Tutorial: using SIFTED to design monomeric TALEs

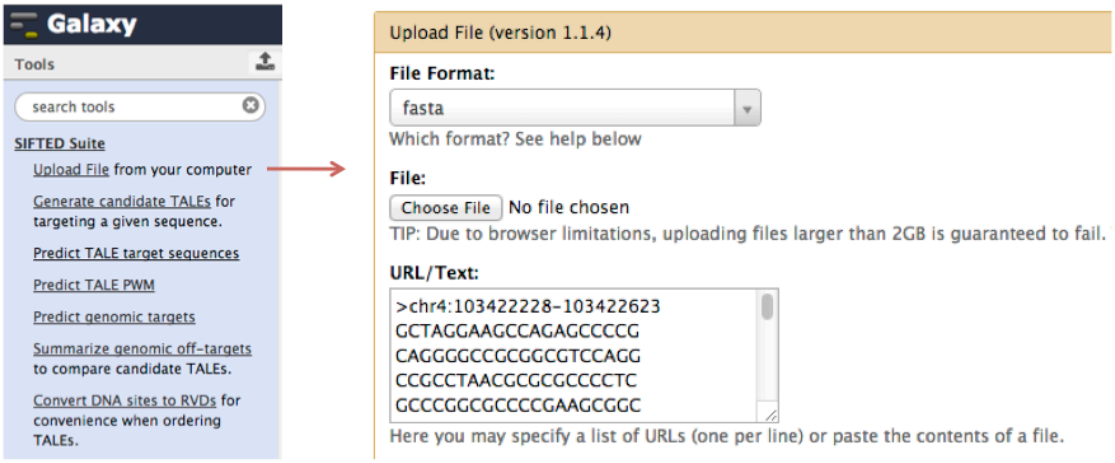
1. Getting started

The SIFTED suite is implemented as a set of online tools within Galaxy, an online bioinformatics platform. If you are already familiar with Galaxy, you should find SIFTED very intuitive. Even if you are not, this tutorial will guide you through all the necessary steps. If you would like to learn more about Galaxy before using SIFTED, you can find an excellent tutorial here:

https://usegalaxy.org/u/aun1/p/galaxy101

2. Loading data into Galaxy

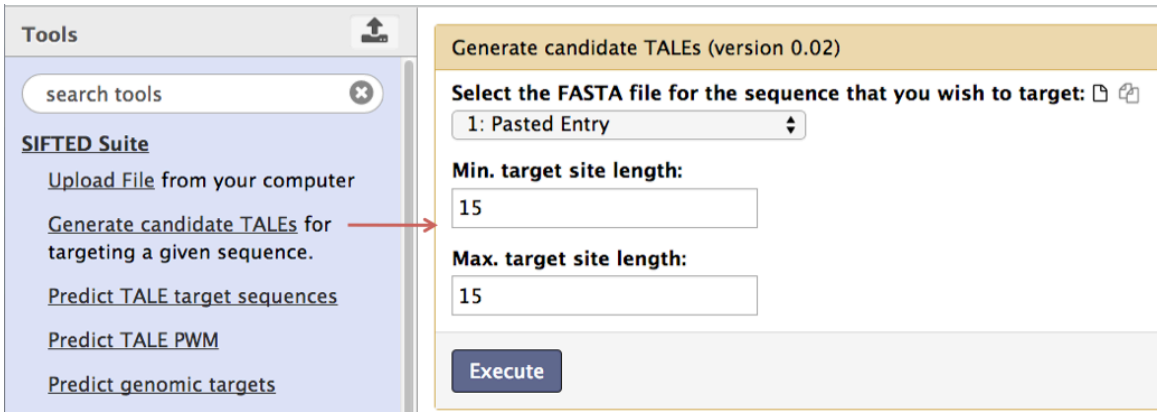
To start using SIFTED, simply click on "SIFTED Suite" at the top left of the screen, which will open the tools panel. Click on "Upload File" and either select the FASTA file containing the sequence you want to target or paste it in the "URL/Text" box. (A FASTA file is just a text file with a header line that starts with ">" and the DNA sequence starting in the next line). In addition, you should select "fasta" in the File Format box. In the example below, we seek to target a genomic region the promoter region of the NFKB1 gene, so we have pasted its corresponding sequence into the box. If your sequence corresponds to a particular genome assembly, you should also enter that information in the "Genome" box.



