Memory $CD4^+$ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus

Title: Memory $CD4^+$ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus

Authors

Nicolas De Neuter[1,2,3*], Esther Bartholomeus[3,4*], George Elias[3,5*], Nina Keersmaekers[3,6], Arvid Suls[3,4], Hilde Jansens[7], Evelien Smits[3,5,8,9], Niel Hens[3,6,10,11], Philippe Beutels[3,6], Pierre Van Damme[3,11], Geert Mortier[3,4], Viggo Van Tendeloo[3,5], Kris Laukens[1,2,3], Pieter Meysman[1,2,3$], Benson Ogunjimi[3,5,6,8,12$]

Affiliations

1. Adrem Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
2. Biomedical Informatics Research Network Antwerp (biomina), University of Antwerp, Antwerp, Belgium
3. AUDACIS, Antwerp Unit for Data Analysis and Computation in Immunology and Sequencing, University of Antwerp, Antwerp, Belgium
4. Center for Medical Genetics, University of Antwerp / Antwerp University Hospital, Edegem, Belgium
5. Laboratory of Experimental Hematology (LEH), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium
6. Centre for Health Economics Research & Modeling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium
7. Department of Laboratory Medicine, Antwerp University Hospital, Edegem, Belgium

8. Center for Cell Therapy and Regenerative Medicine, Antwerp University Hospital, Edegem, Belgium
9. Center for Oncological Research Antwerp, University of Antwerp, Antwerp, Belgium
10. Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
11. Centre for the Evaluation of Vaccination (CEV), Vaccine & Infectious Disease Institute (VAXINFECTIO), University of Antwerp, Antwerp, Belgium
12. Department of Paediatrics, Antwerp University Hospital, Edegem, Belgium

Author contributions

NDN, EB and GE contributed equally to this article*

PM and BO contributed equally to this article[$]

Abstract

Pathogens of past and current infections have been identified directly by means of PCR or indirectly by measuring a specific immune response (e.g. antibody titration). Using a novel approach, Emerson and colleagues showed that the cytomegalovirus serostatus can also be accurately determined by using a T cell receptor repertoire data mining approach. In this study, we have sequenced the $CD4^+$ memory T cell receptor repertoire of a Belgian cohort with known cytomegalovirus serostatus. A random forest classifier was trained on the CMV specific T cell receptor repertoire signature and used to classify individuals in the Belgian cohort. This study shows that the novel approach can be reliably replicated with an equivalent performance as that reported by Emerson and colleagues. Additionally, it provides evidence that the T cell receptor repertoire signature is to a large extent present in the $CD4^+$ memory repertoire.

Introduction

Identification of both past and current infections has long relied on the detection of the pathogen within the host. Currently, numerous molecular assays are being employed that rely on the detection of pathogen DNA/RNA in the host[1,2]. More recently, infectious disease diagnostics has seen the development of novel biotechnologies focused on host RNA signatures derived from patient blood samples[3]. Signatures within the host RNA levels have proven usable for the identification of causative pathogen(s)[4], for example to distinguish between bacterial and viral infections in febrile infants[5,6] or to distinguish tuberculosis from other diseases in children[7]. These approaches with determination of blood RNA signatures achieve accuracies ranging from 85% up to 98%.

Host RNA signatures may not be the only way of accurately identifying the causative pathogen. The adaptive immune system is tasked with the recognition and elimination of invading pathogens. As such, pathogen specific signatures can be expected to be traceable within the immune repertoire[8]. Indeed, identification and quantification of T cell receptor (TCR) sequences associated with a certain pathogen or disease promises to be a fundamentally new approach for clinical diagnosis and monitoring of infectious diseases, autoimmunity and cancer. In this case, RNA or DNA from an individual's blood is selectively sequenced to characterize the TCR beta-chain and/or alpha-chain sequences that represent the individual's T cell repertoire[9]. These TCR sequences can be linked to the epitope that the T cell targets[10]. Signature TCR sequences have been reported for several diseases such as diabetes[11] or multiple sclerosis[12,13] and were suggested to be associated with hepatitis B seroconversion during antiviral treatment[14].

