SOFTWARE TOOL ARTICLE

REVISED iCAT: diagnostic assessment tool of immunological history using high-throughput T-cell receptor sequencing [version 2; peer review: 2 approved]

Ahmad Rajeh[1], Kyle Wolf[2], Courtney Schiebout[1], Nabeel Sait[3], Tim Kosfeld[3], Richard J. DiPaolo[2], Tae-Hyuk Ahn[1,3]

[1]Program in Bioinformatics and Computational Biology, Saint Louis University, St. Louis, MO, 63103, USA
[2]Molecular Microbiology and Immunology, Saint Louis University School of Medicine, St. Louis, MO, 63104, USA
[3]Computer Science, Saint Louis University, St. Louis, MO, 63103, USA

v2 **First published:** 03 Feb 2021, **10**:65
https://doi.org/10.12688/f1000research.27214.1
Latest published: 29 Jun 2021, **10**:65
https://doi.org/10.12688/f1000research.27214.2

Abstract

The pathogen exposure history of an individual is recorded in their T-cell repertoire and can be accessed through the study of T-cell receptors (TCRs) if the tools to identify them were available. For each T-cell, the TCR loci undergoes genetic rearrangement that creates a unique DNA sequence. In theory these unique sequences can be used as biomarkers for tracking T-cell responses and cataloging immunological history. We developed the immune Cell Analysis Tool (iCAT), an R software package that analyzes TCR sequencing data from exposed (positive) and unexposed (negative) samples to identify TCR sequences statistically associated with positive samples. The presence and absence of associated sequences in samples trains a classifier to diagnose pathogen-specific exposure. We demonstrate the high accuracy of iCAT by testing on three TCR sequencing datasets. First, iCAT successfully diagnosed smallpox vaccinated versus naïve samples in an independent cohort of mice with 95% accuracy. Second, iCAT displayed 100% accuracy classifying naïve and monkeypox vaccinated mice. Finally, we demonstrate the use of iCAT on human samples before and after exposure to SARS-CoV-2, the virus behind the COVID-19 global pandemic. We were able to correctly classify the exposed samples with perfect accuracy. These experimental results show that iCAT capitalizes on the power of TCR sequencing to simplify infection diagnostics. iCAT provides the option of a graphical, user-friendly interface on top of usual R interface allowing it to reach a wider audience.

Keywords

T-cell receptor sequencing, diagnostic classification, R-package, biomarkers

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 29 Jun 2021	✓ report	✓ report
version 1 03 Feb 2021	? report	? report

1. **Shuo-Wang Qiao**, University of Oslo, Oslo, Norway
   Ying Yao, University of Oslo, Oslo, Norway
2. **Se-Ran Jun**, University of Arkansas for Medical Sciences, Little Rock, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Richard J. DiPaolo (richard.dipaolo@health.slu.edu), Tae-Hyuk Ahn (taehyuk.ahn@slu.edu)

Author roles: Rajeh A: Methodology, Software, Visualization, Writing – Original Draft Preparation; **Wolf K**: Data Curation, Formal Analysis, Methodology, Software, Writing – Original Draft Preparation; **Schiebout C**: Formal Analysis, Methodology, Software; **Sait N**: Software, Validation, Visualization; **Kosfeld T**: Software, Validation, Visualization; **DiPaolo RJ**: Conceptualization, Funding Acquisition, Investigation, Methodology, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Ahn TH**: Conceptualization, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Funding by a research contract from the Federal Bureau of Investigations (FBI) to R.D.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Rajeh A, Wolf K, Schiebout C *et al.* **iCAT: diagnostic assessment tool of immunological history using high-throughput T-cell receptor sequencing [version 2; peer review: 2 approved]** F1000Research 2021, **10**:65 https://doi.org/10.12688/f1000research.27214.2

First published: 03 Feb 2021, **10**:65 https://doi.org/10.12688/f1000research.27214.1

REVISED Amendments from Version 1

In the new version, revisions in various sections have been made following the reviewers' recommendations.

- In the Introduction section, we updated that the rarity of specific T cells to a specific antigen is also a challenge in addition to the diversity and magnitude of the TCR repertoire as the reviewer commented.
- In the Methods section, we considered an uncertainty metric in the Prediction tab based on the difference between the probability density in the pre-exposure distribution and the probability density in the post-exposure distribution. This new feature has been implemented and now available on GitHub iCAT.
- In the Use Cases, we included computing time for use cases.
- We updated the iCAT program to allow inputting either a space or a colon to delimit range values as the reviewer commented.
- The manuscript has been updated to clarify some ambiguous words and sentences per the reviewers' comments.
- Figure 1 and Figure 4 were updated.
- Several minor typos were corrected.
- We upgraded the iCAT program to avoid any possible installing errors for all possible operating systems including Linux, Mac, and Windows.

Any further responses from the reviewers can be found at the end of the article

Introduction

T- and B-cell responses are responsible for long-lasting immune memory responses to infectious agents, such as bacteria and viruses. Expansion of pathogen specific T-cells provide us with a robust resource for understanding whether an individual has been infected with a pathogen. The T-cell receptor (TCR), located on the surface of T-cells, is responsible for recognizing pathogen-specific peptides, leading to immune response and development of protective immune memory.[1] During T-cell development, the loci that encode TCRs α and β-chains are rearranged by recombination of the variable (TCRV), diversity (TCRD), and joining (TCRJ) gene segments, encoding the complementary determining region 3 (CDR3).[2] These genetic rearrangement events result in a high degree of diversity in the CDR3 regions of individual TCR loci.[3]

During an infection or vaccination, T-cells that carry receptors recognizing pathogen associated peptides become activated and undergo rapid clonal expansion. The clonally expanded T-cells carry the same unique TCR rearrangement and a portion remain in circulation long after the pathogen has been cleared to provide long-lived immunological memory. The persistence of memory T-cells in circulation make the genetically rearranged TCR loci a stable biomarker documenting an individual's immunological history. To utilize the diverse TCR repertoire as a potential biomarker for specific pathogen exposure, pathogen-specific TCR sequences common to different individuals exposed to the same pathogen need to be identified. This poses significant challenges given the diversity and magnitude of the TCR repertoire. On average, $\sim 10^7$ unique TCRβ chains can be identified from the $\sim 10^{12}$ circulating T-cells present in a healthy human adult.[4] A healthy human adult can have 10^{18} mathematically possible TCR recombinations resulting from the genetic rearrangement.[4,5] The potential diversity of the repertoire coupled with the limited number of T-cells present in individuals makes identifying identical TCR sequences among multiple individuals exceptionally challenging. In addition, a specific TCR response to particular antigen can be extremely rare, which can pose an even greater challenge to identifying signals of T cell memory.[6] However, by analyzing the large and diverse TCR repertoire using high-throughput TCRβ sequencing, it is possible to identify pathogen specific TCRs shared among different individuals exposed to the same infectious agent.[7]

Recently, high-throughput next-generation sequencing (NGS) techniques were employed to analyze the diverse immune cell repertoire.[8-10] Additionally, recent publications described an analytical approach for computationally identifying common/public TCR sequences.[5,10,11] However, analyses of the TCR repertoire for diagnostic purposes have remained largely resource- and time-intensive efforts.

Utilizing the diagnostic methodologies described in Wolf, et al. (2018),[5] we have developed an R package with a user-friendly interface, the immune Cell Analysis Tool (`iCAT`), to identify TCR sequences statistically associated with pathogen exposure and to distinguish infected from non-infected samples. By providing an interactive interface through `iCAT`, sample exposure can be assessed and predicted conveniently without requiring command-line skills. `iCAT` has an ability to classify target-associated receptor sequences (TARSs) and diagnose exposure with a high accuracy.