After you hit the "Execute" button, the sequence you uploaded will now appear on the "History" column at the right of the screen, from which you can use the different buttons to view, download and edit files.

3. Generating candidate TALEs

The locus we uploaded contains many potential subsequences that could be targeted by a TALE. The first step in finding the optimal TALE is to enumerate the possible TALE target sequences. This is done by clicking on the "Generate Candidate TALEs" tool.



Here, you can constrain the length of the target site you would like to use by including both minimum and maximum length values (in bp). In this example, we use 15 base pairs, which would correspond to a 13.5 RVD TALE. When ready, hit the "Execute" button. Your history will now contain a file of candidate target sites, which includes all valid subsequences of the length you specified.

Guidelines for selecting binding site length:
There is no definitive consensus for selecting an optimal TALE length. However, based on our PBM experiments, we have observed that TALEs approaching 20bp/18.5 RVDs can have highly degenerate binding. Therefore, we recommend starting with a single binding site length of 15-18 bp, which has been shown to be sufficient to drive TALE activator activity in most cases (Maeder et al., *Nat Biotechnology,* 2013*)*. Although TALEs with longer binding sites (18.5 RVDs/21 bp) drove higher expression, SIFTED predicts that proteins of this length will typically have a large number of off-target binding sites. This trade-off should be kept in mind during experimental design, as different lengths may be preferable depending on the goal (for more details, see Figure 1 in Maeder et al.). If the initial length setting does not lead to an adequate TALE candidate, the range can be expanded. However, be aware that the running time can increase drastically at longer lengths (hours instead of minutes). This can be partially compensated by using a more stringent K_d threshold, as described in the next step.

Modification:
You can use SIFTED while providing your own set of candidate TALEs. For example, if you used another TALE design package and would like to predict the off-targets of your candidate proteins, you can upload a FASTA file (as in step 2)

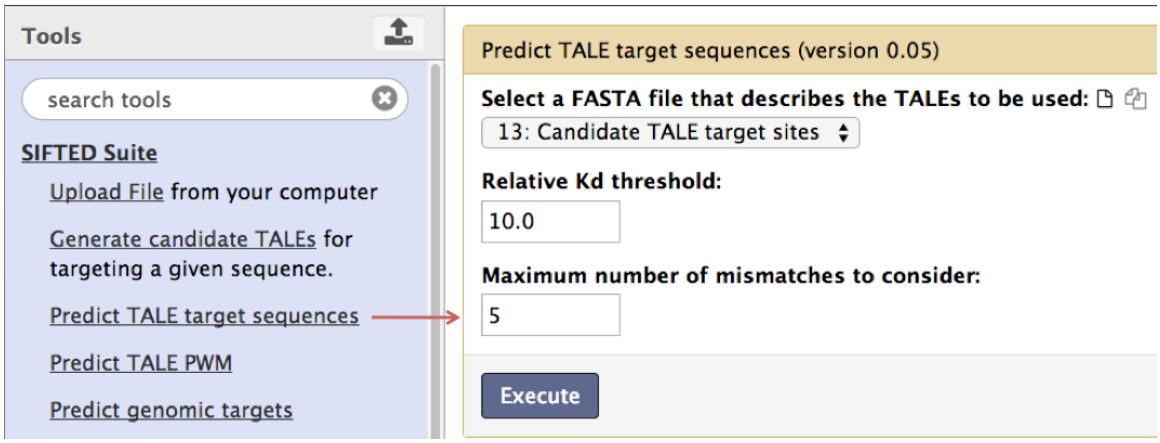
that contains their target sites (as predicted by the TALE code), as in this example:

```
>candTALE-1
TAGGAAGCCAGAGCC
>candTALE-2
TCCAGGCCGCCTAAC
>candTALE-3
TAACGCGCGCCCCTC
```

and continue onto Step 4.

