

# Optimization of RNAi Targets on the Human Transcriptome

Ahmet Arslan Kurdoglu

Computational Biosciences Program

Arizona State University



# my background

- Undergraduate Degree
  - computer systems engineer (ASU 2004)
- Currently
  - Computational Biosciences Program
  - working/interning at TGen – Systems Biology Unit



# today's presentation...

- vocabulary
- RNA Interference background
- Part I – human siRNA library analysis
- Part II – algorithm design
- Q & A (feel free to ask during as well)

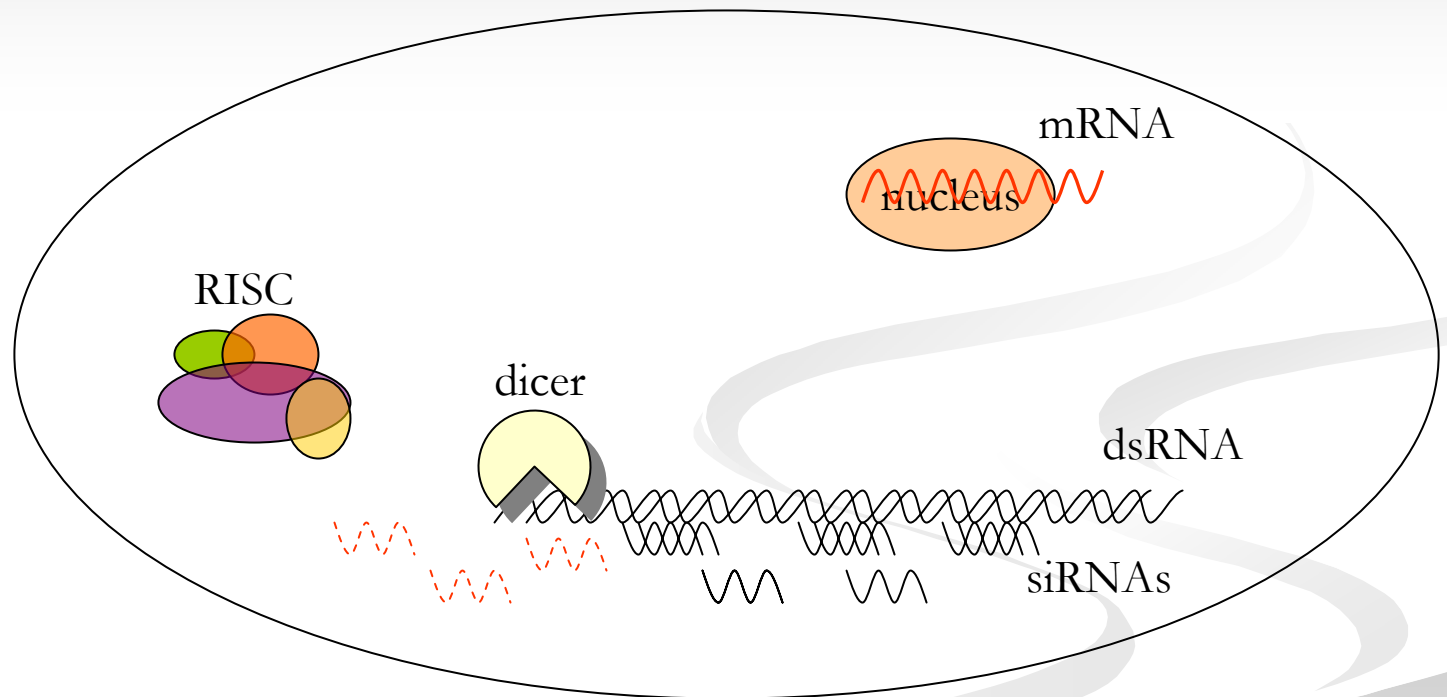
# vocabulary

- target ~ mRNA ~ gene
- silence ~ knockdown
- off-target effect: when an siRNA interferes with an mRNA that it wasn't intended for, caused by sequence similarity

# RNAi background

- RNA Interference (RNAi) is a naturally occurring gene silencing mechanism
- possibly evolved to defend cells from transposable DNA elements and viruses
- Andrew Fire and Craig C. Mello received 2006 Nobel Prize in Physiology or Medicine for their paper (1998) explaining RNAi in *C. elegans*

# how does RNAi work?



video!



