The Journal of Pathology, Original Paper

Classification of intestinal T cell receptor repertoires using machine learning methods can identify patients with coeliac disease regardless of dietary gluten status

Andrew D. Foers[1]†, M. Saad Shoukat[1]†, Oliver E. Welsh[1,2], Killian Donovan[3], Russell Petry[1,2], Shelley C. Evans[1], Michael E. B. FitzPatrick[4], Nadine Collins[5], Paul Klenerman[4,6], Anna Fowler[7]‡, Elizabeth J. Soilleux[1,8,9]‡*

1. Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QP, UK
2. Centre for Mathematical Sciences, University of Cambridge, UK
3. Oxford University Medical School, Oxford, UK
4. Translational Gastroenterology Unit, Nuffield Department of Medicine, University of Oxford, UK
5. Department of Molecular Pathology, Royal Surrey NHS Foundation Trust, Guildford, UK
6. Peter Medawar Building for Pathogen Research, University of Oxford, UK
7. Department of Health Data Science, Institute of Population Health, University of Liverpool, UK
8. Nuffield Division of Clinical Laboratory Sciences, Radcliffe Department of Medicine, University of Oxford, UK

*Correspondence to: Dr Elizabeth Soilleux, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QP, UK. E-mail: ejs17@cam.ac.uk

† Equal first authors

‡ Equal contribution

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/path.5592

Conflicts of interest: Dr E. Soilleux and Dr A. Fowler (inventors) of GB Patent Application No: 1718238.7, for Oxford University Innovation, dated 3 November 2017; International Patent Application No: PCT/GB2018/053198 for Cambridge Enterprise, based on GB Application No: 1718238.7, dated 5 November 2018. Status: pending. Training was provided to Muhammad Saad Shoukat by LabPMM, GmbH, Germany, a subsidiary of Invivoscribe, Inc., USA. No other conflicts of interest were disclosed.

RUNNING TITLE: Machine learning method to identify coeliac disease by TCR repertoires

Main text word count: 3736

ABSTRACT

In coeliac disease (CeD), immune-mediated small intestinal damage is precipitated by gluten, leading to variable symptoms and complications, occasionally including aggressive T-cell lymphoma. Diagnosis, based primarily on histopathological examination of duodenal biopsies, is confounded by poor concordance between pathologists and minimal histological abnormality if insufficient gluten is consumed. CeD pathogenesis involves both CD4+ T-cell-mediated gluten recognition and CD8+ and γδ T-cell mediated inflammation, with a previous study demonstrating a permanent change in γδ T-cell populations in CeD. We leveraged this understanding and explored the diagnostic utility of bulk T-cell receptor (TCR) sequencing in assessing duodenal biopsies in CeD.

Genomic DNA extracted from duodenal biopsies underwent sequencing for TCR-δ (TRD) (CeD, n =11; non-CeD, n=11) and TCR-γ (TRG)) (CeD, n =33; non-CeD, n=21). We developed a novel machine learning-based analysis of the TCR repertoire, clustering samples by diagnosis. Leave-one-out cross-validation (LOOCV) was performed to validate the classification algorithm.

Using TRD repertoire,100% (22/22) duodenal biopsies were correctly classified, with a LOOCV accuracy of 91%. Using TCR-γ (TRG) repertoire, 94.4% (51/54) duodenal biopsies were correctly classified, with LOOCV of 87%. Duodenal biopsy TRG repertoire analysis permitted accurate classification of biopsies from patients with CeD following a strict gluten-free diet for at least 6 months, who would be misclassified by current tests. This result reflects permanent changes to the duodenal γδ TCR repertoire in CeD, even in the absence of gluten consumption. Our method could complement or replace histopathological diagnosis in CeD and might have particular clinical utility in the diagnostic testing of patients unable to tolerate dietary gluten, and for assessing duodenal biopsies with equivocal features.

This approach is generalisable to any TCR/ BCR locus and any sequencing platform, with potential to predict diagnosis or prognosis in conditions mediated or modulated by the adaptive immune response.

Key words: coeliac disease, gluten, T-lymphocyte, T-cell receptor repertoire, machine learning, TRG, TRD, clustering, duodenum.

INTRODUCTION

Coeliac disease (CeD) is a gluten-sensitive enteropathy that develops in genetically susceptible individuals who are exposed to cereal gluten proteins, found in in wheat, rye, and barley. Much of the genetic susceptibility is contributed by possession of the MHC class II molecules HLA-DQ2 and HLA-DQ8. Proteins encoded by these genes bind gluten peptides, particularly those peptides post-translationally modified by tissue transglutaminase (tTG). Recognition of MHC-bound gluten peptide antigens by CD4+ T-lymphocytes induces inflammation, which damages the small intestine, leading to malabsorption, which, in turn, causes many of the symptoms of CeD[1]. The key histopathological changes identifiable in CeD duodenum are mainly consequences of the T-cell response and include villous atrophy, crypt hyperplasia and increased numbers of mucosal lymphocytes, which are a mixture of γδ and CD8+ αβ T-cells[1-3]. While most of the inflammation disappears on a gluten-free diet (GFD), numbers of intraepithelial γδ T-cells remain elevated [2–8]. A recent study elegantly demonstrated that not only do the γδ T-cell numbers increase permanently, once tolerance to gluten is lost, but that there is also an irreversible alteration of the functional subtypes of γδ T-cells present in the duodenum. They demonstrated depletion of naturally occurring, innate-like Vγ4+/Vδ1+ intraepithelial lymphocytes (IELs) with pecificity for the butyrophilin-like (BTNL) molecules BTNL3/BTNL8, expressed in duodenum. In tandem, they observed expansion of gluten-sensitive, interferon-γ-producing Vδ1+ IELs bearing T-cell receptors (TCR) with a shared non-germline-encoded motif that failed to recognize BTNL3/BTNL8 and were phenotypically more akin to adaptive T-cells [4].

