Sequencing TALENs

Sequencing TALENs (or any TAL array) can be challenging due to the repetitive nature of the sequence. In the Voytas lab, we usually are only concerned with the identity of the RVDs in each repeat as the repeats are maintained in E. coli and we very very seldom identify any mutations within the repeats. After sequencing several hundred TAL arrays, we've only seen two or three problems with the arrays themselves. Most problems occur when the user adds the wrong repeat-containing plasmid, thus confirming the identity of each RVD seems to be the most important aspect.

Sequencing strategy: We use primers (TAL_Seq_5-1 and TAL_R3 or TAL_R2) to sequence into the repeat array from both the 5' and 3' directions. TAL_Seq_5-1 and TAL_R3 are very close to the repeat arrays. Our sequencing reads are usually in the 800-1000 base range (see sources, below), thus allow us to 'see' up to 20 repeats using the expression vector as template and the primers above. If a TAL array is >20 repeats, or if your sequencing reactions don't yield enough data, you may need to sequence the pFUS vectors to determine the identity of the RVDs. We currently don't have a way to sequence the middle of a long, completed array in an expression vector.

Sequencing sources: We primarily use two companies for sequencing. ACGT, Inc has given us the longest sequence reads, but we also use Functional Biosciences due to cost. We do not have any connection except as users to either of these sequencing companies. (Websites: https://www.acgtinc.com/ http://functionalbio.com/)

Sequencing primers: Below are sequences of primers and descriptions of their use.
pCR8_F1: 5'-ttgatgcctggcagttccct – sequence pFUS vector, forward direction
pCR8_R1: 5'-cgaaccgaacaggcttatgt - sequence pFUS vector, reverse direction
TAL_Seq_5-1 5'-catcgcgcaatgcactgac – sequence completed TAL array, forward
TAL_R3: 5'-ggctcagctgggccacaatg - sequence completed TAL array, reverse
TAL_R2: 5'-ggcgacgaggtggtcgttgg - sequence completed TAL array, reverse

Sequence analysis: Most software packages we've tried to use to look at sequencing results are at least a little awkward due to the repetitive nature of the sequences. Most members of the Voytas lab prefer to use the free software ApE (http://biologylabs.utah.edu/jorgensen/wayned/ape/ - which runs on PC and Mac). ApE will identify exact matches to a reference sequence even if it occurs multiple times within a sequence. To make ApE find a specific kind of repeat in your sequence file, open 10 new ApE windows. Paste individually the 10 sequences below (and save as the indicated annotation) into each of the windows. Starting (for example) with NG, highlight the entire sequence within the window, go to Features/New Feature In Library… name the feature "NG", pick a color – color each repeat type a different color – then save the changes. Do this

for each of the 10 sequences below. Open your sequencing result (ApE can read most forms of sequence traces), then File/New DNA From Baseballs. Once this is open, you annotate the file with the kinds of repeats: Features/Annotate Features using Library. It can be helpful to scale the window such that each line is 102 bases long (the length of a complete repeat). To read the AA sequence, ORFs/Translate… and then set the line width to 34 – the length of a repeat. Now you can easily read off the RVD sequence. Additionally, the color-coding in the DNA sequence file will immediately show you if any base within a repeat deviates from one of the expected sequences.

```
>HD1
CTGACCCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCG
GCTGTTGCCGGTGCTGTGCCAGGACCATGGC
>HD
ctgactccggaccaagtggtggctatcgccagccacgatggcggcaagcaagcgctcgaaacggtgcagcg
gctgttgccggtgctgtgccaggaccatggc
>NN
ctgaccccggaccaagtggtggctatcgccagcaacaatggcggcaagcaagcgctcgaaacggtgcagcg
gctgttgccggtgctgtgccaggaccatggc
>NK
ctgaccccggaccaagtggtggctatcgccagcaacaagggcggcaagcaagcgctcgaaacggtgcagcg
gctgttgccggtgctgtgccaggaccatggc
>NG
ctgaccccggaccaagtggtggctatcgccagcaacggtggcggcaagcaagcgctcgaaacggtgcagcg
gctgttgccggtgctgtgccaggaccatggc
>NI
ctgaccccggaccaagtggtggctatcgccagcaacattggcggcaagcaagcgctcgaaacggtgcagcg
gctgttgccggtgctgtgccaggaccatggc
>LR-NN
ctgaccccggaccaagtggtggctatcgccagcaacaatggcggcaagcaagcgctcgaaagcattgtggc
ccagctgagccggcctgatccggcgttggcc
>LR-NI
ctgaccccggaccaagtggtggctatcgccagcaacattggcggcaagcaagcgctcgaaagcattgtggc
ccagctgagccggcctgatccggcgttggcc
>LR-NG
ctgaccccggaccaagtggtggctatcgccagcaacggtggcggcaagcaagcgctcgaaagcattgtggc
ccagctgagccggcctgatccggcgttggcc
>LR-HD
ctgaccccggaccaagtggtggctatcgccagccacgatggcggcaagcaagcgctcgaaagcattgtggc
ccagctgagccggcctgatccggcgttggcc
```

Note that the sequence of HD1 repeats (positions 1, 11, and 21 in an array) differs from that of all other HD repeats by a single base.