Bioinformatics, 37(21), 2021, 3938–3940
doi: 10.1093/bioinformatics/btab573
Advance Access Publication Date: 6 September 2021
Applications Note

Sequence analysis

TCRpair: prediction of functional pairing between HLA-A*02:01-restricted T-cell receptor α and β chains

Anja Mösch[1,2] **and Dmitrij Frishman**[1,3,*]

[1]Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technical University of Munich, Freising 85354, Germany, [2]Medigene Immunotherapies GmbH, A Subsidiary of Medigene AG, Planegg 82152, Germany and [3]Department of Bioinformatics, Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg 195251, Russia

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 27, 2020; revised on May 21, 2021; editorial decision on August 3, 2021

Abstract

Summary: The ability of a T cell to recognize foreign peptides is defined by a single α and a single β hypervariable complementarity determining region (CDR3), which together form the T-cell receptor (TCR) heterodimer. In ~30–35% of T cells, two α chains are expressed at the mRNA level but only one α chain is part of the functional TCR. This effect can also be observed for β chains, although it is less common. The identification of functional α/β chain pairs is instrumental in high-throughput characterization of therapeutic TCRs. TCRpair is the first method that predicts whether an α and β chain pair forms a functional, HLA-A*02:01 specific TCR without requiring the sequence of a recognized peptide. By taking additional amino acids flanking the CDR3 regions into account, TCRpair achieves an AUC of 0.71.

Availability and implementation: TCRpair is implemented in Python using TensorFlow 2.0 and is freely available at https://www.github.com/amoesch/TCRpair.

Contact: d.frishman@wzw.tum.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

T cells are a key element of the adaptive immune system because they can detect infected or aberrant cells through their receptors T-cell receptor (TCR). The TCR is a heterodimer composed of one α and one β chain. Each chain contains a hypervariable complementarity determining region (CDR3), which interacts with a peptide bound to the human leukocyte antigen (HLA), the human version of the major histocompatibility complex, expressed on the surface of an antigen presenting cell. Since the CDR3α and CDR3β regions are highly variable due to the V(D)J recombination, peptide recognition is very specific and each TCR only binds to one or just a few peptides presented by an HLA allele (Hughes *et al.*, 2003; Lu *et al.*, 2019). The peptide specificity is controlled by the process of thymic selection, which only allows T cells that do not recognize peptides of the healthy peptide repertoire to circulate in the body. Most of the positively selected T cells express a single unique TCR on the cell surface, for which only one transcript of an α and one of a β chain is present. However, it has been shown that ~30–35% of T cells express two α chains on the mRNA level and some T cells also express two β chains, although their number is significantly lower due to transcriptional allelic exclusion and other mechanisms (Dupic *et al.*, 2019; Redmond *et al.*, 2016; Schuldt and Binstadt, 2019; Stubbington *et al.*, 2016). If two α or two β chains can be detected by RNA sequencing of clones or single cells, two surface TCRs might be present but more often only one of the two chains from the same locus is part of the functional TCR (Schuldt and Binstadt, 2019). Identifying the functional α/β TCR combination is crucial for the assessment of suitable TCRs for cancer immunotherapy (Parkhurst *et al.*, 2017; Shitaoka *et al.*, 2018). Current methods to identify α/β pairing require specific experimental setups and are more geared toward the identification of α/β chain pairs in T-cell repertoires (Egorov *et al.*, 2015; Holec *et al.*, 2019; Howie *et al.*, 2015; Lee *et al.*, 2017). Here, we present TCRpair, a deep learning algorithm to predict functional pairs of α/β TCRs recognizing HLA-A*02:01 restricted peptides. TCRs are reconstructed from the CDR3 sequence and the V/J gene annotation, which represents the minimum annotation of a TCR in publicly available databases (Bagaev *et al.*, 2020; Dhanda *et al.*, 2019; Shugay *et al.*, 2018; Vita *et al.*, 2019). TCRpair can be instrumental in speeding up TCR sequence verification if RNA sequencing data does not yield unequivocal results. Additionally, TCRpair supports input from MiXCR (Bolotin *et al.*, 2015), including filtering for possible α/β combinations by clonotype frequency.

