Computational Biology Project

Introduction

RNA interference (RNAi) is a phylogenetically widespread gene-silencing process triggered by double-stranded RNA (dsRNA). Small interfering RNAs (siRNAs) are an example of RNAi; they recognize, bind, and regulate other complementary cellular RNAs. The introduction of siRNAs to eukaryotic cells frequently results in off-target gene silencing – unintended silencing of related, but not identical, mRNAs. Off-target gene silencing is one of the major concerns during the application of siRNAs-based technologies for gene discovery and the treatment of human disease.

C. elegans is a strong model organism for siRNA research and the investigation of off-target processes. Regarding C. elegans, two distinct populations of small RNAs have been proposed to participate in RNAi: "primary siRNAs" – derived from DICER nuclease-mediated cleavage of the original trigger, and "secondary siRNAs" – additional small RNAs whose synthesis requires RNA-dependent RNA polymerases (RdRPs) [3]. Secondary siRNAs are the result of intracellular RNA amplification and are therefore very abundant when the cell is undergoing RNAi.

The project's main goal is to develop a powerful tool that will allow us to predict what genes could be affected by off-target small RNA regulation when treating the organisms with dsRNA against a specific gene of interest. For optimal detection of off-targets, the understanding of secondary siRNA alignment patterns is crucial.

In the project we used available small RNA data induced by the treatment of C. elegans worms with dsRNA against the dpy-13 gene. These worms were shown to produce many small RNAs against multiple genes other than dpy-13 [1]. We used this data to analyze the homology relations that allow the appearance of off-target small RNA production.

In order to reach our goal, our strategy was to study the constraints of optimal alignments for secondary siRNAs that lead to the production of novel small RNAs in off-targets, and therefore could potentially lead to silencing of off-targets. Subsequently, we intended to translate these findings into a set of optimal parameters for the prediction of off-target effects following different RNAi treatments.

Our research consists of a defined pipeline with three main stages. The first stage was downloading dpy-13 mature mRNA sequence and splitting it to multiple fragments by 25 bp-sized sliding windows. The second stage was executing BLAST by nucleotide for each fragment against the C. elegans DB. The final and most complex stage consisted of studying and analyzing the results with regards to small-RNAs sequencing data depicting off-target small RNA production. Specifically, the BLAST results allow us to generate a set of genes which share homology in their sequence with the dpy-13 gene, hence may be considered as off-target. Since we have a list of off-target which were experimentally detected, we have the opportunity to isolate, out of the full set of genes which share homology with the dpy-13, the ones that are indeed affected by the small RNA against dpy-13. For that purpose, we carefully examined a large spectrum of alignment parameters, such as the extent of identity, the number of matched bases and so on. Then, we determined what was their effect on true-positives and false-positives detection rates, based on the available dpy-13 data. Ultimately, several near-

optimal sets of parameters for off-targets' detection were deduced (since 100% true-positive and 0% false-positive were not found).

Results

Secondary small RNAs are produced on mature mRNA templates, and are therefore complementary to exons only, not introns. Based on this knowledge, we downloaded dpy-13's mature mRNA sequence with its UTR (989 bp).

Stage 1: *splitting by 25 bp-sized sliding windows.*

We split the sequence into 989 fragment repeats of 25bp. The UTR regions were included as well.

Stage 2: *executing BLAST for each fragment and crossing data with the WS220_refSeq* file.

For each fragment, we executed BLAST searches with default parameters and word size of 7bp. BLAST hits do not reveal the subject's gene, therefore we had to cross BLAST hits data with refSeq file data. The refSeq file contains information regarding C. elegans genes' transcripts coordinates, C. elegans genes' chromosome numbers, etc. In order to find the gene of the subject, we considered the genes' chromosome number, transcripts, the subject's strand direction and if the subject's coordinates fall between transcript coordinates. Each subject's gene symbolizes an off-target gene. Our loose BLAST constraints at this stage lead to a huge number of hits – off-targets' surpluses.

Next, we intended to intelligently and rationally refine the constraints, e.g. narrow the optional values for BLAST parameters. We took into consideration the following parameters: Score, E-value, Length, Identity, Matches, D (delta).

```
query_name:Dpy-13:597_621
Score 16 (30 bits), expectation 2.9e-01, alignment length 26
Query:       1 GCACCAGGAAAGCCAGGAGCA-CCAG 25
               |||||||||| |||||| || ||||
Sbjct:12493843 GCACCAGGAAACCCAGGA-CAGCCAG 12493819
```

Figure 2: a BLAST hit example and parameters explanation.

Score weigh the number of matches, mismatches and gaps. **E-value** describes the number of hits one can "expect" to see by chance when searching a database of a particular size, 0.29 in this example. **Length** represent the alignment length. **Identity**=matches/length, equals to 0.92=23/25 in this example. **Matches** is the number of matches in the alignment, 23 in this example. **D** parameter will be discussed in the following paragraph.

D (delta) parameter: Table S2 was taken from [1]. The table here is not complete, and is shown for illustration purpose only.

The table holds off-target RNAi relevant data. It represents the generation of small RNAs against off-target genes with partial homology to the gene that was originally targeted by dsRNA – dpy-13. The first column from the right denotes the number of small RNA reads against a certain (collagen) gene bound to the Argonaute protein NRDE-3 in normal conditions. The second column from the right denotes the same, but after dpy-13 RNAi treatment.

For each gene in the table, D is the first column subtracted from the second. For the first gene in the table, sqt-3, The D parameter, would be 1859–0=1859.

We synthetically decided on some specific cut-offs, below which a certain small RNA phenomenon is considered too weak to have a biologically relevant effect. Those cut-offs are represented as different values of parameter D (7, 19, 59 or 99).

During our research, we used the assumption that a gene with higher D values has a higher probability to be an off-target gene. For the first gene in the table, sqt-3, D value is the largest compared to the rest of the genes in the table. Hence, we will interpret it to have a biologically relevant effect – it is an off-target gene.

<u>stage 3</u>: *analyzing the results and finding an optimal parameters combination.*

We examined all the following parameters combinations:

- Score = {14, 16, 18, 20} each value symbolizes the minimum value that the Score must have in order to be considered as a hit.
- E-Value = {0.001, 0.1, 3, 4}. Each value symbolizes the maximum value for E-value to be in order to be considered as a hit. The typical, default value of E-Value is 10.
- Length = {14, 16, 18, 20}. Each value symbolizes the minimum value of alignment Length that is needed in order to be considered as a hit.
- Identity = {60%, 80%, 100%}. Each value symbolizes the minimum value of alignment Identity needed in order to be considered as a hit.
- Matches = {16, 18, 20, 22, 25}. Each value symbolizes the minimum value of Matches in order to be considered as a hit.
- D = {7, 19, 59, 99}. Each value symbolizes the minimum value that the specific gene delta must have, in order to be counted in *maxD* and *Dgenes* (explained in the following lines).

We will define a few variables/expressions that were used for our calculations:

Table S2 Deep sequencing The number of reads of NRDE-3-associated siRNAs targeting collagen genes [5].

sequence	gene	eri-1(mg366);dpy-13(e458);GFP::NRDE-3;dpy-13(RNAi)	GFP::NRDE-3
F23H12.4a	sqt-3	1859	0
F36A4.10	col-34	454	0
F01G12.5a	let-2	309	1
T11F9.9	col-157	145	1
T28C6.4	col-117	120	0
T28C6.6	col-3	120	0
F41F3.4	col-139	120	0
K04H4.1b	emb-9	105	0
F30B5.1	dpy-13	105	1
C29F4.1	col-125	98	0
EGAP7.1	dpy-3	94	0
F29B9.9	col-111	93	2
F14F7.1	col-98	68	0
Y41C4A.19	col-96	61	1
F55C10.3	col-155	55	1
ZC513.8	col-43	54	0
Y41C4A.16	col-95	39	0
F55C10.2	col-154	38	1
ZK1010.7	col-97	37	0
T07H6.3a	col-166	31	0
T05A1.2	col-122	27	0
D2024.8	col-114	23	0
F57B1.4	col-160	21	0
Y38C1BA.3	col-109	19	0
C34H4.4a	col-107	17	0
F54B11.2	col-44	13	0
F19C7.7	col-110	10	0
F53G12.7	col-45	9	3
ZC373.7	col-176	8	0
F46C8.6	dpy-7	8	5
F57B1.3	col-159	8	0
F22D6.10	col-60	8	1
C09G5.4	col-39	7	0
ZK1290.3a	rol-8	7	0

- ***maxD*** for some d **(**d∈**{**7, 19, 59, 99**})**

  The number of genes (from table S2) experimentally detected, for which D≥d**.**
  Those genes in *maxD* symbolize genes in which the small RNA phenomenon is expected to have a biologically relevant effect, as defined by the synthetic cut-off that we chose – d. Thus, those genes will be predicted as off-target genes for our purpose.

