Exploring TCR TiRP Scores for Treg Identification and Population analysis

Name: Daan Bossche
Studentnumber: 6181406
Examiner: Can Kesmir
Group: Theoretische biologie en bioinformatica

Abstract

This paper examines the application of the TiRP score, a likelihood score for Treg cells, and its underlying features for predicting T cell phenotypes (Treg or Tconv) and T cell population dynamics. These features were investigated using a random forest classifier, clustering and using the Bray-Curtis statistic. Non-overlapping TCR sequences from a reference dataset were used to train and validate the random forest model, but the predictions were not significantly better than random, indicating that the TiRP score and its underlying features are insufficient to act as a definitive classifier. Afterwards, hierarchical clustering was employed to investigate Tregness patterns in the Emerson dataset based on various features such as age, gender, CMV status, and race, but no clear patterns emerged. Additionally, Bray-Curtis (BC) similarity scores between the Emerson dataset and reference datasets were calculated, showing equally high dissimilarity compared to Treg and Tconv populations, indicating that the BC scores were non-informative on the nature of the donor T cell repertoire. In addition, the BC scores did not exhibit significant changes with age or gender. Overall, the TiRP score proved inadequate for population dynamics and classifier models due to overlapping TCR, low diversity and the inherent noise in the scoring.

Layman summary

In this study, we investigated a scoring system called TiRP, which aims to predict the phenotypes of immune cells known as T cells using a specific protein known as the T cell receptor (TCR). This scoring system tries to determine if a T cell is a regulatory T cell (Treg) or a conventional T cell (Tconv). These are two phenotypes with different roles in the immune system. To accomplish this, we used an algorithm called a random forest classifier, which can make predictions based on TCR data. The random forest is a machine learning method as the machine learns to make decisions based on the data it is provided. Unfortunately, the random forest model did not perform well in accurately determining the phenotype of T cells based on the TiRP scores. We therefore also looked at other human factors such as age, gender, and race to see if they had any relationship with the different types of T cells, but we didn't find any clear patterns or connections. To further investigate T cell populations, we

used a statistical method called Bray-Curtis dissimilarity. This method helps compare different groups of T cells using their TiRP score to see how similar or different they are. Unfortunately, even with this method, we couldn't find any significant differences between a group of donors and the known types of T cells.

In summary, the TiRP scoring system did not prove to be effective in predicting the types of T cells or in analyzing changes in T cell populations. This study shows that it's challenging to predict T cell types based only on the TCR sequence. We tried different approaches like machine learning and statistical analysis, but we didn't obtain satisfactory results. Further studies and alternative methods are needed to improve our understanding of T cell populations and their classification using only their TCR.

Introduction

T cells play an important role in the host immune response (Kumar, Connors, en Farber 2018). They coordinate multiple aspects of the adaptive immune system, such as targeting pathogens, and tumors, and induce the activation of B cells. In addition, they are involved in immune responses throughout the body and can be active over long periods. Many different effector T cells perform these functions. Apart from effector T cells, there is a subpopulation of T cells involved in regulating the immune response generated by effector cells; the T regulatory cells (Treg) (Savage, Klawon, en Miller 2020).

Tregs can influence the activation state of effector T cells and are involved in processes such as preventing autoimmunity and maintaining immune homeostasis. A lack of functioning Tregs can have strong adverse effects such as causing several autoimmune diseases, e.g., Systemic Lupus Erythematosus or Rheumatoid Arthritis (Scheinecker, Göschl, en Bonelli 2020). Under normal circumstances Tregs are identified by their Fox3P+ expression, a regulatory T cell marker. In addition, previous studies have been trying to identify Tregs using sequence motifs found in their T cell receptors (TCR) (Lagattuta et. al 2022).

The TCR is made up of two chains: the alpha and the beta chain, or gamma and delta chain. Within the scope of this thesis, we will only focus on T cells with an alpha-beta TCR. The alpha and beta chains together form a heterodimer that can recognize the peptide-MHC complex during T cell activation. During the thymic development, both chains undergo VDJ gene segment recombination (Michie en Zúñiga-Pflücker 2002). The alpha chain undergoes V and J-segment recombination, and the beta chain undergoes V-D-J segment recombination. This V-D-J recombination is a process by which the T cells semi-randomly join different gene segments to generate the highly diverse TCR repertoire in the body. During this period of development, T cell clones undergo positive and negative selection (Klein e.a. 2014). First, positive selection takes place; in this process the immature T cells are tested for their binding capability to the peptide-MHC complexes presented by thymocytes. Immature T cells should be able to have a weak interaction with MHC-self peptide complexes to be protected from death by neglect. If the immature T cells bind strongly to the MHC-self peptide they undergo negative selection as these T cells will be autoreactive. For immature T cells with Treg fate, the process is thought to be slightly different. After undergoing positive selection the immature T cells that bind strongly to self-peptides, within a window of affinity, undergo Treg

differentiation. This process generates Treg cells that have TCRs with an intermediate to high-affinity interaction with a self-peptide presented via an MHC complex.

As the conventional T cells do not recognize self-peptides presented in the thymus, while Tregs do, one can speculate that TCRs of both cell types could contain distinct sequence patterns and underlying physical properties. Lagattuta et al 2022 attempted to identify features of TCRs expressed by Treg and Tconv cells. They developed a score that uses different features of the TCR beta (TCRB) chain of T cells to discriminate between T conventional cells and T regulatory cells. This scoring named TCR-intrinsic regulatory potential (TiRP), takes into account the CDR3Bmr region in addition to other key amino acids of the TCRB, the V-gene and the J-gene on their likeliness of being a TCR from a Treg or Tconv cell. Lagattuta et al showed that the TiRP score could be used to indicate the likelihood of a TCR being from a Treg cell. We wanted to take this one step further and test if it was possible to train a random forest binary classifier model on the TiRP score. This would enable us to use the TiRP score to identify if a TCR from a T cell data set had a Treg origin or a Tconv origin.

To this end, we trained random forest models on two reference T cell subsets made available by Lagattuta et al 2022 and used different components of the TiRP. In addition, we wanted to use the TiRP score to identify patterns in the composition of T cell populations. If the Tregness of a population increases or decreases this should be identifiable using the TiRP score. Known possible influences on the composition of T cell population are ea. age, sex and CMV status. If the amount of Treg cells should increase/decrease due to any of these features logically the TiRP score should reflect this as well and a pattern should be observable.

