes</i>	16
Discussion	21
<i>Use of BLAST parameters</i>	21
<i>Use of HGNC to identify genes from same families</i>	23
<i>About siRNA design guidelines for multiple targets</i>	24
Future Work	25
Acknowledgements	26
References	27
Appendices	28
<i>Appendix A</i>	28
<i>Appendix B</i>	32
<i>Appendix C</i>	34

Abstract

siRNA is a single stranded RNA molecule, usually 21 nucleotides in length. It is the key component in RNA interference (RNAi), which is a recently discovered and very effective method for studying functional genomics as well as a potential therapeutic to block certain disease pathways. The process is based on the fact that siRNAs introduced into a cell will find and degrade (with the help of other cellular cofactors) mRNAs with complimentary sequences thereby preventing translation and consequently protein synthesis. It is believed that sequence similarity between the siRNA and its target gene is the most important factor affecting its specificity. In this study the largest commercially available siRNA library designed for the whole human genome was analyzed (*in silico*) for off-target effects caused by sequence mismatch. Over 70% of the siRNAs were found to have a potential for off-target effects. The library was also analyzed to find ‘unintentional but useful’ off-target effects that may shutdown several genes from the same disease pathways. Over 300 such cases were found where a single siRNA will potentially interfere with multiple genes from the same gene family.

Also a set of guidelines to design an algorithm that will search for siRNAs that can effectively silence multiple genes at once was gathered. In theory this is possible due to the prevalence in most gene families of near-identical protein functional domains.

Introduction

RNAi (RNA Interference) is a naturally occurring gene silencing mechanism in cells. It is thought that this process evolved to defend the cells from transposable DNA elements and viruses [6]. It was discovered that this same mechanism can be used to facilitate systematic analysis of gene function – a project in progress at TGen.

RNAi is mediated by siRNA (short interfering RNA), which is usually 21 to 23 nucleotides in length and is usually formed by double stranded RNA (dsRNA) cut into small pieces or synthesized experimentally [6].

dsRNA is the initiator of the gene silencing mechanism. It is cleaved by an enzyme called *Dicer* into smaller fragments to form siRNAs. These siRNA sequences, together with other proteins, form the **RNA Induced Silencing Complex** usually referred to as RISC. RISC identifies mRNA targets that have the same sequence as the minus strand of the siRNA. RISC then binds to these mRNAs and cleaves them to stop the mRNA from being utilized in translation [7].

At first this process was thought to be very specific. However it is now known that siRNAs can target mRNAs without exact sequence match [6]. This means that an siRNA designed to knock-down a specific gene may have adverse affects and create unwanted

phenotypes. This off-target effect of the RNAi mechanism is of great importance for the study of functional genomics or development of therapeutic drugs based on RNAi.

Currently it is not known exactly how incomplete base pairing between siRNA and the target mRNA influences gene knockdown. There have been studies in which off-target effects have proved to be considerable, ranging from 5 to 80% [6] when using the coding sequence of the target gene as dsRNA ranging from 100 nucleotides to 400 nucleotides, arbitrarily cut into 17 to 28 nt long siRNAs. In fact sequence identity of only 11 contiguous nucleotides has been shown to cause off-target phenotypes [2]. However in more recent work done by Brummelkamp et al. [1], and Miller et al. [4] it was shown that even a single mismatch in the siRNA sequence can stop allele-specific gene silencing.

A considerable amount of effort has been given investigating how to design better siRNAs. Currently there are two main guidelines the science community follows, based on work of two different scientists. One of them is developed by Reynolds et al. [7], and the other by Ui-Tei et al.[10]; both of which are explained in more detail in the Methods and Discussion sections. These papers focus on explaining how to design more efficient siRNAs, but not on how the off-target effects play a role in RNAi.

Even though there have been numerous studies [7,10] on how to design extremely target specific siRNAs; there has been very little or no investigation of how to hit multiple genes with a single siRNA. In particular, no work has been done to see if certain off-target effects can be used to knockdown similar genes that belong to the same pathway.

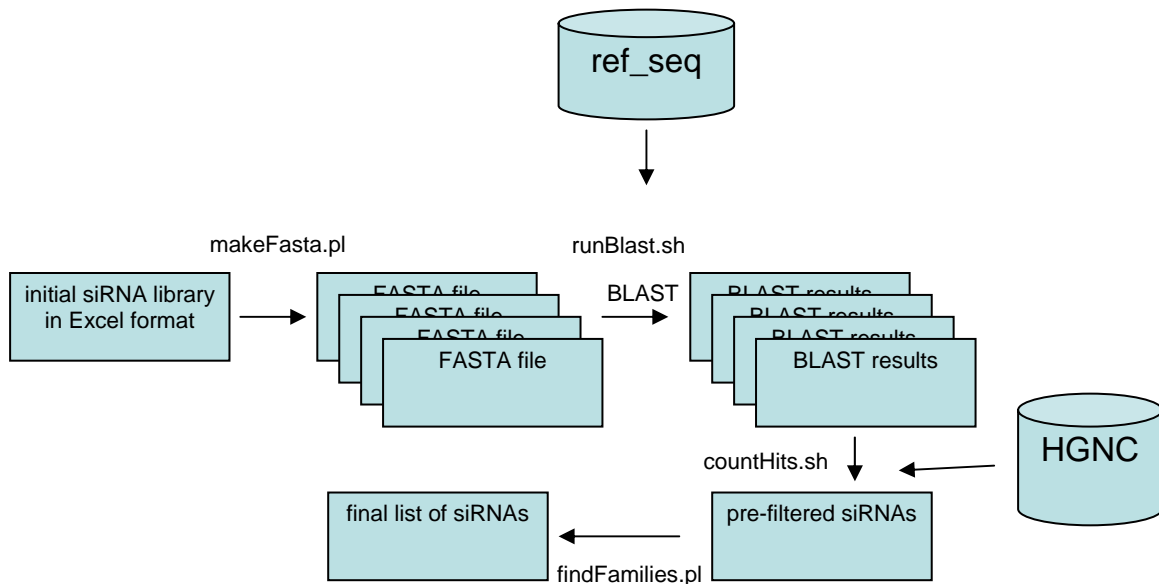
This approach – which basically tries to exploit the off-target effects – brings up several follow-up questions such as: How can one manipulate the existing siRNA design algorithms to optimize for hitting multiple genes, and what is the most important factor that determines specificity?

With a view to an eventual answer to these basic questions, an extensive siRNA library designed by a commercial company for off-target effects is analyzed. A goal is to see if any “useful” but “unintentional” off-targets effects that may be hitting a certain pathway can be found. Additionally, intentional off-target effects that are known to be useful, are created and studied.