Table 1: maxD values for different values of D.

	D=7	D=19	D=59	D=99
maxD	35	24	14	9

- ***Dgenes* for some d**

  Gene numbers experimentally detected for which:
  1. D≥d **.**
  2. We found at least one hit between dpy-13 fragment and them when executing BLAST.

  The genes that constitute *Dgenes* symbolize the predicted off-target genes that BLAST has found. Hence, those genes will be considered as true-positive off target genes.

- ***true_off-targets_score***

  true_off-targets_score = *Dgenes* / *maxD* (for specific value of d)

  For example, if we take the following parameters:
  Score=16, E-Value=4, Length=18, Identity=80%,Matches=25 and D=59;
  we will see that *maxD*=14, because there are 14 genes detected experimentally for which their D≥14. *Dgenes*=8, because in our BLAST results (under the parameters constraints) we found eight hits from table S2 genes for which their D≥59. Therefore, for that combination *true_off-targets_score*=8/14≈57.14%. *true_off-targets_score* helps us to estimate our off-targets detection. *true_off-targets_score*=100% indicates that we found hits to all genes suspected as off-targets. Hence, we will search such parameter combination that will hold a high *true_off-targets_score* value.

- ***Max_off_off_targets***

  The number of Table S2 genes for which D≤0**.** We found that *Max_off_off_targets*=62**.** In our point of view, those genes do not represent off-targets.

- ***off_off_targets***

  The number of Table S2 genes for which D≤0 and have at least one dpy-13 fragment hit.
  D≤0 is below all D values, all cut-offs. We considered such values as too weak to have biologically relevant effect. When executing BLAST, if a gene (present in table S2) result with D≤0 is found, we counted it as a false-positive off-target gene. That is because this gene was found with the parameter combination in the BLAST search, due to some extent of sequence homology, but is not considered as a true hit since we defined the small RNA effect triggered against it to be weak to induce silencing (based on the experimental results). In other words, we consider the gene as a false-positive hit based on the cut-off we chose under this specific set of parameters.

- ***irrelevant_off-targets_score***

  irrelevant_off-targets_score=(off_off_targets)/*max_off_off_targets=off_off_targets/62***.**
  Considering *max_off_off_targets* and *off_off_target* definitions, we will look for a parameter combination that will give us the smallest *irrelevant_off-targets_score* we can possibly get (ideally, the predicted off-target genes **are** relevant).

In light of the above *true_off-targets_score* and *irrelevant_off-targets_score* interpretations, the theoretical optimal BLAST parameters' combination (that lead to 100% true off-target genes detection, and 0% false-positive off-target genes) would give rise to high *true_off-targets_score* value and low *irrelevant_off-targets_score* value.

To discover the best combinations, we analyzed our data in the following ways:

1. **Observing the data from single parameter point of view:**

   A. <u>The BLAST's Score value:</u>

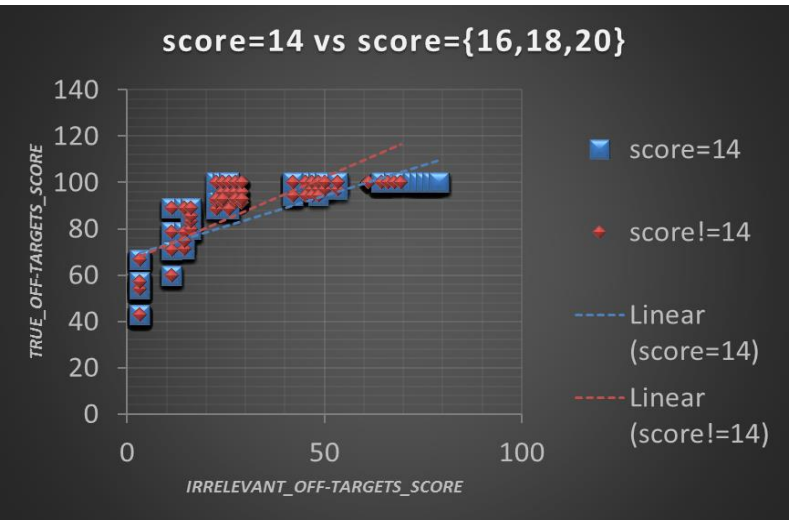
   The X- and Y-axis are *irrelevant_off-targets_score* and *true_off-targets_score*.

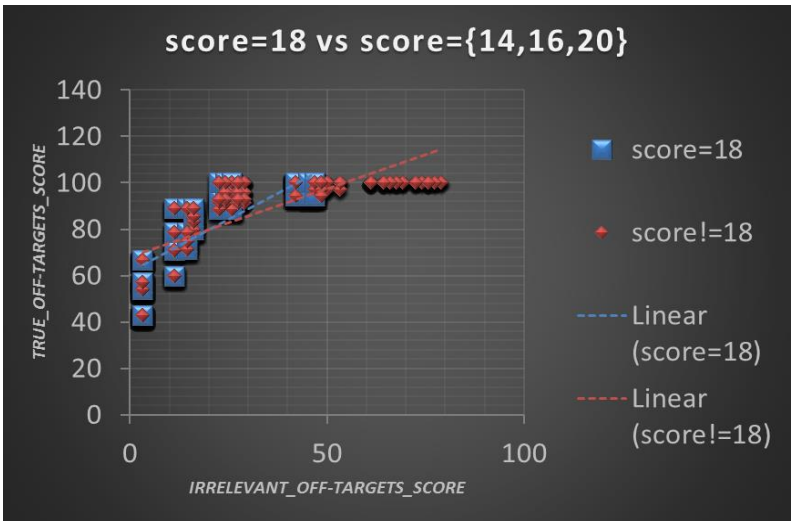
   Each dot in the graphs below represents some values of *true_off-targets_score* and *irrelevant_off-targets_score* values. Those values were calculated after executing BLAST search with specific combination of values for the parameters: Score, E-value, Length, Identity, Matches and D.

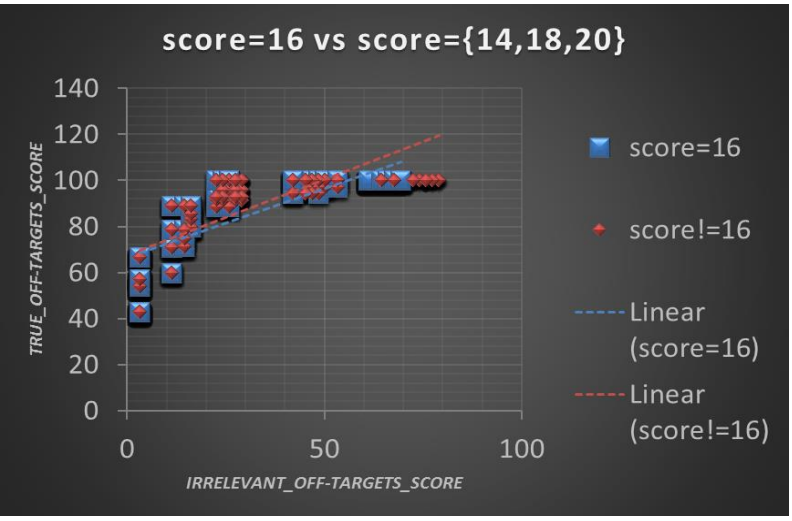
   The same set of coordinates is presented four times – each time the coloring splits the dots differently. The blue color stands for the dots that have a specific score value, as depicted in the title; the red color stands for dots with all others different score values.