CeD treatment is a strict gluten-free diet to avoid complications of malabsorption (vitamin/ mineral deficiency, weight loss, anaemia, infertility, osteoporosis), linked immune phenomena (dermatitis herpetiformis, microscopic colitis), and, rarely, enteropathy-associated T-cell lymphoma. Clinical presentations of CeD are variable, and include non-specific gastrointestinal symptoms (abdominal pain, bloating, diarrhoea), fatigue and cognitive difficulty [1,9,10]. The

estimated prevalence of coeliac disease (CeD) in the UK and US population is 1% [10,11] and rising [12,13]. Screening studies suggest up to 90% of cases remain undiagnosed [11]. An increasing proportion of the population follows a self-imposed gluten-free diet (GFD) without a CeD diagnosis [14].

Adult CeD testing strategies comprise serology for anti-tissue transglutaminase (tTG) and anti-endomysial antibody (EMA), and histopathological examination of duodenal endoscopic biopsies, the latter remaining the diagnostic "gold standard" [10,11]. Biopsy examination by a pathologist is unavoidably subjective, with poor interobserver concordance [15], variable concordance with serology [16], and a high rate of "equivocal" biopsies [2]. Both serology and endoscopic biopsy require patients to eat appreciable amounts of gluten for 6 weeks prior to testing to avoid false negative or equivocal results [17], meaning that many gluten-sensitive patients choose not to seek testing, due to the unpleasant symptoms that follow gluten ingestion. There is an unmet need for a more robust and objective test to diagnose CeD in patients, irrespective of gluten intake, particularly for patients with severe gluten-induced symptoms.

TCRs determine the antigen(s) a T-cell can bind and respond to and are heterodimers of TCR-αβ and TCR-γδ type. A randomly selected and recombined variable (*V*) and joining (*J*) segment encode the antigen binding region of TCR α- and γ-chains (encoded by the TRA and TRG genes, respectively), while TCR β- and δ-chains (encoded by the TRB and TRD genes, respectively) are encoded by *V*, *J* and diversity (*D*) regions [18]. In addition to this somatic recombination, template independent nucleotide insertion and deletion occur, meaning that the small set of TCR genes can theoretically create 10^{15} to 10^{20} unique TCR clonotypes. The most variable part of the TCR, encoded by the V(D)J junction, known as the complementarity determining region 3 (CDR3) is critical in determining antigen specificity [19] and can be used as a genetic 'barcode' to detect,

track and analyse T-cells. The TCR repertoire (TCRR) refers to the range of different TCRs expressed and is shaped by previously encountered antigens [20,21].

Clinically, TCRs are only assessed when PCR and fragment analysis of TCR sequences is used to assess clonal status in suspected T-cell lymphoma [22]. Bulk sequencing of TCRRs is also an important research tool [18], capturing the V(D)J regions in large numbers of T cells. Although this produces large datasets, there are few machine learning algorithms for diagnosing immunological conditions from TCRRs [21], with none in clinical use. Furthermore, many studies use only a fraction of the total information in a TCRR dataset, which comprises multiple closely related, but distinct, TCR sequences. Previous studies of the TCRR of CeD patients have focused on identifying one or a few previously identified TCR sequences, identifying shared motifs between different individuals [23], quantifying sequence diversity in a sample using Shannon diversity or on assessing the magnitude of clonal expansions [4,23, 24, 25–33]. Very few studies have undertaken comparison of patient groups by means of holistic analysis of TCRR sequence data [34,35], due to a lack of bioinformatic tools for doing so. Here, we undertake bioinformatic analysis of the entire TCRR, derived from a duodenal biopsy, in order to classify patients according to their gluten sensitivity status. We show that our approach is successful regardless of whether or not the patient's diet contained gluten prior to biopsy.

MATERIALS AND METHODS

Ethical approval, patient samples, and DNA extraction

Fully anonymised, formalin-fixed, paraffin-embedded (FFPE) duodenal biopsy samples surplus to diagnostic requirements were obtained from the Oxford Radcliffe Biobank via the Oxford Centre for Histopathology Research, or from the Human Tissue Research Biobank, Cambridge University Hospitals NHS Foundation Trust, with full ethical approval (IRAS:162057). Specific

informed consent from individual patients was not required for entry into this study, as all samples used were (a) surplus to diagnostic requirements and (b) fully anonymised to the research team. Patient details and criteria for inclusion are included in **Table 1** and **supplementary material, Table S1**. We obtained duodenal biopsies from patients with active CeD (at least Marsh 3b (n=43): Marsh 2 (n=1)[36]) and from non-CeD individuals (n=32) with no clinical or histological suspicion of CeD (undergoing upper gastrointestinal endoscopy for clinical indications of reflux, dyspepsia and gastritis). We also obtained duodenal biopsies from CeD patients, with a previous biopsy showing at least Marsh 3b features, who had been on a strict GFD for at least 6 months with normal duodenal biopsy histopathology (n=4). DNA was extracted from 10 FFPE scrolls cut at 5 μm per case using the QIAamp DNA FFPE tissue kit (Qiagen, Manchester, UK), as per the manufacturer's instructions.

TCR repertoire sequencing

Bulk amplification of T-cell receptor repertoires was undertaken with the Biomed-2 kit (Invivoscribe, Martinsried, Bavaria, Germany) for TCRδ (TRD), using 150 ng DNA input. An equal amount of all purified amplicons was pooled into a library of 4 nM, denatured, diluted, loaded on to a MiSeq (Illumina, San Diego, CA, USA), and subjected to MiSeq run (Illumina; v3, 2 × 300 cycles). For TCRγ (TRG), the LymphoTrack kit (Invivoscribe) was used following the manufacturer's instructions, with 200 ng of duodenal biopsy DNA as a template. Primers in the LymphoTrack (Invivoscribe) assays are designed with Illumina adapters. Subsequently, each amplicon was purified by AMPure XP beads (Beckman Coulter, Brea, CA, USA) followed by quantification using a 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). Samples were pooled in an equimolar fashion and sequenced on an Ion PGM (Thermo Fisher Scientific, Loughborough, UK).

