Clinical Immunology 222 (2021) 108621

Contents lists available at ScienceDirect

Clinical Immunology

journal homepage: www.elsevier.com/locate/yclim

Full Length Article

T cell receptor repertoire as a potential diagnostic marker for celiac disease

Ying Yao [a,b], Asima Zia [a,1], Ralf Stefan Neumann [a,b], Milena Pavlovic [c], Gabriel Balaban [c], Knut E. A. Lundin [b,2], Geir Kjetil Sandve [b,c], Shuo-Wang Qiao [a,b,*]

[a] Department of Immunology, Institute of Clinical Medicine, University of Oslo, Norway
[b] K.G. Jebsen Coeliac Disease Research Centre, University of Oslo, Norway
[c] Department of Informatics, University of Oslo, Norway

ARTICLE INFO

Keywords:
Celiac disease
CD4+ T cells
TCR repertoire
Disease inference
High-throughput sequencing

ABSTRACT

An individual's T cell repertoire is skewed towards some specificities as a result of past antigen exposure and subsequent clonal expansion. Identifying T cell receptor signatures associated with a disease is challenging due to the overall complexity of antigens and polymorphic HLA allotypes. In celiac disease, the antigen epitopes are well characterised and the specific HLA-DQ2-restricted T-cell repertoire associated with the disease has been explored in depth. By investigating T cell receptor repertoires of unsorted lamina propria T cells from 15 individuals, we provide the first proof-of-concept study showing that it could be possible to infer disease state by matching against a priori known disease-associated T cell receptor sequences.

1. Introduction

T cell plays a central role in cell mediated immune response. The T cell receptor (TCR) is responsible for recognizing antigenic peptides bound to major histocompatibility complex (MHC) molecules. In 95% of human T cells, the TCR heterodimer consists of the α (TCRα) and β chain (TCRβ). During T-cell development, each thymocyte generates its unique TCR variant through recombination of different V, D, J gene segments and random deletion and/or insertion of nucleotides at the junctions. This results in a highly diverse TCR repertoire and the diversity is important for maximizing potential coverage of the protective immunity. Structural studies of TCRs binding to pMHC ligands [1] have shown that although there are some exceptions, as a rule, the variable regions of the TCRα chain largely contact the MHC molecule whereas the TCRβ chain makes most contact with the antigenic peptide. This notion is supported by genetic studies where the usage of V-gene segment of the TCRα is more closely associated with the human leukocyte antigen (HLA) profiles of the person [2].

In cell mediated immune response, naïve T cells are activated and clonally expand after recognition of foreign antigenic peptides presented by MHC. Thus, although the diversity is highest in the naive compartment, in the memory compartment it is skewed towards certain specificities as a result of past antigen exposure and subsequent antigen-driven selection and expansion. Since T cells directed against certain antigen in a disease setting are clonally expanded, a biased repertoire should be observed given enough sequencing power. With the advancement of high-throughput immune receptor sequencing methods, TCR repertoire has the potential to be a diagnostic marker for infections or autoimmune diseases. However, due to the complexity and diversity of individual TCR repertoires, identifying the TCR signatures associated with an antigen is challenging. Somma et al. [3] identified a number of TCRβ clonotypes implicated in the pathogenesis of multiple sclerosis, which were clonally expanded in both the healthy and the affected twin. In contrast, studies on monozygotic twins [4] have demonstrated that different disease settings altered the TCR gene usage and TCR repertoire as a whole. Despite TCR complexities, TCR clonotyping has been used as diagnostic tool in Emerson et al. [5] where the exposure to cytomegalovirus (CMV) of 666 subjects could be inferred by their TCR repertoires. The TCR repertoire data was generated from peripheral blood, since CMV elicits a particularly strong immune response where an unusually

Abbreviations: CD, celiac disease; TCR, T cell receptor; TCRα, T cell receptor α chain; TCRβ, T cell receptor β chain; MHC, major histocompatibility complex; pMHC, peptide:MHC complex; HLA, human leukocyte antigen; UMI, unique molecular identifier; CMV, cytomegalovirus; ROC, Receiver Operating Characteristics; AUC, area under the ROC curve.

* Corresponding author at: Department of Immunology, Institute of Clinical Medicine, University of Oslo, Norway.
E-mail address: s.w.qiao@medisin.uio.no (S.-W. Qiao).
[1] Current address: Living Systems Laboratory, King Abdullah University of Science and Technology, Saudi Arabia
[2] Present address: Department of Gastroenterology, Oslo University Hospital, Oslo, Norway.

https://doi.org/10.1016/j.clim.2020.108621
Received 4 June 2020; Received in revised form 27 October 2020; Accepted 7 November 2020
Available online 13 November 2020

large proportion of the T-cell response in blood is CMV-related. However, the T-cell response is less pronounced for most other diseases. In the CD4 compartment, only 1 to 100 per million CD4 T cells in blood would be expected to be specific to a given pMHC [6,7] whereas in affected tissue the frequency of antigen-specific TCR would be expected to be around 1 to 5 per hundred CD4 T cells, at least in celiac disease [8]. Therefore, it is advantageous to look at the T-cell response in the affected tissue where the frequency of disease-relevant T cells is much higher than in blood.