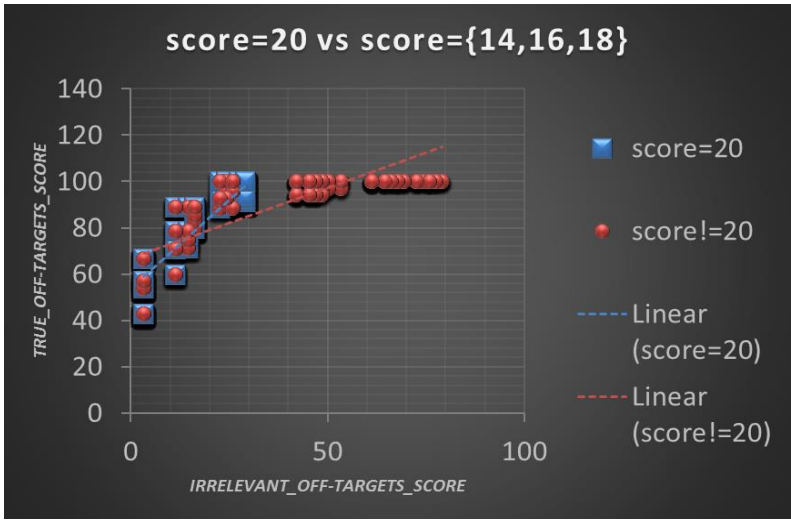
   For instance, let us observe the upper left graph: all the combinations with score=14 are colored in blue, while in red we see the rest of the combinations that have score values of 16, 18 or 20.

   This visual presentation helps to better understand how different score values affect the probability of detecting both true off-targets and false-positive hits.





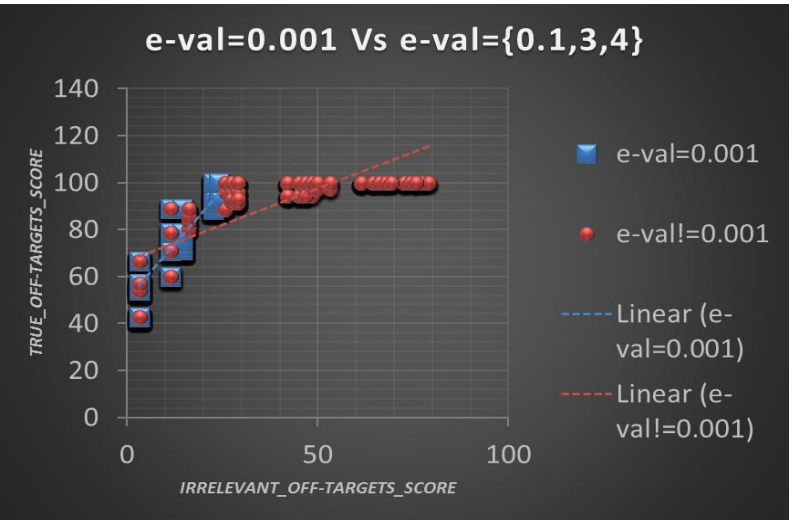


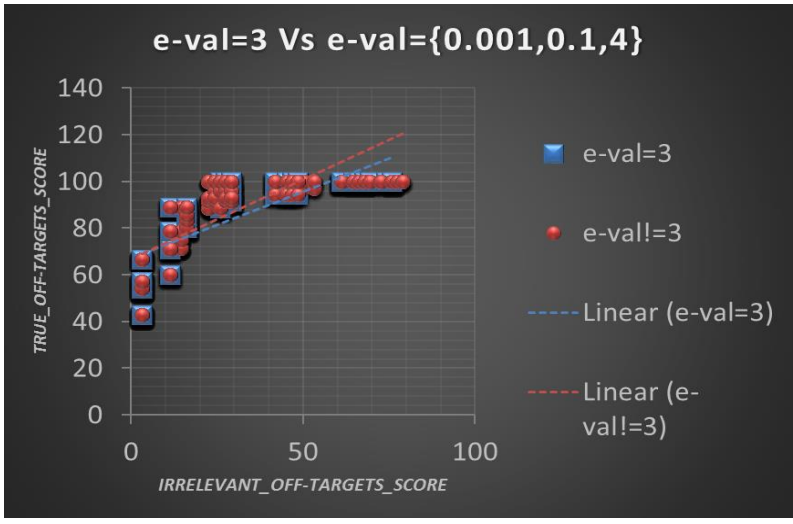


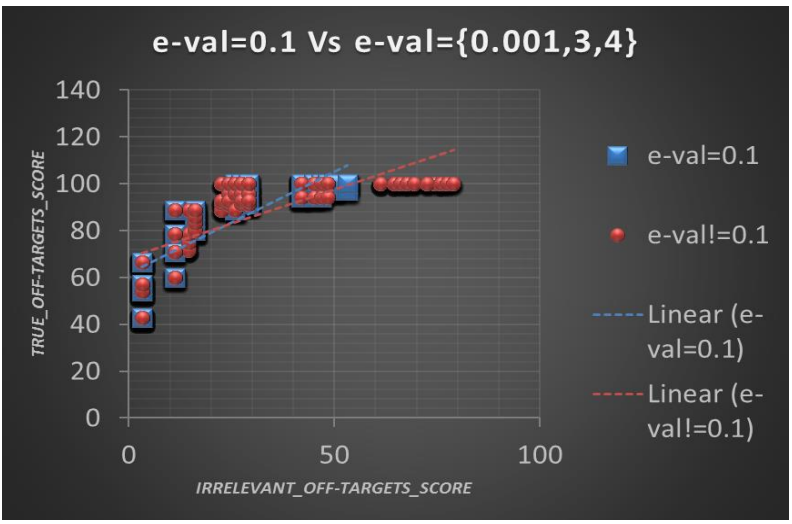
It can be seen that as the score value grows, the blue dots "move" to the left, meaning that the combinations with higher score values have smaller *irrelevant_off-targets_score* values. This result suggest that that greater score values in BLAST search will lead to smaller number of false-positives – incorrect off-targets' detections.

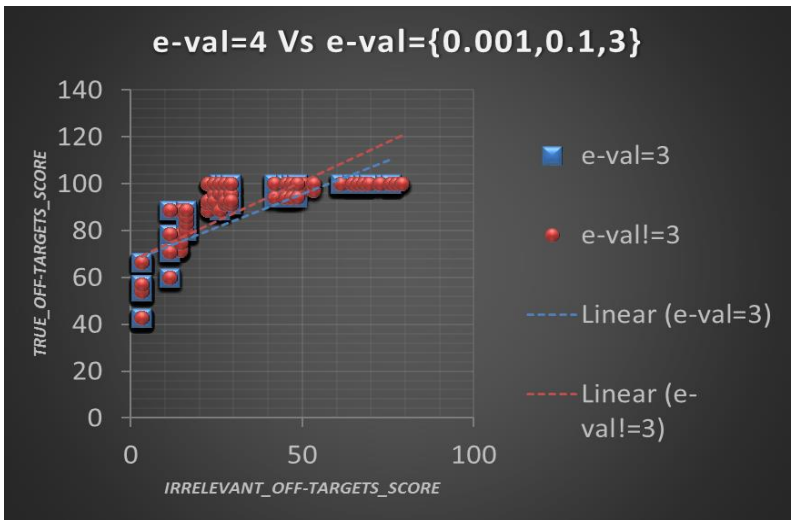
We next made the same visualizations regarding the other parameters, as we did regarding the Score parameter.