4. Predict target sequences

The next step is to predict the sequences that will be targeted by each potential TALE. This is accomplished with the "Predict TALE target sequences" tool, shown below.



Here, it is necessary to pick a threshold for what is considered an off-target. In this example, we use a relative K_d threshold of 10. This means that all sequences predicted to be bound at up to 1/10th the affinity of the canonical target site sequence will be considered as potential-off targets. In addition, the user can select a maximum number of nucleotide mismatches (from the binding site predicted by the TALE code) to consider. If a genomic sequence has more than this number of mismatches, it will not be counted as an off-target. Reducing this value will typically make the SIFTED pipeline run faster, but may cause some off-target sites to be missed. By default, it is set at 10, which is highly permissive (will not exclude almost any sites). Make sure the "Candidate TALE target sites" file is selected and click the "Execute" button.

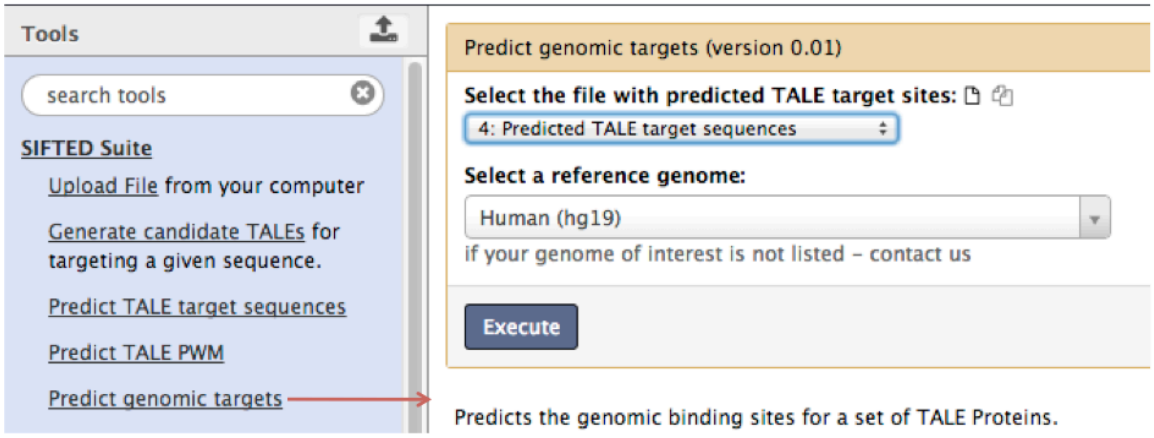
Setting a K_d value

Whether a given genomic sequence will be bound by a TALE depends strongly on the protein's concentration. At a high TALE concentration, off-target sites that

may have remained unoccupied at lower concentrations may be bound. Therefore, we recommend that users carefully compare the results obtained with various K_d thresholds. A threshold of 2 will allow the rapid identification of high affinity off-target sites for hundreds of TALEs. A threshold of 10 will provide a more comprehensive list, which can be used once the candidate sequences have been narrowed down to a few candidates. We recommend that the user start with a threshold of 5-10. If the running time is acceptable, the threshold can be increased to make the off-target list more comprehensive. Once a single candidate is being considered, the tool can be re-run with a high threshold (15-20), which will allow the user to manually inspect the off-target list and ensure that no problematic off-target sites are present.

5. Predicting genomic target sites

To choose the best TALE from the set of candidates, we need to compare the number of genomic off-targets for the different proteins in our set.