Emerson and colleagues demonstrated that the repertoire of T cell receptor beta (TCRβ) sequences in the blood of healthy US bone marrow donors is highly specific for the

cytomegalovirus (CMV) serostatus[15]. They determined the TCRβ repertoire of 641 donors with known CMV serostatus through high-throughput next generation sequencing. Subsequently, they identified TCRβ sequences that were statistically significantly enriched in CMV seropositive donors. These differential TCRβ sequences then formed the basis of a classifier that accurately predicted the CMV serostatus of individuals in an independent cohort. In this work, we show that CMV specific TCR signature is conserved in the $CD4^+$ memory repertoire of 33 Belgian individuals.

Results & Conclusion

In this study, we collected peripheral blood samples from 9 CMV seropositive and 24 CMV seronegative healthy Belgian adults. We sequenced TCRβ sequences from the $CD4^+CD45RO^+$ lymphocyte population only, as opposed to the $CD4^+CD45RO^{+/-}$ and $CD8^+CD45RO^{+/-}$ lymphocyte populations collected in the original study[15], and thus focused solely on the immune signal within the $CD4^+$ memory repertoire. After removal of out of frame TCR sequences, 2 204 828 distinct TCRβ sequences were obtained, with a mean of 66 813 sequences per individual.

In the original study by Emerson et al., 164 TCRβ sequences were found to be differentially associated with CMV seropositive versus CMV seronegative status using the Fisher's exact test. Of these specific CMV TCRβ sequences, 67 could be found within the $CD4^+$ memory repertoire of our Belgian cohort. Each of these CMV specific TCRβ sequences occurred in at least 1 and up to 5 of the 33 individuals and up to 16 CMV specific TCRβ sequences could be found in single individual. Firstly, these results indicate that these CMV associated TCRβ sequences are likely universal as they are present in two geographically distinct populations. Secondly, these sequences are represented within the $CD4^+$ memory repertoire, which supports their long-term nature.

We enumerated for each individual the number of CMV associated TCRβ sequences as well as the total number of productive TCRβ sequences that were sequenced (fig. 1). This figure shows an expected increase in the number of CMV associated TCRβ sequences if more TCRβ sequences were identified. Furthermore, these results already visually show a distinction between CMV+ and CMV- individuals. We implemented the statistical learning framework obtained by Emerson and colleagues. Performance was evaluated on bootstrapped samples of the Belgian cohort and resulted in a median AUC of 0.95 (95% CI: 0.76-1.00)

(fig. 2). For further comparison, we trained a random forest classifier on the cohort data obtained by Emerson and colleagues containing the 641 USA based individuals. Then we applied the classifier on the Belgian cohort and predicted their CMV serostatus based on the number of CMV associated TCRβ sequences present. Training on the US dataset and application of the resulting classifier to our Belgian cohort resulted in a median AUC of 0.91 (95% CI: 0.69 – 1.00) after bootstrapping of the validation set (fig. 3). This result is similar to the AUC of 0.94 obtained by Emerson and colleagues on their own independent dataset[15]. Thus, the classification approach can be transferred to an out of the box random forest with only a slight loss in performance.

Emerson and colleagues presented a novel method for the identification of CMV serostatus based on signatures within the TCR repertoire. Although they validated their classification framework on their own dataset, for adoption in clinical practice it is crucial to further validate this new approach on new TCRβ datasets. We present a study evaluating these results on TCR repertoire data obtained using different experimental set-ups and in another study population.