Celiac disease (CD) is a chronic HLA-associated inflammatory disorder that primarily affects the small intestine. The disease shows strong genetic association with HLA class II alleles encoding HLA-DQ2.5 (*HLA-DQA1*05/HLA-DQB1*02*, expressed by 90% of patients), HLA-DQ8 (*HLA-DQA1*03/HLA-DQB1*03:02*), and HLA-DQ2.2 (*HLA-DQA1*02:01/HLA-DQB1*02*) [9–11]. The epitopes of the causative antigen gluten are well defined and gluten-specific CD4 T cells that are only found in the small intestine of celiac disease patients, but not in healthy controls, have been isolated and extensively studied. All gluten-reactive T cells in the lesions are restricted by the disease-associated HLA-DQ2.5 molecule in HLA-DQ2.5-positive subjects [12]. HLA-DQ-gluten tetramers carrying the immunodominant gluten epitopes have been used to visualize gluten-specific T cells directly from blood or small intestinal tissue [13]. TCR sequencing studies have shown that public features, i.e. identical TCRα, TCRβ, or paired TCRαβ amino acid sequences found in different individuals, are frequently observed among gluten-specific T cells [13].

To explore whether disease state could be assessed from a limited number of tissue-derived cells, we started with around 10,000 T cells taken from the lamina propria of duodenal biopsies to assess the celiac disease state. In order to find the best diagnostic biomarkers, we evaluated the usage of different types of prior information, i.e. all gluten-specific TCRs versus a smaller subset of public gluten-specific TCRs shared across multiple CD patients. This is a proof of principle study and the aim is to show the potential of using TCR-based diagnostics. This is the first step towards biopsy-free diagnostics of CD where TCR information would be collected directly from blood.

2. Materials and methods

2.1. Sample collection

The project was approved by the Regional Committee for Medical and Health Research Ethics South-East Norway (REK 2010/2720) and signed informed consent forms were obtained from all subjects. From each donor, two pieces of duodenal biopsies were collected in ice-cold RPMI-1640. The epithelial layer that largely contain CD8+ intra-epithelial T cells was removed with two 5-min incubation with PBS + 2%FCS + 2 mM EDTA at 37C. After thorough washes with PBS to remove detached epithelial cells, the remaining lamina propria tissue was digested for 45 min with 1 mg/ml Collagenase (Sigma) and 0.1 mg/ml DNAse (Sigma). The resulting lamina propria single-cell suspension was counted and seeded directly in TCL lysis buffer (Qiagen) in four concentrations (108,000, 36,000; 18,000 and 9000 cells per well) and eight biological replicates for each concentration. After thorough mixing to aid cell lysis, the lysates were kept frozen at —70C until processed. After defrosting, the cell lysate in TCL was transferred to 96-well plates precoated with dT-oligos in the TurboCapture 96 mRNA kit (Qiagen). mRNA extraction and cDNA synthesis using the plate-immobolized oligo-dT was carried out in accordance with the manufacturer's instructions with the modification of additional template switch oligo (Bio-d(AAGCAGTGGTATCAACGCAGAGTAGTNNNNNN)-r(GGG),
where N denotes random nucleotides that serve as unique molecular identifier (UMI)). Following cDNA synthesis, two semi-nested TCRα- and TCRβ-specific PCR reactions were carried out as in [13].

2.2. TCR sequencing and data processing

Double indexing was applied in library preparation, thereby every pair of reads had an index composed of two barcodes on the forward and reverse read respectively, encoding the sample origin. Libraries were sequenced on the Illumina MiSeq platform with 250 nt pair-end sequencing at the Norwegian Sequencing Center (Oslo University Hospital).

All paired end reads were de-multiplexed based on the combination of their R1 and R2 barcode sequences. Both of the paired R1 and R2 reads were dropped if any of them had any nucleotide mismatch with the reference barcodes. On the paired reads assigned to each sample, we performed UMI tag extraction and UMI-guided assembly using the MIGEC pipeline [14], where all reads in a sample were grouped by their UMI and then each group with larger than 10 reads were assembled to generate a consensus sequence by multiple alignment. Both consensuses need to be successfully assembled for paired reads, otherwise the pair was dropped. Considering the relatively short UMI length and large expected number of cells in some wells in the study, the probability for a pair of similar UMI caused by sequencing error was relatively low. We have therefore not corrected sequencing errors in the UMI. The consensus sequences of samples from the same patient were then pooled and aligned with mismatches, inserts and deletions to the TCR database following the MiXCR pipeline [15], thereby TCRαβ chain and CDR3 repertoires were extracted from the assembled consensus sequences. Identical sequences were grouped in clonotypes, and the corresponding clonecounts were recorded. Consensus with poor quality were also collected and mapped to the grouped clonotypes for correction of PCR and sequencing errors. The default parameters of MiXCR were applied throughout this process.