Methods

Implementation

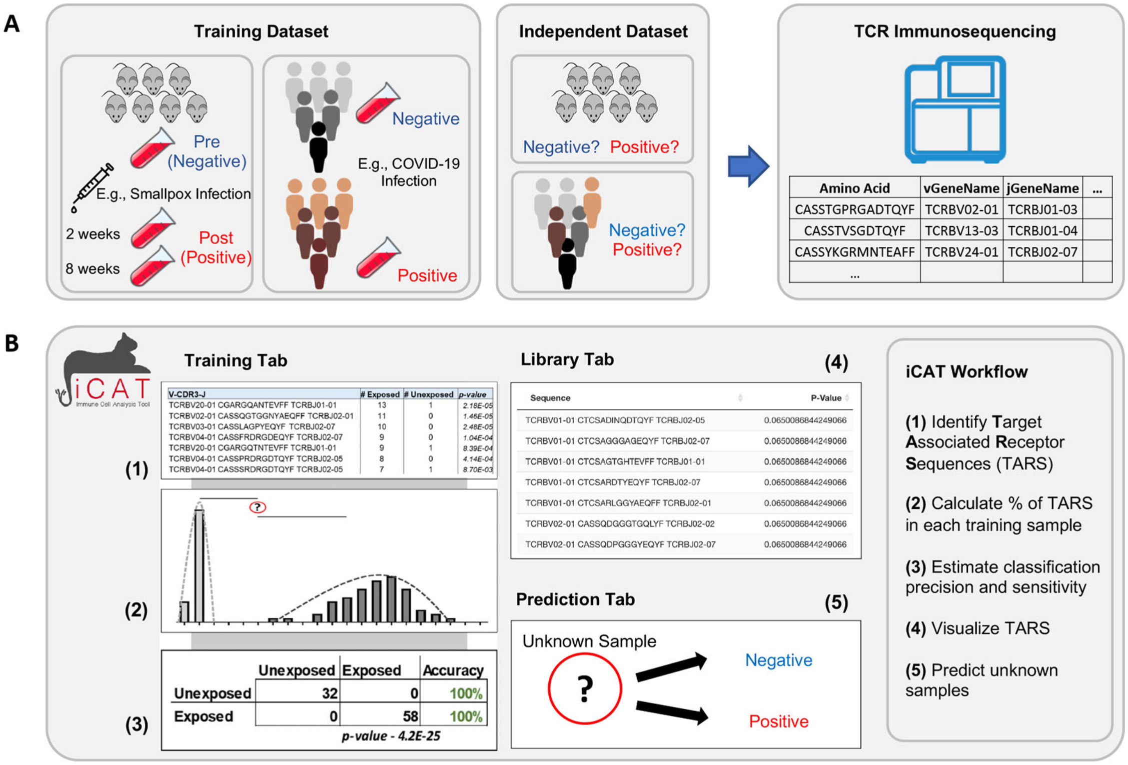
We developed `iCAT`, an R package utilizing high-throughput TCR sequencing data to analyze TCR sequences and to diagnose infection in a user-friendly format (Figure 1). `iCAT` provides both a graphical user interface (GUI) in the form of a web-application utilizing R-Shiny[12] and a command-line R interface for batch processing of large-scale data. The simplest method to install `iCAT` on a system is directly from GitHub using `devtools` and `install_github`:

```
install.packages("devtools")
devtools::install_github("BioHPC/iCAT")
```

In addition to `shiny`, `iCAT` also uses `shinyjs, data.table, ggplot2, DT, hash, and magrittr`. Required packages will be installed through the `install_github` step. Alternatively, users can clone or download the repository from GitHub and run `devtools::install("iCAT/")`.

A user can upload multiple TCR sequence repertoires from negative (control) and positive (experimental) cohorts. `iCAT` accepts tab-delimited files with the size limit of 10 gigabytes per file with multiple options to define TCR clonotypes within samples. The `iCAT` shiny app has three tabs, separating major functionalities: Training, Library, and Prediction. Under the "Training" tab, clicking 'Train Model' will start the pipeline to statistically identify a subset of TARSs that will act as feature selections for training the diagnostic classifier, diagnosing samples as either negative or positive. Upon training, `iCAT`'s main tab provides a table summary of the data, a figure shows the distribution of TARSs between the positive and negative samples, and a classification matrix predicting the exposure status of samples used in the training data (Figure 2). All figures and tables can be downloaded to the user's machine. A progress bar will show on the bottom-right corner to update on the status of training.

A separate tab, "Library", is unlocked upon training and shows a table where each row describes a TARS and its presence in the positive and negative samples. All tables and figures are supplemented with a custom button for easy download (Figure 3).



Amino Acid	vGeneName	jGeneName	...
CASSTGPRGADTQYF	TCRBV02-01	TCRBJ01-03	
CASSTVSGDTQYF	TCRBV13-03	TCRBJ01-04	
CASSYKGRMNTEAFF	TCRBV24-01	TCRBJ02-07	
...			

V-CDR3-J	# Exposed	# Unexposed	*p-value*
TCRBV20-01 CGARGQANTEVFF TCRBJ01-01	13	1	*2.18E-05*
TCRBV02-01 CASSQGTGGNYAEQFF TCRBJ02-01	11	0	*1.46E-05*
TCRBV03-01 CASSLAGPYEQYF TCRBJ02-07	10	0	*2.48E-05*
TCRBV04-01 CASSFRDRGDEQYF TCRBJ02-07	9	0	*1.04E-04*
TCRBV20-01 CGARGQTNTEVFF TCRBJ01-01	9	1	*8.39E-04*
TCRBV04-01 CASSPRDRGDTQYF TCRBJ02-05	8	0	*4.14E-04*
TCRBV04-01 CASSSRDRGDTQYF TCRBJ02-05	7	1	*8.70E-03*

	Unexposed	Exposed	Accuracy
Unexposed	32	0	100%
Exposed	0	58	100%

p-value - 4.2E-25

Sequence	P-Value
TCRBV01-01 CTCSADINQDTQYF TCRBJ02-05	0.0650086844249066
TCRBV01-01 CTCSAGGGAGEQYF TCRBJ02-07	0.0650086844249066
TCRBV01-01 CTCSAGTGHTEVFF TCRBJ01-01	0.0650086844249066
TCRBV01-01 CTCSARDTYEQYF TCRBJ02-07	0.0650086844249066
TCRBV01-01 CTCSARLGGYAEQFF TCRBJ02-01	0.0650086844249066
TCRBV02-01 CASSQDGGGTGQLYF TCRBJ02-02	0.0650086844249066
TCRBV02-01 CASSQDPGGGYEQYF TCRBJ02-07	0.0650086844249066

iCAT Workflow

(1) Identify **T**arget **A**ssociated **R**eceptor **S**equences (TARS)