The paper is organized such that both Methods and Discussion sections contain two parts: First, the analysis of the already existing commercial siRNA library is shown. Second, the guidelines for the construction of an algorithm that would design siRNAs that can knock-down multiple genes belonging to the same pathway are presented.

Methods and Results

Before going into technical details of the analysis methods a brief summary is presented: First a commercially available siRNA library that covered the whole human transcriptome was attained. All the siRNA sequences in this library were converted into a FASTA file so they could be aligned using BLAST. For each BLAST output, the results were filtered so as to only contain human mRNA sequences. At this point each siRNA had a list of mRNAs (referred to as genes throughout the report) which contained the intentional target as well as possible off-target hits. Every such list was checked against the HUGO Gene Nomenclature Committee (HNGC) database to see if they belonged in the same family.



In the second part of the project, literature relating to siRNA design was examined in order to learn what part of siRNA is important for specificity. The same set of rules was used to come up with guidelines to design siRNAs that can silence a predetermined set of genes.

Statistical Analysis of the Whole Human Genome siRNA Library

There are certain commercial life sciences companies that are involved in sample preparation and molecular diagnostics products. TGen has purchased the whole human genome siRNA library from such a company. This data file contains the gene symbol, such as “RIMS1”, reference sequence code such as “NM_014989”, and four different siRNA sequences each of which may look like “TCCAAGTATAACATACATAAA”. The whole dataset contains just more than 10,000 such rows each for a different gene product. A sample few rows of this table can be found in Appendix B.

The complete siRNA library is this company’s commercial product, and their methods for putting together this database is proprietary knowledge. They do not share which algorithms were used for designing these siRNA sequences, or how each one of these 21 nucleotide sequences was validated.

Analysis of this library is the starting point for studying off-target effects of siRNAs. It is known that sequence similarity is a big factor for siRNA specificity. Therefore it is suitable to BLAST each and every siRNA against the whole reference sequence database from PubMed (namely refseq_rna). Since using the web-interface is extremely slow at a large scale (more than 40,000 requests) it is more appropriate to download and install the NCBI toolbox locally. Refseq_rna database which is just over 2GBs was downloaded in the same manner. After the NCBI toolbox was installed and a local copy of the whole

refseq database was downloaded, command prompt access to tools such as ‘blastall’ became available reducing the time to BLAST each sequence to a few seconds.

A program to convert every row of the siRNA library to a FASTA file that BLAST would accept was written. Also another program to automate running of BLAST and save all the outputs into files was created. The following command was used while running BLAST:

```
blastall -p blastn -d $dbPath/refseq_rna -i $dataPath/v0$i/$file -o $resultsPath/v0$i/$file.out -m 2
```

Additionally, another program to analyze this generated output was written. This program parsed the BLAST output, processed every hit. It recorded and counted the ones that were from the *Homo sapiens* genome. It also formatted the results so that it could easily be accessed by MS Excel. The code for all these programs can be found in Appendix A.

To get an idea of the scale of the off-target effects one can simply graph the number of BLAST hits for each gene which produces the following chart:

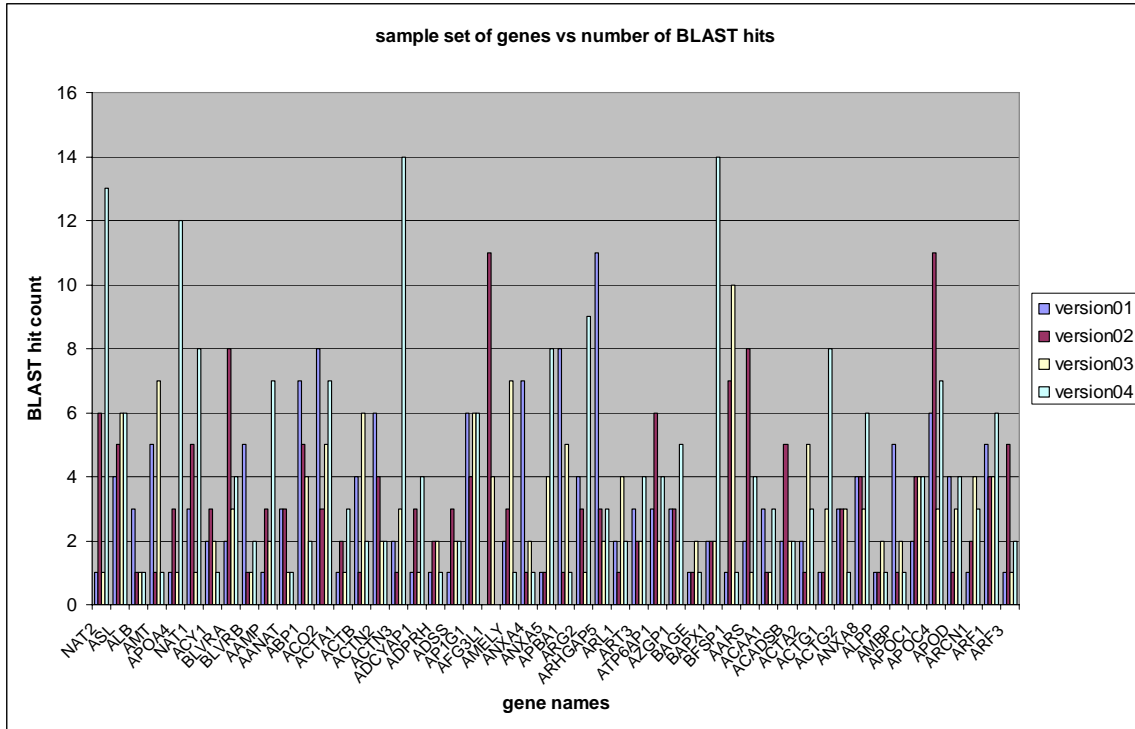


Figure 1 For sake of visibility the number of genes in this chart is greatly reduced since showing ten thousand genes would not be practical.

In this chart, each gene is represented by four columns, since the library contained four different siRNA versions for each gene. An interesting pattern (or lack there of) here is that there is no correlation between each siRNA version and the number of hypothesized off-target hits. It is also important to note that some siRNA sequences result in more than ten hits in the human genome where as some of them are very specific, resulting in only one hit.

As suggested by others [6], the extent of off-target effect is immense if sequence similarity is taken as a measure. More than 70% of all the siRNAs have two or more, more than 50% have three or more, about 18% have five or more, and finally more than

3% have ten or more BLAST hits from the human mRNA database. The results of this *in silico* analysis were found to be similar to that reported by Qiu [6].