Methods

TiRP score calculation

All coding was performed in Rstudio using R version 4.3.1. Lagattuta et al 2022 have made their script for the calculation of the TiRP score available. The script was obtained from 'https://github.com/immunogenomics/TiRP'. The script calculates the TiRP score based on the CDR3 amino acid sequence, the J gene, and the V-gene. The script selects functional chains by excluding internal stop codons and out-of-frame reads. In addition, the chains need to be between 12 and 17 amino acids long to be considered. We used this script to calculate TiRP scores for both the Emerson and the Reference data set. The script generates an output file containing information represented in Table 1.

Table 1. Overview TiRP score characteristics
An example of the output generated by the script provided by Lagattuta et al 2022 .*TiRP score is the sum of the jTiRP, vTiRP and the mTiRP.*

Characteristic			
Sample_name	**vgene**	**CDR3**	**Jmotif**
Keck0001_MC1	TCRBV05-05	CASSLLGQGYNEQFF	NEQFF
CDR3MR	**length**	**length_score**	**pos_score**
LLGQG	15	0	0
perc_score	**vgene_score**	**p107_score**	**p113_score**
0.034	0	0	0.143
Jmotif_score	**total_score**		
-0.023	0.154		
jTiRP	**vTiRP**	**mTiRP**	**TiRP**
0.192	0.617	-0.018	0.791

Furthermore, different components of the TiRP score can be calculated by the utils.R script, also provided by Lagattuta et al 2022. This script takes the same input as the TiRP.R file and generates a detailed scoring file that contains different components of the TiRP score. There was a small mistake in the utils.R file. The outputs of modified utils.R and TRiP.R scripts were used to train random forest models, Bray-Curtis calculation, and clustering/heatmap generation.

Random forest

The random forest model was trained using the ranger package in R. The performance of the models was evaluated by calculating the ROC (Area under the curve) and the Matthews correlation coefficient (MCC). To calculate these values the mltools package was used. Unfortunately, it was not possible to directly use the output of the utils.R script to train a random forest model because this script generated empty rows that contained no data when the CDR3 was shorter than 17 amino acids for some of the features. These features are only assigned a score when the CDR3 is of a certain length, for example, when the CDR3 is 14 amino acids long the features scored for amino acids 15-17 are left with an NA value. As training with non-applicable (ie. NA values) was not possible in the ranger package, these features were removed from the datafiles to train random forest models. Each model consisted of 50 trees.

Bray-Curtis

The TiRP score is continuous ranging from approximately -5 to 7.7. To be able to use this score to calculate Bray Curtis similarity we divided the score range into 1000 bins and counted the occurrences within each bin. The TiRP score distribution was then used to calculate the Bray-Curtis similarity. The formula used for the Bray-Curtis similarity = (2 * S) / (n1 + n2) Where: S is the sum of the absolute differences in species abundances between the two samples, n1 is the total abundance of species in sample 1, n2 is the total abundance of species in sample 2. The BC similarity is a score between 0 and 1, with 1 being identical and 0 being completely dissimilar.

Hierarchical Clustering

Since each donor in the Emerson dataset had a varying number of reads, we calculated the quantiles of the TiRP score for each donor and used them for clustering. The 40% and the 60% quantile were left out for clustering as these had very little deviation in their values between the different donors. This approach accounted for the diversity of TiRP scores by capturing both the lowest and highest values, which could be influenced by factors like age. Normalization was performed on each quantile group to ensure equal contribution to the clustering process, especially important considering the distance-based clustering method employed. Additionally, the standard deviation which was normalized and used in the clustering process. Clusters and a heatmap were generated using the pheatmap package. The following features were taken along in the creation of the heatmap; age range, gender, race and CMV status.

Results

TiRP score analysis and dataset preparation

Two different datasets are used during this analysis. The first dataset contains TCR sequences from healthy donors obtained from peripheral blood mononuclear cells (PBMCs) which have been sorted as either Tconv or Treg based on FoxP3, a Treg marker. In total, this dataset contains 14 Tconv donors and 8 Treg donors. The dataset was published by (Gomez-Tourino et al. 2017)] and will be called the 'Reference dataset'. This dataset was used by Lagattuta et al 2022 to develop the TRiP score. The second data set used contains TCR sequences from 786 healthy donors likewise obtained from PBMCs. This data set contains additional information about the age, gender, ethnic group, racial group, and CMV infection status of the donors. This data set was published by Emerson e.a. 2017 and will be called the 'Emerson dataset' throughout the paper. For both data sets TRiP scores were calculated using TCR sequences, and an overview of the score per CDR length in the Emerson data set can be found in Fig1. A higher TiRP score is indicative of being a Treg and a lower one of Tconv (Lagattuta et. al 2022). A majority of TCRs in the Emerson data set has a TiRP score around 0. Only X% is higher than >3 which can be considered clearly as Treg. There is no clear association to be seen between TCR length and TiRP score. An overview of most characteristics in the Emerson data set can be found in Table 2. We decided to use three different age ranges 0-19, 20- 59, and 60-74 for further analysis as earlier research has shown that thymic activity changes the strongest between these age ranges (Sauce en Appay 2011).