# RNAi research at TGen

- Pharmaceutical Genomics Division at TGen
- involved in cancer diagnostics and therapeutics
- siRNA-mediated selective knockdown of cancer causing genes (oncogene)
- Analyzing whole human transcriptome by high-throughput screening
- ‘whole human genome siRNA library’

# siRNA library analysis

- the siRNA library was created by a commercial company
- it covered a large part of the human genome (10K genes)
- for each mRNA they designed 4 different siRNAs
- pre-validated, usually meaning more than 80% silencing

# siRNA library analysis

	A	B	C	D	E	F	G	H	I	J	K	L
1	GenbankID	Symbol	Description		Plate A	Offset	Plate B	Offset	Plate C	Offset	Plate D	Offset
2	NM_000662	NAT1	N-acetyltransferase	ENSG00000171428	CCCATAGGAGATTCAATTATA	879	AAGGACAATACAGATCTAATA	899	TACTTTCAACTTACTAAGAAA	133	ATCTTGGAAATTGGTGATTTA	1277
3	NM_000015	NAT2	N-acetyltransferase	ENSG00000156006	GAGCAGTATATTACAACAAA	606	CTCCAACATCTTCATTTATAA	742	ACCCAACCTCACTAATTATCAA	1015	ATCAAATACTTTTCATCCATAA	1094
4	NM_001087	AAMP	angio-associate	ENSG00000127837	GAGGAAGAGATACTAGTTAAA	1738	CAGGGAAGCCCTATCCATGTA	820	CTGGATGTGGAAAGTCCCGAA	678	CTGGACTTTGCCCTCAGCAAA	1294
5	NM_001088	AANAT	arylalkylamine N	ENSG00000129673	CAGGGCTAAATAAAGAGGAGA	962	TACCCTTCTATGAGAGTTCA	728	CGCCTTTGAGATCGAGCGTGA	375	GTGGCTTCTCACGGCCTGAAA	987
6	NM_001605	AARS	alanyl-tRNA synt	ENSG00000090861	AAGGACATCATTAAATGAGAA	1224	AAGTGGATGACAGCAGTGAA	1758	CACGCTCGCATCTATAGATAA	3248	CCCAGGCAACATGAAGGATAA	611
7	NM_001091	ABP1	amiloride bindin	ENSG00000002726	CACAACAACGAGAACATTGAA	2034	CACTTTAATCCAACTTTAAA	1359	CTGGATAAAGGTGAAAGGCAT	327	ACCCACCTGATTGGCAACATA	1578
8	NM_001128	AP1G1	adaptor-related	ENSG00000186747	CTGCTGTTAGATGAAAGACAA	585	AAGGAGAATTAATTCATCTAA	5558	CTCATGAATATTCACATCAAA	4112	CTCGATTCACTTGACTGTAA	1927
9	NM_001132	AFG3L1	AFG3 ATPase fa	ENSG00000187540	AAGGGTTTGAGAAATCTTTA	296	CTGGAAGTCGTGAACAACAAA	623	ATCGTTGATGTGTGTGTGTA	2838	CTACGTGATCAGTTTATTTAA	2816
10	NM_000477	ALB	albumin	ENSG00000183631	CCCAAGAGTTTAAATGCTGAA	1606	AAGGAGAGACAATCAAGAAA	1666	AAAGTGTTCGATGAATTTAAA	1225	TGGAAGAGCCTCAGAATTTAA	1253
11	NM_005165	ALDOC	aldolase C, fruct	ENSG00000109107	CAGCAATAAATGGTAGCAAA	1534	CAGAAAGATGATAATGGTGT	312	ATGCCTTGAGTACATACCATA	1513	TCAAACGTTGTCAGTATGTTA	652
12	NM_001630	ANXA8	annexin A8	ENSG00000186807	AAGGAGCGAGATTGACTTTAA	952	CAGCTGAGAAATGAACACGAA	1596	TGCCATTAACATTCATCTAAA	1633	ACCCAAGGACACTGTGTTATA	1723
13	NM_001632	ALPP	alkaline phosph	ENSG00000183283	AAGGAAGTGTGGTAATCCCA	2459	CAGGACACTGGTCGAGAGCCA	2332	CCCGCTGATCTTTGCTTCAGT	2014	CAACTTCTCAGAGCTTCCATA	2249