Bioinformatic analysis

FastX [37] was used to remove low quality sequences. Low quality sequences were defined as individual sequences with more than 10% of nucleotides achieving a quality score below the 3rd quartile (lower 25%). Quartiles were determined using the average nucleotide score for the sequencing output on a per patient basis. Reads were then aligned to the IMGT reference database (www.imgt.org) using IMGT/HighV-Quest (www.imgt.org/IMGTindex/IMGTHighV-QUEST.php[38]), in order to determine V, D and J usage, to identify the CDR3 region, and to determine whether or not a sequence was functional (**supplementary material, Table S2**). Nucleotide sequences were translated to amino acids, which are used in all subsequent analysis, and all sequences predicted to be non-functional were removed at this stage.

Classification algorithm

A novel cluster-based classification algorithm for distinguishing between the TCRR of case and control samples was developed and is described in the Results section (**Figure 1**). The algorithm, implemented in R scripts, is available from Zenodo (http://doi.org/10.5281/zenodo.3964131).

Assessment of classification performance

During clustering, we partition the tree into two clusters by choosing a merge (node) and defining the cluster created at that merge as a single cluster, then assigning all other samples (leaves) into another single cluster. The two clusters are then labelled either CeD or non-CeD according to the true classification of the majority of samples in each cluster. Sensitivity is then calculated as the percentage of CeD samples correctly labelled, specificity as the percentage of non-CeD samples correctly labelled, and (training) accuracy as the percentage of all samples correctly labelled. The optimal partition was chosen as the one with highest accuracy, and in the case of ties, partitions were then ranked based on sensitivity and finally distance between clusters (further details for clustering methodology and parameter optimisation are presented in Supplementary materials and methods).

Leave-one-out cross validation

In order to provide a validation of our methodology and illustrate how the algorithm could be used for diagnosis, algorithm parameters were optimised using *n-1* samples with one sample removed. The removed sample was then re-introduced to determine whether it was assigned to the correct group, on the basis of diagnosis. This process was repeated iteratively until each sample had been removed to estimate the testing accuracy of the algorithm (further details are presented in Supplementary materials and methods, together with more detailed explanation of statistical methodology used in all analyses).

Heatmap and hierarchical clustering analysis based on V and J segment usage

V and J segment usage is visualised by heatmap using the pheatmap R package [39]. V and J segment frequency across patients is represented as standard deviations from the mean. Hierarchical clustering was performed on columns using the complete linkage method with Euclidean distance.

RESULTS

A novel machine learning algorithm for sample classification

We developed an algorithm capable of diagnosing CeD, regardless of gluten consumption, on the basis of TCRR in duodenal biopsies (**Figure 1**). Our approach was based on the hypothesis that there are multiple related γδ TCRs with similar specificities, capable of binding gluten and possibly self-antigens, due to the phenomenon of epitope spreading [40], encoded by closely related TRG or TRD sequences [4,23], both within a single patient's TCRR and between patient TCRRs with CeD. In brief, TCR sequences are translated into amino acids and the hypervariable part of the TCR sequence, complementarity determining region 3 (CDR3) is broken into overlapping kmers (sequences of length k). We produce a set of kmers that is positionally

annotated, by which third of the CDR3 sequence they derive from (start/ middle/ end), and a set that lacks this annotation. We compile a very large (high dimensionality) matrix containing the frequency of each kmer in each patient sample. We reduce the dimensionality of the matrix by principal component analysis (PCA) and cluster the samples using all 1,023 possible combinations of principal components (PCs) 1–10. We then determine which of these PC combinations has the highest sample classification accuracy. Of these PC combinations with high sample classification accuracy, we then select the PC combination that gives the greatest separation between diagnostic groups. We tested our approach on both the positionally annotated and non-positionally annotated sets of kmers, testing parameters, as described above, and determined whether or not positionally annotating the kmers improved samples' classification. Thus, in this machine learning approach, the modifiable parameters are (a) kmer length, (b) whether or not kmers are positionally annotated and (c) the exact PC combination used for sample classification.

TCR delta repertoire analysis of FFPE duodenal biopsy DNA can determine gluten sensitivity status

We applied our algorithm to TRD CDR3 sequences, from DNA extracted from CeD patient duodenal biopsies (n=11) and non-CeD controls (n=11) (**supplementary material, Table S1**). Diagnostic accuracy was optimised (**Figure 2**) by selecting (a) kmer length and (b) kmer type (positionally annotated versus not, as defined in **Figure 1**) and (c) PCs. Non-positionally annotated 4mers gave optimal sample separation by diagnosis, with high accuracy across a broad range of PC combinations, with 14.96% PC combinations (153/1023) giving 100% training accuracy (**Figures 2C,D and supplementary material, Table S3**). With positional annotation, kmer length of 7 was optimal, with 7.53% combinations (77/1,023) giving 100% training accuracy (**Figures 2A,B and supplementary material, Table S4**). Of 153/1023 PC combinations giving 100% accuracy with non-positionally annotated 4mers, the PC combination of 1, 5, 6 and 10 gave greatest separation between diagnostic groups, with greatest vertical distance between cluster plot

branches **(Figure 2C**). From figures 2B,D, it can be appreciated that a wide range of kmer lengths and PC combinations also gave good sample classification, indicating the robustness of our approach. To validate our optimised classification algorithm using non-positional 4mers, we implemented a leave-one-out cross validation (LOOCV) approach (**supplementary material, Figure S3**). 10/11 CeD and 10/11 non-CeD patients clustered correctly, giving a testing accuracy, sensitivity, and specificity of 91%. We excluded the possibility that HLA type or other properties of the TRD data might be confounding our classification methodology (**supplementary material, Figures S1,S2 and Tables S4 and S5–S8, Figure 2A,C**).