One of the fundamental differences with the original study lies in the use of a more specific group of targeted T cells. Whereas the original study analyzed TCR sequences from both the naïve and memory $CD4^+$ and $CD8^+$ repertoires, we restricted the analysis to the memory $CD4^+$ TCR signature. As the TCR sequences derived by Emerson et al. were able to accurately determine the CMV serostatus of individuals in our dataset, results suggest that the TCR signature underlying a positive CMV serostatus is to a large extent present in the $CD4^+$ T cell memory repertoire. This finding is supported by recent reports on the antiviral role of $CD4^+$ effector memory T cells in controlling latent human CMV infections[16] and the influence of CMV on the shape of the $CD4^+$ T cell repertoire[17].

This approach opens potential for new avenues in diagnostic testing where current serological methods fall short. In particular, it could be capable of predicting a personalized infection history from the long term immune memory while remaining agnostic to the pathogen under investigation.

While the initial approach was validated on an independent cohort, both cohorts were US based and the results may therefore be biased by the genetic background. We therefore tested if the same approach was also applicable to a non-US based population. TCR sequences derived from the US population were able to predict CMV serostatus in a Belgian population of healthy individuals. These results show that the genetic background of the population does not affect predictions of CMV serostatus for Belgian individuals and indicates that it is unlikely to play a role in other populations of different origin.

Furthermore, the TCR sequences identified by Emerson and colleagues were predictive for the CMV serostatus independently of the computational approach employed. Both the statistical learning framework used in the original approach and the random forest were able to achieve a AUC value of 0.99 on the US training cohort and produced similar AUC values on their respective validation cohorts.

These results provide an important additional validation step and prove that the approach employed by Emerson et al. remains valid under different experimental conditions. We show the validity of the approach using the $CD4^+$ memory repertoire, a different classification algorithm and a study population of different origin.

Materials & Methods

PBMC acquisition and management

Peripheral blood mononuclear cells (PBMCs) were obtained from 33 healthy Belgian participants. Samples were collected within the scope of another study in which we specifically interrogated the $CD4^+$ T cell memory repertoire. Written informed consent was obtained from all study participants. The study was approved by the ethics board of the Antwerp University Hospital.

PBMC were isolated and frozen following standard operating procedures as detailed elsewhere[18]. After thawing and washing cryopreserved PBMCs, total $CD4^+$ T cells were isolated by positive selection using CD4 magnetic microbeads (Miltenyi Biotech, Bergisch Gladbach, Germany). Memory $CD4^+$ T cells were sorted after gating on single viable $CD3^+CD4^+CD8^-CD45RO^+$ cells. The following fluorochrome-labeled monoclonal antibodies were used for staining: CD3-PerCP (BW264/56) (Miltenyi Biotech), CD4-APC (RPA-T4) and CD45RO-PE (UCHT1) (both from Becton Dickinson, Franklin Lakes, NJ, USA) and CD8-Pacific Orange (3B5) (from Thermo Fisher Scientific, Waltham, MA, USA). Cells were stained at room temperature for 20 minutes and sorted with FACSAria II (Becton Dickinson). Sytox blue (Thermo Fisher Scientific) was used to exclude non-viable cells.

TCR sequencing

DNA was extracted using Quick-DNA™ Microprep Kit (Zymo Research, Irvine, CA, USA) according to manufacturer's instructions. TCRβ DNA from memory $CD4^+$ T cells was sequenced using ImmunoSEQ hsTCRβ kit (Adaptive Biotechnologies, Seattle, WA, USA) on an Illumina Miseq sequencer according to the manufacturer's protocol. Processed TCRβ sequencing data is available at https://clients.adaptivebiotech.com/pub/deneuter-2018-cmvserostatus.

CMV antibody titration

Serum was stored at −80 °C until further processing. The presence of IgGs directed against CMV pp150, pp28, p38, and p52 in thawed serum was determined using a Roche Elecsys assay (Roche, Basel, Switzerland).