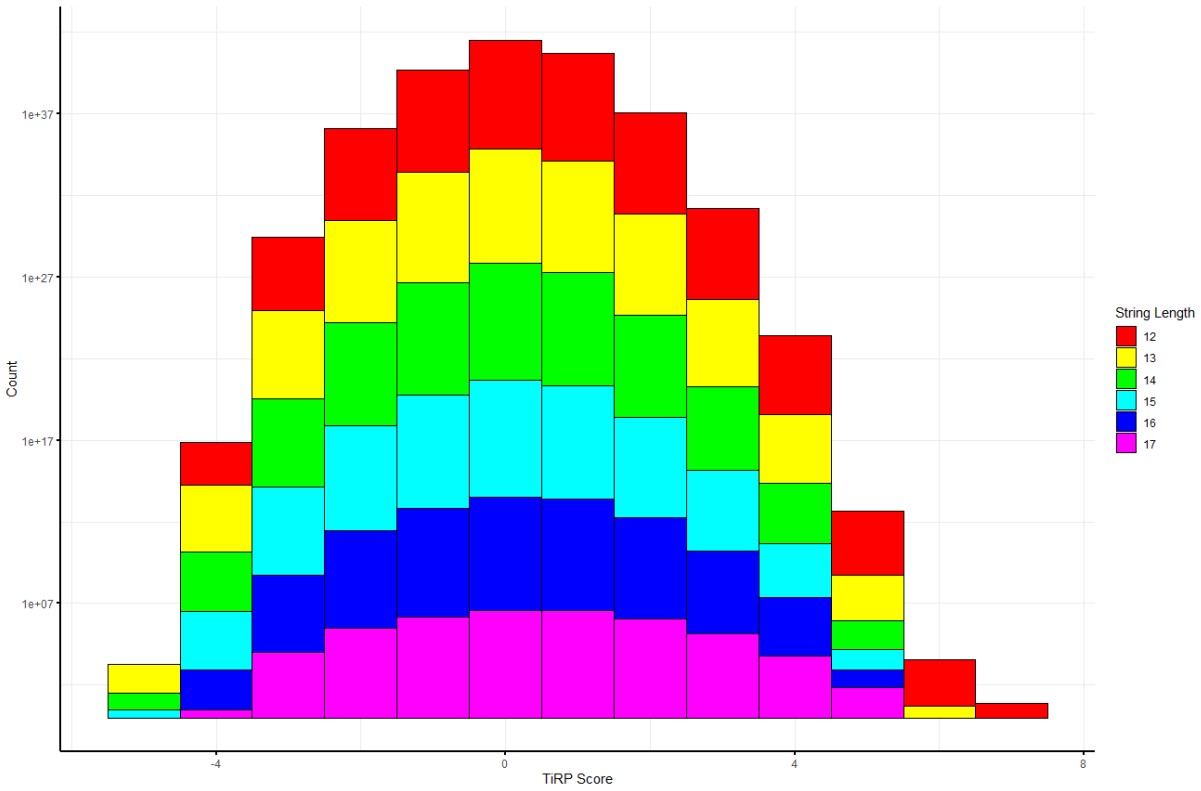

Figure 1: Overview of TiRP score per length of CDR3 of the Emerson dataset.
X-axis represents TiRP score, Y-Axis is total CDR3 count log10 scaled, String Length = CDR3 length.

Table 2. Emerson Characteristics

Characteristic	**Female**, N = 369	**Male**, N = 392
Age_range		
0-19	26 (7.7%)	26 (7.7%)
20-59	288 (86%)	292 (86%)
60-74	22 (6.5%)	20 (5.9%)
Unknown	33	54
Racial_Group		
African Race	5 (1.4%)	6 (1.5%)
Asian or Pacific Islander	26 (7.0%)	20 (5.1%)
Caucasian	227 (62%)	233 (59%)
Native American or Alaska Native	5 (1.4%)	5 (1.3%)
Unknown racial group	106 (29%)	128 (33%)
Virus_Diseases		
Cytomegalovirus -	190 (51%)	230 (59%)
Cytomegalovirus +	179 (49%)	161 (41%)
Unknown	0	1

The total amount of Tconv reads is much greater than the Treg reads (Fig.2). This is to be expected as Tconv cells make up the far majority of T cells in an individual. Moreover, there were more donors for Tconv cells (see above). As reported by Ko et al. almost a third (29.5%) of the Treg CDR3s are shared with Tconv cells (Fig. 2). These shared TCRs are known to have features common with sequences distinct to Tconv but not with sequences distinct to Treg (Owen et al. 2022). This suggests that there are probably 2 distinct Treg populations, one with unique biophysicochemical properties and one which could have been a Tconv but due to the influence of other factors the cell became a Treg.

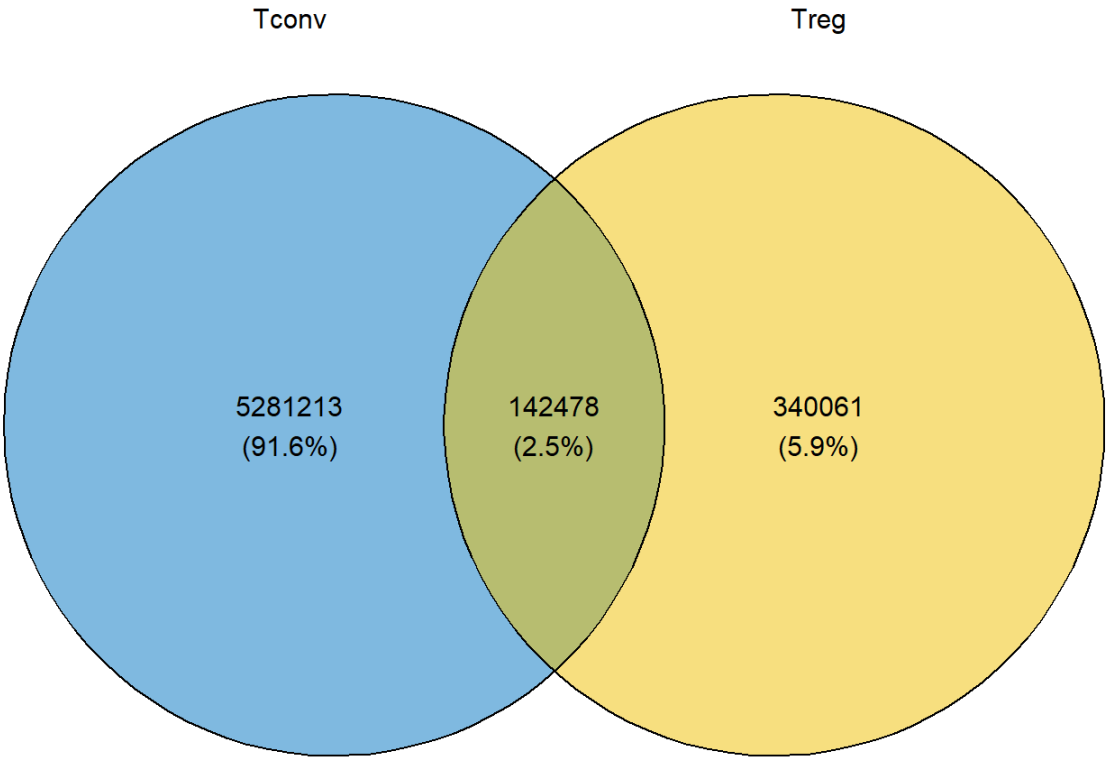

Figure 2: Overview of overlap and richness of Tconv and Treg TCRs within the Reference dataset. *Blue = Tconv TCRs, Yellow = Treg TCRs and Green = overlapping TCRs found in both origins.*

Predicting cell phenotype Using Random Forest Classifier

The TiRP score is a continuous value that estimates the likelihood of a TCR being a Treg cell, and it cannot act as a definitive classifier. Therefore, we made use of a machine learning algorithm, specifically a random forest classifier, to predict the origin of a TCR (Tconv or Treg) using the TiRP score and its underlying components. Treg and Tconv cells sharing the same TCR (Fig. 2) introduces noise in the random forest classifier. To mitigate this, we selected non-overlapping TCR sequences from the reference dataset to train and validate the random forest model. The training set contained 70% of total Treg sequences. Since the original reference dataset contained more Tconv sequences than Treg sequences, a ratio of 1:3 (Treg: Tconv) was chosen to prevent overfitting the model for Tconv sequences. Three different learning approaches were employed: using only the TiRP score (see methods) (Fig. 3), the extended TiRP score, and several data features generated with the utils.R script. Subsequently, the remaining 10% and 20% of the data set were used as a validation and test dataset, respectively, to measure the performance of the trained random forest models.