TCR gamma repertoire analysis of FFPE duodenal biopsy DNA can determine gluten sensitivity status

Our algorithm also performed well in classifying patients' CeD status using TRG repertoires derived from DNA extracted from FFPE duodenal biopsies. In a second, larger patient cohort (n=54), TRG CDR3 sequences were broken into positional 5mers and the PC combination of 4, 6 and 7 was best able to separate the patient cohorts, with 32/33 (97.0% training sensitivity) CeD and 19/21 (90.5% training specificity) non-CeD samples correctly classified, giving a training accuracy of 51/54 (94.4%) (**Figures 3A,B and supplementary material, Table S9)**. **Figure 3B,D** show that a wide range of kmer lengths and PC combinations also gave good sample classification, indicating the broad applicability of our approach to this analysis (**supplementary material, Table S9**). For TRG, in contrast to TRD, positionally annotated kmers outperformed non-positionally annotated kmers (**Figures 3C,D and supplementary material, Table S10).** Neither full-length CDR3 (**Figures 3E,F and supplementary material, Table S11**) nor V/D/J segment usage were able to accurately separate CeD from non-CeD samples (**Figure 3G** and **supplementary material, Figure 4D,E**). We also excluded the possibility that HLA type or other properties of the TRG data might be confounding our classification methodology (**supplementary material, Figures S4, S5, and Tables S12–S15, Figure 3A,C**).

Implementing a LOOCV approach to validate our preliminary findings for TRG, using positional 5mers, 29/33 CeD and 18/21 non-CeD samples were correctly classified, giving a testing accuracy of 87%, a sensitivity of 88% and a specificity of 86% (**supplementary material, Figure S6**).

Gluten intake is not required for correct classification of patient gluten sensitivity status

Finally, we assessed whether our method could classify patients diagnosed with CeD who were adhering to a GFD. Duodenal biopsies were obtained from an additional 4 patients, each comprising one sample at initial CeD diagnosis, when the patient was on a gluten containing diet, (with changes of at least Marsh-Oberhuber [36] grade 3b (**Figure 4**)) and a second sample following at least 6 months on a strict GFD that displayed normal histology (**Figure 4**). We introduced the additional patient samples that were taken at initial CeD diagnosis into the 54 TRG cohort one at a time, using our previously optimised parameters, and all samples clustered with the CeD samples. Remarkably, all 4 GFD samples, when introduced individually into the cohort, also clustered with the CeD samples (**Figure 4**). These data indicate that our algorithm is capable of identifying patients with CeD even in the absence of gluten ingestion.

DISCUSSION

Our novel machine learning approach distinguished samples from patients with and without CeD, when applied to TCRR (TRD and TRG) from duodenal DNA from two independent cohorts, despite using relatively degraded FFPE-derived DNA template, most likely leading to loss of a proportion of the TCR sequences in each sample. The sequencing approaches we selected (Lymphotrack^{TM}/ Biomed-2, Invivoscribe, with Illumina sequencing) are amenable to FFPE-derived DNA and are used clinically in lymphoma/ leukaemia diagnostics. Thus, our approach to CeD diagnosis could easily be incorporated into current pathology department workflows. Following a larger validation study of diagnostic accuracy [41], this algorithm could have the

potential for use in the diagnosis of CeD in cases where current diagnostic techniques do not perform well and pathologists struggle to agree on histological findings. Such as those with an isolated increase in intraepithelial lymphocytes without villous atrophy, those with seronegative villous atrophy and those CeD patients on GFD with normal histology, who are likely to be misclassified by current serological and histopathological testing[2,15–17,42]. Although only a small number of patients on gluten free diets were analysed here, our approach shows promise in eradicating the requirement for patients to ingest gluten for 2–6 weeks prior to testing [43,44], increasing test acceptability to patients.

The current definition of CeD is hampered by imperfect, and sometimes discordant, tests for the condition. Our algorithm has the exciting potential to provide a better definition of CeD based on similarities between patients' TCRRs. Further refinement of diagnostic classification might be achieved by the clustering of patient samples on the basis of a combination of sequence data from several TCR and/ or B-cell receptor (BCR) loci, an approach beyond the scope of the present study.

While the exact features mediating sample clustering are difficult to define with a machine learning method such as ours, the advantage of our method is that it provides a holistic analysis of DNA-based TCRR sequences that considers similar, but non-identical, CDR3 sequences, as being more closely related to each other than any two CDR3 sequences chosen at random. This contrasts with methodologies that simply determine clonal status or search for specific clones or motifs within TCRR, which do not take account of the presence of closely related CDR3 sequences[4,23,25–33]. Furthermore, methodologies that search for one or a small number of specific TCR clones risk generating false negative results, if a critical, but low frequency sequence, is missed, leading to poor sensitivity.

While kmers have been used to analyse analogous BCR sequences, in two previous studies [34,35], neither study employed kmers for holistic CDR3 analysis or as the basis of sample classification. The accurate clustering achieved by our method suggests that there is a greater degree of similarity in the kmer usage, and thus in the CDR3 sequences from which these kmers derive, between patients with CeD than between non-CeD controls. This observation is in keeping with an immune response to a stereotyped set of antigens in the CeD patients. Further work is required to explore the exact kmer patterns underpinning this similarity and identification of these sequences has the potential to provide insight into the underlying immunological mechanisms of CeD. For example, holistic kmer-based TCR analysis could be used as a method to identify consensus sequences within the TCRR of a cohort of CeD patients, with the numbers of separate groups of immunoreceptor consensus sequences giving an indication of the likely numbers of different epitopes being recognised in the condition.