Immunoinformatics

Training data for the classifier described by Emerson and colleagues were obtained through personal communication [Ryan Emerson, April 2017] and consisted of CMV associated TCRβ counts and distinct TCRβ counts as well as the CMV serostatus for each individual in their healthy bone marrow donor cohort. The beta binomial likelihood model trained by Emerson et al. was implemented in the Python programming language. Random forest classifiers were trained using the default parameters as implemented in Scikit-Learn[19]. Bootstrapped samples from the Belgian cohort were used to validate the performance of the classifiers trained on the data from Emerson and colleagues. The median and 95% confidence interval (CI) of the area under the receiver-operator characteristic (ROC) curve (AUC) values were calculated over 10 000 bootstrap iterations. The 95% CI was calculated as the 2.5th and 97.5th percentile over bootstrapped AUC values. 50% and 80% CI were obtained in a similar way. Because AUC values are limited between 0 and 1, they are not normally distributed. Therefore, the median was used together with multiple CI instead of the mean to more accurately reflect their distribution (fig. 2).

Code availability

All code was written in the Python programming language and is available at https://github.com/NDeNeuter/TCR_CMV_pred.

Acknowledgements:

We would like to kindly thank Ryan Emerson for making the necessary training data available.

This research was funded by the University of Antwerp [BOF Concerted Research Action (PS ID 30730), Antwerp Study Centre for Infectious Diseases, Methusalem funding], the Hercules Foundation – Belgium and the Research Foundation Flanders (FWO) [Personal PhD grants to NDN (1S29816N)].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest Statement:

We declare that there are no conflicts of interest to disclose.

References

1 Emmadi R, Boonyaratanakornkit JB, Selvarangan R, Shyamala V, Zimmer BL, Williams L *et al.* Molecular methods and platforms for infectious diseases testing: A review of FDA-approved and cleared assays. J. Mol. Diagnostics. 2011; **13**: 583–604.

2 Maurin M. b. Real-time PCR as a diagnostic tool for bacterial diseases. Expert Rev. Mol. Diagn. 2012; **12**: 731–754.

3 Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C *et al.* Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Immunobiology* 2007; **109**: 1–2.

4 Gliddon HD, Herberg JA, Levin M, Kaforou M. Genome-wide host RNA signatures of infectious diseases: discovery and clinical translation. *Immunology* 2017. doi:10.1111/imm.12841.

5 Mahajan P, Kuppermann N, Mejias A, Suarez N, Chaussabel D, Casper TC *et al.* Association of RNA Biosignatures With Bacterial Infections in Febrile Infants Aged 60 Days or Younger. *Jama* 2016; **316**: 846–57.

6 Herberg JA, Kaforou M, Wright VJ, Shailes H, Eleftherohorinou H, Hoggart CJ *et al.* Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children. *Jama* 2016; **316**: 835–845.

7 Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluka G *et al.* Diagnosis of childhood tuberculosis and host RNA expression in Africa. *N Engl J Med* 2014; **370**: 1712–23.

8 Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med* 2015; **7**: 49.

9 Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol* 2017; **17**: 61.

10 De Neuter N, Bittremieux W, Beirnaert C, Cuypers B, Mrzic A, Moris P *et al.* On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* 2017; : 1–10.

11 Skowera A, Ladell K, McLaren JE, Dolton G, Matthews KK, Gostick E *et al.* Beta-Cell-specific CD8 T cell phenotype in type 1 diabetes reflects chronic autoantigen exposure. *Diabetes* 2015; **64**: 916–925.

12 Utz U, Biddison WE, McFarland HF, McFarlin DE, Flerlage M, Martin R. Skewed T cell receptor repertoire in genetically identical twins correlates with multiple sclerosis. *Nature* 1993; **364**: 243–247.

13 Lossius A, Johansen JN, Vartdal F, Robins H, Šaltyte BJ, Holmøy T *et al.* High-throughput sequencing of TCR repertoires in multiple sclerosis reveals intrathecal enrichment of EBV-reactive CD8+ T cells. *Eur J Immunol* 2014; **44**: 3439–3452.