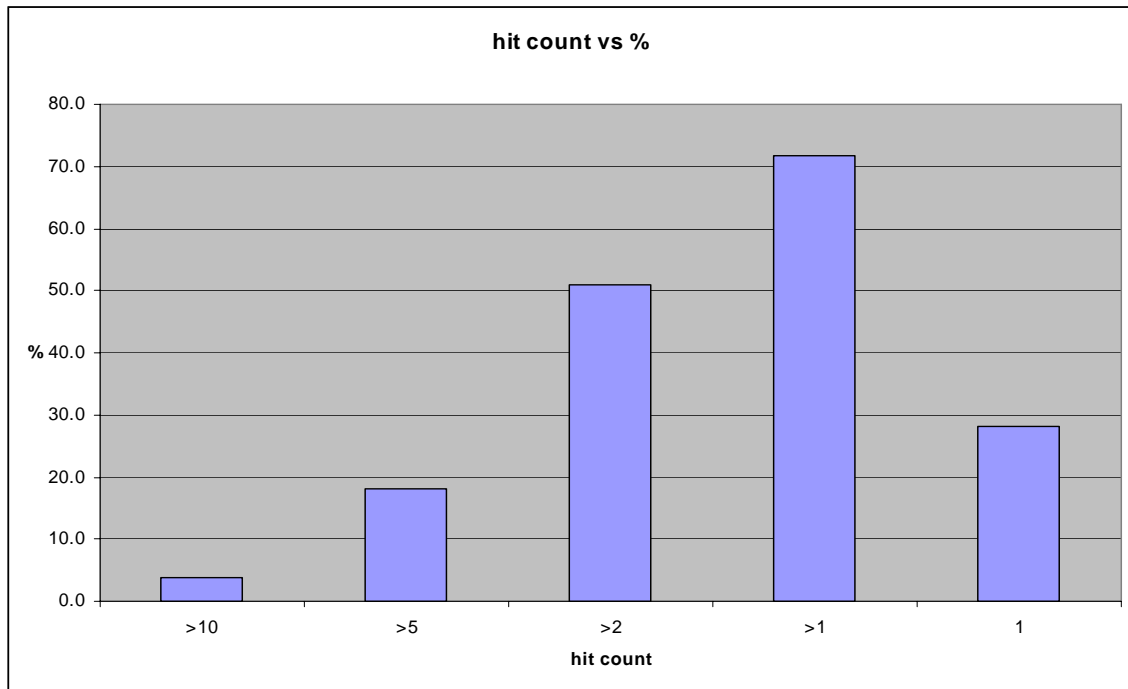


Figure 2 More than 70% of all siRNAs returned multiple BLAST hits, an indicator of off-target effects.

Obviously if the BLAST search is limited to a certain E value, and the results with gaps was omitted in addition to further filtering, the number of hits will be significantly reduced. However it has been shown that even sequence similarity of as few as 11 nucleotides [2] is known to cause gene silencing. Therefore the percentage values reported here should not be a huge over estimate.

After this stage in the analysis the data set was reduced to about 7,000 rows from the initial 10,000 since only genes with multiple BLAST hits were of interest.

Identifying BLAST hits that belong to the same gene family

The HGNC (HUGO Gene Nomenclature Committee) database was used for identifying genes that belonged to the same family. HGNC seemed to be the best among very few options that provided the data in a customizable and downloadable format. This made it suitable for processing with other programs. HGNC constructs the sets of gene families either by sequence similarity, information from publications, specialist advisors for that family, or from other databases. Therefore these gene families may be structural or functional. These gene families and information on how they were compiled together can be accessed at <http://www.gene.ucl.ac.uk/nomenclature/genefamily.html>.

A difficulty using HGNC database is that most of the genes are not associated with a known gene family. In fact out of 26,000 thousand genes (or other known regions) only about 5,000 of them have an entry for a gene family. This lack of information is bound to create some false negatives in the analysis.

To cross reference these set of genes (that were suspected to be suppressed by the same siRNA) to the HGNC database the RefSeq ID's were used. If the RefSeq ID belonged to a gene family this family name was recorded and the same procedure was repeated for each gene in this small set of genes that may be affected by an off-target effect. The Perl program used for this analysis is presented in Appendix A. If a RefSeq ID was not associated with a gene family 'xxx' was recorded as the name to make it easier for the software to recognize. In the end a list that may look as follows was obtained:

symbol	hits	initial refseq											
ZNF326	3	NM_181781	NM_182976.1	NM_182975.1	NM_181781.2	###	ZNF	xxx	ZNF				
ASL	5	NM_000048	NM_0010249	NM_00102494	NM_00048.	NM_00102	###	xxx	xxx	xxx	xxx	xxx	
MYL1	3	NM_079420	NM_079420.2	NM_079422.2	NM_002473.3	###	MYOSIN	xxx	MYOSIN				
KRT17	5	NM_000422	XR_019109.1	XR_015626.1	NM_000422.1	NM_000035.2	###	xxx	xxx	KRT	xxx	CENT	
SLC11A1	3	NM_000578	NM_000578.3	NM_018903.2	NM_0311863.1	###	xxx	xxx	PCDH				
CFHL4	2	NM_006684	NM_012360.1	NR_002169.1	###	OR	OR						

To help clarify, here is the information that can be extracted by reading the top row: The siRNA designed to knock-down the ZNF326 gene resulted in the three human mRNA hits when analyzed by BLAST. These are NM_182976.1, NM_182975.1, and NM_181781.2. The first hit is obviously the gene itself as can be seen from the initial refseq column. The other two hits are subject to possible off-target effects. The '###' characters are just a separator for processing purposes. The remainder of the columns in this first row contain the gene family name associated with these three hits. As can be seen, two of the genes happen to be in the ZNF gene family (zinc-finger family of transcription factors), and one of them (xxx) is not associated with any family.

At this point in the analysis a criterion must be formed to pinpoint siRNAs that can effectively shutdown a family of genes. Therefore siRNAs that returned BLAST hits that belong to the same gene families were separated. To further narrow down this list, genes that resulted in multiple BLAST hits where more than 50% of these came from one family were kept. For example the ZNF326 row that presented above would qualify since two out of three hits (66%) are from the same family. Also the genes that resulted in more than four genes from the same family, no matter how many hits BLAST returned, were collected.

After filtering out the rows that returned no off-target hits or common family names for less than 50% of hits, the final list of genes was reduced to 100 rows. However it should be kept in mind that the siRNA library of interest had four different siRNAs for each gene. Therefore in the end – after removing duplicates – exactly 310 siRNAs remained that could be defined as “likely to have off-target effects on a specific family of genes”. It is very likely that this list would have been a lot longer if the HGNC database had associated each gene with a family. A sample portion of this list can be found in Appendix C.

This list may contain gene families belonging to certain disease pathways and help design better therapeutics. It can also be used to avoid using siRNAs that may have off-target effects on a tumor-suppressing family of genes. These are just a couple of examples of how this information may be utilized. An understanding of how to craft the specificity of siRNAs will greatly improve their effectiveness as potential therapies for human disease.