2.3. Reference database of gluten-specific TCR sequences

To search for the presence of disease-associated TCR sequences in our data, we used a reference database comprised of TCRα- and TCRβ-clonotypes obtained from single-cell TCR sequencing of HLA-DQ2.5-gluten-tetramer-sorted cells from 59 celiac disease patients (manuscript in preparation). Sequences belonging to donors in the present study were excluded. Overall, there were 2929 TCRα- and 2662 TCRβ-clonotypes originating from 6808 HLA-DQ:gluten-tetramer-sorted cells. A clonotype is defined throughout the study as a unique amino acid sequence of the re-arranged variable regions of the TCRα (VJ) or TCRβ (VDJ). Within this large reference database that includes almost all known gluten-specific TCR clonotypes to date, there is a smaller subset that consists of public clonotypes, defined as identical amino acid sequences observed in at least two CD patients. This public TCR subset contains 151 TCRα and 226 TCRβ clonotypes that have been collapsed from 1150 TCRα sequences and 1436 TCRβ sequences from 2003 gluten-specific T cells. The collapse of TCR sequences to clonotypes was caused by both in vivo clonal expansion (multiple cells expressing identical TCRαβ sequences in the same patient) and convergent recombination (different nucleotide sequences encoding identical amino acid sequence).

2.4. Inferring disease state

We used logistic regression to infer the CD status of the donors based on the presence of the aforementioned antigen-specific TCR sequences. Logistic regression was performed by sklearn.linear model.LogisticRegression function from scikit-learn v0.20.4 Python module [16], where either normalized unique match or normalized clonecount match was used a single predictor. All the other parameters were set as default except for the C (inverse of regularization strength) that was set at 1E+5 to eliminate the effect of penalty term since no simplified model was preferable with a single predictor. We also employed the R package ROCR1.0–7 to calculate the sensitivity and specificity while using the

same single predictor ranged from 0 to the maximum in different experimental settings, as well as the corresponding area under the curve (AUCs). Test for association between the prevalence of a clonotype and its frequency in our data using Kendall's tau was done with R package stats 3.4.4.

3. Result

3.1. Data acquisition

We collected intestinal biopsies from seven HLA-DQ2.5 (*HLA-DQA1*05/HLA-DQB1*02*) positive untreated celiac disease patients and eight non-celiac controls or HLA-DQ2-negative patients. Since TCR recognize peptide-HLA complexes, for the purpose of this study where we look for signature sequences of gluten:HLA-DQ2-reactive TCRs, we do not expect to find these TCR sequences in HLA-DQ2-negative patients whose gluten-reactive T cells are HLA-DQ8-restricted. We sampled one million cells from the lamina propria of two duodenal biopsies from each of the 15 donors and sequenced the rearranged TCRα and TCRβ variable region. Flow cytometric analysis showed that approximately 1% of the sampled unsorted lamina propria cells were T cell, of which >80% were CD4 T cells. Thus, we have sampled and sequenced approximately 10,000 T cells from each donor. The number of sequencing reads generated from each donor varied from 0.1 million to 2.7 million, with an average of 1.7 million, representing on average 5821 TCR mRNA molecules per donor after deduplication. The number of unique clonotypes we observed in each donor ranged from 861 to 8778. Basic information of the donors and the libraries were summarised in Table 1.

3.2. TCR clonotypes in our data matched preferentially public gluten-specific TCRs

By collapsing TCRs with the same V gene, J gene and CDR3 amino acid sequences from repertoires from all donors, we had in total 17,261 unique TCRα and 26,820 unique TCRβ clonotypes in our dataset (Fig. 1). To find an optimal set of disease-associated TCR clonotypes for inferring disease state, we employed a reference database consisting of data from multiple single-cell TCR sequencing projects where HLA-DQ2.5:gluten tetramers were used to stain T cells from 59 CD patients, followed by sorting and sequencing of the sorted gluten-specific TCRs (manuscript in preparation). Among the total 5591 gluten-specific TCRα and TCRβ amino acid sequences in the reference database, 377 of them were observed in at least two CD patients and were thus defined as public

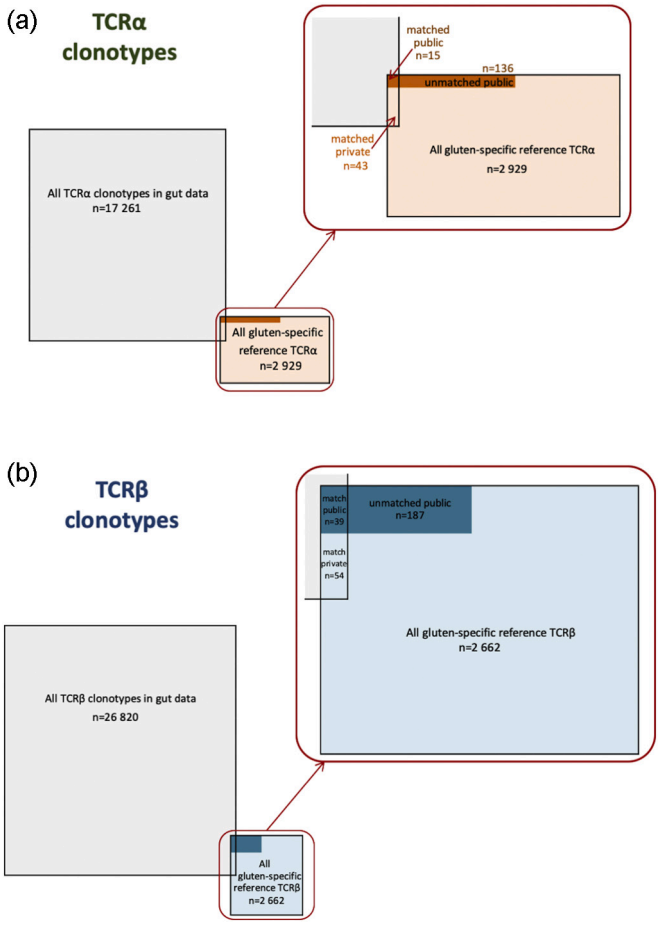

Fig. 1. Number of unique TCRα (A) and TCRβ clonotypes (B) that matched either public or non-public disease associated TCR sequences. The sizes of the boxes and shaded areas are scaled to reflect the number of clonotypes.

clonotypes. When we compared our dataset from the gut tissue with this reference database, we found that 93 TCRβ clonotypes in our data matched gluten-specific TCR sequences in the reference database, of which 39 matched the subset of public gluten-specific TCRβ sequences. While 58 of the TCRα clonotypes in our dataset matched gluten-specific

Table 1
Basic information of the donors and the libraries.