14 Yang J, Sheng G, Xiao D, Shi H, Wu W, Lu H *et al.* The frequency and skewed T-cell receptor beta-chain variable patterns of peripheral CD4(+)CD25(+) regulatory T-cells are associated with hepatitis B e antigen seroconversion of chronic hepatitis B patients during antiviral treatment. *Cell Mol Immunol* 2016; **13**: 678–87.

15 Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 2017; : 1–10.

16 Jackson SE, Sedikides GX, Mason GM, Okecha G, Wills MR. Human Cytomegalovirus (HCMV)-Specific CD4 $^{+}$ T Cells Are Polyfunctional and Can Respond to HCMV-Infected Dendritic Cells *In Vitro*. *J Virol* 2017; **91**: e02128-16.

17 Pera A, Vasudev A, Tan C, Kared H, Solana R, Larbi A. CMV induces expansion of highly polyfunctional CD4 + T cell subset coexpressing CD57 and CD154. *J Leukoc Biol* 2017; **101**: 555–566.

18 Ogunjimi B, Van den Bergh J, Meysman P, Heynderickx S, Bergs K, Jansen H *et al.* Multidisciplinary study of the secondary immune response in grandparents re-exposed to chickenpox. *Sci Rep* 2017; **7**: 1077.

19 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011; **12**: 2825–2830.

Figures

Figure 1: Total number of CMV associated TCR sequences (y axis) versus the size of the distinct, productive $CD4^+$ memory T cell repertoire (x axis) for all individuals in the US cohort (A) and Belgian cohort (B). Individuals were labelled according to their CMV serostatus with either a blue circle (CMV seropositive) or a green cross (CMV seronegative). CMV seropositive individuals typically had a higher number of CMV associated TCR sequences within their repertoire than CMV seronegative individuals. The plots essentially represent the feature space that the classifier was trained on, as the total number of CMV associated TCR sequences and the size of the distinct, productive $CD4^+$ memory T cell repertoire are the only features used to discriminate between CMV seronegative and CMV seropositive individuals. Visual inspection reveals a distinction in number of CMV associated TCR sequences between the two classes. Though a smaller number of individuals was assessed in the Belgian cohort, a similar separation of CMV seropositive and seronegative individuals can be observed as in the US cohort.

Figure 2: (A) Median ROC curve and confidence intervals (50%, 80% and 95%) for the beta binomial model trained by Emerson and colleagues and tested on the Belgian cohort. ROC curves were bootstrapped after bootstrapping individuals from the Belgian cohort. As the ROC curve is obtained by plotting the true positive rate versus the false positive rate, the ideal curve would lie in the top left corner of the plot, indicating a perfect true positive rate that is independent of the false positive rate (AUC value = 1). The dashed diagonal line indicates an equal increase in both the true positive and false positive rate, indicating random performance (AUC value = 0.5). The classifier was validated on a cohort of 33 healthy volunteers and shows an AUC value of 0.915, indicating better performance than an out of the box random forest classifier. (B) Histogram of 10 000 bootstrapped AUC values obtained on the Belgian cohort.

Figure 3: (A) Median ROC curve and confidence intervals (50%, 80% and 95%) for the random forest classifier trained on TCR data from the 641 US individuals and tested on the Belgian cohort. ROC curves were obtained after bootstrapping individuals from the Belgian cohort. As the ROC curve is obtained by plotting the true positive rate versus the false positive rate, the ideal curve would

lie in the top left corner of the plot, indicating a perfect true positive rate that is independent of the false positive rate (AUC value = 1). The dashed diagonal line indicates an equal increase in both the true positive and false positive rate, indicating random performance (AUC value = 0.5). The classifier was validated on a cohort of 33 healthy volunteers and shows an AUC value of 0.91, indicating excellent performance. (B) Histogram of 10 000 bootstrapped AUC values obtained on the Belgian cohort.