# T-Cell Receptor (TCR) Analysis Pipeline

## Table of Contents

* [Introduction](#introduction)
* [Features](#features)
* [Dataset](#dataset)
* [Installation](#installation)
* [Usage](#usage)
  + [Script 1: Clone Fraction Analysis](#script-1-clone-fraction-analysis)
  + [Script 2: Diversity Metrics Calculation](#script-2-diversity-metrics-calculation)
* [Outputs](#outputs)
* [Dependencies](#dependencies)
* [Example](#example)
* [Contributing](#contributing)
* [License](#license)
* [Contact](#contact)

## Introduction

Welcome to the **T-Cell Receptor (TCR) Analysis Pipeline**! This repository contains two Python scripts designed to process and analyze TCR sequencing data. The scripts facilitate the extraction, filtering, statistical analysis, and visualization of TCR gene segments, enabling researchers to gain insights into T-cell diversity and clonality.

## Features

* **Data Loading & Merging:** Efficiently load and merge multiple CSV files based on specified keywords.
* **Data Filtering:** Filter TCR data based on clone fraction and alignment score thresholds.
* **Gene Segment Analysis:** Count and analyze TRAV and TRBV gene segments across different timepoints.
* **Statistical Testing:** Perform statistical tests (e.g., Mann-Whitney U) to compare gene segment distributions.
* **Diversity Metrics:** Calculate diversity indices such as Gini Coefficient, D50 Index, and Gini-Simpson Index.
* **Visualization:** Generate box plots and diversity metric plots with customized aesthetics.
* **Excel Reporting:** Save statistical results and metrics into Excel files for further analysis.

## Dataset

The pipeline expects input CSV files containing TCR sequencing data with the following columns:

Copy code

cloneId, cloneCount, cloneFraction, targetSequences, targetQualities, allVHitsWithScore, allDHitsWithScore, allJHitsWithScore, allCHitsWithScore, allVAlignments, allDAlignments, allJAlignments, allCAlignments, nSeqFR1, minQualFR1, nSeqCDR1, minQualCDR1, nSeqFR2, minQualFR2, nSeqCDR2, minQualCDR2, nSeqFR3, minQualFR3, nSeqCDR3, minQualCDR3, nSeqFR4, minQualFR4, aaSeqFR1, aaSeqCDR1, aaSeqFR2, aaSeqCDR2, aaSeqFR3, aaSeqCDR3, aaSeqFR4, refPoints

### Sample Data

csv

Copy code

cloneId,cloneCount,cloneFraction,targetSequences,targetQualities,allVHitsWithScore,allDHitsWithScore,allJHitsWithScore,allCHitsWithScore,allVAlignments,allDAlignments,allJAlignments,allCAlignments,nSeqFR1,minQualFR1,nSeqCDR1,minQualCDR1,nSeqFR2,minQualFR2,nSeqCDR2,minQualCDR2,nSeqFR3,minQualFR3,nSeqCDR3,minQualFR3,nSeqFR4,minQualFR4,aaSeqFR1,aaSeqCDR1,aaSeqFR2,aaSeqCDR2,aaSeqFR3,aaSeqCDR3,aaSeqFR4,refPoints

0,42862,0.0499035390739518,TGTGCAGCAAGTATGAGGCAATGCAGGCAAATCAACCTTT,GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG,TRAV13-1\*00(1272.7),,TRAJ27\*00(263.9),TRAC\*00(153.1),591|605|624|0|14||70.0,,26|48|79|18|40||110.0,,,,,,,,,,,,TGTGCAGCAAGTATGAGGCAATGCAGGCAAATCAACCTTT,38,,,,,,,,CAASMRQ\_AGKSTF,,:::::::::0:1:14:::::18:-6:40:::

1,25054,0.0291699703224018,TGTGCCAGCAGTTACCGAGGACTAGCGGGGGGCGGCGATGAGCAGTTCTTC,GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG,TRBV6-3\*00(1324.9),TRBD2\*00(55),TRBJ2-1\*00(209.1),TRBC2\*00(270.5),516|531|553|0|15||75.0,17|28|48|18|29||55.0,28|42|70|37|51||70.0,,,,,,,,,,,,TGTGCCAGCAGTTACCGAGGACTAGCGGGGGGCGGCGATGAGCAGTTCTTC,38,,,,,,,,CASSYRGLAGGGDEQFF,,:::::::::0:-2:15:18:-1:-4:29:37:-8:51:::

2,16625,0.01935622082741,TGTGCCAGCAGTCCCGGACTCGGAACAGATACGCAGTATTTT,GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG,TRBV27\*00(1207),"TRBD1\*00(25),TRBD2\*00(25)",TRBJ2-3\*00(229.2),TRBC2\*00(272.4),556|568|593|0|12||60.0,9|14|36|12|17||25.0;13|18|48|12|17||25.0,23|41|69|24|42||90.0,,,,,,,,,,,,TGTGCCAGCAGTCCCGGACTCGGAACAGATACGCAGTATTTT,38,,,,,,,,CASSPGLGTDTQYF,,:::::::::0:-5:12:12:3:-10:17:24:-3:42:::