Another approach to finding siRNAs that silence multiple genes at once is specifically designing them, as opposed to looking for unintentional targets.

Guidelines for designing siRNAs to silence multiple genes

The existing guidelines for siRNA design in the literature are limited. One of them is the paper by Reynolds [7] that identifies an eight-component method. Some of the criteria are low percentage of G/C content, presence of A/U at positions 15-19, lack of internal repeats, the absence of G or C at position 19 and a few other similar points.

Another widely accepted set of guidelines was published by Ui-Tei [10]. These guidelines are also similar in structure: A/U at the 5' end of the antisense strand, at least five A/U residues in the 5' terminal one-third of the antisense strand, and the absence of any GC stretch of the more than 9 base-pairs long.

Most common features in both guidelines are that the ends of the siRNA sequences are of great importance but the central regions do not have any strict limitations except having a low GC content. In fact this central region, if it can show a sequence similarity of about 11-12 nucleotides, is bound to cause off-target effects as defined by Pei [5]. When all these criteria are combined the following figure becomes evident:

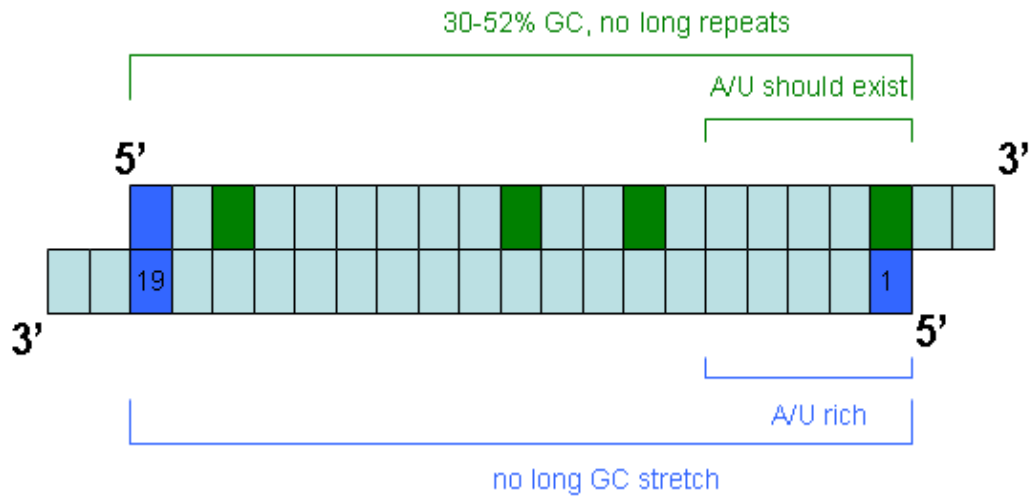


Figure 3 A 21mer dsRNA. One of these strands will enter RISC.

As it can be seen from the figure a region of about 13 nucleotides long (positions 2 – 14) that are not under strict requirements is available. In fact if siRNAs that were 23 base-pairs long were used this “flexible” region would be 15 nucleotides long as seen below:

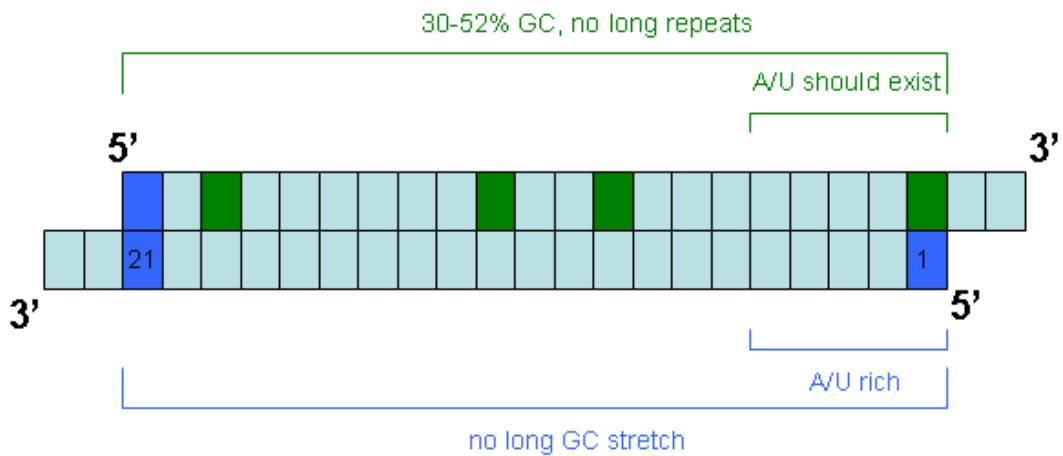


Figure 4 A 23mer dsRNA.

Even though these guidelines will have to be followed to some extent, it must be remembered that these are related to efficiency of an siRNA rather than its specificity. Sacrifice of maximum efficiency in order to optimize specificity (or lack of specificity in this case) may be necessary.

There are no readily published guidelines for siRNA specificity, however there are a couple of studies that primarily focus on what regions of siRNA determine specificity. It is suggested that target binding specificity is defined by a region – called the “seed” region – that is 6-7 nucleotides in length and at the 5’ end of the anti-sense strand [3, 9].

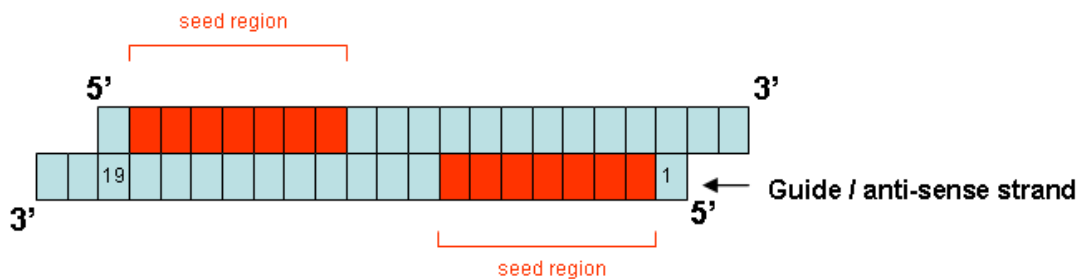


Figure 5 Illustration of seed region.

This “seed” region was first defined by Schwarz [9]. A single mismatch of a nucleotide in this region prevented the design of siRNAs that can discriminate between a wild-type and mutant-type alleles. Moreover, the last two nucleotides of an siRNA do not contribute to binding specificity.

The effects of mismatch in this seed region were also investigated by Qiu [6] where it was shown computationally that base-pair difference in positions 2 through 9 reduced off-target chances by a great margin.