Subject ID	Age group	HLA	Histology (Marsh)	Serology (IgA-TG2)	CD status	Group	Reads	cDNA molecules (TCRα)	cDNA molecules (TCRβ)	Clonotypes (TCRα)	Clonotypes (TCRβ)
CD1357	50–54	DQ2	3a	n.a.	UCD	UCD	1966714	7788	8701	1721	2563
CD1358	35–39	DQ2	3c	93	UCD	UCD	1783764	3857	6945	1866	3249
CD1364	20–24	DQ2	3b	10	UCD	UCD	2727644	6913	9956	3218	4726
CD1368	40–44	DQ2	3c	66	UCD	UCD	2519295	6293	13086	2938	5840
CD1370	50–54	DQ8	1	<1	control	control	838353	2206	5729	692	1757
CD1386	30–34	DQ2	0	<1	control	control	1696311	4302	4548	677	1033
CD1390	18–19	DQ8	3c	40	UCD	control	106758	567	1236	318	657
CD1393	25–29	DQ2	3b	9	UCD	UCD	2269730	9746	6693	2041	1999
CD1408	25–29	n.a	0	n.a.	control	control	2001839	8792	4298	2051	1615
CD1409	30–34	DQ2	0	<1	control	control	1760049	7526	4002	1341	1071
CD1422	30–34	DQ2	3a	6	UCD	UCD	1901639	3881	2998	500	532
CD1428	20–14	DQ2	0	<1	control	control	506210	924	1877	270	591
CD1450	35–39	DQ2	0	<1	control	control	1559849	2158	4035	471	1468
CD1451	65–69	DQ2	3c	70	UCD	UCD	1783685	2884	2673	461	842
CD1453	30–34	DQ8	0	5	Potential*	control	1880982	4225	3728	808	1148

DQ2: HLA-DQA1*05/HLA-DQB1*02.
DQ8: HLA-DQA1*03/HLA-DQB1*03:02.
n.a.: not available.
* Potential CD is defined as positive seology but normal histology.

TCRα sequences in the reference database, 15 out of these matched the public TCRα sequences (Fig. 1). Since the public clonotypes account for 5% and 8% of the total TCRα and TCRβ reference dataset, respectively, it is interesting to note that among the matches we found in our gut-derived TCR data, 26% and 42% of them were matched to the public TCRα and TCRβ sequences, respectively. This finding indicates that our TCR data acquired from unsorted gut samples preferentially matched public gluten-specific clonotypes shared between two or more CD patients.

From published studies of gluten-specific TCR sequences, it is known that some TCR clonotypes such as the TRBV7-2/TRBJ2-3 clonotype with the CDR3 amino acid sequence ASSxRxTDTQY (x denotes any amino acid residue) are found in virtually all CD subjects [13,17]. On the other hand, many public CD clonotypes were found in only two subjects of total 59 CD patients from whom the reference database was made. We hypothesized that highly public clonotypes found in many individuals in the reference database were more likely to be observed in our test dataset derived from unsorted T cells from the gut. For all TCR clonotypes from CD patients that matched the reference gluten-specific TCR database, we calculated the Kendall's rank correlation (Supplementary Table 1). Result of the test showed that the frequency of a given TCR in our data was indeed positively correlated with the number of patients that shared this same TCR in the gluten-specific reference database, with tau of 0.338 and *P* value of 0.0026. Therefore, we found that the most publicly used clonotypes found in many CD patients were also more frequently observed in our gut repertoire dataset.

3.3. Matching TCRβ alone was sufficient for predicting celiac disease state

We next set out to explore whether the CD status could be inferred based on the number of TCR clonotypes in our dataset that matched the reference database, either the complete gluten-specific database or the subset of public clonotypes. We counted the matches in two different ways, either counting only the number of unique TCR clonotypes that matched, or counted the number of times we found matching clonotypes by taking into account the clonal expansion. Thus, for each TCR-repertoire of the 15 donors, we summed up the number of unique TCR clonotypes that matched the gluten-specific reference TCR database, this is referred to as 'sum unique disease associated TCRs'. Since clonal expansion is a feature associated with gluten-specific T cells in earlier studies, we took into account the clone size of each matched sequence, measured by the 'clonecount'. Therefore, we also calculated the sum of the clonecount of the same matched TCR sequences. In order to adjust for the variable sequencing depth and the variable number of total TCR sequences retrieved from each donor, we normalized the sum unique disease associated TCRs by dividing it with the total number of unique clonotypes found in that individual. We name this normalized output as 'unique match'. Similarly, we calculated for each individual the 'clonecount match' where the total clonecounts of all matched sequences were divided by the total number of clonecounts in the repertoire (Fig. 2, Supplementary Table 2).