14	NM_006492	ALX3	aristaless-like h	ENSG00000156150	TTGGATGTTGGTAAGAATAAA	1762	CTCGCTCAGGGTAAAGCCCAA	978	CATGATGAGATGGAAACCAA	1455	CAGGACTCTTCTACCACAAA	1265
15	NM_016519	AMBN	ameloblastin, er	ENSG00000178522	TACCAGCATGACACTATTATA	1733	CACCATCAGATAAGCCACCAA	578	TGGGAAGTCTGCAGAGATTA	224	AAGCATATATAAATAATGCA	1536
16	NM_001633	AMBP	alpha-1-microgl	ENSG00000106927	CAGGGCAAACTGAGGTCAA	85	CACCATTACTGCCAAGCTCTA	658	TGGGAAGTTTCTCTATCAAA	538	TCGGATCTATGGGAAGTGTA	340
17	NM_001143	AMELY	amelogenin, Y-li	ENSG00000099721	AAGAAGTTTCTCTGAATATA	29	AGGGATGACACAAGCACAAA	697	ATGCCTGTTCCTGGCCAGCAA	372	GAGCATGATAAGACCACCATA	197
18	NM_000481	AMT	aminomethyltran	ENSG00000145020	CTCCAGTTTGTCTTCCGTA	1774	CACAACTACTATACCCTCAA	1317	CACCTTTGTGGAGAGGATAAA	1675	CTGGCAACAGCTATTCTGAAA	856
19	NM_020978	AMY2B	amylase, alpha	ENSG00000197839	TTCGGTTATTATCACCTTAAA	82	TAAGCGTTTATTAGATAAA	193	CTAGGACAACCTAGACTTCAA	533	AGGGCTGACTATATACCCAT	448
20	NM_005139	ANXA3	annexin A3	ENSG00000138772	AACCAAGAAAGATAATCTCCAA	1018	CTGATTGTTAAGGAATATCAA	209	CAGGACAAGCAGGCAATGAA	403	TTCCTATATTACAGCAATTA	937
21	NM_001153	ANXA4	annexin A4	ENSG00000196975	AACCAATTTATCTGAACTAAA	1653	CTGAATTTAGATGATATAAA	1829	CACAGACATTGAATATTTAA	1328	AACAGAGTATAAATCTGAAA	741
22	NM_001154	ANXA5	annexin A5	ENSG00000184111	CGGCTTTATGATGCTTATGAA	457	TCCATTTATATTACATTTGTA	1458	CTGGATGACCTGAAATCAGAA	388	AACCATGATACTTTAAATAGA	1478
23	NM_001163	APBA1	amyloid beta (A4	ENSG00000107282	CAGGAAGAGAAGAATCCACAA	2856	CAGGATCAAGATGGCCAGAA	1696	CAGGAGGTGCAGACTTCTTAA	3069	CACAGAGAGCTTGGCAATTC	3396
24	NM_005503	APBA2	amyloid beta (A4	ENSG00000034053	CACGCAGTAACAAAGCTTTA	3515	TTCGCGCATATCAATAATA	3591	CAGGACAGACTTGTAAATGAA	2945	AAGGCTGCTAAGATCAAGAAA	1419
25	NM_173075	APBB2	amyloid beta (A4	ENSG00000183697	CAGCTCATATTTCAAATTTA	426	CTGGAAGACATCAAAATTTA	4023	CAGGAGTAATGACTGTTTAA	3947	CTGGCTGTCAAGTGAACATGAA	2289
26	NM_000482	APOA4	apolipoprotein A	ENSG00000110244	CCACCTCTCAATATTTCAATAA	1425	GCCGTGGAACATCTCCAGAAA	247	CAAGGACTCGGAGAACTGAA	390	CACCTGTCTGTCTGTCCAAA	1357
27	NM_001645	APOC1	apolipoprotein C	ENSG00000130208	CTGAAGGAGTTTGGAAACACA	199	CTGGAGGACAAGGCTCGGGAA	220	AGAGACATTTCAGAAAGTAA	297	CCCAACCAAGCCCTCCAGCAA	42
28	NM_001646	APOC4	apolipoprotein C	ENSG00000130207	AAGTTCATACTTCTCAATAA	578	CAGAGGGACAGAGGCACGGAA	11	ACGGGTGGCAATGGTCTGGA	240	CAGCCTCTTGAAGAAGACCCA	358
29	NM_001647	APOD	apolipoprotein D	ENSG00000189058	AACGGAAAGTCAAAAGTGTTA	272	CTGATGGAACGTGAATCAAA	309	CAGAAACAGTGGACTCTCTAA	531	ATGCCTGTCTTCACTTGTAAA	1