Biological understanding and computational methods for γδ T-cells are not yet well enough developed to predict likely epitopes/ antigens bound from the TCR CDR3 sequences alone. Indeed, relatively little is known about γδ T-cells' antigen binding mechanisms, unlikely αβ T-cells, which are known to recognise short peptide antigens bound to MHC molecules, with the TCR generally contacting both the peptide and the MHC protein. The ability of short kmer sequences, derived from the CDR3 sequences of γδ T-cells, to separate CeD from normal biopsies indicates that CDR3 sequences are likely to be very important in γδ T-cell-antigen binding or other γδ T-cell interactions in CeD. This observation fits well with a recent study of CeD duodenum, demonstrating depletion of resident duodenal Vγ4+/Vδ1+ intraepithelial lymphocytes (IELs), with semi-invariant TCRs, and their replacement with gluten-sensitive, interferon-γ-producing Vδ1+ IELs bearing T-cell receptors (TCR) with CDR3 motifs that are shared both within and between CeD patients' TRD repertoires. Canonical sets of kmers from these shared

CDR3 motifs are likely to be a key feature in the CeD patient TCRR, detected by our classification algorithm [4].

It likely that a major reason for the success of our algorithm is the fact that it compares non-CeD and CeD duodenum, rather than simply looking for predefined features of CeD duodenum. It is thus able to detect a signal based both on the loss of the semi-invariant Vγ4+/Vδ1+ TCRs of the innate-like γδ T-cell population and on the development of the antigen driven Vδ1+ γδ T-cell population. The loss of these semi-invariant Vγ4+/Vδ1+ TCRs may contribute to the fact that we see an unexpected increase in TCR diversity in CeD, rather than the decrease one might expect if there is an evolving clonal response to gluten. The relatively poorly characterised, innate-like Vγ4+/Vδ1+ IELs are thought to maintain homeostasis in the local small intestinal microenvironment, either by eliminating virus-infected or malignant cells in response to innate signals, or by promoting tissue healing via the production of growth factors[4]. Their presence appears to be key in maintaining a gluten tolerant IEL population. Therefore, detecting loss of Vγ4+/Vδ1+ IELs may be as important as detecting the novel gluten-sensitive Vδ1+ IELs, in our ability to classify samples as CeD or non-CeD on the basis of TCRR.

Commensurate with our observation that duodenal biopsies from CeD patients on GFD are correctly classified on the basis of TRG TCRR, exclusion of dietary gluten in the study by Mayassi *et al*. was insufficient to reconstitute the physiological Vγ4+/Vδ1+ IEL population [4]. Our ability to classify duodenal biopsies from CeD patients on GFD correctly is also in keeping with reports of persistent elevation of intraepithelial γδ T-cells [2–6] in CeD patients, even on GFD. These data show that CeD-associated γδ T-cells do not recirculate away from the small intestine in the absence of dietary gluten. The observed persistence of CeD-associated γδ T-cells in the duodenum, without gluten ingestion, indicates that TRG and TRD may be the most appropriate TCR loci to analyse to determine gluten sensitivity (CeD) status, tests which are otherwise

confounded by insufficient gluten intake. This biological phenomenon appears to underpin the success of our potential novel diagnostic approach to the condition.

In summary, we have developed a machine-learning algorithm that, following further testing on a larger cohort, could be used for CeD diagnosis, regardless of dietary gluten status, which uses TCR sequencing methodology amenable to current histopathological and molecular diagnostic workflows. Our novel, machine learning-based bioinformatic approach is generalisable to sequences from all 4 TCR and all 3 BCR loci, which are obtained using any sequencing platform. Thus, this approach might similarly be applied to the prediction of diagnosis or prognosis in other conditions mediated by the adaptive immune response. These might include other autoimmune or immune-mediated inflammatory conditions, allergic reactions, and possibly immune responses to both infections, such as SARS-CoV-2, and cancers.

Acknowledgements

We are grateful to Oxford Radcliffe Biobank and Cambridge University Hospitals NHS Foundation Trust Human Tissue Research Biobank for the provision of tissue. We thank Professor M. Arends and Dr S. Eglen for critical reading of this manuscript.

Funding

This work was funded by Coeliac UK (ES; INOV01-18; ES01-14), Medical Research Council (ES; 28901; MC_PC_17185), Biotechnology and Biological Sciences Research Council (ES; PKFM.GAAB), Cancer Research UK (ES; C30885/A29312), Oxford Health Services Research Committee Fund (ES; 1125) and Celgene (ES/MEBF; R39207/CN011).

Author contributions statement

EJS and AF designed and conceived the study. KD and SCE selected samples. MSS, SCE and NC performed experiments. OE, RP, MSS, ADF and AF analysed the data. ADF, AF, MEBF and PK and EJS wrote the manuscript. All authors reviewed and approved the final manuscript.

Data availability statement

The classification algorithm, implemented in R scripts, is available from Zenodo (http://doi.org/10.5281/zenodo.3964131)