The unique match and clonecount match were then used as a predictor in a logistic regression model to infer the status of celiac disease. We performed the analysis by using all TCR sequences or by using only TCRβ repertoires. We chose not to use TCRα alone since the TCRα

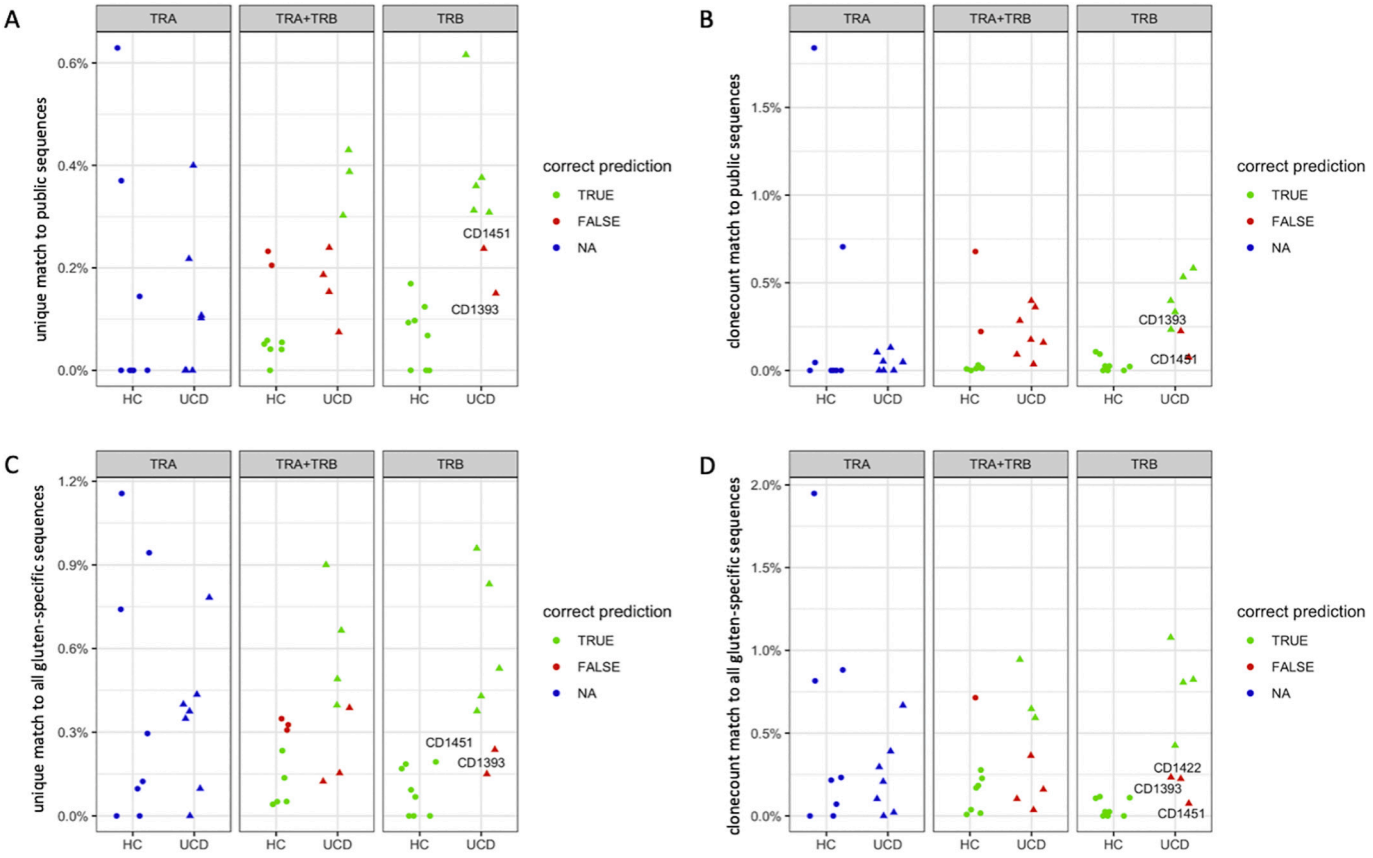

Fig. 2. Normalized number of TCR clonotypes of each donor that matched to reference database. The donors were grouped by disease status. The donors on the left side of each frame were controls, while donors on the right were untreated HLA-DQ2.5-positive celiac disease patients (UCD). Colours indicate if a donor was correctly predicted by logistic regression models trained on the others. (A) Using unique match and the public sequences as reference (B) Using clonecount match and the public sequences as reference (C) Using unique match and all gluten-specific sequences as reference (D) Using clonecount match and all gluten-specific sequences as reference.

repertoire is much smaller than the TCRβ repertoire in our dataset, and preliminary results showed poor performance of TCRα as predictor, where only six correct predictions were made of the 15 donors. In each combination of experimental settings, i.e. repertoire data, reference data and ways for counting the match, a balanced predictive accuracy was evaluated by applying LOOCV (leave one out cross validation) (Fig. 3A, B). In each iteration, one donor was set aside for prediction while a model was trained on the remaining 14 donors. The process was repeated 15 times so that every donor has been set aside once and received a predicted disease status, and finally the balanced predictive accuracy was evaluated based on the results of all 15 donors. In order to enhance the robustness of the result, we used a single nonparametric classifier which was essentially every possible cut-off point along the predictor (unique match or clonecount match). Donor with a value of predictor above the cut-off point was predicted as CD and vice versa. Both sensitivity and specificity of the prediction at all valid cut-off points were comprehensively delineated in the ROC (receiver operating characteristic) curve, and the corresponding AUC (area under the curve), similar to the balanced accuracy, provides a general metric evaluating the predictive performance through balancing sensitivity and specificity. (Fig. 3C, D). While comparing the prediction performance in the eight scenarios, the result measured by AUC was concordant with the one measured by the balanced accuracy for logistic regression.