(2) Calculate % of TARS in each training sample

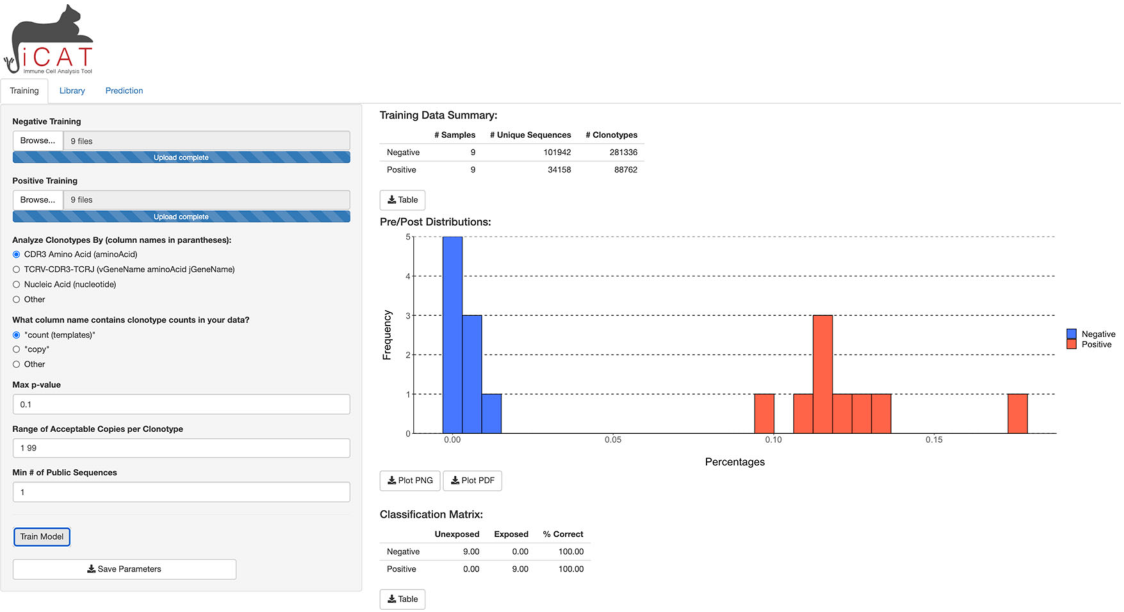
(3) Estimate classification precision and sensitivity

(4) Visualize TARS

(5) Predict unknown samples

Figure 1. Workflow for TCR repertoire sequencing and diagnostic assessment of prior antigen exposure using iCAT. A) Flow chart depicting the purification of DNA from blood samples and the production of TCR repertoires after TCR-specific amplification and sequencing. B) Visual representation of the iCAT methodology.

The third tab of iCAT, “Prediction”, also unlocks after training and allows the user to upload one or more independent TCR-sequencing samples for classification (Figure 4).



Negative Training

9 files

Positive Training

9 files

Analyze Clonotypes By (column names in parantheses):

● CDR3 Amino Acid (aminoAcid)
○ TCRV-CDR3-TCRJ (vGeneName aminoAcid jGeneName)
○ Nucleic Acid (nucleotide)
○ Other

What column name contains clonotype counts in your data?

● "count (templates)"
○ "copy"
○ Other

Max p-value: 0.1

Range of Acceptable Copies per Clonotype: 1 99

Min # of Public Sequences: 1

Training Data Summary:

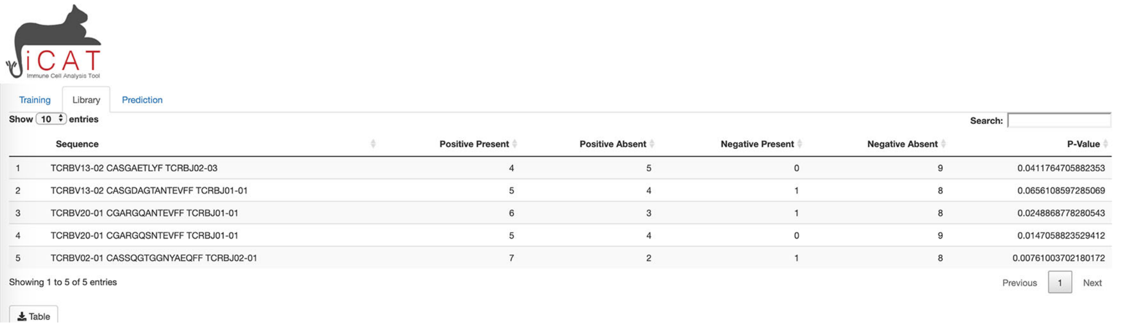
	# Samples	# Unique Sequences	# Clonotypes
Negative	9	101942	281336
Positive	9	34158	88762

Pre/Post Distributions:

Classification Matrix:

	Unexposed	Exposed	% Correct
Negative	9.00	0.00	100.00
Positive	0.00	9.00	100.00

Figure 2. iCAT Training tab. After samples are uploaded, clicking “Training” will start training to select features for the diagnostic classifier from the negative and positive samples.

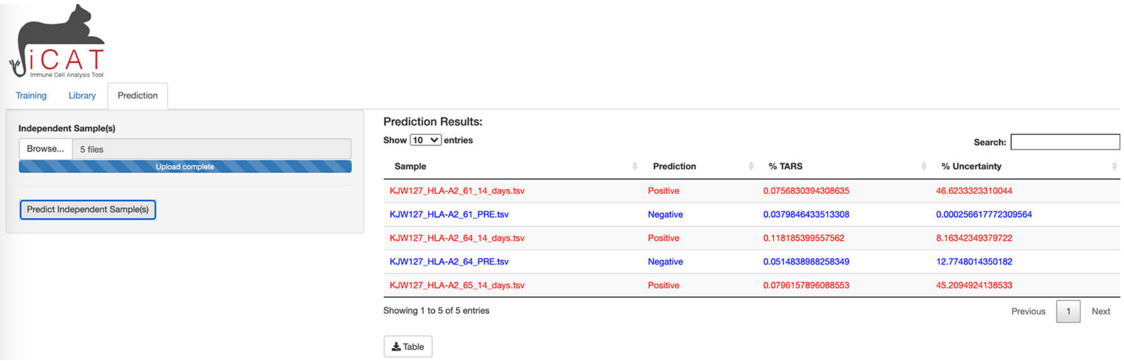


Show 10 entries

	Sequence	Positive Present	Positive Absent	Negative Present	Negative Absent	P-Value
1	TCRBV13-02 CASGAETLYF TCRBJ02-03	4	5	0	9	0.0411764705882353
2	TCRBV13-02 CASGDAGTANTEVFF TCRBJ01-01	5	4	1	8	0.0656108597285069
3	TCRBV20-01 CGARGQANTEVFF TCRBJ01-01	6	3	1	8	0.0248868778280543
4	TCRBV20-01 CGARGQSNTEVFF TCRBJ01-01	5	4	0	9	0.0147058823529412
5	TCRBV02-01 CASSQGTGGNYAEQFF TCRBJ02-01	7	2	1	8	0.00761003702180172

Showing 1 to 5 of 5 entries

Figure 3. iCAT Library tab. The library tab shows a table of target-associated receptor sequences (TARS).



Independent Sample(s)

5 files

Prediction Results:

Show 10 entries

Sample	Prediction	% TARS	% Uncertainty
KJW127_HLA-A2_61_14_days.tsv	Positive	0.0756830394308635	46.6233323310044
KJW127_HLA-A2_61_PRE.tsv	Negative	0.0379846433513308	0.000256617772309564
KJW127_HLA-A2_64_14_days.tsv	Positive	0.118185399557562	8.16342349379722
KJW127_HLA-A2_64_PRE.tsv	Negative	0.0514838988258349	12.7748014350182
KJW127_HLA-A2_65_14_days.tsv	Positive	0.0796157896088553	45.2094924138533

Showing 1 to 5 of 5 entries

Figure 4. iCAT Prediction tab. The prediction tab allows the user to upload one or more independent TCR-sequencing samples for classification.