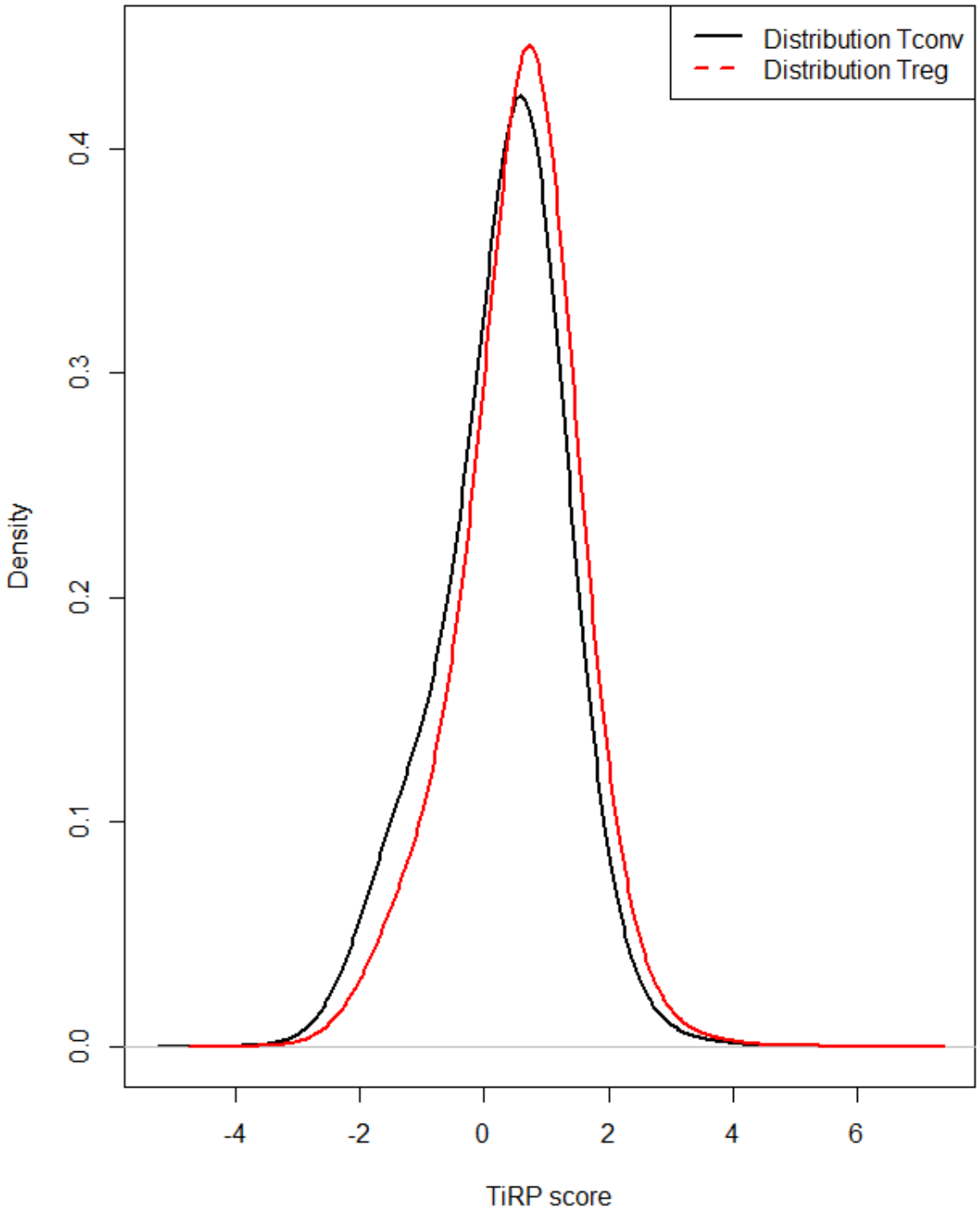

Figure 3: Distribution of Reference dataset TiRP score
Black: TiRP score for T conventional cells, Red: TiRP score for T regulatory cells

We trained three random forest models (as described in methods) using i) TRiP scores, ii) the extended TiRP score, and iii) several features generated with the utils.R script and calculated their performance on the test sets (through the ROC and Matthew correlation coefficients, MCC). Both performance measures indicated that the predictions were hardly better than the random predictions (ROC= 0.54 for all three models and the MCC between 0.0884 – 0.151). Therefore neither the TiRP score alone nor the extended scores were able to generate a model that can predict cell phenotype in a non-random manner. This could be due to the large overlap of TiRP scores between Treg and Tconv cells (Fig 3). Still, TRiP scores and TCR features might be useful to identify population patterns such as changes in Tregness with lifetime and with gender.

Investigating Tregness Patterns and Feature Impact on Emerson Dataset Using Hierarchical Clustering

The data made available by Emersson et al 2017 contains the T cell repertoire from individuals between 0 and 74 age and therefore it allows us to study the change in Treg fraction during a lifetime. Since our attempts to develop a predictor to classify Treg/Tconv cells failed, we can not predict exact fractions of Treg and Tconv cells in the Emerson data set. However, as the TRiP score indicates the tendency of a T cell to be a Treg, we might still see a difference in T cell repertoires of different age groups. To this end, we (re)analyzed TRiP scores of T cell repertoires in Emerson data.

Previously, it was found that the median TRiP score of an individual's repertoire is not influenced by gender, age, or CMV status of a donor in the Emerson dataset (By Dana, results not shown). Obviously, using a median value of the TRiP score to describe the whole repertoire composition of Treg and Tconv cells results in severe information loss. We, therefore, decided to re-investigate repertoire composition using more parameters to describe an individual's T cell repertoire.

Hierarchical clustering is an unsupervised machine-learning technique used to group similar data points based on similarity or dissimilarity. This technique is valuable for identifying patterns in a dataset. In our study, we aimed to determine if the Tregness of the Emerson dataset changes over time as individuals aged. Additionally, we investigated whether other features, such as gender, CMV status, and racial group, had an impact on the population's Tregness. The clustering was performed to identify possible TiRP clusters in the Emerson dataset based on the different features. The clustering using several quantiles, the median and standard deviation to describe TRiP scores of a TCR population did not result in clear patterns (Fig. 4). Only the minimum and maximum values of TRiP scores (Q0 and Q100 values) cluster clearly into two big classes. However, gender, age range, CMV status, and racial group of the individuals in the Emerson data set, as indicated in vertical bars in Fig. 4, do not differentiate between the two classes. The median column corresponds to the previous analysis.