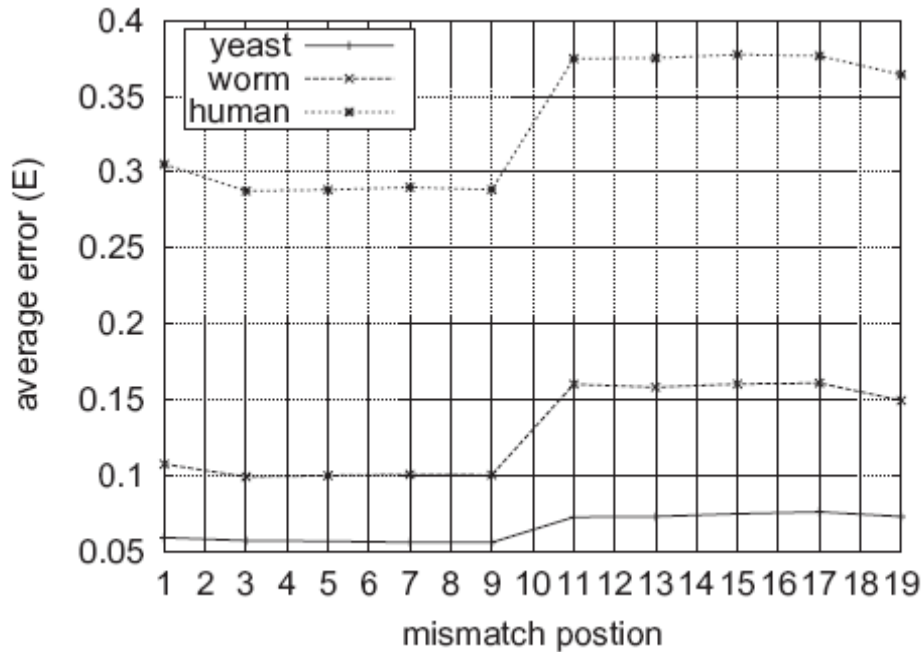


Figure 6 A graph [1] representing the position of a mismatch along the siRNA and off-target error it causes.

Another pointer from Qui's paper that may be of help to build an algorithm to design siRNAs with multiple targets is that shorter siRNAs have a greater chance for off-target effects. This seems intuitive, since the longer the siRNA is, the less likely it will have a near-perfect complementarity with unintended targets.

One final paper [5] that will be taken into consideration distinguishes off-target effects into two classes: a) Those who share near perfect match in the seed region – as demonstrated by others [3, 6, 9] – b) Those who share contiguous and centrally located sequences for more than half the length of siRNA. This suggests that the central region plays a role even though it is not as important as the seed region.

After having considered previous work in siRNA design and specificity, enough information is now present to form a set of guidelines for an algorithm. The most important rule is to have sequence similarity. **(i)** The set of genes that are of interest for silencing must contain at least 11 base-pairs of near-perfect match. It has been shown that [6] the location of this stretch is insignificant so it does not matter how these mRNAs align. **(ii)** The 6-7 matching base-pairs of this aligned region should correspond to the seed region of the siRNA. This is where RISC recognizes its target, therefore it is critical that this region be identical. **(iii)** If the aligned region can be extended, include as much as possible in the opposite direction from the seed region. The longer the siRNA, the more likely it will bond to its target and the less likely it will have off-targets. Effective siRNAs can be as long as 23 nucleotides. **(iv)** As possible, apply the guidelines for siRNA efficiency as outlined by Reynolds and Ui-Tei.

Before the above guidelines can be used to construct an algorithm a few other parameters must be considered. One of these is the use of a multiple alignment program. Whichever program is utilized, a decision must be made whether gaps are allowed. The region and the number of maximum base-pairs become of interest if this is the case.

Another issue will arise when the siRNA efficiency guidelines are being applied as mentioned in **(iv)**. These algorithms were written so that given one target mRNA sequence, they seek for the best region that fits the criteria. In contrast, the algorithm proposed here will find a region common to all target mRNAs and will try to apply efficiency guidelines around this region.

Discussion

Use of BLAST parameters

One of the biggest assumptions in this analysis is the fact that similarity of an siRNA sequence to other mRNA sequences will cause off-target effects and therefore may knock-down these unintentional mRNA targets. It is widely accepted that this is true, however no solid evidence exists that shows to what extent the sequence similarity is important.

Both Ui-Tei [10] and Reynolds [7] have worked on optimizing the correct position of each nucleotide in a short sequence, proving the importance of nucleotide position. Their works resulted in widely accepted guidelines that mostly focus on which nucleotides should appear on each ends of the siRNA and the density of G/C base pairs along certain regions. However both these papers only focus on designing efficient siRNAs usually described by more than 80% silencing for the targeted gene.

Another simplification in this analysis was the 'naïve' use of BLAST. The results of BLAST were chosen not to be filtered in any way. The most obvious filter that could have been used is the E value. However the default parameter for the E value which is set at 10 was not changed. Even a perfect match of 21 nucleotides between the siRNA and the mRNA sequence only resulted in a relatively high E value of 0.002 and an exact sequence match of 15 base pair stretches resulted in E values close to 5.

Another advanced search feature in BLAST to filter the results can be done by manipulating the cost for opening a gap and extending a gap. However the default options for these parameters were not changed as well for it is still unknown how stretches of base-pair complementarity play a role in siRNA specificity. In fact, Sachse [8] observed in his studies that stretches of base-pair similarity have not exhibited recognizable thresholds in length, composition, or any other readily recognizable patterns.

Therefore even though top BLAST hits with low E values do not necessarily mean silencing of unintended targets, high E values also do not let us eliminate these hits as immune to off-target effects. Especially since it has been shown that siRNA can be effective even when as short as 17 base-pairs long and a sequence similarity of 11 nucleotides can result in RNA interference.

In addition it is also notable that BLAST is designed to optimize speed and only initiates a hit with 11 consecutive base-pair matches are found. Considering a 23mer siRNA, a hit will not be reported if there is perfect match on all 23 nucleotides except positions say 7 and 16 since two sequences will not share 11 consecutive nucleotides. Therefore it would be safe to say all the above numbers are an underestimate for the extent of off-target effects.

Use of HGNC to identify genes from same families

As mentioned in the Methods section there were not many options that presented gene families in a format that made it easy to process. HGNC (HUGO Gene Nomenclature Committee) was chosen, which is an entity maintained by Department of Biology and University College of London. It would be out of the scope of this project to try to validate how and if their gene families are assumed correct. However HGNC reports that they are always in touch with specialists in the area, and for each gene family they have listed these advisors on this page: <http://www.gene.ucl.ac.uk/nomenclature/advisors.html>. The negative side of using HGNC database is that only 1/5 of the all the genes are associated with a family. The rest of them do not have an entry.