Operation

Samples, such as from blood or lymph tissue, are collected and genetic material is purified. TCR sequences present in the sample are selectively amplified and then sequenced (Figure 1A). The first step of iCAT is the "`Training`" step. A user should provide multiple negative training samples (naïve, unexposed, uninfected, etc.) using the `Browse` button. Then, repeat the step for positive training samples (exposed, infected, vaccinated, etc.). The user should select the type of training feature. iCAT provides three options: (1) `CDR3 Amino Acid Sequence` (TCRs will need the same CDR3 region to be called "Identical"), (2) `TCRV-CDR3-TCRJ` (TCRs will need the same TCRBV segment, CDR3 region, and TCRJ segment to be called "Identical"), (3) `Nucleic Acid (DNA)` (TCRs will need the exact same DNA rearrangements/sequence across TCRBV, CDR3, and TCRJ). Selecting `TCRV-CDR3-TCRJ` is recommended as a balance between sensitivity and specificity and this option has been used for all the use cases in this paper. In addition, users can customize the range acceptable of copies per clonotype, and the minimum threshold of public sequences, which determines the minimum samples a TCR sequence must be observed in to be considered for analysis.

One important option of this "`Training`" tab is `Max p-value (default: 0.1)`, which determines the minimal degree of statistical significance that iCAT will accept as being potentially "associated" with the positive group. The statistical methodology of iCAT is based on identifying a subset of TARSs that informs classification.[5] TCR sequences significantly associated with positive samples as opposed to negative samples are identified by performing a one-tailed Fisher's exact test. iCAT determines the optimal p-value cutoff to generate the TARS library based on the idea of coverage ratio. To determine an optimal p-value threshold for identifying vaccine-associated TCRβ sequences, we applied a heuristic test that selected the optimal p-value threshold based on the "coverage" provided by the library for both vaccinated (C_v) and naïve samples (C_n).[5] "Coverage" is defined as the summation of the number of samples containing each TARS divided by the number of samples. In the equations below, x_i denotes the number of vaccinated samples a single TCRβ is identified and n_v represents the number of positive samples in the training data. y_i also denotes for naïve samples and n_n represents the number of naïve samples.

$$C_v = \frac{\sum_{i=1} x_i}{n_v}, C_n = \frac{\sum_{i=1} y_i}{n_n}. \tag{1}$$

The ratio of C_v to C_n is determined for each p-value. The p-value with the largest C_v:C_n ratio and offers significant coverage to distinguish vaccinated (v) from naïve (n) samples was chosen.

For the classification of vaccinated and naïve samples, iCAT calculates the percentage of TARS (% TARS) present in a sample. The % TARS for each sample in the training data is compared against the % TARS normal distribution for each group to predict if each sample is "positive" or "negative", determined by which group a sample is more closely associated with.[5] Such a normal distribution has been adopted to calculate the distance of a sample to the mean. In detail, the normal distributions for the naïve and vaccinated populations in our training data were calculated based on a function of the difference between a single sample value (x) and the mean of a set of data (μ) over the standard deviation of that set of data (σ) from the below equation.

$$f(x|\mu,\sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}}) \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2}$$

If the value is bigger, we can conclude that the sample is more associated with the training group. By comparing a sample against the normal distribution of vaccinated and naïve training groups, we can determine which group a sample is more statistically associated with. A progress bar will show on the bottom-right corner to update on the status of training. After finishing, the "`Training`" tab will show some exploratory tables and a figure regarding the training data and the model built, which can all be downloaded to the user's machine easily. In addition, the "`Library`" and "`Prediction`" tabs will unlock.

The "`Library`" tab displays a table consisting of the TARS, determined to be statistically associated with exposure to the target/agent/pathogen (Figure 3). The table displays each sequence, number of positive and negative training samples the sequence is present in/absent from, and how statistically associated the sequence is to the positive training data (p-value). The table can be downloaded to the user's computer for further analysis (Figure 3).

To allow the diagnostic classifier to test independent cohorts of samples, a third tab in iCAT, "`Prediction`" tab, also unlocks after training and allows the user to upload one or more independent TCR-sequencing samples for classification using the parameters generated by the training data (Figure 4). The "`Prediction`" tab allows the user to diagnose unknown samples (e.g. not included in the previous training data) for classification as "Positive" or "Negative" and determining the accuracy of the diagnostic assay. Use the `Browse` button to upload such independent samples for prediction. Multiple samples can be uploaded simultaneously. Click `Predict Independent Sample` will analyze the dataset. A table will appear after analysis is complete. The table displays sample names along with the prediction “Positive” (red) or “Negative” (blue), and displays the %TARS that is the percent of individual sequences from the sample that are included in the TARS library. The prediction results can be downloaded as a table.

iCAT requires R 3.4.0 or upper and can be run on any operating system with common specifications (1 GB disk space, 4 GB memory, and multicore CPU is recommended).

Use cases

To evaluate the efficacy of iCAT, we used three TCR sequencing data sets that are publicly available. The first viral data set consists of 148 training and 20 independent mouse samples. The training set has 32 mouse pre-treatment naïve group and 116 vaccinated samples, each cohort inoculated intranasally with the ACAM2000 smallpox vaccine. The second viral data set consists of 133 (27 negative and 96 positive) training and 15 (5 negative and 20 positive) monkeypox virus. The two mouse datasets are publicly available from https://doi.org/10.17632/cf92gt44zf.1. The third data set is human TCR samples exposed to the novel SARS-CoV- 2 virus which is the cause of the ongoing COVID-19 global pandemic.[12] The sample size is small (two cohorts), but those two cohorts’ TCR repertoires were obtained for other projects one and two years prior to infection. Therefore, this negative and positive SARS-CoV-2 human TCR sequencing data set from same cohorts can be a great example to show the great potential of iCAT as a diagnostic assay. This data is available at https://doi.org/10.5281/zenodo.3835956.

Use case 1: smallpox mouse data

32 pre-exposure (naïve) samples were analyzed, which included 2,049,383 unique TCR sequences (clonotypes). We setup iCAT options to analyze by either the CDR3 amino acid sequence or the V-gene and J-gene names. 714,522 amino acid sequence-gene name combinations were found in the naïve samples. 58 samples taken 2- and 8-weeks post-vaccination for smallpox were analyzed, which included 1,581,619 clonotypes and 573,612 unique amino acid sequence-gene name combinations. After training, iCAT accurately generated the same virus-associated TCR library (314 TCR sequences) identified in Wolf, et al., 2018.[5] When applied to the training data as a baseline check, iCAT correctly classified 32 of 32 naïve samples as ”unexposed” and 58 of 58 vaccinated samples as ”exposed” (100% accuracy). We utilized TCR-sequencing files from 10 mice pre- and post-smallpox vaccination that were not involved in the training of the diagnostic classifier to act as independent cohorts to test the diagnostic accuracy of the iCAT generated classifier. The classification results are displayed in the “Prediction” tab and can be downloaded as a .txt file. From a total of 20 samples, 90% of pre-vaccination samples (9 of 10) were correctly classified as “negative” and 100% of samples post-smallpox vaccination (10 of 10) were classified as ”positive”. Training time was 2.36 minutes and classification time was 30.6 seconds. Overall, this data displays that the iCAT platform computationally identifies target-associated public TCRs, utilized to train a diagnostic classifier capable of distinguishing between exposed and unexposed samples with a high degree of accuracy.