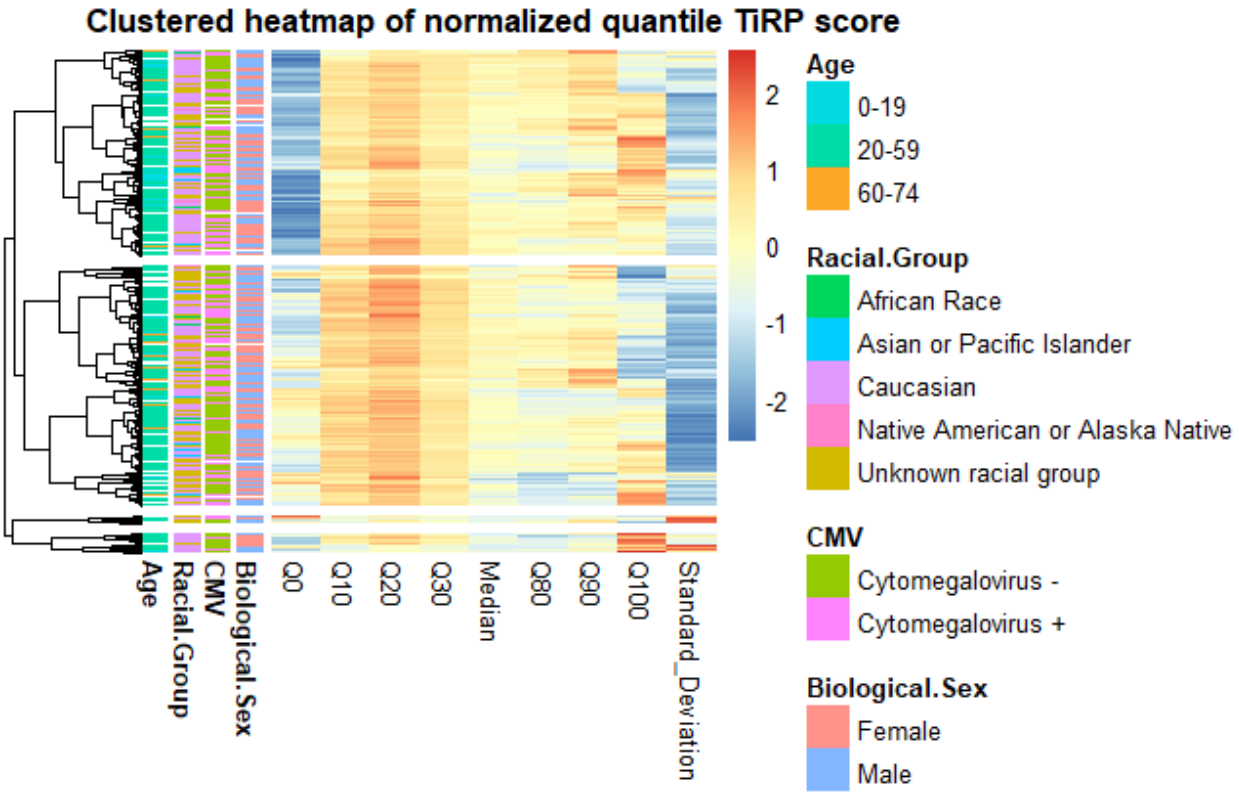

Figure 4: Clustered heatmap of normalized TiRP score quantiles
Heatmap of normalized TiRP score quantiles from the Emerson dataset. Shown are the Q0, Q10, Q20, Q30, Median, Q80, Q90 , Q100 and the standard deviation for each Emerson donor's TiRP scores. Heatmap is row-clustered and four different features are shown for each donor; Age, Racial group, CMV status and Biological sex.

Next, we calculated the Bray-Curtis similarity between the two reference datasets from Lagattuta et al 2022 and the Emerson dataset. The Bray-Curtis dissimilarity is a statistical method that quantifies the similarities between two populations based on the counts of each occurrence type in the respective populations. In our case, the total TiRP scores per donor/origin in the Reference or Emerson dataset can be considered as populations. Therefore, the Bray-Curtis dissimilarity can be employed to quantify whether the TiRP score can indicate distinct populations of different T cells, namely Treg and Tconv. If the TiRP score is indeed more similar between two Tregs (or two Tconvs) than one Treg and one Tconv, the Bray-Curtis (BC) statistic can be used to identify Treg dynamics in a donor population. Using this approach, we describe a repertoire by using the distribution of TRiP scores. That is, we enhance the number of data points per individual from 9 as in Figure 4 to 1000 data points (total number of bins in Bray Curtis statistics, see methods).

In Fig. 5A-B demonstrates that the Emerson data is highly dissimilar for both the Tconv and the Treg Reference data sets. This can be observed by the scores ranging between 0.05-0.22 for the Treg BC scores and 0.05 - 0.30 for the Tconv BC scores. Moreover there is no apparent change in the BC score compared to Tconv with an increase in age for both male and female donors (Fig6). The BC score compared to the Tconv reference set doesn't change over time, indicating that the TiRP score of a repertoire does not change over someone's lifetime. When looking at the Treg male/female boxplot we see a similar pattern, the median BC score remains under 0.3 indicating that the Emerson TiRP populations are highly dissimilar compared to the Treg reference set and that gender has no impact (Fig7).