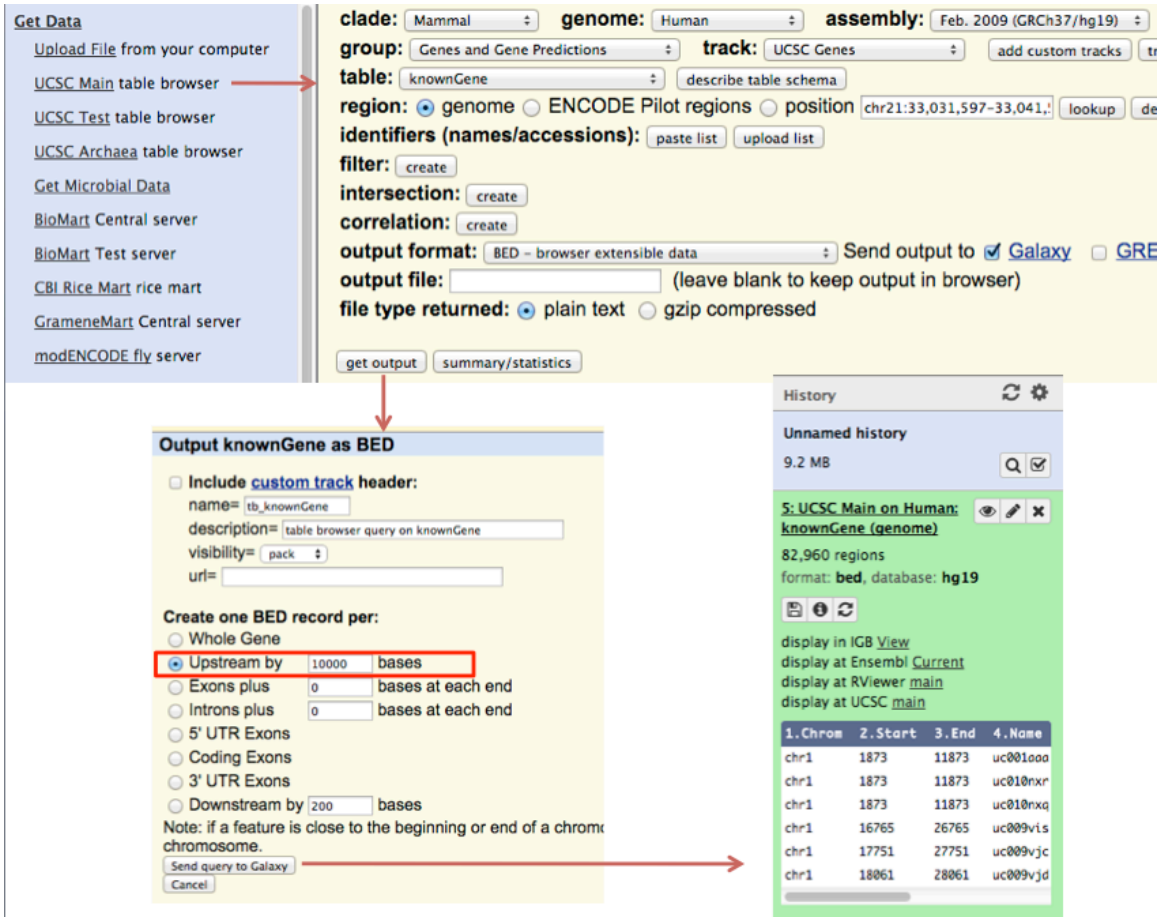


To do this, click on the "Predict genomic targets" tool. All you need to do is select the reference genome you wish to use. We provide many commonly used genomes as part of the SIFTED server. If you do not find the one you are looking for, please contact us and we will install it for you. Make sure you have selected "Predicted TALE target sequences" at the top and hit the "Execute" button.