S

cts

- creat
- need
- BL
- do
- fo
- filter
- end

```
#!/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;

#!/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/"

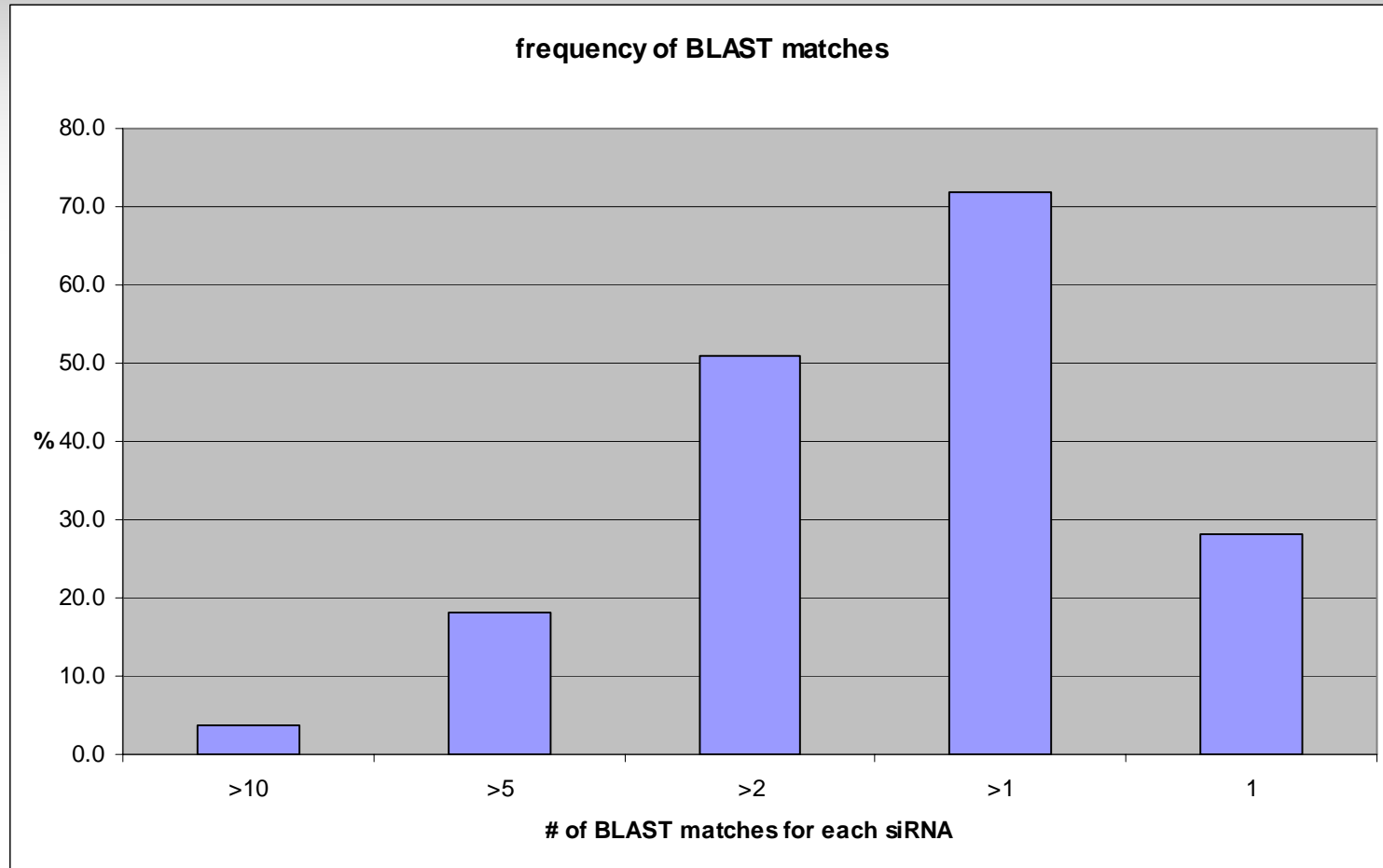
#removing non NM_ files
cd $dataPath
find . -not -name "NM*" -type f -exec rm {} \;

#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

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);
```



# extent of off-target effects



# about using BLAST

- what is my E value?
- word size?