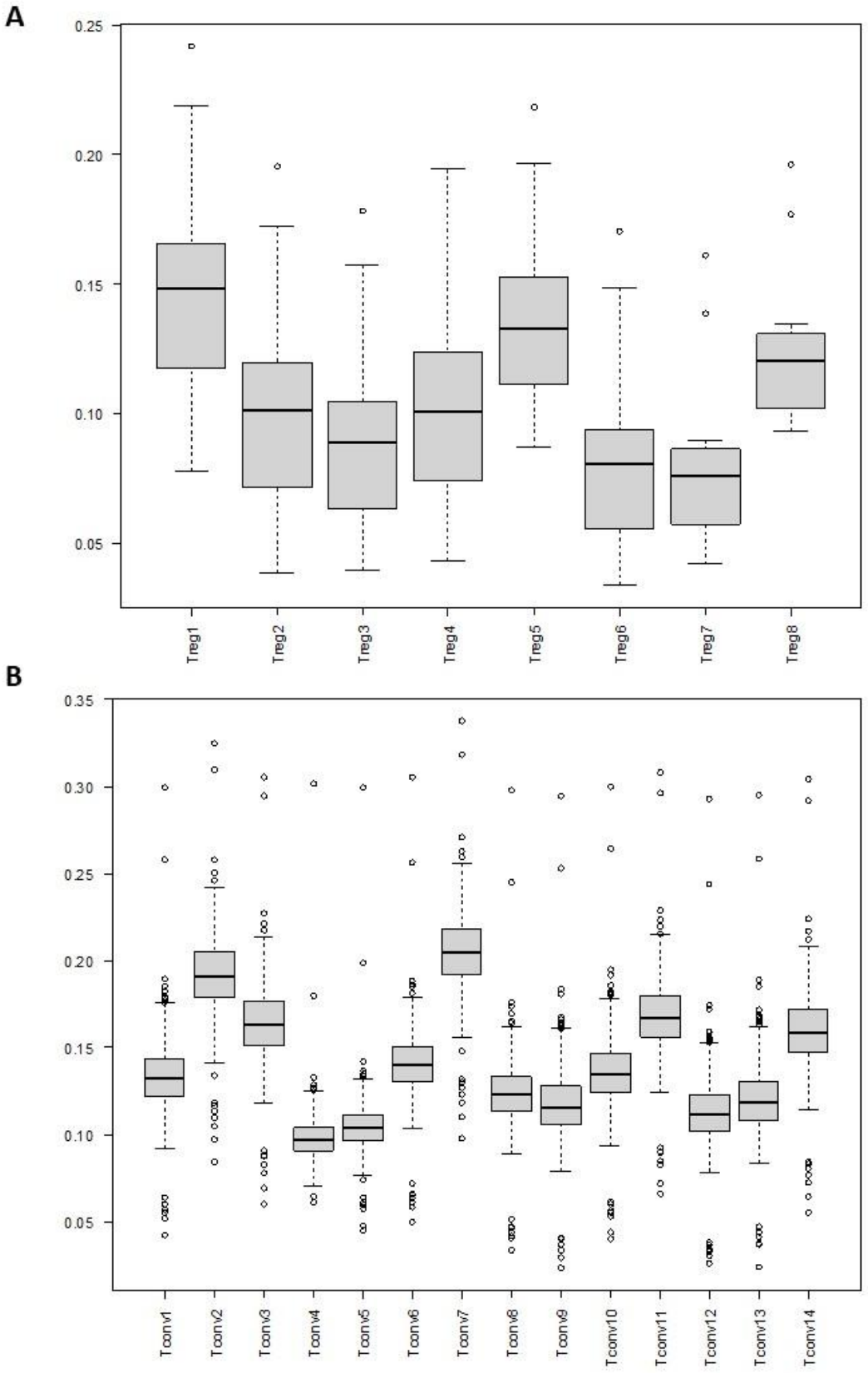

Figure 5: Bray-Curtis scores generated against the Reference dataset.
(A): Bray-Curtis scores generated against the Treg entries of the Reference dataset. (B) Same as A however scores are generated against the Tconv entries of the Reference dataset.

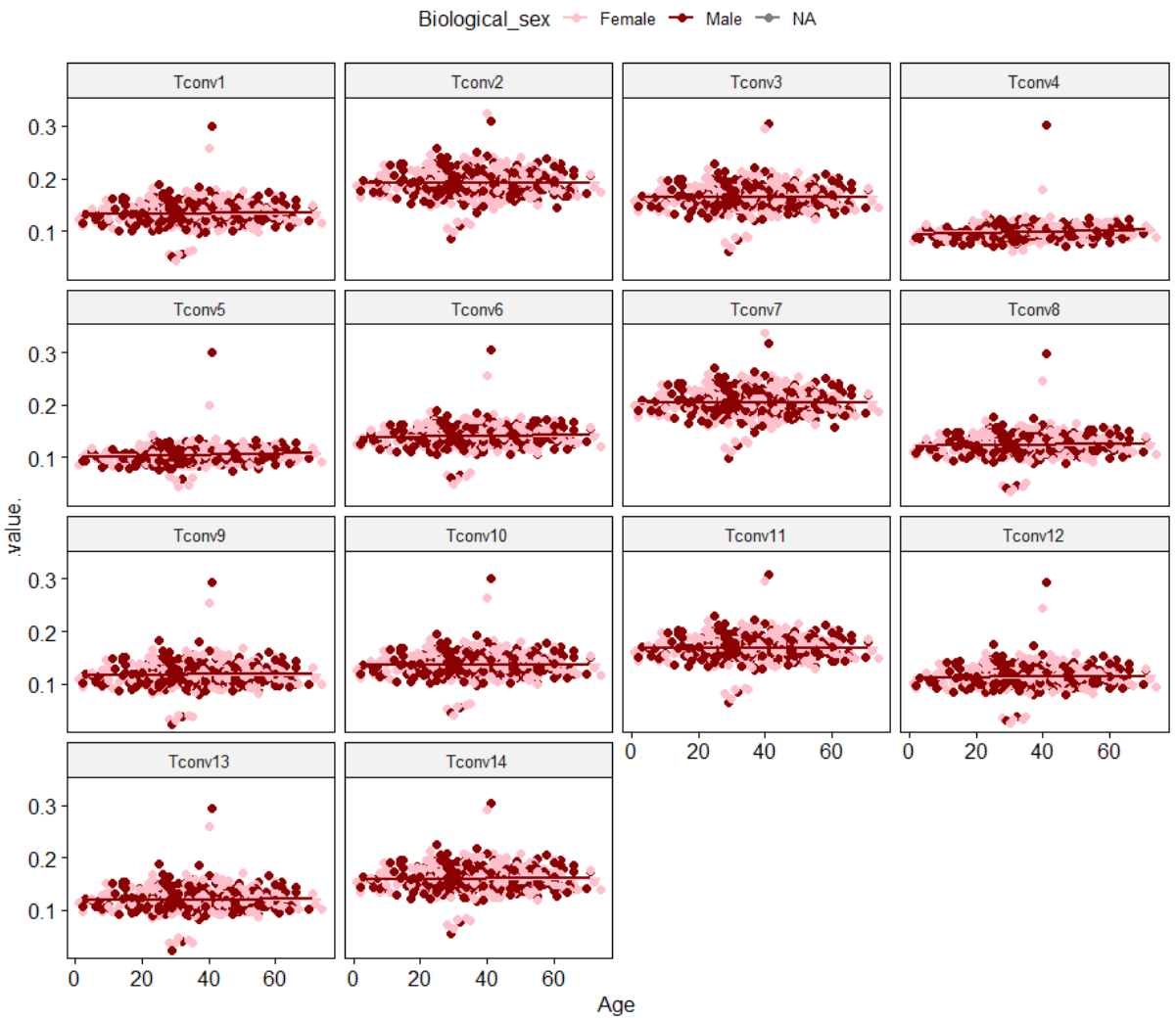

Figure 6: Plotting of Bray-Curtis scores against Tconv data entries separated by age. *Scatterplots generated from Bray-Curtis scores from each Tconv data entry, splitted by age and gender. Trendline per gender is shown.Pink: Female , Red: Male and Grey: Not Applicable (NA).*