We did not observe any clear and consistent differences in the predictive performance between matching against all gluten-specific TCR sequences in the reference database, compared with matching against the public TCR sequences only (Fig. 3). Also, the predictive performance was similar whether information of clonal expansion was used or not. We did in all experimental settings observe higher predictive performance when using only TCRβ repertoire data (AUC between 0.94 and 0.98; 12–13 correct predictions) compared to using the sum of TCRα and TCRβ matches (AUC between 0.66 and 0.88; 6–10 correct predictions). Thus, the inclusion of TCRα data worsened the predictive performance.

Considering the fact that the individuals in the diseased group were older than those in the control group as a whole (see Supplementary Fig. S1 online), we checked if age could be a confounding factor in the predictions. In each scenario, we calculated the Pearson correlation coefficient of age and the predictor, either the unique match or the clonecount match. In seven out of eight scenarios, including all four scenarios of matching TCRβ sequences only, we got a low correlation coefficient, ranged from −0.13 to 0.15 (Supplementary Fig. S2 online), showing that age was most likely not a confounding factor.

In conclusion, we showed that by matching gut-derived TCRβ sequences against a reference database of gluten-specific TCRs, we could correctly predict the CD status in 13 out of the 15 donors.

4. Discussion

With the rapid advances in sequencing technology and in particular the ability to sequence a large number of TCRs, the TCR signatures associated with the recognition of a particular antigen, and in its extension a particular disease, can conceivably be used to infer the disease state. In this paper, we have used celiac disease in which extensive information about the disease-specific TCRs exist, to do a proof-of-principle study showing that disease state can be inferred based on TCR sequences derived from the diseased tissue. Using a small set of a few thousand clonotypes sequenced from around 10,000 T cells sampled from each of the 15 donors, the CD status was correctly predicted for 13 out of the 15 donors by matching against known gluten-specific TCRs.