CCGUUAUAUAUUAGGUGAUUA

| | | | | | | | | | | | | | | |

CCGUUGUAUAUAUUAGGAGAUUA

- overestimate or underestimate?

whi

■ it could  
genes

■ which s  
family?

■ cross-re  
databas

■ HUG

■ assign

```
#!/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
```

eful?

nily of

n one

th HGNC

e



# which ones could be useful?

- criteria:
  - if HGNC associated more than half of my hits for a certain siRNA
  - or more than 4 no matter how many hits
- 300 siRNAs that could be identified as:
  - “likely to have off-target effects on a specific family of genes”
- how to use this list?

# today's presentation...

- vocabulary
- RNA Interference background
- Part I – human siRNA library analysis
- **Part II – algorithm design**
- Q & A (feel free to ask during as well)

# algorithm design

- previous section: searched for unintentional but useful off-target effects using a commercial library
- this section: let's create off-target effects that we know will be useful
- **how to design siRNA's that can silence multiple genes?**
  - functional redundancy exists with a lot of proteins – that's why it is important to knock out multiple gene family members
  - delivery is an issue

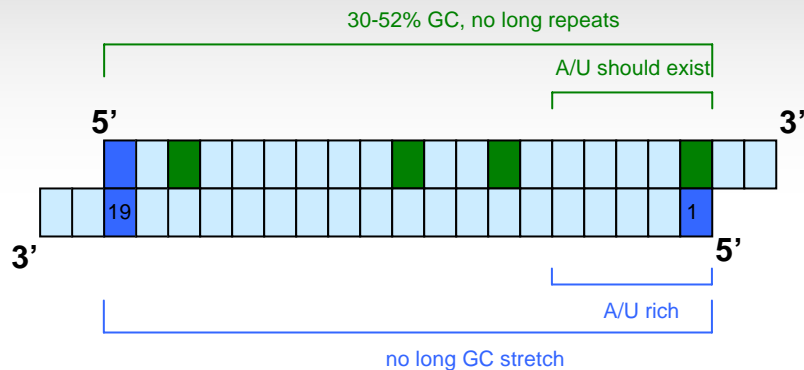
# algorithm design

- most work has been done on eliminating off-target effects
- two sets of widely accepted guidelines
  - Reynolds et. al (8 items)
  - Ui-Tei et. al (4 items)

# Reynolds vs. Ui-Tei

- 30%-52% GC content
- Three or more A/Us at positions 15-19 (sense)
- A at position 19 (sense)
- A at position 3 (sense)
- U at position 10 (sense)
- No G/C at position 19 (sense)
- No G at position 13 (sense)
- A/U at the 5' end of the antisense strand
- G/C at the 5' end of the sense strand
- AU-richness in the 5' terminal one-third of the antisense strand
- the absence of any GC stretch over 9bp in length.

# algorithm design guidelines

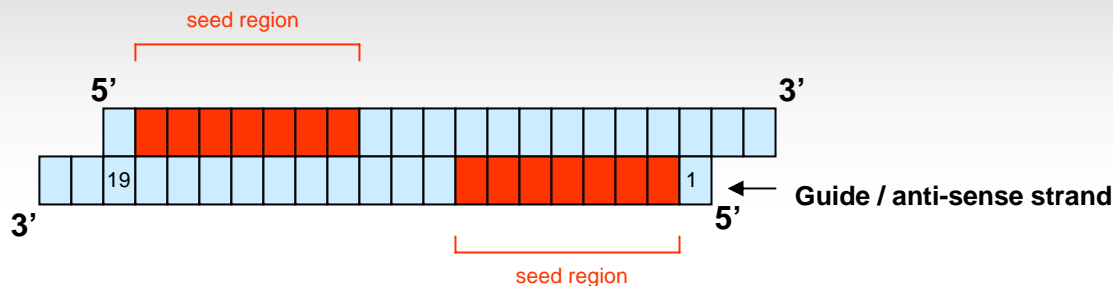


•Reynolds

•Ui-Tei

- these are for increasing efficiency
- we may have to sacrifice efficiency for specificity