Use case 2: monkeypox mouse data

We tested iCAT using another TCR-sequencing mouse dataset which included 27 naïve samples and 48 samples 2- and 8-weeks post infection with monkeypox. We chose to analyze based on CDR3 amino acid sequence in addition to V-gene and J-gene names. The p-value cutoff was set to 0.1 and the minimum number of public sequences was set to 1. Those parameters produced the best separation experimentally. Naïve samples included 1,772,085 clonotypes and 630,381 unique amino acid sequence-gene name combinations. Exposed samples included 1,070,615 clonotypes and 382,906 unique amino acid sequence-gene name combinations. iCAT correctly classified this training data with 100% accuracy. When tested on an independent monkey pox data set – set up by excluding 5 samples from the naïve group and 10 samples from the exposed group pre-training – iCAT correctly classified the 5 naïve samples as negative and the 10 exposed samples as positive. Thus, we demonstrated a 100% classification accuracy using iCAT on this monkeypox dataset. Training time was 47.81 seconds and classification time was 6.83 seconds.

Use case 3: human SARS-CoV-2 data

We further tested iCAT on TCR-sequencing data from two human individuals exposed to the novel SARS-CoV-2 virus.[13] Data included 4 naïve samples from 2018 and 2019, and 4 samples collected 15- and 30-days post-infection. We chose the iCAT option to analyze by CDR3 amino acid sequences only. The p-value cutoff was set to 0.1 and the minimum number of public sequences was set to 1. The naïve data included 2,935,893 clonotypes and 1,120,606 unique CDR3 amino acid sequences. The exposed data included 1,987,608 clonotypes and 541,111 unique amino acid sequences. iCAT achieved a perfect classification accuracy on the training data, correctly assigning the 4 naïve and 4 exposed samples. Further, iCAT correctly classified 4 independent exposed samples as positive. Training time was

1.50 minutes and classification time was 33.12 seconds. This demonstrates the wide utility of iCAT and the methodology it implements.

Conclusions

In this article, we have presented iCAT, a powerful software tool for determining pathogen exposure through TCR sequencing data. It has significant clinical applications in disease diagnosis, surveillance, as well as for determining potential vaccine efficacy. Once data interpretation is fully automated, the TCR sequencing analysis and other types of NGS will likely become a standard tool for diagnosis and management of disease. Our current datasets are from pre- and post- exposure to viruses, and serve as a proof of principle that TCR sequencing analysis can be utilized to identify individuals exposed to infectious agents or vaccines with great accuracy, speed, and accessibility. We demonstrated the use of iCAT for accurately detecting exposure to the SARS-CoV-2, the virus behind COVID-19. Although this use case was based on a few number of samples, it shows the immense potential of our software the utilization of TCRs as a biomarker. This type of analysis may be used to distinguish between two different but highly related infections, such as Zika virus and Dengue, which is one of the global concerns considering Zika virus's association with fetal complications in infected pregnant women, and current laboratory testing cannot distinguish between the two. Parallel endeavors in our group show promising results in identifying virus-associated TCR sequences uniquely associated with a prior Zika versus Dengue virus infection in mice using iCAT. Further, the iCAT platform may prove useful for diagnosing individuals in the early stages of autoimmunity, by identifying auto-reactive TCRs before symptoms and significant tissue damage occurs. Earlier diagnosis may allow for preventative measures, better treatment, and better outcomes. Broadly, our approach can be used to diagnose autoimmune disease and possibly immune responses to cancer before or after immunotherapy.

Data availability

Mendeley Data: Identifying and Tracking Low Frequency Virus-Specific TCR Clonotypes Using High-Throughput Sequencing, https://doi.org/10.17632/cf92gt44zf.1.[14]

This project contains the raw sequencing data from HLA-A2 transgenic mice before and after infection with either the ACAM2000 smallpox virus or highly releated monkeypox virus.

Zenodo: Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T cell memory formation after mild COVID-19 infection, https://doi.org/10.5281/zenodo.3835956.[15]

This project contains the third human TCR samples exposed to the novel SARS-CoV-2 virus is available from zenodo in mixcr format.

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

Source code available from: https://github.com/BioHPC/iCAT.

Archived sourced code as at time of publication: http://doi.org/10.5281/zenodo.4436485.[16]

License: MIT

References

1. Rosati E, Dowds CM, Liaskou E, *et al.*: **Overview of methodologies for t-cell receptor repertoire analysis.** *BMC Biotechnol* 2017; **17**(1): 61. ISSN 1472-6750.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Venturi V, Kedzierska K, Turner SJ, *et al.*: **Methods for comparing the diversity of samples of the t cell receptor repertoire.** *J Immunol Methods* 2007; **321**.
PubMed Abstract | **Publisher Full Text**

3. Cabaniols JP, Fazilleau N, Casrouge A, *et al.*: **Most alpha/beta t cell receptor diversity is due to terminal deoxynucleotidyl transferase.** *J Exp Med.* 2001; **194**(9): 1385–1390. ISSN 0022-1007 1540-9538.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

4. Robins HS, Campregher PV, Srivastava SK, *et al.*: **Comprehensive assessment of t-cell receptor beta-chain diversity in alphabeta t cells.** *Blood* 2009; **114**(19): 4099–107. ISSN 1528-0020 (Electronic) 0006-4971 (Linking).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Wolf K, Hether T, Gilchuk P, *et al.*: **Identifying and tracking low-frequency virus-specific tcr clonotypes using high-throughput sequencing.** *Cell Rep* 2018; **25**(9): 2369–2378, e4, ISSN 2211-1247.
Reference Source | **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

6. Pogorelyy MV, Fedorova AD, McLaren JE, *et al.*: **Exploring the pre-immune landscape of antigen-specific T cells.** *Genome Med* 2018; **10**: 68.
Publisher Full Text

7. Emerson RO, DeWitt WS, Vignali M, *et al.*: **Immunosequencing identifies signatures of cytomegalovirus exposure history and hla-mediated effects on the t cell repertoire.** *Cell Rep* 2017; **49**(5): 659–665. ISSN 1546-1718.
PubMed Abstract | **Publisher Full Text**

8. DeWitt WS, Emerson RO, Lindau P, *et al.*: **Dynamics of the cytotoxic t cell response to a model of acute viral infection.** *J Virol* 2015; **89**(8): 4517–4526.
Reference Source | **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

9. Kirsch I, Vignali M, Robins H: **T-cell receptor profiling in cancer.** *Mol Oncol* 2015; **9**(10): 2063–2070. ISSN 1574-7891.
Reference Source | **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

10. Gerritsen B, Pandit A, Andeweg AC, *et al.*: **Rtcr: a pipeline for complete and accurate recovery of t cell repertoires from high throughput sequencing data.** *Bioinformatics* 2016; **32**(20): 3098–3106, ISSN 1367-4811 (Electronic) 1367-4803 (Linking).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

11. Nazarov VI, Pogorelyy MV, Komech EA, *et al.*: **tcr: an r package for t cell receptor repertoire advanced data analysis.** *BMC Bioinformatics* 2015; **16**: 175, ISSN 1471-2105 (Electronic) 1471-2105 (Linking).
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Chang W, Cheng J, Allaire J, *et al.*: ***shiny: Web Application Framework for R.*** 2018.
Reference Source

13. Minervina AA, Komech EA, Titov A, *et al.*: **Longitudinal high-throughput tcr repertoire profiling reveals the dynamics of t cell memory formation after mild covid-19 infection.** *bioRxiv* 2020.
Reference Source | **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

14. Wolf K: **Identifying and Tracking Low Frequency Virus-Specific TCR Clonotypes Using High-Throughput Sequencing.** *Mendeley Data* 2018; **V1**.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

15. Minervina AA, Komech EA, Titov A, *et al.*: **Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T cell memory formation after mild COVID-19 infection (Version 1.0) [Data set].** Zenodo; 2020
Reference Source | **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

16. Rajeh A, Ahn TH: (2021, January 13). BioHPC/iCAT: First release of iCAT (Version v1.0.0). Zenodo.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 12 July 2021

https://doi.org/10.5256/f1000research.57705.r88494

✓ **Se-Ran Jun**
Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

Thank you so much for the improved manuscript. I have minor comments:

- Please add a sentence(s) on how you calculated uncertainty.
- Please check the following sentence: "The p-value with the largest C_v:C_n ratio and offers significant coverage to distinguish vaccinated (*v*) from naïve (*n*) samples was chosen."
- Is the following sentence correct: "If the value is bigger, we can conclude that the sample is more associated with the training group."?