Another possibility to investigate if two genes are related would be to use the GO (Gene Ontology) Browser (<http://www.geneontology.org>) However one would need to make more assumptions about what distance in the tree – also the depth would have to be considered – two items have to be apart to qualify them as in the same family. Evaluating this for more than two genes would amount to even more complications. For these reasons GO Browser was not utilized in this analysis.

About siRNA design guidelines for multiple targets

One of the first questions that come to mind while trying to design siRNAs for multiple targets is to find the necessary identical domains among the genes that are to be knocked down. If this identical or near-identical domain does not exist for these set of genes, then designing an siRNA would be impossible. In pure statistical terms the chances of finding a 21 base-pair long identical domain in only three genes each 1000 base-pairs long would be close to 10^{-30} . However it should be kept in mind that exact sequence match is only required in the seed region and gaps and mismatches are allowed to some extent for the rest of the siRNA. But even more importantly, it is known that a family of genes will contain protein domains that are conserved within the members of the family increasing the likelihood of finding sequence similarity. One can also take advantage of the fact that siRNAs as short as 17 nucleotides have proven to be effective if finding stretches of identical sequences becomes an issue.

If the designed siRNA is more likely to bind to only one of the targets than the others, then only partial silencing of the less-preferred genes can be expected. Another similar concern that comes to mind is that all intended targets must exist in the cytosol at the same time when siRNA is delivered. However these issues are slightly out of the scope of this project.

Future Work

This analysis can be expanded in several different ways. One improvement would be to filter the BLAST output in Part I of the analysis in accordance with guidelines defined by Reynolds et al. [7], and Ui-Tei et al.[10]. By implementing such a filter a portion of BLAST hits could be eliminated because of their low efficiency as siRNAs. In addition the quality of this study could be improved by using a more comprehensive gene family database.

Also the guidelines for the algorithm defined here can be integrated into software to streamline the process of designing siRNAs that can target multiple genes. However *in vitro* confirmation of some the results would also be beneficial.

siRNA design is still in its early years. As more effective and efficient methods are developed in this field, the guidelines for designing siRNAs that can hit multiple genes will be more apparent.

Acknowledgements

I'd like to acknowledge and thank Dr. Jeff Touchman for his support and guidance and for allowing me the time to work on this project. I would like to thank TGen for creating an environment where students can work with other scientist and for trusting me to join their team.

I'd also like to thank people at TGen – Translational Drug Development (TD2) for sharing their data with me.

References

1. Brummelkamp, T.R., R. Bernards, and R. Agami, *Stable suppression of tumorigenicity by virus-mediated RNA interference*. *Cancer Cell*, 2002. **2**(3): p. 243-7.
2. Jackson, A.L., et al., *Expression profiling reveals off-target gene regulation by RNAi*. *Nat Biotechnol*, 2003. **21**(6): p. 635-7.
3. Jackson, A.L., et al., *Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity*. *Rna*, 2006. **12**(7): p. 1179-87.
4. Miller, V.M., et al., *Allele-specific silencing of dominant disease genes*. *Proc Natl Acad Sci U S A*, 2003. **100**(12): p. 7195-200.
5. Pei, Y. and T. Tuschl, *On the art of identifying effective and specific siRNAs*. *Nat Methods*, 2006. **3**(9): p. 670-6.
6. Qiu, S., C.M. Adema, and T. Lane, *A computational study of off-target effects of RNA interference*. *Nucleic Acids Res*, 2005. **33**(6): p. 1834-47.
7. Reynolds, A., et al., *Rational siRNA design for RNA interference*. *Nat Biotechnol*, 2004. **22**(3): p. 326-30.
8. Sachse, C., et al., *High-throughput RNA interference strategies for target discovery and validation by using synthetic short interfering RNAs: functional genomics investigations of biological pathways*. *Methods Enzymol*, 2005. **392**: p. 242-77.
9. Schwarz, D.S., et al., *Designing siRNA That Distinguish between Genes That Differ by a Single Nucleotide*. *PLoS Genet*, 2006. **2**(9).
10. Ui-Tei, K., et al., *Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference*. *Nucleic Acids Res*, 2004. **32**(3): p. 936-48.

Appendices

Appendix A

makeFasta.pl is a simple Perl program that converts the siRNA library given in tabular format into a separate FASTA files.

```
#!/usr/bin/perl
use strict;
use File::Find;
print "This program will create fasta files from the siRNA library\n";

my($siRNAList)="siRNAlibrary.txt";
my($dataPath)="/users06/akurdogl/rnai/data/";

#attempt to open file
open(THISFILE, $siRNAList) || die;

#read all lines from this file
while(<THISFILE>)
{
    my($thisLine)=$ ;
    writeFasta($thisLine);
}

sub writeFasta
{
    my($inputLine) = $_[0];
    #split line into words using tab as a delimiter
    my(@words) = split /\t/, $inputLine;

    #open file to write
    open V1, ">$dataPath/v01/@words[3].fasta.v1" or die("Cannot open file");
    open V2, ">$dataPath/v02/@words[3].fasta.v2" or die("Cannot open file");
    open V3, ">$dataPath/v03/@words[3].fasta.v3" or die("Cannot open file");
    open V4, ">$dataPath/v04/@words[3].fasta.v4" or die("Cannot open file");

    print V1 ">ref|@words[3]|symbol|@words[4]|@words[5]\n";
    print V1 "@words[15]\n";
    close(V1);
    print V2 ">ref|@words[3]|symbol|@words[4]|@words[5]\n";
    print V2 "@words[17]\n";
    close(V2);
    print V3 ">ref|@words[3]|symbol|@words[4]|@words[5]\n";
    print V3 "@words[19]\n";
    close(V3);
    print V4 ">ref|@words[3]|symbol|@words[4]|@words[5]\n";
    print V4 "@words[21]\n";
    close(V4);
}
#close the file
close(THISFILE);
```

runBlast.sh is a simple Bash script that will run BLAST on all the FASTA files.

```
#!/bin/bash
echo "This program will run blastall on the fasta files"

#define some paths
dataPath="/users06/akurdogl/rnai/data"
resultsPath="/users06/akurdogl/rnai/results"
dbPath="/users06/akurdogl/refseq/"

#removing non NM files
cd $dataPath
find . -not -name "NM*" -type f -exec rm {} \;

#loop to go through all 4 directories
for (( i=1 ; i <= 4 ; i++ ))
do
    echo "version0$i"
    for file in $( ls $dataPath/v0$i/ )
    do
        echo "blasting $file..."
        blastall -p blastn -d $dbPath/refseq rna -i $dataPath/v0$i/$file -
o $resultsPath/v0$i/$file.out -m 2
        echo "..... $file done!"
    done
done
done
```