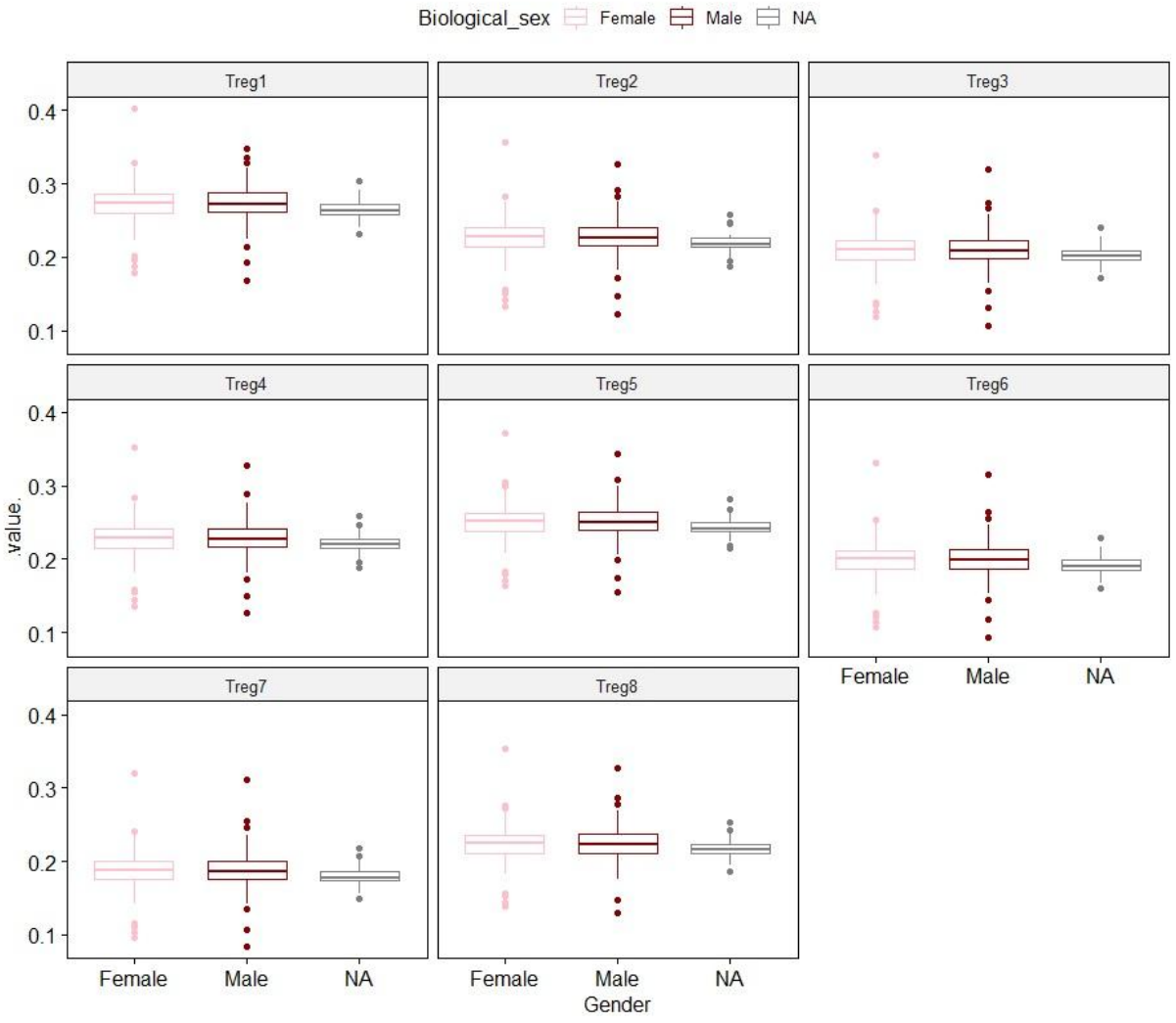

Figure 7: Plotting of Bray-Curtis scores against Treg data entries separated by gender.
Boxplots generated for Bray-Curtis scores against Treg data split by gender. Pink: Female , Red: Male and Grey: Not Applicable (NA). None of the differences are statistically significant.

Discussion

This paper examines the application of the TiRP score and its underlying features. According to Lagattuta et al. (2022), the TiRP score is developed to quantify the propensity of a T cell to become a Treg solely based on its TCR. To validate this, we attempted to train random forest models using the TiRP score and the extended scores/data generated. Firstly, since all randomly generated models were unable to accurately assign the phenotype of T cells based on TCR, we can conclude that the TiRP score and its derivatives are unsuitable to train classifiers for determining the origin (Treg or Tconv) of a TCR. Notably, even with a high TiRP score of approximately 5.91, the Treg: Tconv chance ratio remained around 1:3 (Lagattuta et al., 2022).

The TiRP score is based on the biophysicochemical properties of the CDR3, known to influence Treg fate, such as the presence of hydrophobic or larger amino acids like tryptophan and tyrosine. Enhanced hydrophobicity increases the affinity to self-pMHC, promoting Treg development. Another research group (Ko et al., 2020) utilized this biophysicochemical information, along with V+ J gene information, to train a random forest classifier model for mouse Treg TCRs. Their model accurately predicted the Treg or Tconv fate of a TCR, contrasting the likelihood scoring approach of the TiRP score by Laguttuta et al. Therefore, alternative approaches such as the one employed by Ko et al. or other machine learning methods (Katayama et al., 2022) may be more effective in predicting Treg TCRs.

Considering that the Emerson dataset includes donor-related data, which could impact Treg development, we sought to investigate this through the usage of clustering analysis using TiRP scores. However, the clustering results did not reveal any distinct patterns based on age, gender, racial group, or CMV status within the dataset. While this may indicate that these features do not contribute significantly to the composition of the T cell repertoire, previous research suggests that gender and age do influence Tregness (Robinson et al., 2022; Emerson et al., 2017; Afshan, Afzal, & Qureshi, 2012), making this interpretation less likely. Alternatively, The TiRP score may lack the capacity to provide meaningful information on quantitative population dynamics. This observation is further supported by the failure of the random forest model and the BC scoring of the donors.