2 Materials and methods

Pairs of CDR3α and CDR3β sequences with their respective V and J allele annotation as well as information on the recognized peptide and the HLA-A allele were downloaded from IEDB (Dhanda *et al.*, 2019; Vita *et al.*, 2019) and VDJdb (Bagaev *et al.*, 2020; Shugay *et al.*, 2018), which predominantly consists of single cell sequencing data. In total we obtained 21 715 unique TCRs, of which 3250 HLA-A*02:01 restricted TCRs were used for model training/testing and validation (Supplementary Table S1). A negative dataset ($n = 2209$) was generated by randomly combining CDR3α and CDR3β chains and then selecting only those chain pairs, for which the CDR3α chain originates from a TCR recognizing a different peptide as the TCR from which the CDR3β chain originates (Supplementary Fig. S1). For each TCR, the full TCR sequence was reconstructed by aligning CDR3 sequences to the sequences of their respective V and J alleles from the IMGT/LIGM database (Giudicelli, 2006). Nine different sequence types were used as model inputs: CDR3 region only, CDR3 region with 3, 5, 7, 9, 11, 13 or 15 flanking amino acids and the full TCR sequence (Fig. 1A). For each TCR, α and β chain sequences were concatenated to be used as single sequence input and BLOSUM62 encoded (Henikoff and Henikoff, 1992; Nielsen *et al.*, 2003) (Fig. 1B and Supplementary Information S1). The dataset was randomly split into 80% training and 20% validation data. For each input type, a model was trained for 20 epochs with batch size 50 using the Adam optimization algorithm (Shao *et al.*, 2020).

An independent dataset of 11 HLA-A*02:01-restricted TCRs with two α or two β chains detected at the RNA level was used to test whether TCRpair can identify the functional chain by comparing likelihood scores. RNA sequencing data of T-cell clones was processed by MiXCR (Bolotin *et al.*, 2015), whereas for 10 clones two CDR3α chains and for 1 clone two CDR3β chains showed a clone fraction of at least 0.35. The functional chain for each TCR was experimentally identified by expressing the β chain in combination with both α chains (or in one case the α chain in combination with both β chains) and comparing their cytotoxicity *in vitro* by coculturing with peptide-presenting cells. All 11 TCRs recognize peptides for which no TCRs are present in the training data. The differences between validation dataset and independent dataset are described in Supplementary Information S2.

3 Results and discussion

TCRpair can predict whether a pair of α and β chains has the tendency to form a functional TCR and assists with the identification of the chain that is part of the functional TCR if two α or two β chains are detected at the RNA level for HLA-A*02:01 restricted TCRs (Supplementary Fig. S2). The models using flanking amino acid sequences performed better than models using only the CDR3 sequence or the full TCR sequence, which includes CDR1 and CDR2 sequences that show a limited diversity compared to the highly variable peptide binding CDR3 sequence (Arden, 1998). On the validation dataset, the models with 5 and 7 flanking amino acids both achieved an area under the receiver operating characteristic curve (AUC) of 0.71 and an average precision of 0.80 (Fig. 1C). The model with 7 flanking amino acids correctly identified 7 out of 11 TCRs from the independent dataset (Fig. 1D and Supplementary Table S2). Both models with 9 and 11 flanking amino acids performed comparably well. All these four models showed improved prediction

