Comparing RNA-seq vs DNA-seq methods for identifying T cell repertoires

Michelle Miron

12/2/2017

The goal of this project

Individual T cells can be uniquely identified by the sequence of their T cell receptor gene. Unlike most cells in the body, T cells have rearragned segment of DNA that is esentially a barcode to identify that T cell and all subsequent daughter cells originating from that cell.

There is newly developed software to extract T cell receptor (TCR) sequences from bulk RNA-seq data.

The novelty here, is that usually in order to detect T cell receptor sequences, you design specific primers to sequence the T cell receptor locus within the DNA. However, recently people have designed algorithms to extract TCR sequences from bulk-RNA seq data.

A major limitation in analyzing TCR sequences that were extracted from bulk RNA-seq data, is that you will have very limited sampling. How limited is this sampling? What portion of the reperotire is this capturing compared to is you used DNA based methods to detect TCR sequeces? Equivalent to about about many cells? That is the question I will be addressing here.

I am uniquely positioned to answer this questions, with experimenta data of TCR sequencing from both RNA and DNA based methods.

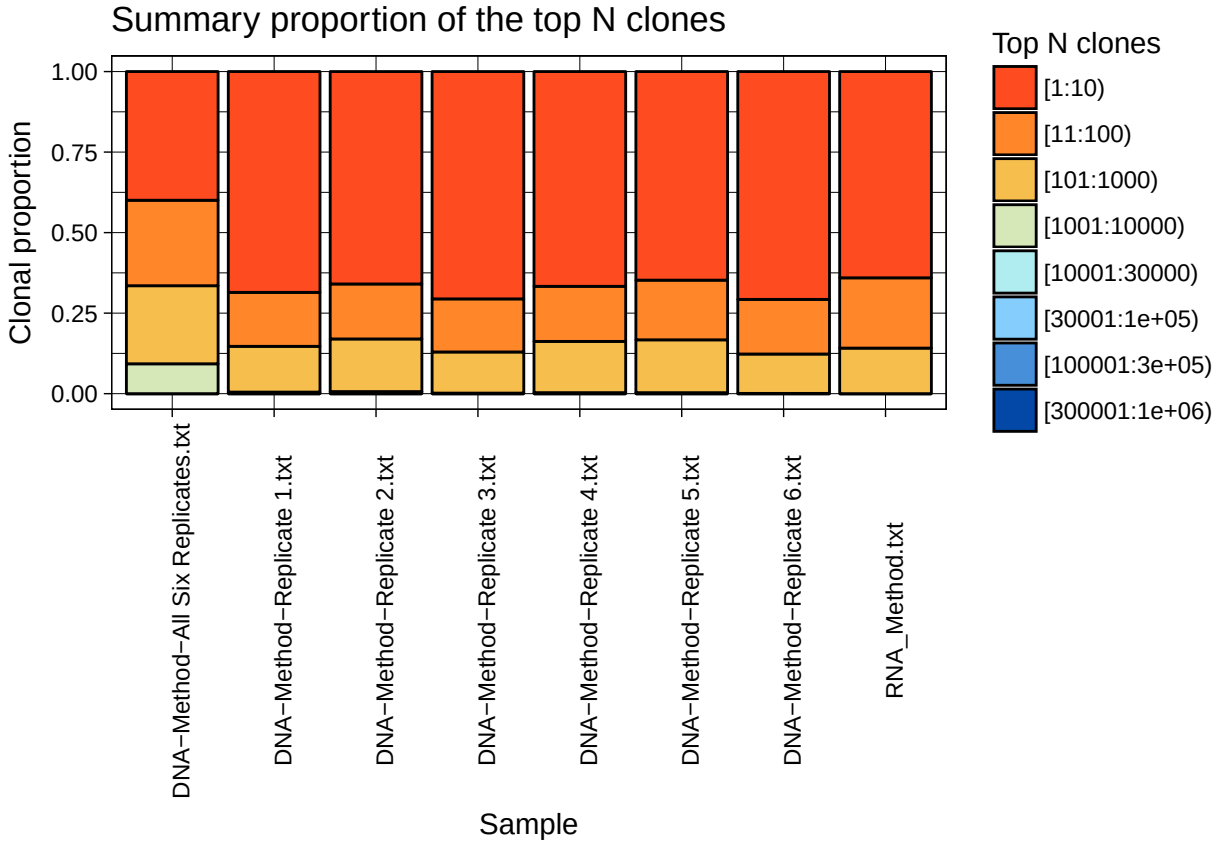
Two methods for identifying T cell receptor repertoires:

1. DNA-based TCR sequencing: Here we sorted T cells and isolated all genomic DNA and subsequently did PCR to amplify the TCR region. For this data we have sequenced 6 replicates. This means that one DNA sample was split into 6 and sequenced independetly.
2. RNA-based TCR sequencing: Here frmo the same T cells we sorted above, we also isolated RNA from these T cells. Subsequent RNA was sent for bulk RNA-seq. FRom total RNA-seq data we will run the alignment again and specifically look for TCR sequences. This is paired-end RNA-seq.

Methods

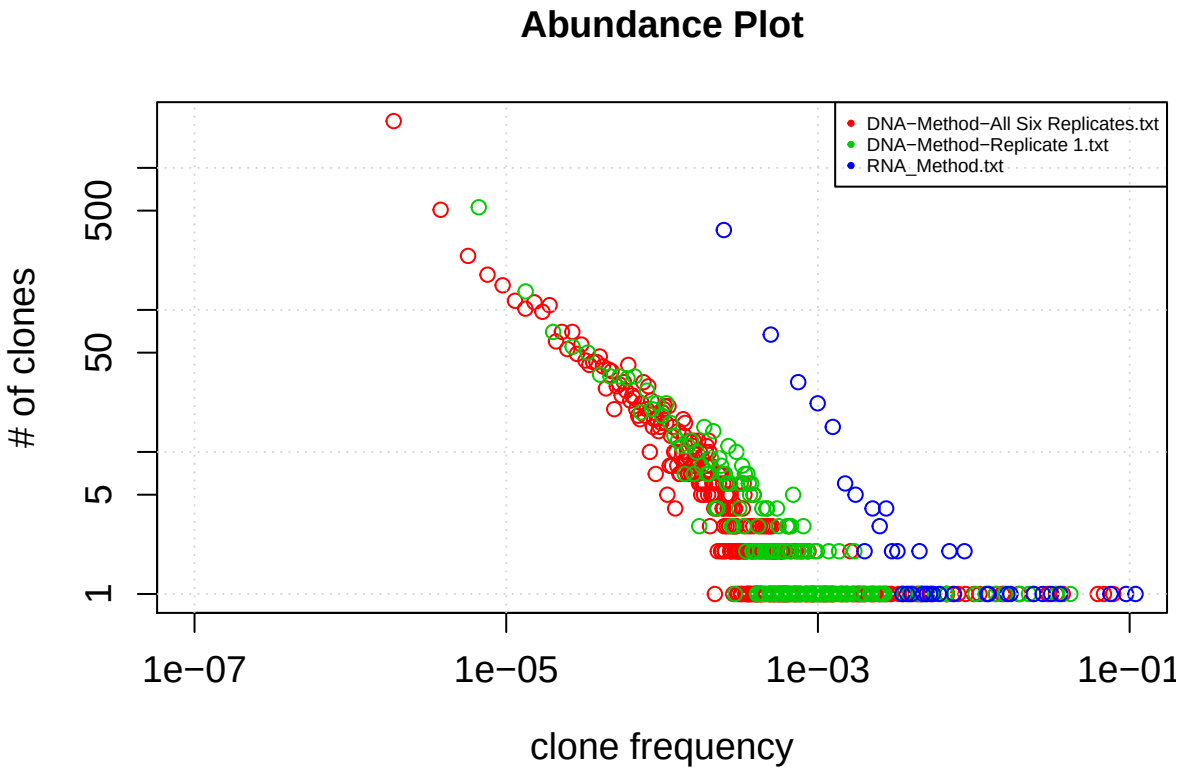
I will be using MiXCR https://milaboratory.com/software/mixcr/ to extract TCR sequences from the RNA-seq data. I will be using tcR pacakge and VDJ tools to analyze the data and plot. I followed the directions outiined in the MiXCR package documentation.