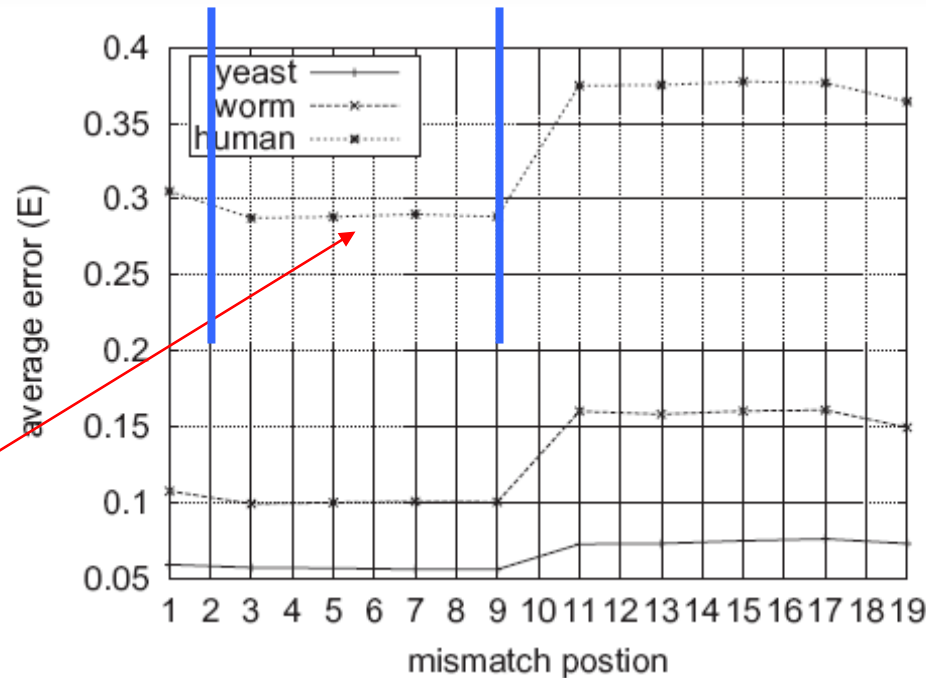
# what effects specificity?



- seed region; 6-7 nucleotides at the 5' end of the anti-sense strand
- this region is known to initiate binding to the mRNA

# seed region

- Qiu showed that a base-pair difference in this region reduced off-target chances by a great margin