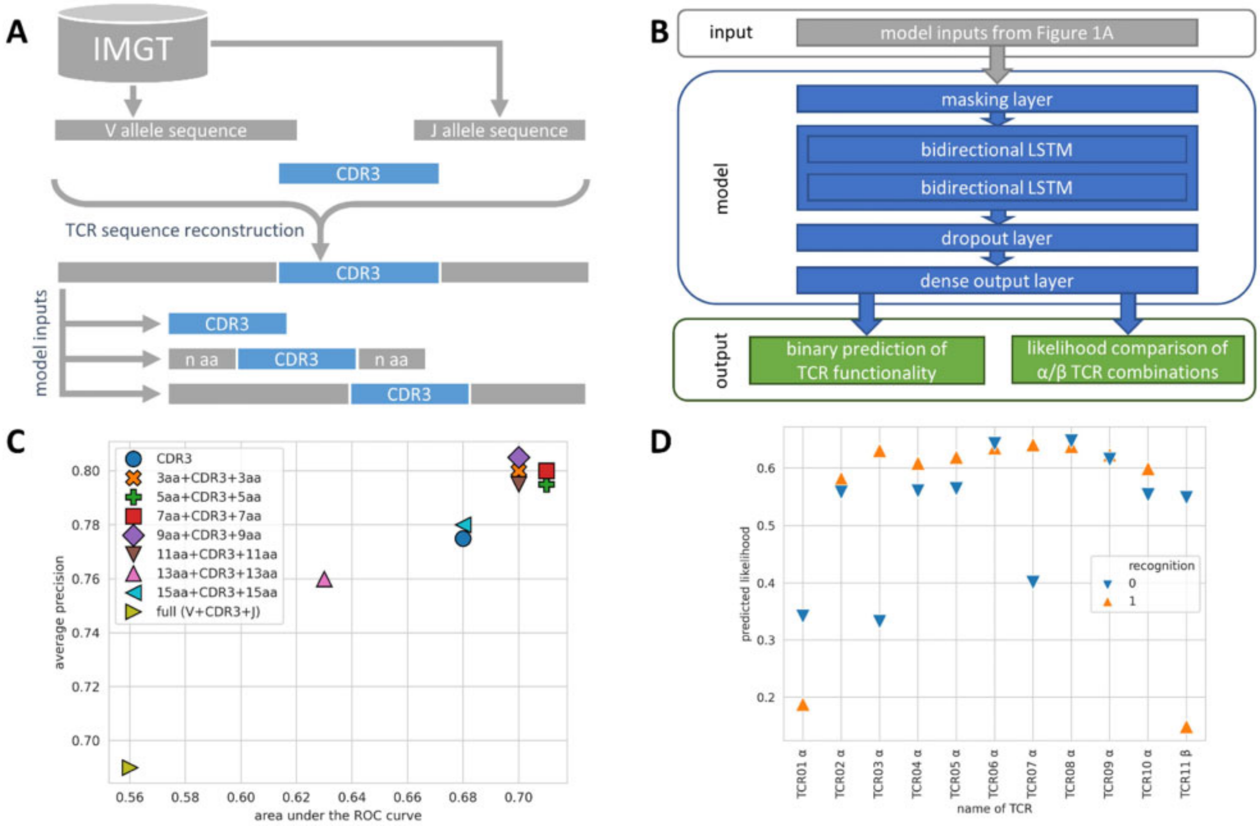

Fig. 1. Workflow and performance of TCRpair. (**A**) TCR reconstruction with IMGT V and J allele sequences and varying model inputs: CDR3 regions only, CDR3 regions with *n* flanking amino acids, where *n* is 3, 5, 7, 9, 11, 13 or 15, as well as full TCR sequences. (**B**) TCRpair model workflow with input as described in A, model with layers and two possible types of output. (**C**) Area under the ROC curve and average precision performance of models using different input sequences (CDR3 only, CDR3 with flanking sequences and full TCR sequence) on the validation dataset. (**D**) Comparison of the likelihood scores for the model using as input CDR3 regions with flanking sequences of length 7 on the independent dataset. For each TCR, likelihood scores of α/β chain pairs to be functional are represented by orange triangles facing up while scores of nonfunctional α/β chain pairs are represented by blue triangles pointing down

performance compared to the model using only the CDR3 sequences and the model using the full TCR sequence. These observations also hold true when comparing AUCs for individual peptides (Supplementary Table S3). Furthermore, we observed higher differences between the likelihood scores of real and perturbed amino acid input vectors for regions with a higher amino acid variation such as the V region compared to more conserved positions such as the first two positions of the CDR3 regions (Supplementary Information S3 and Table S5) (Yu *et al.*, 2019). These results demonstrate that TCRpair learned to identify the features of the TCR's α and β chain sequences, which ultimately determine functional pairing and thus TCR specificity, without the need to know the sequence of the recognized peptide. TCRpair performs comparably to NetTCR 2.0 (https://services.healthtech.dtu.dk/service.php?NetTCR-2.0; Jurtz *et al.*, 2018), which in contrast requires one of three possible peptides as additional input (Supplementary Table S4). Additionally, TCRpair demonstrates that sequence context can improve performance for sequence-based machine learning algorithms using LSTM layers, which might apply to similar prediction problems.