countHits.sh is Bash script that analyzes the BLAST output and filters out any results but hits for Homo Sapiens.

```
#!/bin/bash
echo "This program will count the number of Homo sapiens hits for each sequence"

#define some paths
dataPath="/users06/akurdogl/rnai/data"
resultsPath="/users06/akurdogl/rnai/results"
dbPath="/users06/akurdogl/refseq/"

#loop to go through all 4 directories
for (( i=1 ; i <= 4 ; i++ ))
do
    #remove previous version0x.out files
    rm -f $resultsPath/v0$i/version0$i.out

    echo "version0$i"
    for file in $( ls $resultsPath/v0$i/ )
    do
        nmName=$( echo $file | cut -d'.' -f1 )
        geneSymbol=$( grep '>' $dataPath/v0$i/$nmName.fasta.v$i | cut -
d'|' -f 4)
        hitString=$nmName", "
        hitCount=0
        for eachResult in $( grep 'Homo sapiens' $resultsPath/v0$i/$file |
cut -d'|' -f 2 )
        do
            ((hitCount++))
            hitString=$hitString$eachResult", "
        done
        echo "$geneSymbol,$hitCount,$hitString" >>
$resultsPath/v0$i/version0$i.out
    done
done
```

findFamilies.pl is a Perl program that checks if a set of mRNAs are in the same family.

```
#!/usr/bin/perl
use strict;
use File::Find;
print "This program finds if the genes are in the same family\n";

my($resultsPath)="/users06/akurdogl/rnai/results/";

for (my($i)=1; $i<=4 ; $i++)
{
    #attempt to open the source file
    open THISFILE, "$resultsPath/v0$i/version0$i.out" || die;

    #attempt to open the new results file
    open NEWFILE, ">$resultsPath/v0$i/commonFamilies0$i.out" or die("Can not
open new results file");

    #read all lines from this file
    while(<THISFILE>)
    {
        my($thisLine)=$ ;
        $thisLine =~ s/\n//g;
        findFamilyName($thisLine);
    }
    #close the source file
    close(THISFILE);
    #close the new results file
    close(NEWFILE);
}
#counts the common family names
sub countFamilies
{
    my($inputLine) = $_[0];
    my($hitCount) = $_[1];
    my($return) = "NO";

    #parse the line with commas
    my(@words) = split /,/ , $inputLine;
    my($word);
    for (my($i)=0; $i<$hitCount; $i++)
    {
        my($same)=0;
        for (my($j)=0; $j<$hitCount; $j++)
        {
            if ( (@words[$i] eq @words[$j]) && (@words[$i] ne "xxx") )
            {
                $same++;
            }
        }
        my($measure) = $same/$hitCount;

        if ( ($measure > 0.50) || ($same >= 4) )
        {
            $return="YES";
        }
    }
    ($return);
}
#for each of the hit it searches HGNC database for a family name
#prints the name if it exists, prints xxx if not.
sub findFamilyName
{
    #open the file only once put it in array
    open (HGNCFILE , "gdlw.pl") || die;
    my(@hgncContent) = <HGNCFILE>;
    close (HGNCFILE);

    my($inputLine) = $_[0];
```

```

#parse the line with commas
my(@words) = split /,/, $inputLine;

#first token is the gene symbol, second is the hit count
my($geneSymbol) = @words[0];
my($hitsCount) = @words[1];

#parse using a dot to remove version number from refseq id
my(@initialRefSeq) = split /\./, @words[2];
my($theRefSeq) = @initialRefSeq[0];

#first four things are not actual hits
my($loopEnd) = $hitsCount + 3;

if ( $hitsCount > 1 )
{
    print "$geneSymbol\n";
    my($familyString) = "";
    #this loop searches for each refseq id in the hgncContent
    for (my($i)=3; $i<$loopEnd; $i++)
    {
        #parse using a dot to remove version number from refseq id
        my(@refSeqs) = split /\./, @words[$i];
        my($thisRefSeq) = @refSeqs[0];

        #line match comes from the hgncContent, not my results file
        my($lineMatch) = grep /$thisRefSeq/, @hgncContent;
        my(@lineWords) = split /\t/, $lineMatch;

        #11th column in the HGNC file is the family name
        my($familyName) = @lineWords[10];
        $familyName =~ s/,//g;
        $familyName =~ s/ //g;

        #if there's no family associated with a certain refseq just
call it xxx
        if ( !$familyName )
        {
            $familyName = "xxx";
        }
        $familyString = "$familyString$familyName,";
    }
    my($tooCommon) = countFamilies($familyString, $hitsCount);
    $familyString = "$geneSymbol,$hitsCount,$theRefSeq,$familyString";
    print NEWFILE "$tooCommon, # $inputLine # $familyString \n";
}
}

```

Appendix B

A representative section of the human siRNA library:

A	B	C	D	E	F	G	H	I	J	K	L
GenbankID	Symbol	Description		Plate A	Offset	Plate B	Offset	Plate C	Offset	Plate D	Offset
1	NM_000632	NAT1	N-acetyltransferase	ENS0300000171428	879	CCCAATAGAGAGATTCATTTATA	899	TACTTCACTTACTTAGAGAA	133	ATCTG9AAATTTGTGATTTA	1277
2	NM_000015	NAT2	N-acetyltransferase	ENS0300000156006	606	GAAGACATATTCAGAAAGAAA	742	ACCCAACTGACTATATTATCA	1015	ATCAATATCTTCAATCCATTA	1094
3	NM_001087	AAAP	angio-associate	ENS0300000121837	1738	CAAGGAAAGCCCTATCCATGTA	820	CTGGATGTGGAAAGTCCCGAA	678	CTGGACTTGGCCCTCAGCCAA	1294
4	NM_001088	AAAT	arylsulfonamide	ENS0300000128673	962	TACCTCTGTATGAAAGGTTGA	728	CCCTTGTAGATTCAGAGCTGA	375	GTGGCTTCTCAGG9CCTGAAA	987
5	NM_001605	AARS	alanine-RNA synthetase	ENS0300000190861	1224	AAAGTGTGATGACAGAGCTGAA	1758	CAAGCTCCGATCTATAGATTA	3248	CCCAAGGCAACTGAAAGGATTA	611
6	NM_001091	ABP1	amiloride binding	ENS0300000102726	2034	CACTTATTCAGACTTTAA	1359	CTGGATTAAGGTGAAAGGCAAT	327	ACCCAACTGATTTG9CAACATA	1578
7	NM_001128	AP1G1	adenosine-related	ENS0300000166747	585	AAAGAGATTAATTCATCTTA	5558	CTCATGATATTCAGATCGAAA	4112	CTCGATTCAGCTGTGATCGTAA	1927
8	NM_001132	AF3L1	AF3 ATPase	ENS0300000167540	296	CTGGAGTCTGTGAAAGCCAA	623	ATCGTTGATGTGTGTGTGTA	2838	CTAAGTGTAGGTTTATTAA	2816
9	NM_000477	ALB	albumin	ENS0300000163631	1534	AAAGAGATTAATTCAGACTTTAA	1666	AAAGTGTTCAGTGAATTTAA	1225	TGGAAAGCCTCAGAAATTTA	1253
10	NM_001630	ALDOC	aldolase C, fruct	ENS0300000109107	952	CAAGAAATTAATTCAGACTTTAA	1534	CAAGAAAGATTAATTCAGACTTTAA	1513	TCAAAGTGTGAGTATGTTA	652
11	NM_001632	ALP	alkaline phosph	ENS0300000163283	2459	AAAGAAAGTGTGTTGATCCCA	2459	CAAGAAAGTGTGTTGATCCCA	2232	CCCGTGTATCTTGGCTCACT	2014
12	NM_001633	ALX3	aristaless-like h	ENS0300000156150	1782	CTCGCTCAGG6GTAAAGCCCAA	978	CATGATGAGTGTGAAAGCCCAA	1455	CAGAGCTCTTCTACAGCCAAA	1265
13	NM_001633	ALX3	aristaless-like h	ENS0300000178522	1733	CAACATCAAGATTAAGCCCAA	578	TGGAAAGTCTGCAAGAGTTAA	224	AAAGCATATTAATTAATGCA	1536
14	NM_001633	AMBP	alpha-1-microgl	ENS0300000108927	85	CAAGCTTACTGCAAGCTCTA	658	TGGAAAGTCTGCAAGAGTTAA	538	TCGGATCTATG99AA9TG9GTA	340
15	NM_001143	AMELY	amelogenin, v1	ENS0300000145020	29	AGGGATGACAGCAAGCCAA	697	ATGCCCTTCTG9CCAGCCAA	372	GAAGCATGATTAAGCCAACTA	197
16	NM_000481	AMT	amionomethylar	ENS0300000145020	1774	CAAGAAAGTGTCTCTGATATA	1317	CAACCTTGTGGAAAGATTA	1675	CTGGCAACAGCTATCTGAAA	856
17	NM_001633	AMY2B	amylose, alpha	ENS0300000197839	82	TTCGGTTATATCACTTTAA	1018	CTAGGGAGCTTACTAGATTA	533	AG99GCTGACATATACCAT	448
18	NM_001633	AMY4	amylose, alpha	ENS0300000138772	1018	CTGATGTGTAAGGATATCA	209	CAAGCAAGGCAAGCCAAATGAA	403	TTCCTATATTCAGCAATTA	937
19	NM_001633	AMY4	amylose, alpha	ENS0300000198975	1653	CTGAAATTAATTCAGATATA	1829	CAAGCAATTAATTCAGATATA	1328	AMCAGAGTATTAATTCGAAA	741
20	NM_001633	AMY4	amylose, alpha	ENS0300000164111	457	TCGATTAATTAATTCAGATATA	1458	CTGGATGACTGAAATCGAAA	388	AACTGATTAATTCAGATATA	1478
21	NM_001633	AMY4	amylose, alpha	ENS0300000107282	2866	CAAGAAAGTGTGCAAGCTTTAA	1886	CAAGAAAGTGTGCAAGCTTTAA	3069	CAAGAAAGTGTGCAAGCTTTAA	3386
22	NM_001633	AMY4	amylose, alpha	ENS0300000130533	3515	TTCGGTTATATCACTTTAA	3591	CAAGCAAGGCAAGCTTTAA	2945	AAAGGCTGCAAGTAAAGGAAA	1419
23	NM_001633	AMY4	amylose, alpha	ENS0300000163867	426	CAAGCTCTCAATTCAGAAATTA	4023	CAAGAAAGTGTGCAAGCTTTAA	3947	CTGGATGTGCAAGTAAAGGAAA	2288
24	NM_001633	AMY4	amylose, alpha	ENS0300000110244	1425	GCCTCTCAATTCAGAAATTA	247	CAAGCAAGGCAAGCTTTAA	390	CAAGCTGTGCTGTGCTCCAAA	1357
25	NM_001645	APC1	apolipoprotein	ENS0300000130208	199	CTGGAAAGGTTTGGAAAGCCAA	220	AGAGCAATTCAGAAAGGTTAA	297	CCCAAGCAAGCCTCTCAGCCAA	42
26	NM_001646	APC4	apolipoprotein	ENS0300000130207	578	CAAGGAGGCAAGGAGGCAAGGAAA	11	ACGGAGTGTGCAATGTGTTGTA	240	CAGCCTCTTGGAAAGGAAAGCCAA	358
27	NM_001647	APD3	apolipoprotein	ENS0300000189058	272	CTGATGAACTGTGATGATA	309	CAAGAAAGTGTGCAAGCTTTAA	531	ATGCCCTGTCTTCACTTTGAAA	1

Appendix C

A sample section of siRNA's bound to cause off-target effects in a family of genes.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1 YES	SEMA6C	3 NM_030913	NM_030913.3	NM_024421.1	NM_004948.2	###	###	3 NM_03091	SEMA	CDH	CDH				
2 YES	WBSR20C	3 NM_032158	NM_001039487	NM_032158.3	NM_148936.2	###	###	3 NM_03215	NSUN	NSUN	xxx				
3 YES	ZNF528	4 NM_032423	NM_032423.2	NM_006956.2	NM_016285.3	###	###	ZNF528		4 NM_03242	ZNF	ZNF	ZNF	xxx	
4 YES	KIAA1984	3 NM_032874	NM_001039374	NM_018538.3	NM_001017922.1	###	###	3 NM_03287	4	xxx	bloodgroup	bloodgroup			
5 YES	KRTAP4-10	3 NM_033060	NM_033060.2	NM_001004719.1	NM_001006500.1	###	###	3 NM_03306		KRTAP	OR	OR			
6 YES	NALP12	2 NM_033297	NM_033297.1	NM_144697.1	###	###	2 NM_03329	NLR	NLR						
7 YES	LOC113251	3 NM_052879	NM_199190.1	NM_199188.1	NM_052879.3	###	###	3 NM_05287	LARP	xxx	LARP				
8 YES	C21orf63	3 NM_058187	NM_058187.3	NM_018833.2	NM_000544.3	###	###	3 NM_05818	xxx	ABC	ABC				
9 YES	MYO18A	3 NM_078471	NM_078471.3	NM_203318.1	NM_002006.3	###	###	3 NM_07847	MYOSIN	MYOSIN	xxx				
10 YES	DEFB104	3 NM_080389	NM_001040702	NM_080389.2	NM_145284.3	###	###	3 NM_08038	DEFB	DEFB	xxx				
11 YES	ACY3	3 NM_080658	NM_080658.1	NM_001002029.1	NM_007293.2	###	###	3 NM_08065	xxx	bloodgroup	bloodgroup				