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomic Epidemiology, Microbiome Epidemiology, Multiomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 06 July 2021

https://doi.org/10.5256/f1000research.57705.r88495

✓ **Shuo-Wang Qiao**
Department of Immunology, Oslo University Hospital, Rikshospitalet, University of Oslo, Oslo,

Norway

Ying Yao

Department of Immunology, University of Oslo, Oslo, Norway

No further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Immunology, T cells, TCR, single-cell receptor sequencing,

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 13 April 2021

https://doi.org/10.5256/f1000research.30069.r82161

? **Se-Ran Jun**

Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

This article is a valuable resource for T-cell receptor (TCR) sequencing data-based infection diagnostics. The authors provided a R software package named iCAT which provides a step-by-step analysis of TCR sequencing data to predict if individual subjects are infected or not. First, iCAT takes positive and negative pathogen-specific TCR datasets to estimate parameters involved in in model. Second, iCAT provides the Library tab which summarizes the target associated receptors sequences with p-values. Third, iCAT provides prediction tab which takes test TCR datasets and make diagnostic decision. I enjoyed reading this article and github tutorial.

Major comments:

It seems like that the performance of this tool does depend on the number of samples. Please provide discussion or guideline on the number of negative and positive samples relating to the performance.

It seems like that this is the first tool available for TCR data-based diagnostic tool. Please clarify if this is the first tool which uses TCR sequencing data for the purpose of diagnostic. If this is not the first tool, then comparison results with other tools should be included.

Minor comments:

First, I have a question for clarification purpose. Before TCR sequencing data is uploaded, does TCR sequencing data need any preprocessing step to make input compatible with iCAT?

Install.packages(“devtools”) does not work with R4.0.3 on Mac OS X. The error I got is as follows:

```
Downloading GitHub repo BioHPC/iCAT@HEAD
Error in utils::download.file(url, path, method = method, quiet = quiet,  :
download from 'https://api.github.com/repos/BioHPC/iCAT/tarball/HEAD' failed
```

Each input file is limited to 10Gigabyte in size. Where does this limitation come from?

I downloaded one of test cases which contains many files. It was not easy to figure out which files to be uploaded. It would be nice that this R package provides example datasets along with the package so that they can be played in R directly.

There is an option named ‘Other’ for ‘analyze clonotypes by (column names in parantheses)’. What is other?

Please include information of computational time for training and prediction steps for each case.

Although feature from an option of ‘TCRV-CDR3-TCRJ’ is recommended, please include the performance for other features for each test case which could help readers understand and compare features.

What does TCRBV stand for?

What does VATS stand for?

A correction constant is not shown in equation (2).

Whole genome sequences can distinguish DENGU from ZIKE with 100% accuracy. For the sentence ‘current laboratory testing cannot distinguish between the two’, are you excluding whole genome sequencing data?

Is it possible to make change range of acceptable into something like 1:99 or 1~99 instead of 1 99 to indicate range?

Does the max p-value change automatically according to the training data provided? Or should it be changed manually based on the selected one by Fisher’s t-test?

Please revise typo in Figure 1(B)

Please revise typo with ‘the minimum samples’

Please revise the following sentences:
- The ratio of Cv to Cn is determined for each p-value. The p-value with the largest Cv:Cn ratio and offers significant coverage to distinguish (v)accinated from (n)aïve samples was chosen.
- iCAT was configured to analyze by CDR3 amino acid sequence only

If the value is bigger, we can conclude that the sample is more associated to the the training group. Is this a correct sentence?

From the training data, iCAT correctly classified 32 of 32 naïve samples as ”unexposed” and 58 of 58 vaccinated samples as ”exposed” (100% accuracy). Is this a correct sentence?

Is the rationale for developing the new software tool clearly explained?
Partly

Is the description of the software tool technically sound?
Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomic epidemiology, Multiomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 10 Jun 2021

Tae-Hyuk Ahn, Saint Louis University, St. Louis, USA

Major comments:

1. It seems like that the performance of this tool does depend on the number of samples. Please provide discussion or guideline on the number of negative and positive samples relating to the performance.

Response: We really appreciate all the great and thoughtful comments from the reviewer. The sample size is of course very important in determining the power and accuracy of the diagnostic. We have generated positive and negative datasets of various sample sizes to determine the impact of sample numbers on accuracy. The conclusion was that the number of samples required to generate an accurate diagnosis is determined by the nature of the samples. For example, a dataset from Zika infected mice produced a small but conserved virus-specific TCR repertoire due to the small number of immunogenic epitopes (Hassert and Wolf et. al. 2020) ; this allowed for the generation of a powerful diagnostic assay using relatively small numbers of samples. Comparatively, using a dataset from mice before and after smallpox vaccination resulted in a large pool of virus-specific TCRs due to a large number of potential epitopes; the less focused response required a greater number of samples in each group to generate sufficient coverage and diagnostic accuracy (Wolf et. al. 2018).

2. It seems like that this is the first tool available for TCR data-based diagnostic tool. Please clarify if this is the first tool which uses TCR sequencing data for the purpose of diagnostic. If this is not the first tool, then comparison results with other tools should be included.

Response: To our limited knowledge, this is the first publicly available tool for diagnostic classification using TCR sequencing data.

Minor comments:

1. First, I have a question for clarification purpose. Before TCR sequencing data is uploaded, does TCR sequencing data need any preprocessing step to make input compatible with iCAT?

Response: iCAT takes in TCR data in tab-delimited clonotype abundance tables. These tables are the result of mapping of reads from Variable, Diversity, and Joining segments then the assembly of clonotypes. iCAT does not perform these steps so they may be considered pre-processing. Clonotype tables are provided through the immunoSEQ assay from Adaptive Technologies, which is the most commonly used immunosequencing pipeline in recently published studies. While iCAT input-parsing was built around immunoSEQ’s format, it also has a flexible interface that allows handling of different formats with different user-selected column names.

2. Install.packages(“devtools”) does not work with R4.0.3 on Mac OS X. The error I got is as follows:

This issue is resolved and updated on iCAT.

3. Each input file is limited to 10Gigabyte in size. Where does this limitation come from?

Response: This is an arbitrary limitation that we coded into the tool’s interface. It is not a required limitation for any of the underlying functions in iCAT. The reason we put this size limit in place is because one of our future goals with iCAT is to host it on a website where users can upload their data for analysis, and thus data size can become a limiting factor.

4. I downloaded one of test cases which contains many files. It was not easy to figure out which files to be uploaded. It would be nice that this R package provides example datasets along with the package so that they can be played in R directly.

Response: Example negative, positive, and independent datasets are included in the R package under the path “inst/extdata/”.