REFERENCES

1. Abadie V, Sollid LM, Barreiro LB, *et al.* Integration of genetic and immunological insights into a model of celiac disease pathogenesis. *Annu Rev Immunol* 2011; **29**: 493-525.
2. Lonardi S, Villanacci V, Lorenzi L, *et al.* Anti-TCR gamma antibody in celiac disease: the value of count on formalin-fixed paraffin-embedded biopsies. *Virchows Arch* 2013; **463**: 409-413.
3. Leon F. Flow cytometry of intestinal intraepithelial lymphocytes in celiac disease. *J Immunol Methods* 2011; **363**: 177-186.
4. Mayassi T, Ladell K, Gudjonson H, *et al.* Chronic Inflammation Permanently Reshapes Tissue-Resident Immunity in Celiac Disease. *Cell* 2019; **176**: 967-981 e919.
5. Bhagat G, Naiyer AJ, Shah JG, *et al.* Small intestinal CD8+TCRgammadelta+NKG2A+ intraepithelial lymphocytes have attributes of regulatory cells in patients with celiac disease. *J Clin Invest* 2008; **118**: 281-293.
6. Calleja S, Vivas S, Santiuste M, *et al.* Dynamics of non-conventional intraepithelial lymphocytes-NK, NKT, and gammadelta T-in celiac disease: relationship with age, diet, and histopathology. *Dig Dis Sci* 2011; **56**: 2042-2049.
7. Halstensen TS, Brandtzaeg P. Activated T lymphocytes in the celiac lesion: non-proliferative activation (CD25) of CD4+ alpha/beta cells in the lamina propria but proliferation (Ki-67) of alpha/beta and gamma/delta cells in the epithelium. *Eur J Immunol* 1993; **23**: 505-510.
8. Kutlu T, Brousse N, Rambaud C, *et al.* Numbers of T cell receptor (TCR) alpha beta+ but not of TcR gamma delta+ intraepithelial lymphocytes correlate with the grade of villous atrophy in coeliac patients on a long term normal diet. *Gut* 1993; **34**: 208-214.
9. Rubio-Tapia A, Hill ID, Kelly CP, *et al.* ACG clinical guidelines: diagnosis and management of celiac disease. *Am J Gastroenterol* 2013; **108**: 656-676.
10. Lebwohl B, Sanders DS, Green PHR. Coeliac disease. *Lancet* 2018; **391**: 70-81.
11. Murch S, Jenkins H, Auth M, *et al.* Joint BSPGHAN and Coeliac UK guidelines for the diagnosis and management of coeliac disease in children. *Arch Dis Child* 2013; **98**: 806-811.
12. Kang JY, Kang AH, Green A, *et al.* Systematic review: worldwide variation in the frequency of coeliac disease and changes over time. *Aliment Pharmacol Ther* 2013; **38**: 226-245.
13. West J, Fleming KM, Tata LJ, *et al.* Incidence and prevalence of celiac disease and dermatitis herpetiformis in the UK over two decades: population-based study. *Am J Gastroenterol* 2014; **109**: 757-768.
14. Rubio-Tapia A, Ludvigsson JF, Brantner TL, *et al.* The prevalence of celiac disease in the United States. *Am J Gastroenterol* 2012; **107**: 1538-1544; quiz 1537, 1545.
15. Arguelles-Grande C, Tennyson CA, Lewis SK, *et al.* Variability in small bowel histopathology reporting between different pathology practice settings: impact on the diagnosis of coeliac disease. *J Clin Pathol* 2012; **65**: 242-247.
16. Schyum AC, Rumessen JJ. Serological testing for celiac disease in adults. *United European Gastroenterol J* 2014; **1**: 319-325.
17. Downey L, Houten R, Murch S, *et al.* Recognition, assessment, and management of coeliac disease: summary of updated NICE guidance. *BMJ* 2015; **351**: h4513.
18. Laydon DJ, Bangham CR, Asquith B. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philos Trans R Soc Lond B Biol Sci* 2015; **370**.
19. Lefranc MP. Unique database numbering system for immunogenetic analysis. *Immunol Today* 1997; **18**: 509.
20. Greiff V, Bhat P, Cook SC, *et al.* A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med* 2015; **7**: 49.
21. Emerson RO, DeWitt WS, Vignali M, *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 2017; **49**: 659-665.
22. van Dongen JJ, Langerak AW, Bruggemann M, *et al.* Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* 2003; **17**: 2257-2317.
23. Han A, Newell EW, Glanville J, *et al.* Dietary gluten triggers concomitant activation of CD4+ and CD8+ alphabeta T cells and gammadelta T cells in celiac disease. *Proc Natl Acad Sci U S A* 2013; **110**: 13073-13078.
24. Ritter J, Zimmermann K, Johrens K, *et al.* T-cell repertoires in refractory coeliac disease. *Gut* 2018; **67**: 644-653.
25. Dahal-Koirala S, Ciacchi L, Petersen J, *et al.* Discriminative T-cell receptor recognition of highly homologous HLA-DQ2-bound gluten epitopes. *J Biol Chem* 2019; **294**: 941-952.
26. Hussein S, Gindin T, Lagana SM, *et al.* Clonal T cell receptor gene rearrangements in coeliac disease: implications for diagnosing refractory coeliac disease. *J Clin Pathol* 2018; **71**: 825-831.