B. E-value:







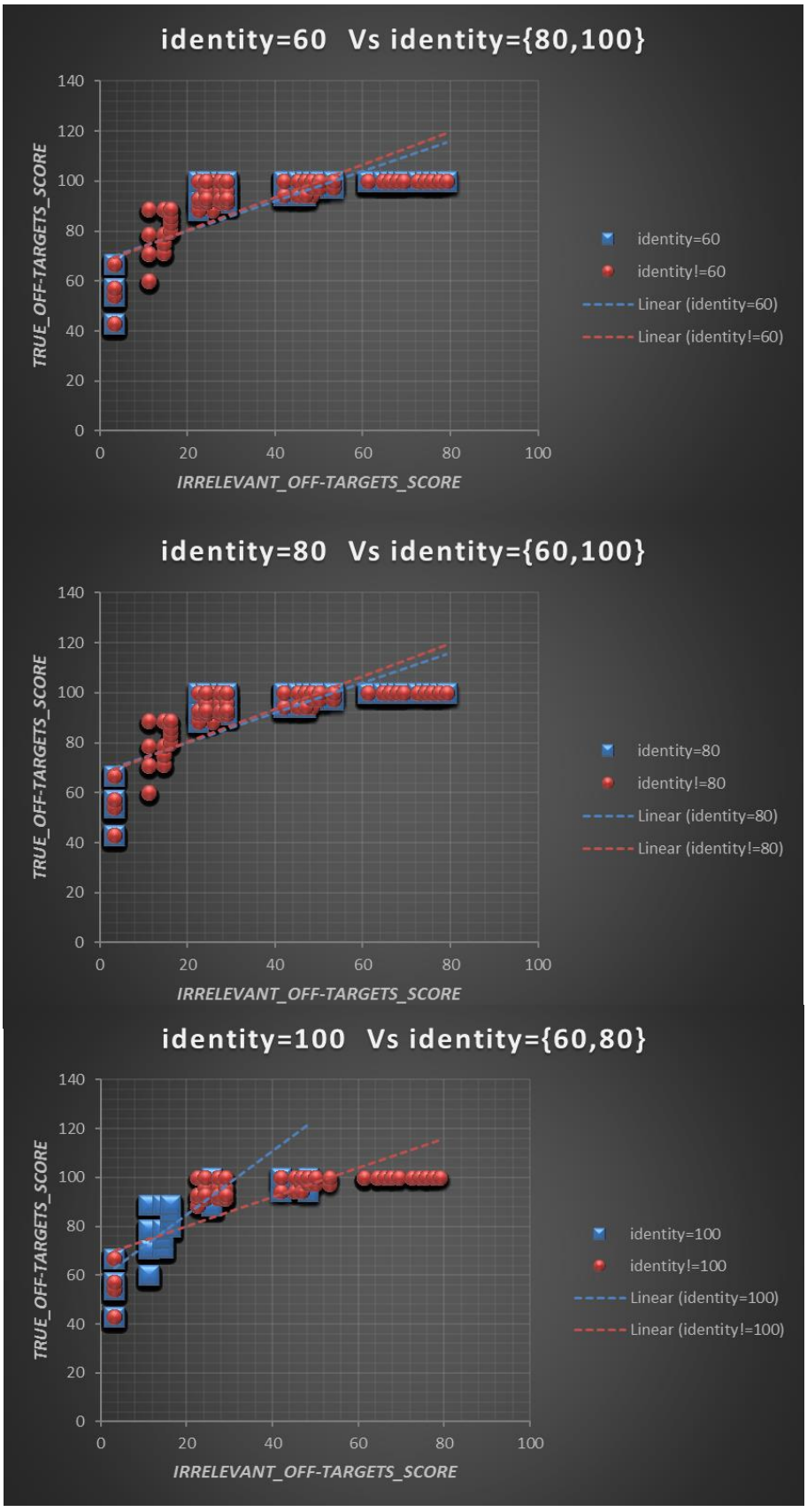


In the above graphs, as opposed to the Score graphs, we can see that as the E-value grows, we get points with higher *irrelevant_off-target's_score* values. This is expected, since the E-value describes the number of hits one would "expect" to see by chance when searching a database of a particular size. In addition, we can see that trendlines of greater E-value constitute of more points for which *true_off-targets_score*=100, i.e. they represent off-targets' correct detection. The objective is to find such values that would minimize the first and maximize the latter.

Length: Length graphs were not added, since the graphs that were produced by the data had minor differences between one another (we checked for the values:

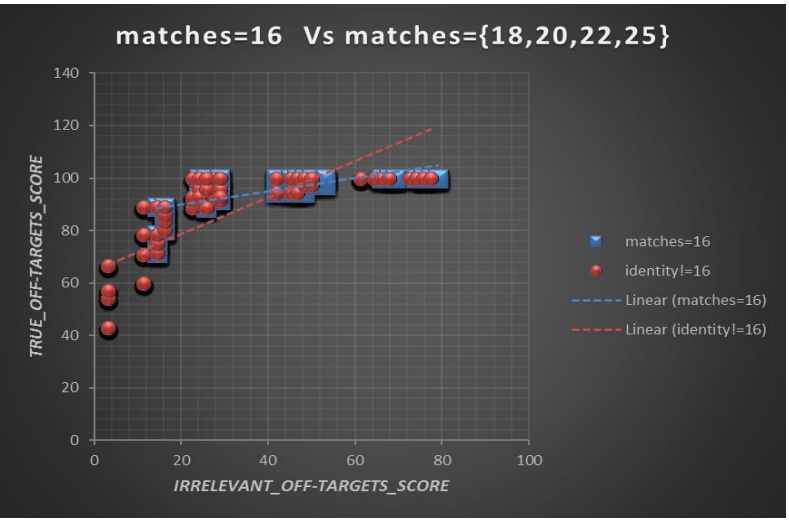
{14,16,18,20}). In essence, Length value of 20 cause slightly less false positive off target detection (lower *off_off_targets* values).

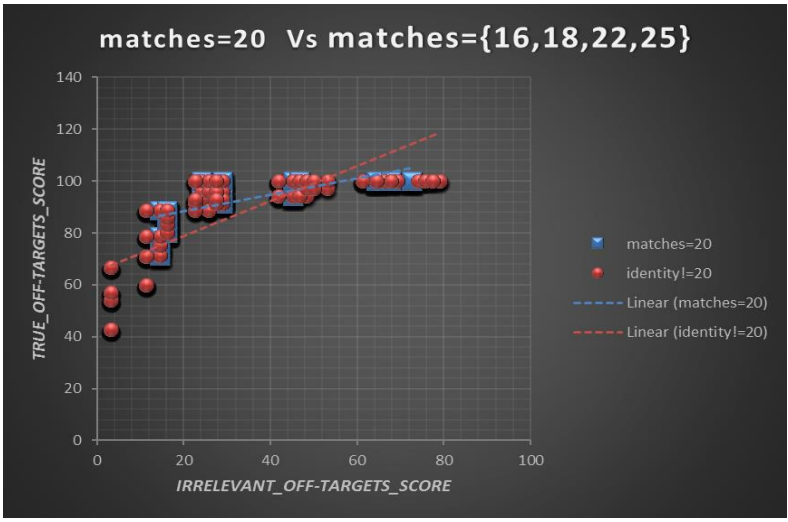
C. Identity:

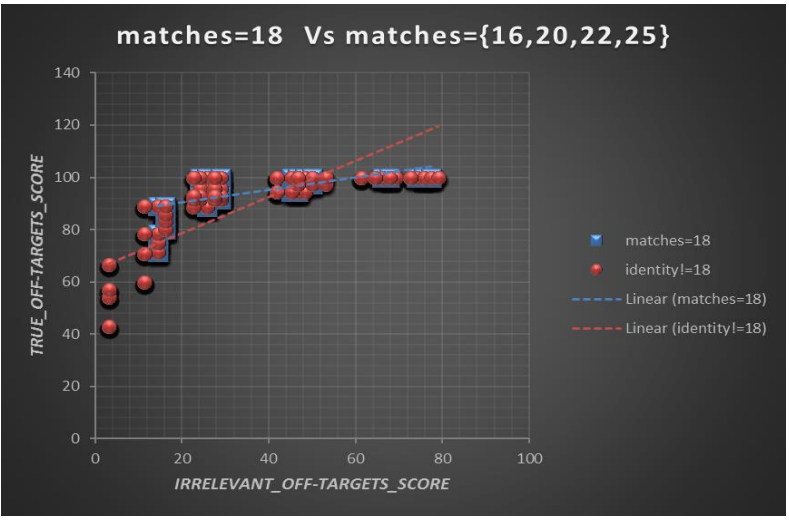


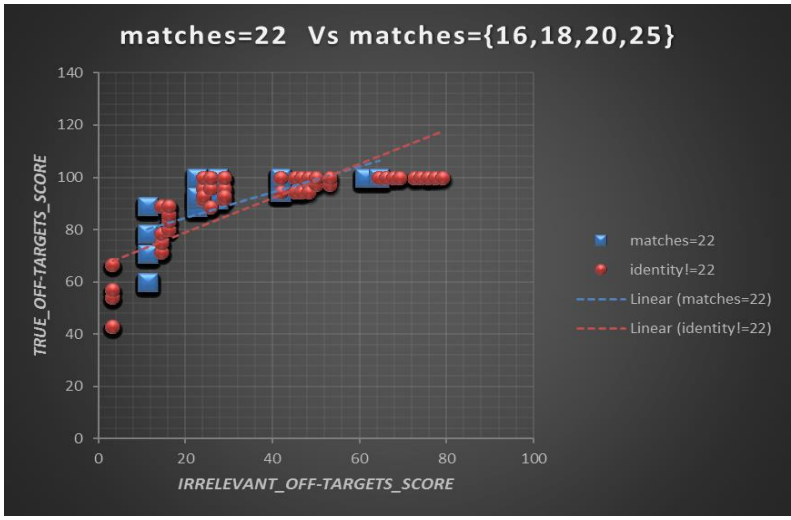
Here we can see that the population of Identity=100% spreads differently among the graph, compared to the other population. Identity=100% promises smaller number of false-positive detections.

D. Matches:



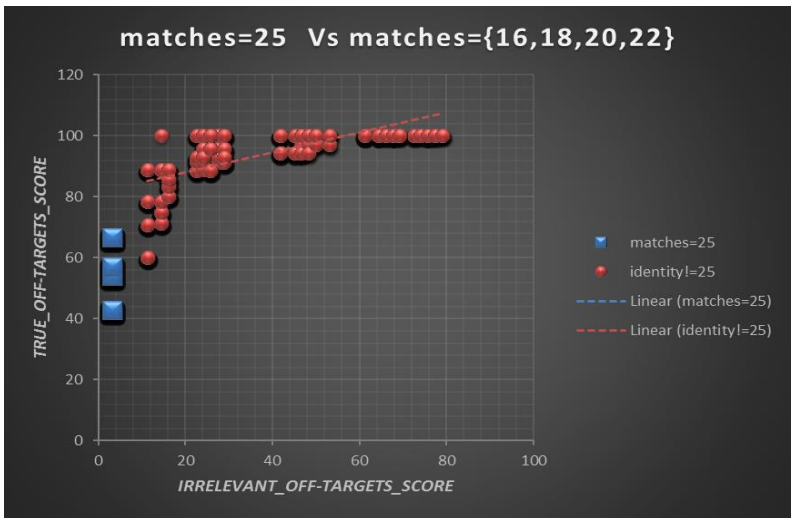






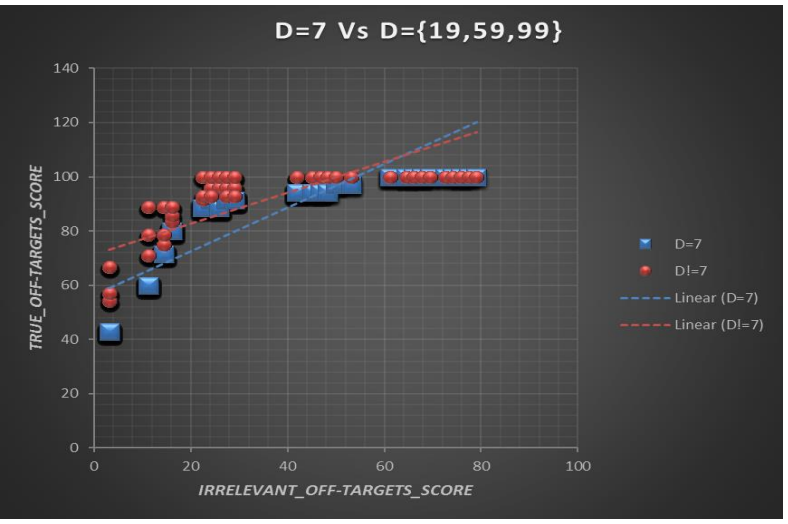
In general, we see that as the Matches value grows, we get points with smaller *irrelevant_off-targets_score* values.

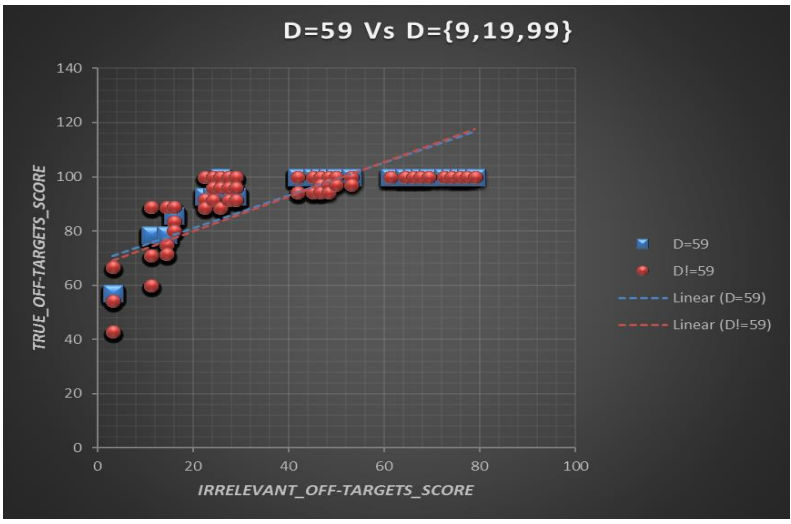
Matches=25 is an extreme value for match. On the one hand, it gives a low *irrelevant_off-targets_score*, but on the other hand, it also gives low *true_off-targets_score* values. Those values can suggest that using the Match value of 25 in order to detect off-target genes using BLAST might cause very limited off-target detection.