5. There is an option named ‘Other’ for ‘analyze clonotypes by (column names in parantheses)’. What is other?

Response: This option exists to allow users to input a custom column name to analyze by. TCR data can vary in formatting (e.g. some files might describe the amino acid column as “aminoAcid” while others can describe it as “aa”) so this option can help users load different data formats into iCAT.

6. Please include information of computational time for training and prediction steps for each case.

Response:

Use case 1:
Training time: 2.36 minutes
Classification time: 30.6 seconds

Use case 2:
Training time: 47.81 seconds
Classification time: 6.83 seconds

Use case 3:
Training time: 1.50 minutes
Classification time: 33.12 seconds

The manuscript was updated to reflect this information.

7. Although feature from an option of ‘TCRV-CDR3-TCRJ’ is recommended, please include the performance for other features for each test case which could help readers understand and compare features.

Response: In general, we found that using the J-gene and V-gene info improves accuracy in our testing, but other features are available in case researchers to believe they might get better results with them. Using the amino acid sequence for the monkeypox data, 1/5 negative samples were identified correctly, while 6/10 positive samples were identified correctly. Using the DNA sequence, 5/5 negative samples were identified correctly and 3/10 positive samples were identified correctly.

8. What does TCRBV stand for?

Response: It stands for T cell receptor beta chain variable region.

9. What does VATS stand for?

Response: It stood for vaccine-associated target sequences. This is now changed to target-associated receptor sequences (TARSs) for consistency in the manuscript.

10. A correction constant is not shown in equation (2).

Response: That is correct We updated the sentence for the equation in the manuscript.

11. Whole genome sequences can distinguish DENGU from ZIKE with 100% accuracy. For the sentence ‘current laboratory testing cannot distinguish between the two’, are you excluding whole genome sequencing data?

Response: We are limiting that comment to typical laboratory diagnostics which typically does not include whole-genome sequencing.

12. Is it possible to make change range of acceptable into something like 1:99 or 1~99 instead of 1 99 to indicate range?

Response: We reflected this comment and iCAT now allows inputting either a space or a colon to delimit range values.

13. Does the max p-value change automatically according to the training data provided? Or should it be changed manually based on the selected one by Fisher’s t-test?

Response: The max p-value can be changed manually. The purpose is to allow the user to modify the stringency of the diagnostic. This could be done after generating the diagnostic library and the user determines they would prefer more (or less) stringent parameters.

14. Please revise typo in Figure 1(B)

Response: Fixed typo in the new manuscript version.

15. Please revise typo with ‘the minimum samples’

16. Please revise the following sentences:
- The ratio of Cv to Cn is determined for each p-value. The p-value with the largest Cv:Cn ratio and offers significant coverage to distinguish (v)accinated from (n)aïve samples was chosen.
- iCAT was configured to analyze by CDR3 amino acid sequence only

Response: Sentences were revised in the new manuscript version to make them more clear.

17. If the value is bigger, we can conclude that the sample is more associated to the the training group. Is this a correct sentence?

Response: The sentence was revised in the new manuscript version.

18. From the training data, iCAT correctly classified 32 of 32 naïve samples as ”unexposed” and 58 of 58 vaccinated samples as ”exposed” (100% accuracy). Is this a correct sentence?

Response: The sentence was revised in the new manuscript version.

Competing Interests: No competing interests were disclosed.

Reviewer Report 12 April 2021

https://doi.org/10.5256/f1000research.30069.r82160

?

Shuo-Wang Qiao
Department of Immunology, Oslo University Hospital, Rikshospitalet, University of Oslo, Oslo, Norway
Ying Yao
Department of Immunology, University of Oslo, Oslo, Norway

Diagnostic methods based on the detection of disease-specific or disease-associated T-cell receptor (TCR) sequences are novel approaches whose development is still in its infancy. This paper introduces an intuitive, easy-to-use, software package for the analysis of TCR deep sequencing data with the purpose of grouping data into ‘naïve/healthy’ or ‘diseased/exposed’ groups. The statistical methods used is rather rudimentary, but serve the purpose. It demonstrates remarkably high accuracy in three use cases, including a small dataset of COVID-19. However, the choice of parameter settings and how data is split into training and test datasets, is not well explained. Thus, one is left with a feeling that some over-fitting and over-tuning may have occurred. Nevertheless the shortcomings, overall this software package is first of its kind and will be welcomed by the growing community of scientists working with immune receptors and disease prediction.

Major comments:

1. Introduction: Aside from the immensity of TCR diversity, both theoretical and within each individual, by far the largest challenge to identifying signals of T-cell memory to a particular antigen, is the rarity of specific T cells to any given antigen. In circulation during the homeostatic memory phase, this number can be as low as 1 in 100 000 for CD4+ T cells, and may be up to 1% in the CD8 compartment.

2. Operation: Does iCAT support both TCRB and TCRA data, separately, or in the same dataset? Or does it only support TCRB data?

3. Setting (2) during training: at which level of TRBV and TRBJ allelic similarity is the tool tuned?

In other words, would human TRBV7-2*01 and TRBV7-2*02 be considered the same, or different? How about human TRBV4-2 and TRBV4-3 that are very similar, especially at the 3'-end where the last 40 amino acids are identical. Many HTS short reads would not be able to distinguish between these two gene segments.

4. Use Case 1: The authors wrote that they analysed this dataset either by the CDR# amino acid sequence only, or with the V- and J-gene identity in addition. Did these two options behave differently?

5. In the use cases, it is not common to report the accuracy on training dataset. it is not clear how the data was split into the training and testing dataset. More common practice is to do cross validation, and report the averaged precision, sensitivity, maybe F-score, especially for unbalanced samples. For Use case 2 for example, how were the 5 naïve and 10 exposed samples in the test group chosen, randomly? Would one get the same 100% accuracy if another random set of 5 naïve and 10 exposed samples was chosen?

6. In the training tab, the default Max p-value is 0.1. Usually in disease associated studies, a more stringent cut-off p value is recommended. Considering the main purpose of iCAT is for diagnosis, the specificity is not as crucial, 0.1 might be a good choice. Is there any analysis supporting the choice, e.g. in the use cases or on other public available data which antigen specific sequences is known?

7. In the prediction tab, it would be better to provide a measurement of uncertainty in addition to %TARS

8. Some studies such as, Schneider-Hohendorf et al. (2018)[1], Britanova et al. (2014)[2], and DeWitt et al. (2018)[3] suggested that MHC, age and sex also effect the immune receptor repertoire. Incorporating such meta information should be useful, even if in cases where the sample size is too limited for these factors to be modelled in the training, they are better to be randomized in training and testing datasets.

9. How about computing time? How long did it take to do the training, and classification of for instance the use case 2 data?

10. There are substantial public resources including some antigen-specific database and repertoire database. It might be useful to allow loading some commonly used datasets as controls in case the user has limited sample size.

Minor comments:

1. Figure legend 1: TCR sequencing of blood samples can be done with, but not limited to, genomic DNA. Many groups also use cDNA.

2. Why does the tip of the iCAT's tail has an antibody? If this were a TCR analysis tool, would it not be more appropriate with a cartoon drawing of TCR?

3. What is VATS in last paragraph in Page 5, is it the same as TARS?

4. In the library tab, pattern for significant TARS would make more sense, e.g. are they all similar or there were numbers of clusters based on the similarity?