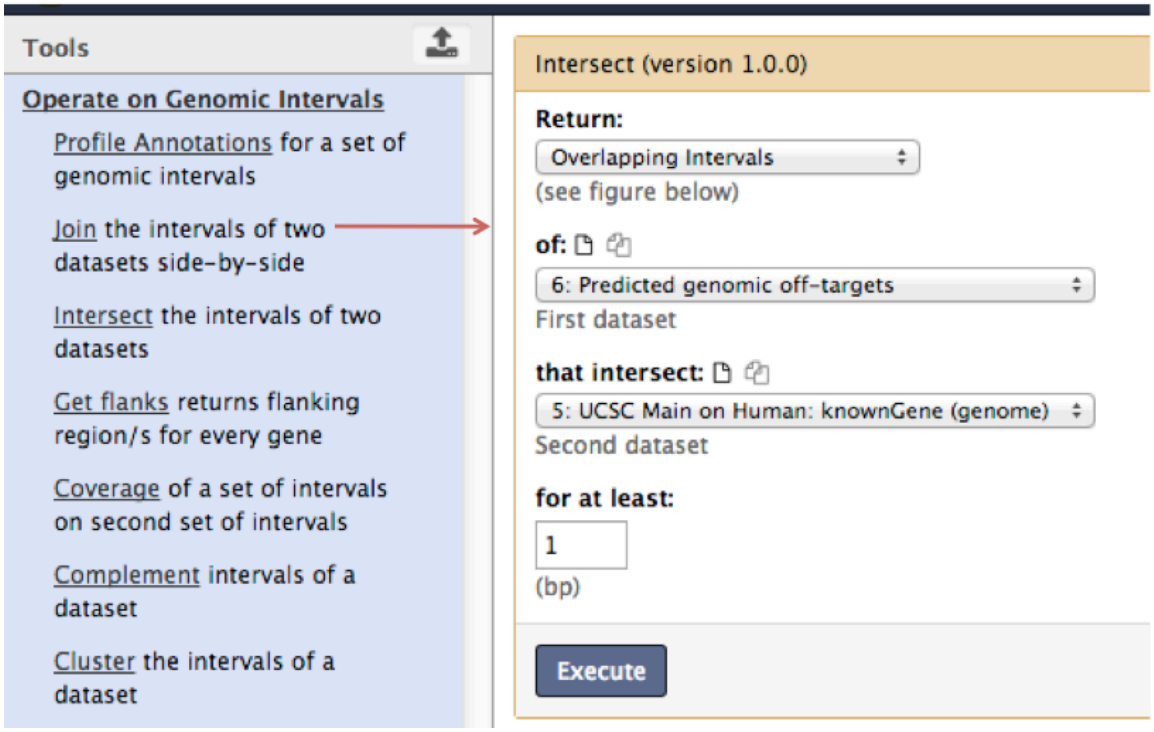
6. (Optional) Filtering off-target sites

Not all genomic off-target sites are created equal. For example, you may only be interested in avoiding off-targets for a set of genomic regions of particular biological relevance. Here, we show how Galaxy can be used for this. If this does not apply to you, proceed to step 7.
The first step is to obtain a BED file that defines the regions we want to consider. You can either upload one yourself, or use the Galaxy interface for downloading data. Here, we show how you can limit your search to regions that are 10 kb upstream of annotated genes.

First, select the "UCSC Main" option from the "Get Data" section. In the example below, we use a human gene track and obtain the regions as described.

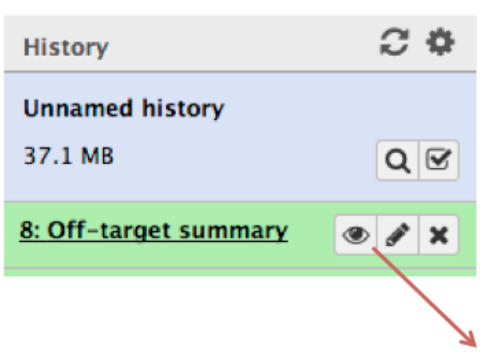


Next, we use the "Intersect" tool in the "Operate on Genomic Intervals" section to determine which of our predicted target sites fall within our regions of interest (10 kb upstream of all genes, in this case). Simply select the file containing the off-targets and the intervals of interest you defined and click "Execute." An example is shown below.