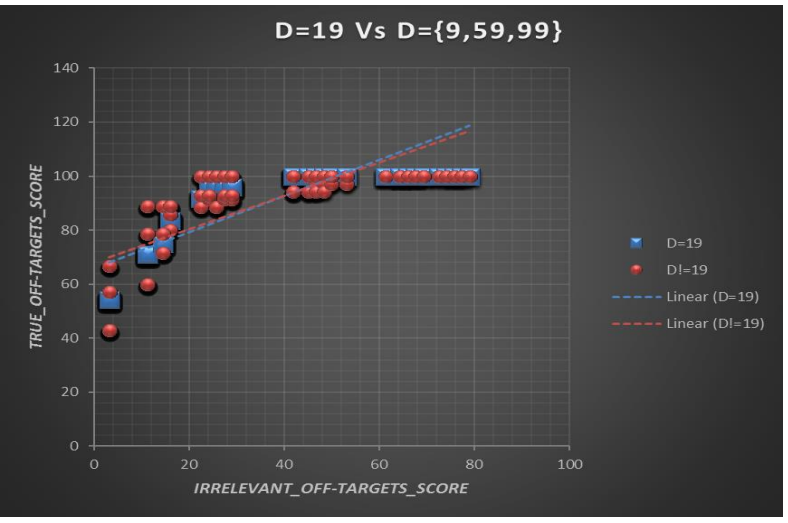


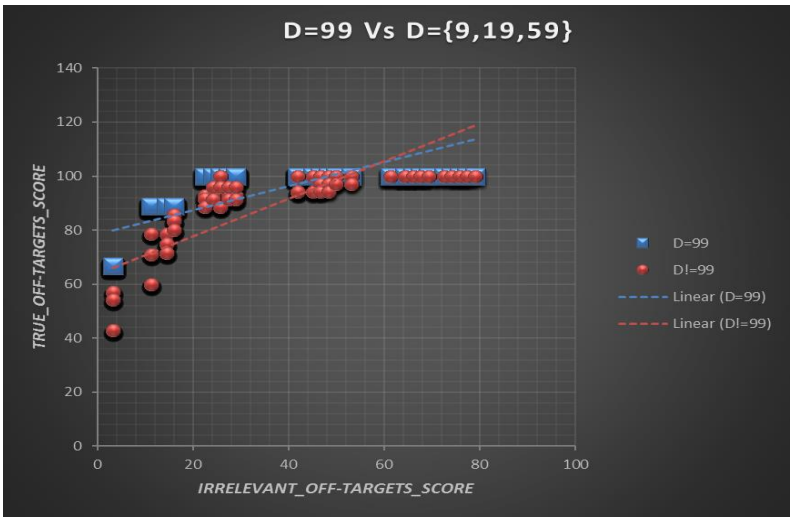
E. D:

D refers to the minimum delta value that a gene in table S2 (page 3) must have in order to be considered as a true-positive off-target. That effects the *true_off-targets_score* calculation (as described in page 4).









Based on the blue dots, graphs that show higher D values exhibit 'better' trendlines. Better, meaning that it meets the Y-axis higher and its slope is smaller. That is better, since we estimate the upper left points as ideal points that symbol good and reliable off-targets' detection.

After observing the data by populations for one parameter at a time, we wanted to study and better understand how the different parameters interact with each other. To do so, we first checked how many combinations in total have specific ranges of values:

2. Number of combinations for specific *true_off-targets_score* and *irrelevant_off-targets_score* range values

The X-axis represents the *irrelevant_off-targets_score*, while the Y-axis represents the *true_off-targets_score*. As previously discussed, each parameter combination has a single *true_off-targets_score* score and a single *irrelevant_off-targets_score* score. The yellow squares hold relatively high *true_off-targets_score* values and small *irrelevant_off-targets_score* values. Thus, we would want to explore which parameter values construct those combinations, aspiring to find conclusive values for each parameter type (Score, E-value, Length, Identity, Matches, D).

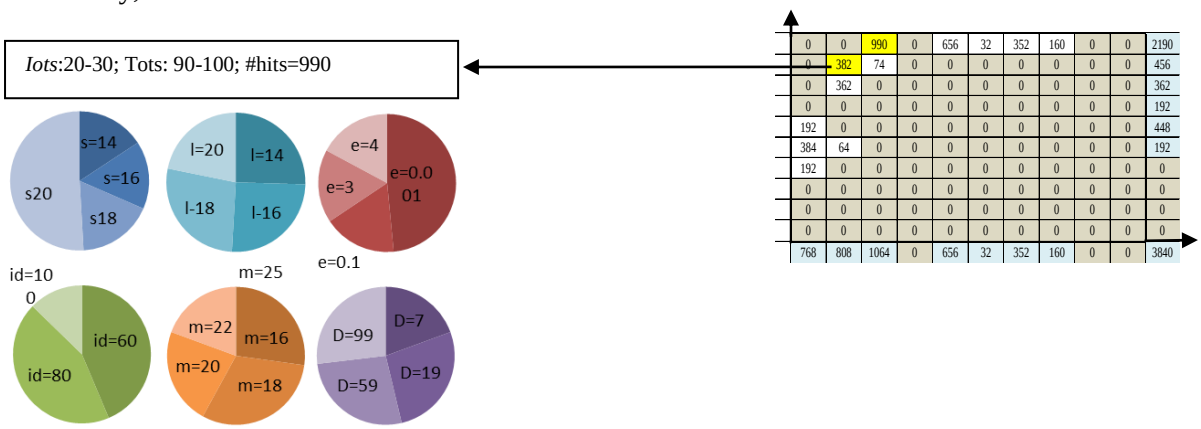
true_off-targets_score (Tots)

Tots \ Iots	0	10	20	30	40	50	60	70	80	90	100	
100	0	0	990	0	656	32	352	160	0	0	2190	
90	0	382	74	0	0	0	0	0	0	0	456	
80	0	362	0	0	0	0	0	0	0	0	362	
70	0	0	0	0	0	0	0	0	0	0	192	
60	192	0	0	0	0	0	0	0	0	0	448	
50	384	64	0	0	0	0	0	0	0	0	192	
40	192	0	0	0	0	0	0	0	0	0	0	
30	0	0	0	0	0	0	0	0	0	0	0	
20	0	0	0	0	0	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	0	0	0	0	
0	768	808	1064	0	656	32	352	160	0	0	3840	

irrelevant_off-targets_score (Iots)

Table 2: Each square represents the number of combination we found (out of our 3840 combinations) with the square specific range of *true_off-targets_score* (*Tots*) and the *irrelevant_off-targets_score* (Iots).

Diving in: next, we examined the yellow squares (representing high detection of true positive and low detection of false positive), in order to carefully observe which parameter values dominate there. In the upper line, from left to right, we can see the distributions of Score, Length and E-value. At the lower line we can see from left to right the Identity, Matches and D distributions.



More than half of this square's combinations were with **Score** value of 20. In addition, **E-value** of 0.001 constitutes considerable amount of the combinations, **Length** is balanced between the different values, **Identity** values of 60 and 80 are the major Identity values for the combinations in the square, and both have the same amount of domination. As we have previously seen in the three Identity graphs, when comparing combinations by those two identity values, difference had not been

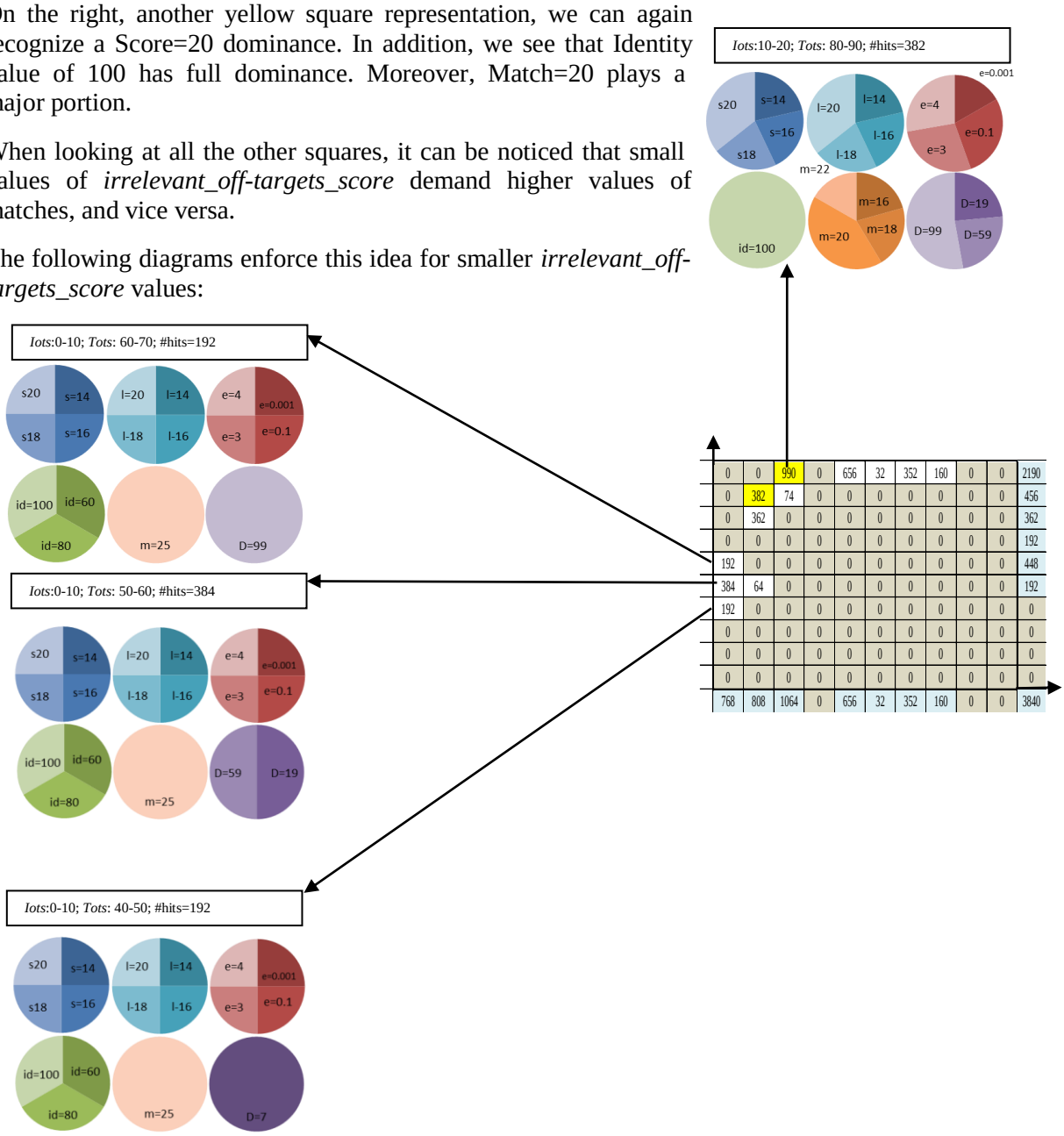
seen. Thus, we will expect both Identity values to produce the same or similar results. Regarding **Matches**, we see zero combinations with Matches=25 in the square. Matches=18 is a significant value in those combinations. **D** values are pretty balanced with disadvantage for the value of 7.