27. Risnes LF, Christophersen A, Dahal-Koirala S, *et al.* Disease-driving CD4+ T cell clonotypes persist for decades in celiac disease. *J Clin Invest* 2018; **128**: 2642-2650.
28. Yohannes DA, Freitag TL, de Kauwe A, *et al.* Deep sequencing of blood and gut T-cell receptor beta-chains reveals gluten-induced immune signatures in celiac disease. *Sci Rep* 2017; **7**: 17977.
29. Gunnarsen KS, Hoydahl LS, Risnes LF, *et al.* A TCRalpha framework-centered codon shapes a biased T cell repertoire through direct MHC and CDR3beta interactions. *JCI Insight* 2017; **2**.
30. Petersen J, Montserrat V, Mujico JR, *et al.* T-cell receptor recognition of HLA-DQ2-gliadin complexes associated with celiac disease. *Nat Struct Mol Biol* 2014; **21**: 480-488.
31. Qiao SW, Christophersen A, Lundin KE, *et al.* Biased usage and preferred pairing of alpha- and beta-chains of TCRs specific for an immunodominant gluten epitope in coeliac disease. *Int Immunol* 2014; **26**: 13-19.
32. Qiao SW, Raki M, Gunnarsen KS, *et al.* Posttranslational modification of gluten shapes TCR usage in celiac disease. *J Immunol* 2011; **187**: 3064-3071.
33. Broughton SE, Petersen J, Theodossis A, *et al.* Biased T cell receptor usage directed against human leukocyte antigen DQ8-restricted gliadin peptides is associated with celiac disease. *Immunity* 2012; **37**: 611-621.
34. Ostmeyer J, Christley S, Rounds WH, *et al.* Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics* 2017; **18**: 401.
35. Greiff V, Weber CR, Palme J, *et al.* Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. *Journal of immunology* 2017; **199**: 2985-2997.
36. Oberhuber G. Histopathology of celiac disease. *Biomed Pharmacother* 2000; **54**: 368-372.
37. Assaf G, Hannon GJ. FASTX-toolkit. Available from http://hannonlab.cshl.edu/fastx_toolkit/ [last accessed 12/11/2020]
38. Alamyar E, Duroux P, Lefranc MP, *et al.* IMGT® tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* 2012; **882**: 569-604.
39. Kolde R. Pheatmap v1.0.12 https://www.rdocumentation.org/packages/pheatmap: Available from: https://www.rdocumentation.org/packages/pheatmap
40. Freitag T, Schulze-Koops H, Niedobitek G, *et al.* The role of the immune response against tissue transglutaminase in the pathogenesis of coeliac disease. *Autoimmun Rev* 2004; **3**: 13-20.
41. Bossuyt PM, Reitsma JB, Bruns DE, *et al.* STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; **351**: h5527.
42. Aziz I, Peerally MF, Barnes JH, *et al.* The clinical and phenotypical assessment of seronegative villous atrophy; a prospective UK centre experience evaluating 200 adult cases over a 15-year period (2000-2015). *Gut* 2017; **66**: 1563-1572.
43. Leffler D, Schuppan D, Pallav K, *et al.* Kinetics of the histological, serological and symptomatic responses to gluten challenge in adults with coeliac disease. *Gut* 2013; **62**: 996-1004.
44. Sarna VK, Lundin KEA, Morkrid L, *et al.* HLA-DQ-Gluten tetramer blood test accurately identifies patients with and without celiac disease in absence of gluten consumption. *Gastroenterology* 2018; **154**: 886-896 e886.
45. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995; **57**: 289-300.
46. Oj D. Multiple comparison among means. *Journal of the American Statistical Association* 1961; **56**: 52-64.
47. Cebula A, Seweryn M, Rempala GA, *et al.* Thymus-derived regulatory T cells contribute to tolerance to commensal microbiota. *Nature* 2013; **497**: 258-262.
48. Oksanen J, Blanchet FG, Friendly M, *et al.* vegan: Community Ecology Package. R package version 2.4–2. 2017. Documentation available from https://www.rdocumentation.org/packages/vegan/versions/2.4-2. [Last accessed 12 November 2020].
49. Hothorn T, Hornik K, van de Wiel MA, *et al.* Implementing a Class of Permutation Tests: The coin Package. *J Statist Soft* 2008; **28**: issue 8.
50. Miqueu P, Guillet M, Degauque N, *et al.* Statistical analysis of CDR3 length distributions for the assessment of T and B cell repertoire biases. *Mol Immunol* 2007; **44**: 1057-1064.
51. Dimitrova DS, Kaishev VK, Tan S. Computing the Kolmogorov-Smirnov Distribution when the Underlying cdf is Purely Discrete, Mixed or Continuous. 2017. Available from https://openaccess.city.ac.uk/id/eprint/18541/

References 45–51 are cited only in the supplementary material.

Table 1. Patient demographics and clinical parameters

TRD (DNA)			
	Coeliac (n=11)	*Normal (n=11)*	*P value*
Age, mean (± SD)	29.2 (20.8)	53.8 (13.3)	0.002
Sex (M : F)	4 : 7	3 : 8	1.00
HLA-DQ2 and/ or HLA-DQ8	11 (100%)	3 (27%)	0.001
Anti-TTG (>laboratory ULN)	11 (100%)	0 (0%)	<0.0001
Anti-EMA (>laboratory ULN	11 (100%)	0 (0%)	<0.0001
Marsh Grade (3;2;1;0)	10;1;0;0	0;0;1;10	<0.0001*

TRG (DNA)			
	Coeliac (n=33)	*Normal (n=21)*	*P value*
Age, mean (± SD)	16.0 (9.2)	54.8 (17.9)	<0.0001
Sex (M : F)	8 : 25	12 : 9	0.021
HLA-DQ2 and/ or HLA-DQ8	32 (100%)^	11 (52%)	<0.0001
Anti-TTG (>laboratory ULN)	33 (100%)	0 (0%)	0.140
Anti-EMA (>laboratory ULN	28 (85%)	0 (0%)	<0.0001
Marsh Grade (3;2;1;0)	33;0;0;0	0;0;0;21	<0.0001*

Gluten-free diet longitudinal patients (n=4)			
	Baseline	*6 months GFD*	*P value*
Age, mean (± SD)	39.5 (13.6)	Same patients	-
Sex (M : F)	0 : 4	Same patients	-
Anti-TTG (>laboratory ULN)	4 (100%)	0 (0%)	-
Anti-EMA (>laboratory ULN	3 (75%)	0 (0%)	-
Marsh Grade (3;2;1;0)	4;0;0;0	0;0;0;4	-

Age and Marsh Grade analysed with Student's *t*-test.

Sex, HLA status, Anti-TTG and Anti-EMA analysed with Fisher's exact test.