7. Summarizing off-targets

Now, click on "Summarize genomic off-targets" and select either the output of step 5 or 6, depending on which one you performed last. Then, click the "Execute" button. This will create an item in your history labeled "Off-target summary". Here, the candidate proteins are ranked by the total number of observed off-targets and a summary score. Each off-target site adds 1/(relative K_d) to the total score. In other words, the lower the score, the fewer high-affinity off-targets. In this example, the protein "candTALE-44" seems like the best candidate, based on its low off-target score.



Protein	# off-targets	Off-target score
candTALE-44	5	0.78
candTALE-22	12	1.54
candTALE-21	11	1.88
candTALE-78	11	1.99
candTALE-20	14	2.05
candTALE-79	9	2.13
candTALE-48	20	2.95
candTALE-19	22	3.41
candTALE-3	20	3.77
candTALE-15	27	6.57

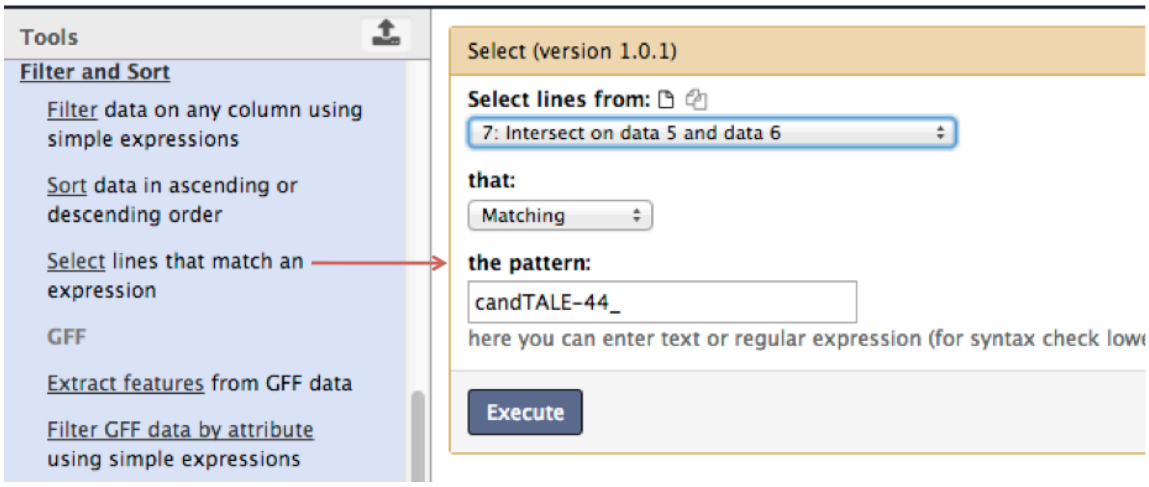
8. Generate RVD sequences for candidate proteins

Click on "Convert DNA sites to RVDs" and find your protein of interest in the output file. This is the RVD sequence you should use for constructing your candidate TALE protein.

candTALE-44	NN-NN-HD-NN-NN-HD-HD-HD-NN-NI-HD-HD-NN-NI

9: (Optional) Determine genomic coordinates of predicted off-targets

You can easily check the genomic locations of the predicted off-targets for your chosen protein to spot any potential problems. To do this, click on the "Select" tool in the "Filter and Sort" section. Then enter the name of your protein followed by an undescore ("_") character as the pattern. Finally, click "Execute."



This will show the positions of all of the predicted off-targets sorted by their affinity, with their predicted K_d in the fifth column.

chr12	6677257	6677272	candTALE-44_Seq43_Off-target	4.14	+
chr2	241078529	241078544	candTALE-44_Seq43_Off-target	4.14	+
chr12	24103378	24103393	candTALE-44_Seq53_Off-target	4.70	+
chr2	237475457	237475472	candTALE-44_Seq136_Off-target	8.30	-
chr6	28645178	28645193	candTALE-44_Seq162_Off-target	9.58	+
chr3	8610513	8610528	candTALE-44_Seq165_Off-target	9.61	-