We will notice few more squares:

On the right, another yellow square representation, we can again recognize a Score=20 dominance. In addition, we see that Identity value of 100 has full dominance. Moreover, Match=20 plays a major portion.

When looking at all the other squares, it can be noticed that small values of *irrelevant_off-targets_score* demand higher values of matches, and vice versa.

The following diagrams enforce this idea for smaller *irrelevant_off-targets_score* values:



All of the following graphs are characterized by *irrelevant_off-targets_score* of 0-10. It can be noticed that in all of the graphs Match=25 is the single value for Matches in the combinations. Because of that, parameters such as Score and Length with maximum value of 20 and Identity become irrelevant, because an alignment with 25 matches guarantees Score and Length of above 20 and probably a 100% Identity. The reason for the balanced E-values are not clearly known.

When we watch the diagrams for squares with *irrelevant_off-targets_score*>40 we find zero combinations that hold E-value of 0.001.

After analyzing the various squares, here are some key notes:

- Ideally, we wanted the highlighted squares' diagrams to have one or two dominating values for each parameter that is distinct when comparing to other unattractive squares' diagrams (with high *irrelevant_off-targets_score* and low *true_off-targets_score*). That is not the case, so we have to comprehensively study the data in order to extract the best parameters' values.

- Score value of 20 (which is determined by matches, mismatches and gaps) seems to be ideal in the best (yellow) squares. Squares with higher *irrelevant_off-targets_score* values (higher than 30%) displayed 0% combinations with Score=2*0*. Its seems there is a correlation between low Score value presence in a square (when the combinations majority in the square have low score value) to high *irrelevant_off-targets_score*: the square with *true_off-targets_score* range of 90%-100% and *irrelevant_off-targets_score* range of 40%-50% consists of more than 50% combinations of Score=18. The square with *true_off-targets_score* range of 90%-100% and *irrelevant_off-targets_score* range of 50%-60% consists only of Score=18 and Score=16 divided evenly. The square with *true_off-targets_score* range of 90%-100% and *irrelevant_off-targets_score* range of 60%-70% is consists only of Score=14 and Score=16. The square with *true_off-targets_score* range of 90%-100% and *irrelevant_off-targets_score* range of 70%-80% is consists only of Score=14.

  That gives us a hint for the recommended score choice for off-targets' detection. The testing of additional greater Score values, such as 22 and 25, could be beneficial for examining the balance between Score and Matches (mainly in squares in which *irrelevant_off-targets_score* is smaller than 30%).

- An E-value range between 0.001 to slightly more than 0.1 seems to be ideal to guarantee few false-positives, and yet find correct off-targets. Nonetheless, another value between 0.1 and 3 could be helpful to get more precision. Indeed, this range of values dominates the upper yellow squares combination (assuming that many of the E-value=3 could be caught with smaller values)

- Length did not provide us any information; testing additional Length parameter values are required in order to understand the parameter's influence on off-target genes detection.

- In each square diagram, the Identity values 60 and 80 have equal number of combinations. On the upper yellow square Identity=80 (and of course 60) was dominant, while in the other yellow square Identity=100 was completely dominant. Therefore, we will recommend to choose Identity values in the range of 80-100, depending on the weight given to off-targets' detection against false-positive detection. When detecting maximum number of off-targets is the most important goal, we will aspire to take the 80 value, and vice versa.

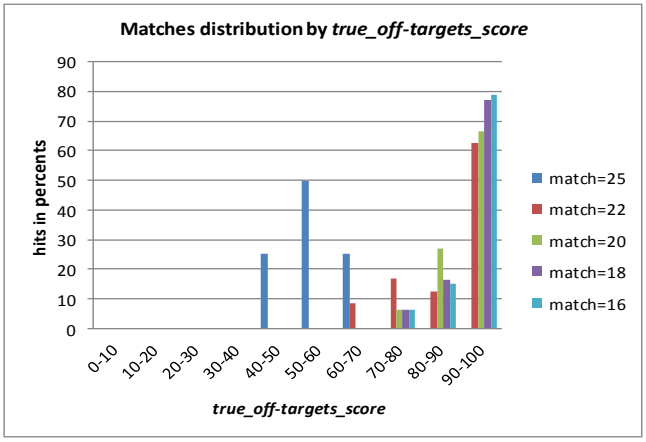
- When analyzing the D parameter in different squares, we did not see any clear tendency. We tried to find a correlation between D values and other parameter values, but did not find any noteworthy results.

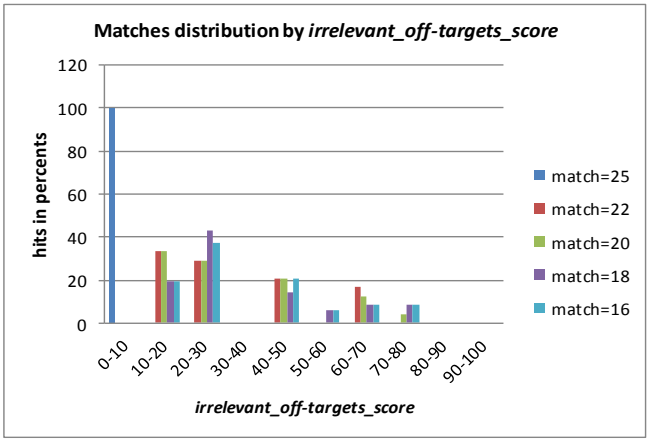
- Matches values in range of 18–20 seem to produce the most beneficial results. Here we will again recommend choosing values by judging our major interest: if the major goal is finding as many off-targets as possible, permissive Match

value of 18 would be recommended; otherwise, in case one puts emphasize on reliability, the stringent Match value of 20 would be suggested.