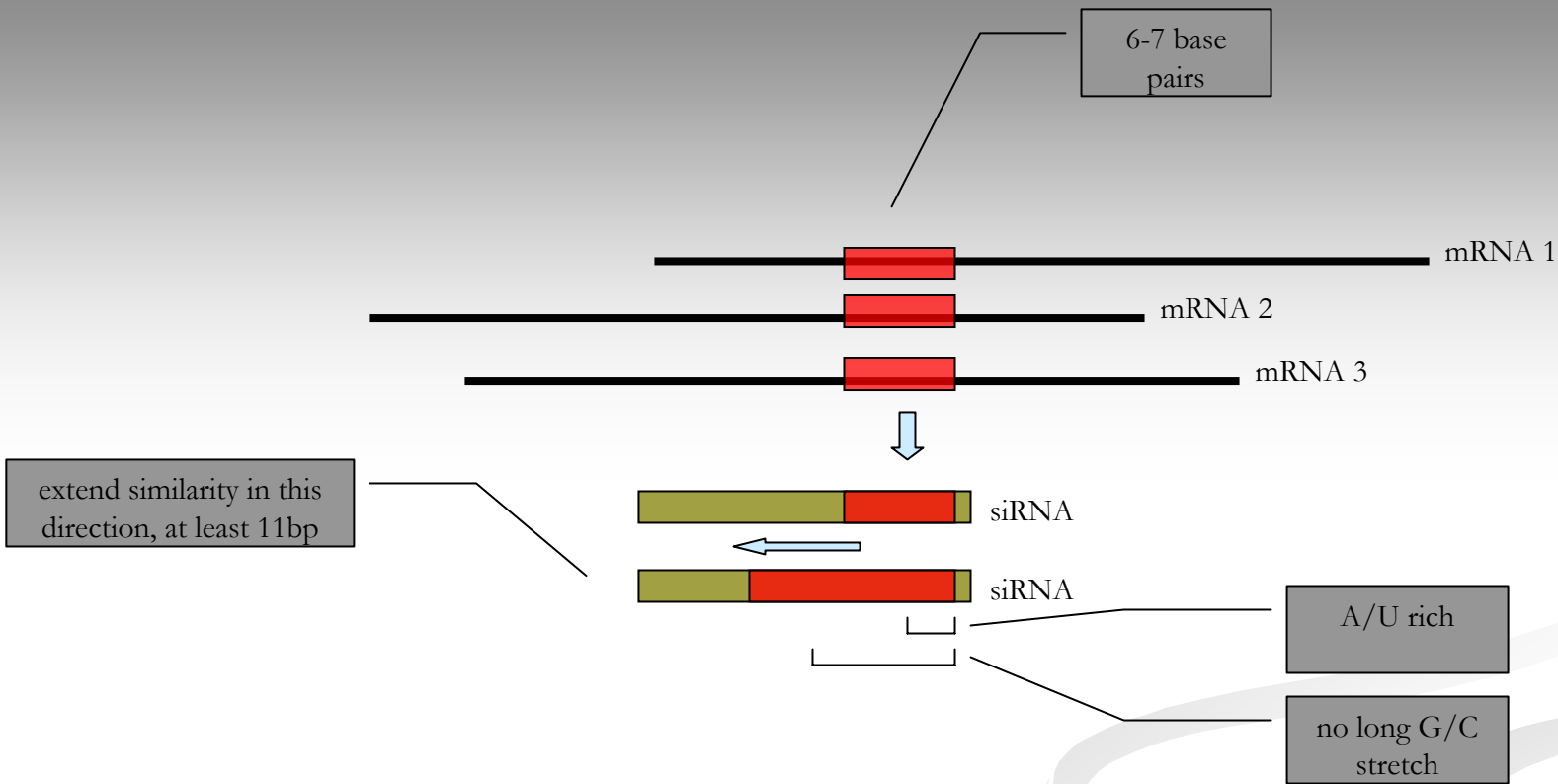


# other facts to consider

- shorter siRNAs have greater chance for off-target effects
- though not as important as seed region; contiguous centrally located complementarity more than half the length of siRNA is also a factor

# algorithm design

- guidelines for designing siRNA to silence multiple genes:
  - The set of genes we wish to silence must contain at least 11 base-pairs of near-perfect match
  - The 6-7 matching base-pairs of this aligned region should correspond to the seed region of the siRNA
  - If the aligned region can be extended include as much as possible in the opposite direction from the seed region
  - Apply the guidelines for siRNA efficiency as outlined by Reynolds and Ui-Tei if possible



- efficiency guidelines?
- multiple alignment program parameters?