TCRα clonotypes performed rather poorly in our study both when

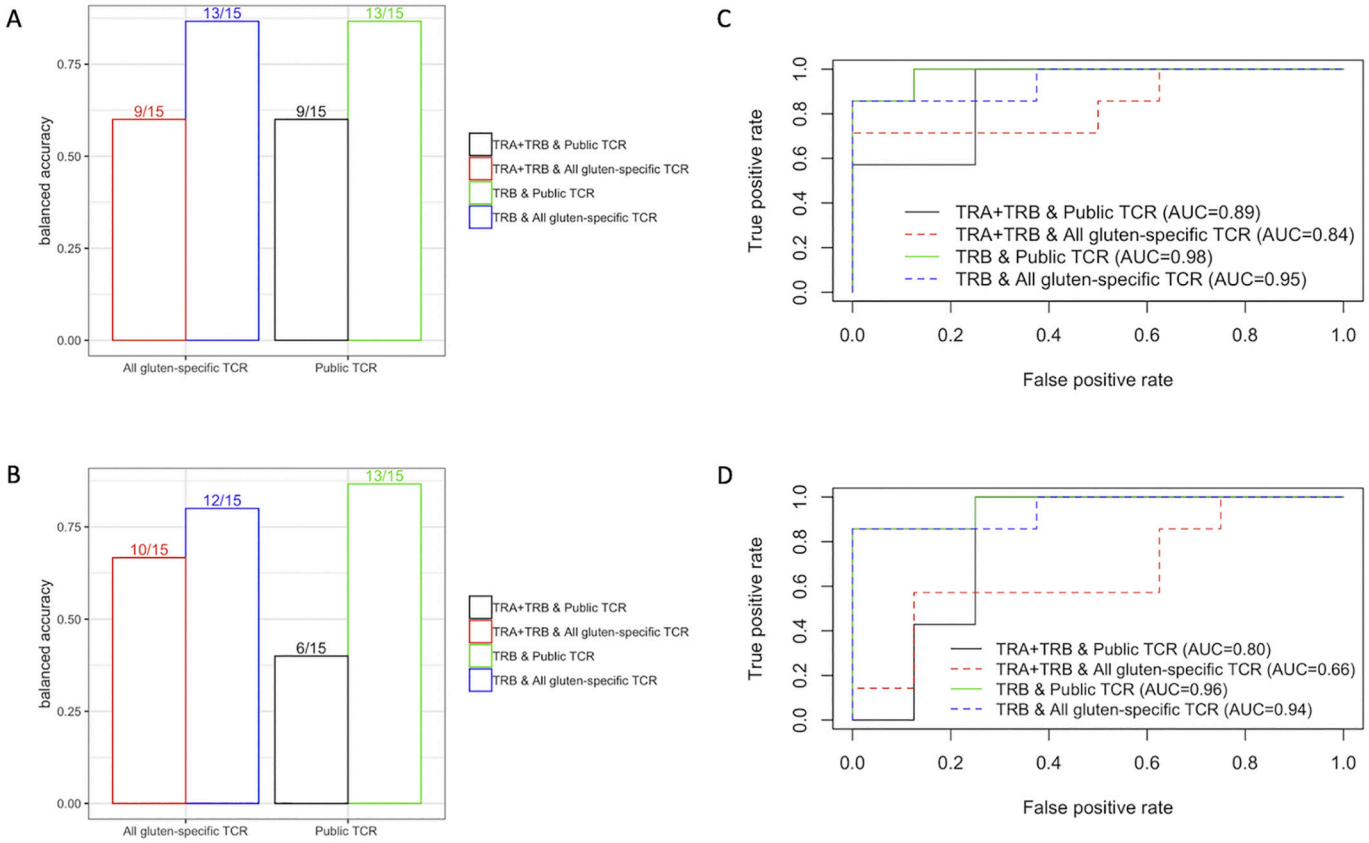

Fig. 3. Predictive performance in all experimental settings. (A) using unique match as a single predictor in logistic regression, balanced accuracy was evaluated by leave one out cross-validation (B) using clonecount match as a single predictor in logistic regression, balanced accuracy was evaluated by leave one out cross-validation (C) Receiver operating characteristic curve (ROC) and the corresponding area under the curve (AUC) by using unique match as classifier (D) Receiver operating characteristic curve (ROC) and the corresponding area under the curve (AUC) by using clonecount match as classifier.

used on its own and in combination with TCRβ. Two controls, CD1428 (healthy HLA-DQ2.5+ control) and CD1390 (HLA-DQ8+ UCD), had relatively large expanded TCRα clones that matched public CD associated TCRα clonotypes. It is possible that these TCRα clonotypes may originate from gluten-specific TCRαβ cells that have expanded in these two control subjects, either as a pre-clinical manifestation of latent CD (in the case of CD1428) or as Treg cells (in CD1390). However, we suspect that these expanded TCRα clones we have observed in our two controls most likely are paired with non-CD-associated TCRβ chains that confer the complete TCRαβ some celiac-unrelated specificities. This notion is supported by the fact that structural and genetic studies show TCRα preference for binding to MHC[1] [2][,]. Thus these two TCRα clonotypes may represent HLA-DQ restriction rather than gluten-specificity.

We used two types of disease associated TCRs as reference; the complete reference database with 5591 clonotypes of gluten-specific T cells, and its small subset of 377 public clonotypes that were observed in two or more CD patients. These two alternative choices of disease associated TCRs present a trade-off between the quantity and specificity for finding matches for disease associated TCRs, since the database of public TCR sequences is considered to be more reliable. For the TCRβ repertoires, despite that the reference database of all gluten-specific TCRβ was 14 times larger than the subset of public TCRβ, roughly the same number of clonotypes in our dataset matched non-public versus public TCRβ reference sequences. The prediction performance was not improved by including the non-public reference TCRβs suggesting that the non-public TCRβ sequences are not as powerful as the public ones for predicting celiac disease state. In addition, it is advantageous to use the public database since its considerable smaller size would save computational power.

Although the majority of TCRs in this study that matched the public reference TCR sequences were from untreated CD patients, a few of them were also observed occasionally in the controls, which is similar to findings in [18]. We therefore believe that specificity of the public clonotypes could be further improved by involving more TCR repertoires from controls for training and purifying. On the other hand, public TCR sequences among CD patients can be continually accumulated by including a larger CD cohort over time. In a previous published study [18], only five of the 39 public reference TCRβ sequences were detected in 10 active CD patients. In comparison, 70 public TCRβ sequences were detected in eight CD patients in our study where 226 public reference TCRβ sequences were used. The positive association between the prevalence and frequency for the public TCR clonotypes indicates that the both the sequencing depth and number of individuals included can be optimized.

In this study, the state of CD was successfully inferred for the majority of the donors. In CD, the antigen specific CD4 T cells are restricted by the disease-associated HLA-DQ molecules, which facilitates the prediction task focusing on distinguishing the HLA-DQ2.5-positive untreated CD patients from the others as controls. As the number of antigen specific CD4 T cells varies in different tissue for different infectious diseases, the repertoire size should be carefully validated when using T cell repertoire information as diagnostic tool.

The number of subjects included in this study is rather small, and only a few thousand clonotypes were sequenced from unsorted lamina propria from each donor. Our results indicate that even with these limitations, it might be possible to infer disease state by matching against known disease-associated TCR sequences. It is to our knowledge the first time this was shown for CD.

Ultimately, in CD, we would like to infer the disease state from blood samples such that diagnosis can be given without the need of endoscopic biopsy. This study is a step towards that ultimate goal, where a larger TCR repertoire needs to be sampled from the blood since the frequency of gluten-specific CD4 T cells in blood is thousand-fold less than in diseased tissue. In addition, with the advances in the knowledge of specific TCRs and TCR sequencing, it is conceivable that TCR repertoire could be used for the diagnosis of other chronic immune-mediated inflammatory diseases.

Data availability

TCR sequencinge reads are uploaded to the NCBI Sequence Read Archive https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?with the project ID: PRJNA678347.