The current version of TCRpair is limited to the TCRs recognizing peptides presented by HLA-A*02:01, which is the most common allele in Caucasian populations (Gonzalez-Galarza *et al.*, 2015). It does not work for other HLA restrictions (see Supplementary Table S2) or naïve T-cell repertoires (see Supplementary Information S1), for which frequency-based methods relying on the distribution of T-cell clones over multiple samples of the same repertoire are more suitable (Holec *et al.*, 2019; Howie *et al.*, 2015; Lee *et al.*, 2017). However, the growing amount and quality of TCR sequencing data especially from single cells will allow the addition of further HLA alleles and the training of a general HLA-independent model in the future.

Acknowledgements

The authors wish to thank Dr Silke Raffegerst for carefully reading the manuscript and many useful comments.

Financial Support: none declared.

Conflict of Interest: A.M. is an employee of Medigene Immunotherapies GmbH, a subsidiary of Medigene AG, Planegg, Germany.

References

Arden,B. (1998) Conserved motifs in T-cell receptor CDR1 and CDR2: implications for ligand and CD8 co-receptor binding. *Curr. Opin. Immunol.*, **10**, 74–81.

Bagaev,D.V. *et al.* (2020) VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Res.*, **48**, D1057–D1062.

Bolotin,D.A. *et al.* (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, **12**, 380–381.

Dhanda,S.K. *et al.* (2019) IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res.*, **47**, W502–W506.

Dupic,T. *et al.* (2019) Genesis of the $\alpha\beta$ T-cell receptor. *PLoS Comput. Biol.*, **15**, e1006874.

Egorov,E.S. *et al.* (2015) Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J. Immunol.*, **194**, 6155–6163.

Giudicelli,V. *et al.* (2006) IMGT/LIGM-DB, the IMGT(R) comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.*, **34**, D781–D784.

Gonzalez-Galarza,F.F. *et al.* (2015) Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.*, **43**, D784–D788.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

Holec,P.V. *et al.* (2019) A Bayesian framework for high-throughput T cell receptor pairing. *Bioinformatics*, **35**, 1318–1325.

Howie,B. *et al.* (2015) High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.*, **7**, 301ra131.

Hughes,M.M. *et al.* (2003) T cell receptor CDR3 loop length repertoire is determined primarily by features of the V(D)J recombination reaction. *Eur. J. Immunol.*, **33**, 1568–1575.

Jurtz,V.I. *et al.* (2018) NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv*, doi:10.1101/433706

Lee,E.S. *et al.* (2017) Identifying T cell receptors from high-throughput sequencing: dealing with promiscuity in TCRα and TCRβ pairing. *PLoS Comput. Biol.*, **13**, e1005313.

Lu,J. *et al.* (2019) Molecular constraints on CDR3 for thymic selection of MHC-restricted TCRs from a random pre-selection repertoire. *Nat. Commun.*, **10**, 1019.

Nielsen,M. *et al.* (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.*, **12**, 1007–1017.

Parkhurst,M. *et al.* (2017) Isolation of T-cell receptors specifically reactive with mutated tumor-associated antigens from tumor-infiltrating lymphocytes based on CD137 expression. *Clin. Cancer Res.*, **23**, 2491–2505.

Redmond,D. *et al.* (2016) Single-cell TCRseq: paired recovery of entire T-cell alpha and beta chain transcripts in T-cell receptors from single-cell RNAseq. *Genome Med.*, **8**, 80.

Schuldt,N.J. and Binstadt,B.A. (2019) Dual TCR T cells: identity crisis or multitaskers? *J. Immunol.*, **202**, 637–644.

Shao,X.M. *et al.* (2020) High-throughput prediction of MHC class I and class II neoantigens with MHCnuggets. *Cancer Immunol. Res.*, **8**, 396–408.

Shitaoka,K. *et al.* (2018) Identification of tumoricidal TCRs from tumor-infiltrating lymphocytes by single-cell analysis. *Cancer Immunol. Res.*, **6**, 378–388.

Shugay,M. *et al.* (2018) VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.*, **46**, D419–D427.

Stubbington,M.J.T. *et al.* (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods*, **13**, 329–332.

Vita,R. *et al.* (2019) The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.*, **47**, D339–D343.

Yu,K. *et al.* (2019) Comparative analysis of CDR3 regions in paired human $\alpha\beta$ CD 8 T cells. *FEBS Open Bio*, **9**, 1450–1459.