- picked
- took
- clust
- chec
- chec
- BLA

gi 53759124 ref NM_001295.2	Homo sapiens chemokine (C-C motif)	40.1	4e-04
gi 117190191 ref NM_001077442.1	Homo sapiens heterogeneous n...	32.2	0.093
gi 117190173 ref NM_004500.3	Homo sapiens heterogeneous nucl...	32.2	0.093
gi 117189974 ref NM_031314.2	Homo sapiens heterogeneous nucl...	32.2	0.093
gi 61966710 ref NM_001013631.1	Homo sapiens heterogeneous nu...	32.2	0.093
gi 30581169 ref NM_178329.1	Homo sapiens chemokine (C-C moti...	32.2	0.093
gi 30581168 ref NM_001837.2	Homo sapiens chemokine (C-C moti...	32.2	0.093
gi 45545404 ref NM_205860.1	Homo sapiens nuclear receptor su...	30.2	0.37
gi 45505141 ref NM_003822.3	Homo sapiens nuclear receptor su...	30.2	0.37
gi 37620205 ref NM_198077.1	Homo sapiens chromosome 1 open r...	30.2	0.37
gi 118136293 ref NM_031371.3	Homo sapiens AT rich interactiv...	28.2	1.5
gi 118136292 ref NM_016374.5	Homo sapiens AT rich interactiv...	28.2	1.5
gi 93102362 ref NM_015045.2	Homo sapiens wings apart-like ho...	28.2	1.5
gi 32967281 ref NM_003337.2	Homo sapiens ubiquitin-conjugati...	28.2	1.5
gi 34101287 ref NM_017437.1	Homo sapiens cleavage and polyad...	28.2	1.5
gi 19923910 ref NM_138386.1	Homo sapiens hypothetical protei...	28.2	1.5
gi 40316929 ref NM_014410.4	Homo sapiens clusterin-like 1 (r...	28.2	1.5
gi 40316925 ref NM_199167.1	Homo sapiens clusterin-like 1 (r...	28.2	1.5
gi 71725366 ref NM_001029996.1	Homo sapiens similar to hypot...	28.2	1.5
gi 7657151 ref NM_014606.1	Homo sapiens hect domain and RLD ...	28.2	1.5
gi 7657268 ref NM_015032.1	Homo sapiens androgen-induced pro...	28.2	1.5
gi 50658085 ref NM_032621.2	Homo sapiens brain expressed X-link	28.2	1.5
gi 39653318 ref NM_013377.2	Homo sapiens PDZ domain containing	28.2	1.5
gi 118600974 ref NM_007269.2	Homo sapiens syntaxin binding prot	26.3	5.8
gi 118572593 ref NM_000579.2	Homo sapiens chemokine (C-C motif)	26.3	5.8
gi 117422441 ref NM_001002258.3	Homo sapiens ATP synthase, H...	26.3	5.8
gi 116805328 ref NM_002958.3	Homo sapiens RYK receptor-like ...	26.3	5.8
gi 116805325 ref NM_001005861.2	Homo sapiens RYK receptor-li...	26.3	5.8
gi 116686121 ref NM_012310.3	Homo sapiens kinesin family member	26.3	5.8
gi 116536086 ref NM_001077197.1	Homo sapiens phosphodiesterase...	26.3	5.8
gi 116536084 ref NM_016953.3	Homo sapiens phosphodiesterase ...	26.3	5.8
gi 46411153 ref NM_024867.2	Homo sapiens KPL2 protein (FLJ23...	26.3	5.8
gi 45580699 ref NM_001776.3	Homo sapiens ectonucleoside trip...	26.3	5.8
gi 115385976 ref NM_080922.2	Homo sapiens protein tyrosine p...	26.3	5.8
gi 115385975 ref NM_002838.3	Homo sapiens protein tyrosine p...	26.3	5.8
gi 115385973 ref NM_080921.2	Homo sapiens protein tyrosine p...	26.3	5.8
gi 110224460 ref NM_000648.2	Homo sapiens chemokine (C-C mot...	26.3	5.8
gi 110224459 ref NM_000647.4	Homo sapiens chemokine (C-C mot...	26.3	5.8
gi 108773800 ref NM_000170.2	Homo sapiens glycine dehydrogen...	26.3	5.8
gi 101943239 ref NM_001520.2	Homo sapiens general transcript...	26.3	5.8
gi 93277100 ref NM_144722.3	Homo sapiens KPL2 protein (FLJ23...	26.3	5.8
gi 93141213 ref NM_001040143.1	Homo sapiens sodium channel, ...	26.3	5.8
gi 93141211 ref NM_001040142.1	Homo sapiens sodium channel, ...	26.3	5.8
gi 93141209 ref NM_021007.2	Homo sapiens sodium channel, vol...	26.3	5.8
gi 93141044 ref NM_198196.2	Homo sapiens CD96 molecule (CD96),	26.3	5.8
gi 93141043 ref NM_005816.4	Homo sapiens CD96 molecule (CD96),	26.3	5.8
gi 37588851 ref NM_003381.2	Homo sapiens vasoactive intestin...	26.3	5.8
gi 37588852 ref NM_194435.1	Homo sapiens vasoactive intestin...	26.3	5.8

```

CAATCGATAGGTAACCTGGCTGTGCT 457
CAATCGATAGATACTGGCTATTGT 788
CGATTGACAGGTAACCTGGCCATGCT 478
* * * * *
AGGACGGTCAACCTTTGGGGTGGTGA 507
AGGACGGTCAACCTTTGGGGTGGTGA 838
CGGACCGTCACTTTGGTGTCAATCA 528
* * * * *

```



■ Questions?