How much of reperotire do top clones make up?



As seen above, it looks like in general, the top 10 clones make up similar portion of the repertoire of the replicates, suggesting some degree of similarity.

Comparing distribution of number of clones detected by their frequency



It looks like in general, the RNA method detected many fewer cloens, however the slope of the abundance plot look somewhat similar.

Future Work

Next we will investigate how likely it is to detect the same sequence in the different samples. From first look here are the top sequences from the different samples:

DNA method: All the replicates combined:

```
head(file_2)
```

```
##                                CDR3.nucleotide.sequence Umi.count
## 1           TGTGCCAGCAGTTTACCCCAGGTACCCACTGAAGCTTTCTTT     70567
## 2           TGCAGTGCCGAAAGGACAGGGGTGACTGAAAAACTGTTTTTT      6248
## 3 TGTGCCAGCAGTCCCCCCCGTGTGCCGCGGCTCCTACAATGAGCAGTTCTTC      5345
## 4              TGTGCCAGTAGTACCGATTTCGACTACGAGCAGTACTTC      5121
## 5           TGTGCCAGCAACATGGGGGGGGGAAATCAGCCCCAGCATTTT      3390
## 6                TGTGCCAGCAGTTCAGCTAACTATGGCTACACCTTC      2973
```

DNA method: One of the six replicates:

```
head(file_3)
```

```
##                          CDR3.nucleotide.sequence Umi.count
## 1       TGTGCCAGCAGTTTACCCCAGGTACCCACTGAAGCTTTCTTT       837
```

```
## 2 TGTGCTGTGGAGTTAATGAATAATAATGCAGGCAACATGCTCACCTTT       438
## 3                      TGTGCCACCTGGGATAAGAAACTCTTT       381
## 4                TGTGCTCTGAACACCGGTAACCAGTTCTATTTT       302
## 5   TGTGCTGCGTGGGATCCCGCCGGCCACTGGTTGGTTCAAGATATTT       144
## 6 TGTGCCAGCAGTACACCGGCGGGACAGGGGATCAATGAGCAGTTCTTC       126
```

RNA method:

```
head(file_3)
```

```
##                          CDR3.nucleotide.sequence Umi.count
## 1       TGTGCCAGCAGTTTACCCCAGGTACCCACTGAAGCTTTCTTT       837
## 2 TGTGCTGTGGAGTTAATGAATAATAATGCAGGCAACATGCTCACCTTT       438
## 3                      TGTGCCACCTGGGATAAGAAACTCTTT       381
## 4                TGTGCTCTGAACACCGGTAACCAGTTCTATTTT       302
## 5   TGTGCTGCGTGGGATCCCGCCGGCCACTGGTTGGTTCAAGATATTT       144
## 6 TGTGCCAGCAGTACACCGGCGGGACAGGGGATCAATGAGCAGTTCTTC       126
```

From above, by quick glance you can see the top clone from the RNA method is the same as the top clone from the combined replicates. This suggests the algorithm is working, since high frequency clones are the ones likely to be detected in another prep of the same sample.

Conclusions and Additional Directions:

This method is easy to use to extract TCR sequences from RNA-seq data. Future directions include:

1. Comparing single-end vs. paired-end RNA-seq and how many TCR clones can be retrieved.
2. Comparing known bioloical differences (ie. CD4 T cells vs CD8 T cell reperotires) to see if TCR data extracted from RNA-seq is robust enough to detect these known differences between repertoires. Specifically, the diversity aspect, since CD8 T cells are less diverse than CD4.
3. Compare diversity (via ecological measures of diversity like shannon entropy and others) of samples between DNA and RNA based methods.