Consequently, after the limited success of the random forest model and the clustering approach, we calculated the BC scores of the Emerson dataset in comparison to the reference set to examine population dynamics. As mentioned earlier, the BC score measures population similarity, in this case, the ratio of TiRP distributed across 1000 bins. The results indicated that the Emerson dataset exhibited similar dissimilarity to both Treg and Tconv data. This can be attributed to the dataset containing TCRs from both Treg and Tconv cells. However, since Tconv cells are more prevalent in the T cell repertoire, we expected the Emerson dataset to be more similar to the Tconv set. Furthermore, when analyzing the BC scores in relation to age and gender, no discernible patterns emerged, suggesting that these features do not significantly influence the BC score. In conclusion, our attempt to predict Treg and Tconv population dynamics using BC scores did not yield the expected results.

The scoring system of the TiRP score, based on the biophysicochemical properties of amino acids in the CDR3 and the V/J gene, poses challenges in transforming a likelihood score into a classifier. Exploring potential solutions remains unknown; however, the aforementioned machine learning approach could be considered. Nevertheless, the issue of overlapping Treg-Tconv TCR populations will inevitably impact such an approach (Wolf et al., 2016; Ko et al., 2020). One possibility could be to initially focus on the non-overlapping TCRs from both populations and introduce a third classification option that identifies the TCR as ambiguous, thereby minimizing misclassifications.

In summary, our findings demonstrate that the likelihood score generated by the TiRP script is inadequate for population dynamics or the development of a classifier model, primarily due to the presence of overlapping TCRs and their associated inherent noise.

References

Afshan, Gul, Nadeem Afzal, en Sadia Qureshi. 2012. 'CD4+CD25(Hi) Regulatory T Cells in Healthy Males and Females Mediate Gender Difference in the Prevalence of Autoimmune Diseases'. *Clinical Laboratory* 58 (5-6): 567-71.

Emerson, Ryan O., William S. DeWitt, Marissa Vignali, Jenna Gravley, Joyce K. Hu, Edward J. Osborne, Cindy Desmarais, e.a. 2017. 'Immunosequencing Identifies Signatures of Cytomegalovirus Exposure History and HLA-Mediated Effects on the T Cell Repertoire'. *Nature Genetics* 49 (5): 659-65. https://doi.org/10.1038/ng.3822.

Gomez-Tourino, Iria, Yogesh Kamra, Roman Baptista, Anna Lorenc, en Mark Peakman. 2017. 'T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes'. *Nature Communications* 8 (november): 1792. https://doi.org/10.1038/s41467-017-01925-2.

Katayama, Yotaro, Ryo Yokota, Taishin Akiyama, en Tetsuya J. Kobayashi. 2022. 'Machine Learning Approaches to TCR Repertoire Analysis'. *Frontiers in Immunology* 13: 858057. https://doi.org/10.3389/fimmu.2022.858057.

Klein, Ludger, Bruno Kyewski, Paul M. Allen, en Kristin A. Hogquist. 2014. 'Positive and Negative Selection of the T Cell Repertoire: What Thymocytes See (and Don't See)'. *Nature Reviews Immunology* 14 (6): 377-91. https://doi.org/10.1038/nri3667.

Ko, Annette, Masashi Watanabe, Thomas Nguyen, Alvin Shi, Achouak Achour, Baojun Zhang, Xiaoping Sun, e.a. 2020. 'TCR Repertoires of Thymic Conventional and Regulatory T Cells: Identification and Characterization of Both Unique and Shared TCR Sequences'. *Journal of Immunology (Baltimore, Md.: 1950)* 204 (4): 858-67. https://doi.org/10.4049/jimmunol.1901006.

Kumar, Brahma V., Thomas J. Connors, en Donna L. Farber. 2018. 'Human T Cell Development, Localization, and Function throughout Life'. *Immunity* 48 (2): 202-13. https://doi.org/10.1016/j.immuni.2018.01.007.

Lagattuta, Kaitlyn A., Joyce B. Kang, Aparna Nathan, Kristen E. Pauken, Anna Helena Jonsson, Deepak A. Rao, Arlene H. Sharpe, Kazuyoshi Ishigaki, en Soumya Raychaudhuri. 2022. 'Repertoire Analyses Reveal T Cell Antigen Receptor Sequence Features That Influence T Cell Fate'. *Nature Immunology* 23 (3): 446-57. https://doi.org/10.1038/s41590-022-01129-x.

Michie, Alison M, en Juan Carlos Zúñiga-Pflücker. 2002. 'Regulation of Thymocyte Differentiation: Pre-TCR Signals and β-Selection'. *Seminars in Immunology*, The Development and Function of Antigen Receptors in Lymphopoiesis, 14 (5): 311-23. https://doi.org/10.1016/S1044-5323(02)00064-7.\

Regulatory T Cell Development in the Thymus Using Single-Cell RNA Sequencing/TCR Sequencing'. *Journal of Immunology (Baltimore, Md.: 1950)* 209 (7): 1300-1313. https://doi.org/10.4049/jimmunol.2200089.

Robinson, George A., Junjie Peng, Hannah Peckham, Gary Butler, Ines Pineda-Torra, Coziana Ciurtin, en Elizabeth C. Jury. 2022. 'Investigating Sex Differences in T Regulatory Cells from Cisgender and Transgender Healthy Individuals and Patients with Autoimmune Inflammatory Disease: A Cross-Sectional Study'. *The Lancet. Rheumatology* 4 (10): e710-24. https://doi.org/10.1016/S2665-9913(22)00198-9.

Sauce, Delphine, en Victor Appay. 2011. 'Altered Thymic Activity in Early Life: How Does It Affect the Immune System in Young Adults?' *Current Opinion in Immunology*, Host pathogens/Immune senescence, 23 (4): 543-48. https://doi.org/10.1016/j.coi.2011.05.001.

Savage, Peter A., David E.J. Klawon, en Christine H. Miller. 2020. 'Regulatory T Cell Development'. *Annual Review of Immunology* 38 (1): 421-53. https://doi.org/10.1146/annurev-immunol-100219-020937.

Scheinecker, Clemens, Lisa Göschl, en Michael Bonelli. 2020. 'Treg Cells in Health and Autoimmune Diseases: New Insights from Single Cell Analysis'. *Journal of Autoimmunity* 110 (juni): 102376. https://doi.org/10.1016/j.jaut.2019.102376.

Wolf, Kyle J., Ryan O. Emerson, Jeanette Pingel, R. Mark Buller, en Richard J. DiPaolo. 2016. 'Conventional and Regulatory CD4+ T Cells That Share Identical TCRs Are Derived from Common Clones'. *PLoS ONE* 11 (4): e0153705. https://doi.org/10.1371/journal.pone.0153705.