The parameter Matches is the most detailed among our parameters (with five optional values). Finding the ideal value for this parameter will significantly help us detect off-targets, as opposed to D. Therefore, let us view this parameter from a different perspective:

3. Matches density distribution diagrams

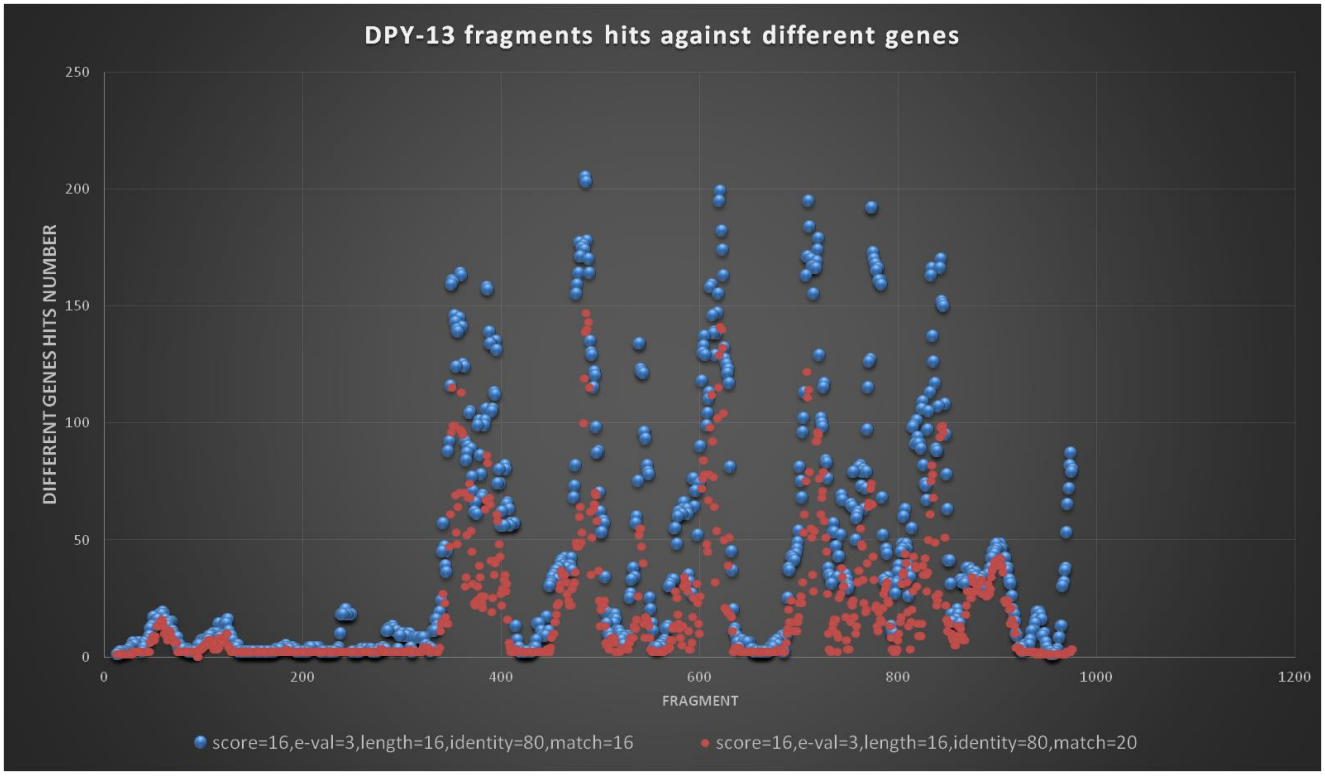




For each Matches value, we checked how all the combinations with that value distribute on *irrelevant_off-targets_score* and *true_off-targets_score* correspondingly; both axes' units are percentages. On the one hand, Match=25 guarantees 0-10 *irrelevant_off-targets_score* (by the left graph); that is: a small amount of false-positives. On the other hand, it has a bad chance for off-target detection. Those two graph show us the interplay between *irrelevant_off-targets_score* and *true_off-targets_score*.

We also prepared some additional visualizations, to understand what was Matches' effect on dpy-13 off-targets:

4. dpy-13 fragments hits against different gene



In the present graph, the X-axis represents fragments of 25bp-long sequences of dpy-13. E.g., the value 1 on the X-axis represents the dpy-13 subsequence of 1bp-25bp (notice, that 1, and not 0, is the first coordinate). The Y-axis represents the number of different genes' hits that have been found for each 25bp-long sequence.

The algorithm could be used to visualize which part of a certain gene shares common homology with other genes in the genome. This can be a helpful view when planning an experiment in which one wants to amplify RNAi to as much genes as possible with minimal resources or minimum work. We can see an example for this use in the graph above.

The graph demonstrates dpy-13's hot-spot areas (with homology to many other genes) that could potentially be responsible for the small RNA targeting of other genes consecutive to the RNAi amplification process. The only difference between the red and the blue dots is the Match parameter. This is another observation for the Match parameter values' influence.

Here, the visualization of two different Match values shows that for the value Match=16, the 3' prime UTR region (position 980 in the X-axis) of dpy-13 is a hot spot to ≈85 genes, while for the value Match=18, the region is not considered a hot spot. If one wants to research the implications of 3' UTR region off-targets, this data could be very meaningful.

Discussion

In this section, we describe the pipeline for developing a new prediction tool for off-target RNAi effects. Our main procedures in the project were splitting the dpy-13 mature mRNA with UTR regions by 25bp-sized sliding windows, executing BLAST search (with loose parameters) for each fragment, and then researching various parameters' impacts. We utilized published data [1] which details off-targets of dpy-13 which were experimentally identified in order to examine our predictions . Our study combined various methods of data visualizations. From those, we infer number of insights (mentioned at page 12). We will emphasize that the ideal parameter combination according to our tool is experiment-dependent.

The most powerful parameters for detection were Matches and E-value. The ideal values were 18-20 for Matches, and 0.001 to x where $x \in (0.1, 3)$ for E-value. The E-value parameter typically serves as a filter. A clever choice of E-value can have a great impact on avoiding false-positive off-targets. On the other hand, E-value which is too low can harm off-target genes' detection. Therefore, Additional E-value values in range of 0.1-3 could have been very helpful to give an explicit suggestion for E-value selection. Moreover, performing the pipeline with additional values for Length (Length>20) and Identity (Identity in the range of 80-100) can make them useful.

A successful prediction tool would be tremendously beneficial biologically. The tool could help understand how an RNAi effect against one gene could affect the expression of other homology-related genes, creating new regulatory connections between genes. Given the known ability of small RNAs to be exported to other cells, such regulatory networks could impact gene expression patterns across tissues and generations (mediated by regulation in the germ line). By combining tissue-specific gene expression data with the power of this new tool, the predictions could be refined in a tissue-specific resolution. The predictions could be an aid for the design of *in vivo* experiments by anticipating potential off-targets effects or focusing the researchers studying off-target phenomena on specific targets predicted via the algorithm.

We will recommend to refine and further study the different parameter combinations with the same pipeline procedures, extra parameters values and additional experimental data. The chosen combinations should then be tested by RNAi experiments, coupled with small RNA and mRNA sequencing. These experimental validations would allow detecting the optimal parameter set with the best prediction power for further use.

References

[1] https://www.ncbi.nlm.nih.gov/pubmed/24532782
Nuclear RNAi Contributes to the Silencing of Off-Target Genes and Repetitive Sequences in Caenorhabditis elegans.
Xufei Zhou,1 Fei Xu,1 Hui Mao, Jiaojiao Ji, Meng Yin, Xuezhu Feng,2 and Shouhong Guang2
School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230027, People's Republic of China

[2] https://www.ncbi.nlm.nih.gov/pubmed/15644189

RNA interference spreading in c.elegans; May RC1, Plasterk RH.

[3] http://science.sciencemag.org/content/315/5809/241

Distinct Populations of Primary and Secondary Effectors During RNAi in *C. elegans;* Julia Pak, Andrew Fire

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1855192/

A Biochemist's Guide to *C. elegans ;* Ann K. Corsi

Supporting Information

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4012473/bin/supp_197_1_121__index.html
Table S2 The number of reads of NRDE-3-associated siRNAs targeting collagen genes.