Author contributions

SQ contributed to conceptualization, funding acquisition and project administration, KEAL contributed with resources. AZ and SQ contributed to methodology and data curation, YY, RN and MP contributed to formal analysis, GS contributed to supervision and formal analysis. YY and SQ wrote the original draft whereas GS and GB contributed to review and editing of the manuscript.

Funding

This work is funded by Research Council of Norway (project 179573/V40 through the Centre of Excellence funding scheme and project 233885), and grants from the Stiftelsen Kristian Gerhard Jebsen (SKGJ-MED-017).

Declaration of Competing Interest

None.

Acknowledgments

The authors would like to thank Shiva Dahal-Koirala, Louise F Risnes and Ludvig M Sollid for access to the reference TCR database; and Victor Greiff for helpful comments and suggestions that improved the data processing.

Appendix A. The following are the supplementary data related to this article

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clim.2020.108621.

References

[1] K.W. Wucherpfennig, E. Gagnon, M.J. Call, E.S. Huseby, M.E. Call, Structural biology of the T-cell receptor: insights into receptor assembly, ligand recognition, and initiation of signaling, Cold Spring Harb. Perspect. Biol. 2 (2010) a005140.
[2] E. Sharon, L.V. Sibener, A. Battle, H.B. Fraser, K.C. Garcia, J.K. Pritchard, Genetic variation in MHC proteins is associated with T cell receptor expression biases, Nat. Genet. 48 (2016) 995–1002.
[3] P. Somma, G. Ristori, L. Battistini, S. Cannoni, G. Borsellino, A. Diamantini, et al., Characterization of CD8+ T cell repertoire in identical twins discordant and concordant for multiple sclerosis, J. Leukoc. Biol. 81 (2007) 696–710.
[4] C. Fozza, S. Contini, G. Corda, P. Virdis, A. Galleu, S. Bonfigli, et al., T-cell receptor repertoire analysis in monozygotic twins concordant and discordant for type 1 diabetes, Immunobiology 217 (2012) 920–925.
[5] R.O. Emerson, W.S. DeWitt, M. Vignali, J. Gravley, J.K. Hu, E.J. Osborne, et al., Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire, Nat. Genet. 49 (2017) 659–665.
[6] J.J. Moon, H.H. Chu, M. Pepper, S.J. McSorley, S.C. Jameson, R.M. Kedl, et al., Naïve CD4+ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude, Immunity 27 (2007) 203–213.
[7] A. Christophersen, M. Ráki, E. Bergseng, K.E. Lundin, J. Jahnsen, L.M. Sollid, et al., Tetramer-visualized gluten-specific CD4+ T cells in blood as a potential diagnostic marker for coeliac disease without oral gluten challenge, United European Gastroenterol J 2 (2014) 268–278.
[8] M. Bodd, M. Ráki, E. Bergseng, J. Jahnsen, K.E.A. Lundin, L.M. Sollid, Direct cloning and tetramer staining to measure the frequency of intestinal gluten-reactive T cells in celiac disease, Eur. J. Immunol. 43 (2013) 2605–2612.
[9] R. Tosi, D. Vismara, N. Tanigaki, G.B. Ferrara, F. Cicimarra, W. Buffolano, et al., Evidence that celiac disease is primarily associated with a DC locus allelic specificity, Clin. Immunol. Immunopathol. 28 (1983) 395–404.

[10] L.M. Sollid, G. Markussen, J. Ek, H. Gjerde, F. Vartdal, E. Thorsby, Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer, J. Exp. Med. 169 (1989) 345–350.
[11] K. Karell, A.S. Louka, S.J. Moodie, H. Ascher, F. Clot, L. Greco, et al., HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European genetics cluster on celiac disease, Hum. Immunol. 64 (2003) 469–477.
[12] B. Jabri, L.M. Sollid, T Cells in Celiac Disease, J. Immunol. 198 (2017) 3005–3014.
[13] L.F. Risnes, A. Christophersen, S. Dahal-Koirala, R.S. Neumann, G.K. Sandve, V. K. Sarna, et al., Disease-driving CD4+ T cell clonotypes persist for decades in celiac disease, J. Clin. Investig. 128 (2018) 2642–2650.
[14] M. Shugay, O.V. Britanova, E.M. Merzlyak, M.A. Turchaninova, I.Z. Mamedov, T. R. Tuganbaev, et al., Towards error-free profiling of immune repertoires, Nat. Methods 11 (2014) 653–655.
[15] D.A. Bolotin, S. Poslavsky, I. Mitrophanov, M. Shugay, I.Z. Mamedov, E. V. Putintseva, et al., MiXCR: software for comprehensive adaptive immunity profiling, Nat. Methods 12 (2015) 380–381.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825-–2830.
[17] S.-W. Qiao, M. Ráki, K.S. Gunnarsen, G.-Å. Løset, K.E.A. Lundin, I. Sandlie, et al., Posttranslational modification of gluten shapes TCR usage in celiac disease, J. Immunol. 187 (2011) 3064–3071.
[18] J. Ritter, K. Zimmermann, K. Jöhrens, S. Mende, A. Seegebarth, B. Siegmund, et al., T-cell repertoires in refractory coeliac disease, Gut (2017), https://doi.org/10.1136/gutjnl-2016-311816.