## Installation

1. **Clone the Repository**

bash

Copy code

git clone https://github.com/yourusername/tcr-analysis-pipeline.git

cd tcr-analysis-pipeline

1. **Create a Virtual Environment (Optional but Recommended)**

bash

Copy code

python3 -m venv venv

source venv/bin/activate # On Windows: venv\Scripts\activate

1. **Install Required Dependencies**

bash

Copy code

pip install -r requirements.txt

If requirements.txt is not provided, install the necessary packages manually:

bash

Copy code

pip install pandas numpy matplotlib seaborn scipy statsmodels openpyxl matplotlib-venn

## Usage

This repository contains two Python scripts:

1. **Script 1: Clone Fraction Analysis**
2. **Script 2: Diversity Metrics Calculation**

### Script 1: Clone Fraction Analysis

**Filename:** clone\_fraction\_analysis.py

This script processes TCR data to analyze clone fractions of TRAV and TRBV gene segments, perform statistical tests, and generate box plots with significance annotations.

**Key Features:**

* Filters data based on clone fraction and alignment score thresholds.
* Counts TRAV and TRBV gene segments in two different timepoints (D0 and D6).
* Generates box plots comparing clone fractions between timepoints.
* Performs statistical tests (Mann-Whitney U) and annotates plots with significance levels.
* Saves statistical results into Excel files.

**Usage:**

1. **Ensure Your Data is Organized**

Place all relevant CSV files in the CSVs directory. The script looks for files containing 'D0' and 'D6' in their filenames.

1. **Run the Script**

bash

Copy code

python clone\_fraction\_analysis.py

1. **Inputs:**
   * **Folder Path:** The script expects CSV files to be located in the "CSVs" directory by default. Ensure your CSV files are placed accordingly.
2. **Outputs:**
   * **Plots:** Saved in the Plots directory as PNG files.
   * **Excel Reports:** Detailed statistics saved as Excel files in the Plots directory.

**Customization:**

* Adjust threshold values for clone fraction and alignment score by modifying the variables at the beginning of the script:

python

Copy code

CF\_Alpha\_threshold\_low = 1e-4

AS\_Alpha\_threshold = 0

CF\_Beta\_threshold\_low = 2e-5

AS\_Beta\_threshold = 0

### Script 2: Diversity Metrics Calculation

**Filename:** diversity\_metrics\_calculation.py

This script calculates diversity metrics such as the Gini Coefficient, D50 Index, and Gini-Simpson Index for TRAV and TRBV gene segments, and visualizes these metrics through box plots.

**Key Features:**

* Calculates Gini Coefficient, D50 Index, and Gini-Simpson Index for each gene segment.
* Aggregates metrics across different timepoints (CTL and P3).
* Generates box plots to visualize diversity metrics.
* Saves calculated metrics into an Excel file.

**Usage:**

1. **Ensure Your Data is Organized**

Place all relevant CSV files in the CSVs directory. The script looks for files containing 'D0' and 'D6' in their filenames.

1. **Run the Script**

bash

Copy code

python diversity\_metrics\_calculation.py

1. **Inputs:**
   * **Folder Path:** The script expects CSV files to be located in the "CSVs" directory by default. Ensure your CSV files are placed accordingly.
2. **Outputs:**
   * **Plots:** Saved in the Plots directory as PNG files.
   * **Excel Reports:** Diversity metrics saved as an Excel file in the Plots directory.

**Customization:**

* Adjust threshold values for clone fraction and alignment score by modifying the variables at the beginning of the script:

python

Copy code

CF\_Alpha\_threshold\_low = 0

AS\_Alpha\_threshold = 0

CF\_Beta\_threshold\_low = 0

AS\_Beta\_threshold = 0

## Outputs

Both scripts generate output files saved in the Plots directory:

* **Box Plots:** Visual representations of clone fractions and diversity metrics.
* **Excel Files:** Detailed statistical results and calculated diversity metrics for further analysis.

## Dependencies

Ensure you have the following Python packages installed:

* pandas
* numpy
* matplotlib
* seaborn
* scipy
* statsmodels
* openpyxl
* matplotlib-venn

You can install them using:

bash

Copy code

pip install pandas numpy matplotlib seaborn scipy statsmodels openpyxl matplotlib-venn

## Example

**Running Clone Fraction Analysis:**

bash

Copy code

python clone\_fraction\_analysis.py

**Running Diversity Metrics Calculation:**

bash

Copy code

python diversity\_metrics\_calculation.py

After running the scripts, check the Plots directory for generated plots and Excel reports.

## Contributing

Contributions are welcome! Please fork the repository and submit a pull request for any enhancements or bug fixes.

## Contact

For any questions or suggestions, please open an issue or contact:

[atena.nemati@umontreal.ca](mailto:atena.nemati@umontreal.ca)or

[nemati.atena.ap@gmail.com](mailto:nemati.atena.ap@gmail.com)