References

1. Schneider-Hohendorf T, Görlich D, Savola P, Kelkka T, et al.: Sex bias in MHC I-associated shaping of the adaptive immune system. *Proceedings of the National Academy of Sciences*. 2018; **115** (9): 2168-2173 Publisher Full Text
2. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, et al.: Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling.*J Immunol*. 2014; **192** (6): 2689-98 PubMed Abstract | Publisher Full Text
3. DeWitt W, Smith A, Schoch G, Hansen J, et al.: Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife*. 2018; **7**. Publisher Full Text

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Immunology, T cells, TCR, single-cell receptor sequencing,

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 10 Jun 2021

Tae-Hyuk Ahn, Saint Louis University, St. Louis, USA

Major comments:

1. Introduction: Aside from the immensity of TCR diversity, both theoretical and within each individual, by far the largest challenge to identifying signals of T-cell memory to a particular antigen, is the rarity of specific T cells to any given antigen. In circulation during the

homeostatic memory phase, this number can be as low as 1 in 100 000 for CD4+ T cells, and may be up to 1% in the CD8 compartment.

Response: We really appreciate all the great and thoughtful comments from the reviewer. As the reviewer commented, the rarity of specific T cells to a specific antigen is also a challenge in addition to the diversity and magnitude of the TCR repertoire. We updated the Introduction from the comment cited a reference for it.

2. Operation: Does iCAT support both TCRB and TCRA data, separately, or in the same dataset? Or does it only support TCRB data?

Response: iCAT has been developed and tested to analyze TCR beta chain of the T-cell receptor in the alpha/beta T cells. The column call for the V and J are vGeneName and jGeneName and do not distinguish between alpha and beta. So, as long as the remaining formatting is the same, there is no reason we couldn't do either TCRA or TCRB. However, it is not designed to do so in tandem.

3. Setting (2) during training: at which level of TRBV and TRBJ allelic similarity is the tool tuned? In other words, would human TRBV7-2*01 and TRBV7-2*02 be considered the same, or different? How about human TRBV4-2 and TRBV4-3 that are very similar, especially at the 3'-end where the last 40 amino acids are identical. Many HTS short reads would not be able to distinguish between these two gene segments.

Response: We do not look that deeply into the different V and J chains. TRBV7-2*01 and TRBV7-2*02 are in the column vMaxResolved or jMaxResolved, which we do not use for the V and J calling. We use the vGeneName and jGeneName, which for this example, both would be called as TRBV7-2. But TRBV4-2 and TRBV4-3 will be called differently as those would be distinguished in the vGeneName column. You can see this in the data table of the generated library, the selected V and J regions do not include the *01 or *02 details. This is also true when you look in the TSV file.

4. Use Case 1: The authors wrote that they analysed this dataset either by the CDR# amino acid sequence only, or with the V- and J-gene identity in addition. Did these two options behave differently?

Response: Using the V- and J-gene identity in addition to the CDR3 amino acid sequence results in significantly higher prediction accuracy in our testing, likely due to identifying more unique markers. Specifically, it results in fewer false positives. For example, in use case 1, using the amino acid sequence only resulted in correctly identifying only 1 / 10 of the pre-vaccination samples, while still correctly identifying 10/10 of the post-vaccination samples.

5. In the use cases, it is not common to report the accuracy on training dataset. it is not clear how the data was split into the training and testing dataset. More common practice is to do cross validation, and report the averaged precision, sensitivity, maybe F-score, especially for unbalanced samples. For Use case 2 for example, how were the 5 naïve and 10 exposed samples in the test group chosen, randomly? Would one get the same 100%

accuracy if another random set of 5 naïve and 10 exposed samples was chosen?

Response: We agree it is not common to report the accuracy on the training data and it is not as useful as the accuracy on the independent testing data in evaluating a classification tool. However, we chose to include it to demonstrate this feature of the iCAT interface. This feature may be useful as a baseline-check after training in the usual scenarios where the ground truth for the testing data is not available.
The independent samples in use case 2 were chosen randomly before training and excluded from the training set. We also repeated the test three times after receiving this question using a script to randomly select 15 samples for testing. The accuracy was at a 100% for all the new random tests.

6. In the training tab, the default Max p-value is 0.1. Usually in disease associated studies, a more stringent cut-off p value is recommended. Considering the main purpose of iCAT is for diagnosis, the specificity is not as crucial, 0.1 might be a good choice. Is there any analysis supporting the choice, e.g. in the use cases or on other public available data which antigen specific sequences is known?

Response: The max P-value is set at default to 0.1, it can be changed to any value. We found through our own testing groups that it was a reasonable cut-off (for a default value) as p-values above this tended to generate libraries that were not stringent enough. However, the user can manually alter the p-value to more stringent conditions. It is worth noting that while the max p-value of the assay may be defaulted at 0.1, iCAT includes a feature for determining the optimal p-value threshold for library generation, which is likely to be well below the max allowed p-value.

7. In the prediction tab, it would be better to provide a measurement of uncertainty in addition to %TARS

Response: This is a great comment and we totally agree that a measurement of uncertainty in addition to %TARS might be useful for iCAT users. We considered an uncertainty metric based on the difference between the probability density in the pre-exposure distribution and the probability density in the post-exposure distribution. This new feature has been implemented and now available on GitHub iCAT and we updated the README screenshot of the prediction tab.

9. Some studies such as, Schneider-Hohendorf et al. (2018)1, Britanova et al. (2014)2, and DeWitt et al. (2018)3 suggested that MHC, age and sex also effect the immune receptor repertoire. Incorporating such meta information should be useful, even if in cases where the sample size is too limited for these factors to be modelled in the training, they are better to be randomized in training and testing datasets.

Response: We agree this meta information can provide insightful context for different use cases. This is one of our goals for future versions of iCAT (iCAT 2.0). For example, we are especially excited to study the effect of clustering data based on HLA subtypes. In our opinion, this would be better explored in a separate paper.

*manuscript modification*10. How about computing time? How long did it take to do the training, and classification of for instance the use case 2 data?

Response: Following are the results from use case 2.

- Training time: 47.81 seconds
- Classification time: 6.83 seconds

We updated the manuscript to reflect this information.

11. There are substantial public resources including some antigen-specific database and repertoire database. It might be useful to allow loading some commonly used datasets as controls in case the user has limited sample size.

Response: Providing commonly used datasets as default controls would be a convenient feature for some use cases. We considered including such datasets with the iCAT software package, but decided against it in the current version due to size consideration. We aimed to create a versatile and easy to use software tool. Using the instructions in this paper and on GitHub, we believe loading public datasets as needed can be performed conveniently by the average iCAT user. We will consider this feature in iCAT 2.0.

Minor comments:

1. Figure legend 1: TCR sequencing of blood samples can be done with, but not limited to, genomic DNA. Many groups also use cDNA.

Response: Fixed in the new manuscript version.

2. Why does the tip of the iCAT's tail has an antibody? If this were a TCR analysis tool, would it not be more appropriate with a cartoon drawing of TCR?

Response: While it is true that a TCR tail would be more descriptive of our tool, it is our opinion that an antibody makes a more aesthetic cat tail than a TCR. Its shape is also more widely recognizable and can give a general idea to non-specialists about what iCAT deals with at first glance.

3. What is VATS in last paragraph in Page 5, is it the same as TARS?

Response: Yes. Fixed in the new manuscript version.

4. In the library tab, pattern for significant TARS would make more sense, e.g. are they all similar or there were numbers of clusters based on the similarity?

Response: The current version of iCAT does not perform clustering of the target-associated receptor sequences based on similarity, but this can be developed and incorporated in future versions. We agree that it would be a useful addition for exploratory analyses. iCAT currently focuses on using the sequences for prediction, but provides a starting point for further analyses by allowing the download of those significant sequences and some statistical information about their frequency in the training data.

Competing Interests: No competing interests were disclosed.