** Calculated from average of Marsh grades*

^ HLA typing inconclusive for patient C31

ULN = upper limit of normal

Figure legends

Figure 1. Flow chart of our novel bioinformatic approach

(A) We translate the nucleic acid sequence to amino acid sequences, remove any non-functional sequences and identify the most variable of the three hypervariable regions in each TCR, the CDR3 region, using the IMGT database[38]. To take account of similar, but not clonotypically identical TCR sequences, we break the entire CDR3 sequence of each TCR into short overlapping segments, designated kmers. We reasoned that the same kmer occurring at substantially different positions within the CDR3 is likely to differ in its effect on antigen binding and so tested positionally annotated kmers (start/ middle/ end) in the functional CDR3s. For example, CALGE (start) is regarded as distinct from CALGE (end). (B) We calculate the frequency of each full length CDR3, unique kmer, and positional kmer identified in step A. (C) Steps A and B are repeated for each patient, so that sample classification results using each of these three different input types can be compared. (D) Frequencies are normalised for each sample and combined for all samples into a single frequency matrix. (E) The frequency matrices are particularly high dimensional, due to the very large number of possible positional kmers (4.8×10^5 to 1.5×10^{12} for 4-9 amino acid kmers, respectively). Therefore, principal component analysis (PCA) is used to reduce this dimensionality, while retaining major sources of variation, simplifying downstream computational steps. (F) To classify samples, we apply hierarchical clustering to the dimensionality reduced data set, which iteratively groups together samples. Samples for which the true underlying disease status is known are used to select optimal parameter sets, consisting of the value of k and principal components (PCs), as well as input type (full length CDR3, non-positional kmers, and positional kmers), by means of a machine learning approach (i.e., to train the model). The optimal parameters generate clusters that correspond best with disease state (see Supplementary materials and methods).

Figure 2. Application of our algorithm to TRD sequence data obtained from formalin fixed paraffin embedded duodenal biopsies of coeliac disease patients (n=11) and non-coeliac disease controls (n=11)

Diagnostic classification accuracy was optimised using all possible input types (positional kmers, non-positional kmers and full length CDR3 sequences) and PC combinations. (A and B) With positional annotation, a kmer length of 7 achieved greatest accuracy, with 77/1023 PC combinations giving 100% accuracy (**supplementary material, Table S4**). Of the 77 PC combinations giving 100% accuracy, PCs 3, 4, 5, 6 and 8 gave the greatest separation between diagnostic groups, with the greatest vertical distance between branches on the cluster plot (known technically as the Mutual Reachability Distance). HLA type (DQ2 and/ or DQ8 or other) does not explain the classification, with non-CeD samples from DQ2 or DQ8 positive subjects clustering on the basis of disease (see also **Figures 2C,E**). (C and D) Without positional annotation, a kmer length of 4 was optimal and 153/1023 PC combinations gave 100% accuracy (**supplementary material, Table S3**). Of the 153 PC combinations giving 100% accuracy, PCs 1, 5, 6 and 10 gave the greatest separation between diagnostic groups, with the greatest vertical distance between branches on the cluster plot. The high classification accuracy across a broad range of parameters indicates the robustness of this approach. (E and F) Using full length CDR3 sequences, no PC combinations gave 100% accuracy, although 1021 PC combinations permitted 21/22 (95.5%) samples to be classified correctly (**supplementary material, Table S5**). (G) Hierarchical clustering on the basis of combinations of V-J segments usage in the sequence data could not classify patient samples by diagnosis. For patient details and inclusion criteria for TRD sequencing cohort, please see **Table 1** and **supplementary material, Table S1**. Sequence data parameters are summarised in **supplementary material, Table S2**. Further validation of these results is included in **supplementary material, Figures S1–S3** and **Tables S6–S8.**

Figure 3. Application of our algorithm to TRG sequence data obtained from formalin fixed paraffin embedded duodenal biopsies of coeliac disease patients (n=33) and non-coeliac disease controls (n=21)

Diagnostic classification accuracy was optimised using all possible input types (positional kmers, non-positional kmers and full length CD3 sequences) and PC combinations. (A and B) With positional annotation, a kmer length of 5 achieved greatest accuracy, with 1 PC combination (PCs 4, 6 and 7) classifying samples with 94.4% accuracy (**supplementary material, Table S9**). (C and D) Without positional annotation, a kmer length of 4 was optimal and 4 PC combinations giving 92.6% accuracy (**supplementary material, Table S10**). HLA type (DQ2 and/ or DQ8 or neither) did not explain the classification, with non-CeD samples from DQ2 or DQ8 positive subjects clustering on the basis of disease (see also **Figure 3A**). (E and F) Using full length CDR3 sequences, 24 PC combinations gave 90.7% accuracy (**supplementary material, Table S11**). (G) To exclude the possibility that other properties of the data might be confounding our classification methodology, we undertook analysis of V and J segment usage, shown as a heat map and showed that hierarchical clustering on the basis of combinations of V–J segments usage in the sequence data could not cluster patient samples by diagnosis. For patient details and inclusion criteria for TRG sequencing cohort, please see **Table 1 and supplementary material, Table S1**. Further validation of these results is included in **supplementary material, Figures S4–S6** and **Tables S12–S15.**

Figure 4. Applicability of TRG analysis to patient samples on a gluten-free diet

Four additional formalin fixed paraffin embedded duodenal biopsy samples from patients on a gluten containing diet (i.e., at the time of initial diagnosis of coeliac disease), were obtained. Histology (haematoxylin and eosin stain) of each sample is shown, all with severe features of coeliac disease (at least Marsh grade 3a). TRG sequence data obtained from biopsy samples shown were added one by one into the cohort and analysed by means of our algorithm. Each

patient sample correctly clustered with the coeliac samples. An additional duodenal biopsy sample was taken from each patient following at least 6 months on a GFD. Histology (haematoxylin and eosin stain) of each sample is shown and all would be classified as histologically normal. TRG sequence data obtained from biopsy samples were added one by one into the cohort and analysed by means of our algorithm. Again, each patient sample correctly clustered with the coeliac samples.