**divBCR.R and divBCR\_isotype.R scripts** calculates different biodiversity metrics, as well as metrics that characterize the different aspects of the BCR repertoire. This script depends on the R package “plyr”.

**Usage:** Rscript divBCR.R inputFile.txt outputFile.txt

**Overview:**

Step 1) BCR counts are normalized to total RNA-seq counts.

Step 2) Calculations of biodiversity and other metrics at the sequence level

Step 3) Sums of clusters

Step 4) Calculations of biodiversity at the cluster level

Step 5) Executing the functions, merging the two data frames and saving the output to a file.

Below are the calculations for each sample (not in column order in final output, column header listed, if by\* then calculated both at the sequence and cluster level):

*Biodiversity metrics:*

1. Shannon entropy (shannon\_entropy\_by\*): -Σ (count proportion\* log(count proportion))
2. Evenness (evenness\_by\*): Shannon entropy / log(species richness)
3. True Diversity (true\_div\_by\*): Exp(Shannon entropy)
4. Gini-Simpson (gini\_simpson\_by\*): 1-(sum(count proportion)^2)

*BCR repertoire characteristics:*

1. Top clone proportion (top\_clone\_prop\_by\*): max(“species” count) / total counts
2. Second top clone proportion (second\_top\_clone\_by\*): second max (“species” count) / total counts
3. Number of unique sequences or clusters (num\_unique\*): species richness, number of unique sequences or number of unique clusters
4. Total counts (total\_counts): sum(count of each sequence)
5. Average CDR3 length (mean\_cdr3\_aa\_len): mean(CDR3 amino acid length)
6. Weighted average CDR3 length (w\_mean\_cdr3\_aa\_len): sum(CDR3 length\* count of each sequence) / sum(count of each sequence)
7. Mean V-region identity (mean\_v\_region\_ident): mean(V-region identity)
8. Weighted mean V-region identity (w\_mean\_v\_region\_ident): sum(V-region identity\* count of each sequence) / sum(count of each sequence)