



A systematic approach for comprehensive T-cell epitope discovery using peptide libraries

Tim Beißbarth^{1,*}, Jason A. Tye-Din^{1,2}, Gordon K. Smyth¹,
Terence P. Speed¹ and Robert P. Anderson^{1,2}

¹Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade and

²The Royal Melbourne Hospital, Parkville VIC 3050, Australia

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: T-cell response to peptides bound on MHC Class I or Class II molecules is essential for immune recognition of pathogens. T-cells are activated by specific peptide epitopes that are determined within the antigen processing pathways and presented on the surface of other cells bound to MHC molecules. To determine which part of allergenic or pathogenic proteins can stimulate T-cells is important for the treatment of diseases. We sought to take advantage of the falling cost of synthetic, screening grade peptides, and devise a comprehensive, non-hypothesis-driven screen for T-cell epitopes. We were interested in the study of celiac disease (CD) and used the ELISPOT technique to perform a large number of T-cell assays. We therefore needed to compensate for the lack of statistical data analysis methods for ELISPOT assays.

Results: We describe a method to comprehensively screen for T-cell epitopes within a family or a group of proteins. We have implemented an algorithm to generate a set of unique short peptide sequences that incorporate all possible epitopes within a group of proteins. T-cell assays were performed in 96-well plates using the ELISPOT assay to screen for responses in CD patients against any epitopes in gluteins. We describe a statistical model to fit the data and an Expectation Maximization algorithm to estimate the parameters of interest and analyze the resulting data.

Availability: Implementations of our algorithms in R or Perl are available at <http://bioinf.wehi.edu.au/folders/immunol>.

Contact: beissbarth@wehi.edu.au

1 INTRODUCTION

The human immune system is designed to recognize, eradicate and ‘remember’ invading pathogens. T-cells play a critical role in sensing and triggering immune responses. Every individual has a pool of T-cells, each with a distinct T-cell receptor relatively specific for certain 9-amino acid long sequences bound to specialized cell surface proteins, MHC Class I for viral and intracellular proteins or MHC Class II on specialized antigen-presenting cells (APC) for extracellular

proteins (Townsend *et al.*, 1986). A formidable challenge of human immunology is the identification of peptides (epitopes) that activate T-cells causing protective or destructive immune responses that could be used to develop protective vaccines or to induce tolerance to antigens causing inappropriate immune responses, e.g. autoimmune diseases (Elsawa *et al.*, 2004).

The simplest and earliest approach was to incubate peripheral blood T-cells with overlapping peptides (15–20mers overlapping by 10–12 amino acids) spanning a protein of interest and measure proliferation of T-cells (Cease *et al.*, 1987). More recently, approaches such as combinatorial peptide libraries have been used to address specificity of T-cells when there is little knowledge of the antigen (Sospedra *et al.*, 2003) or to elute peptides from specific MHC molecules following incubation of an APC with an antigen or an organism of interest (Lemmel and Stevanovic, 2003).

Since peptides are expensive and therefore may limit the number of proteins that can be tested, several algorithmic attempts to predict T-cell epitopes have been suggested (Bian *et al.*, 2003). Most of these methods can be described as machine learning algorithms relying on the knowledge of peptides previously found to bind to MHC or to contribute T-cell epitopes, and they try to find patterns within those sequences and subsequently use these patterns to predict novel epitopes (Srinivasan *et al.*, 2004; Jung *et al.*, 2001; Vader *et al.*, 2002a,b). However, these methods are limited, by the availability of data on previously determined epitopes.

Recent advances have allowed us to considerably scale up the number of candidate epitopes that can be tested using T-cell assays (Anderson *et al.*, 2000). The use of the ELISPOT reader to count spots in 96 well plates has been essential to provide high-throughput, reproducible data (Janetzki *et al.*, 2004), and has facilitated increasingly large-scale testing of T-cell epitopes (Anthony *et al.*, 2002; Keating *et al.*, 2002). However, as Hobeika *et al.* (2005) pointed out, there is no common agreement on what can be called a positive response. To our knowledge, no statistical method to analyze such data has been proposed so far.

We have generated data to study the T-cell epitopes responsible for celiac disease (CD). CD is a T-cell mediated disease

*To whom correspondence should be addressed.

caused by gluten proteins from wheat, rye and barley in many foods (Dicke *et al.*, 1953). CD affects 1% of Western populations and is over 90% associated with a specific MHC Class II molecule, HLA-DQ2. T-cells specific for gluten peptides presented by HLA-DQ2 are present in intestinal tissues. CD is also heavily dependent on deamidation [i.e. glutamine (Q) being modified to glutamate (E)] of glutens by the enzyme transglutaminase (Sollid, 2000). Epitopes derived from gluten have been identified using T-cell clones and chromatographically separated gluten hydrolysates or gluten peptides spanning a restricted number of gluten proteins (Arentz-Hansen *et al.*, 2000; van de Wal *et al.*, 1998). Such techniques are limited by the molecular complexity of gluten, whereas T-cell clones are susceptible to *in vitro* artifacts and may not be representative or quantitative measures of *in vivo* T-cell responses. In June 2003, GENBANK contained 407 'gluten' genes encoded by *Triticum aestivum* (bread-making wheat), rye, barley and oats.

We sought to take advantage of the falling cost of synthetic, screening grade peptides, and the opportunity to use gluten challenge in CD to develop an unambiguous and comprehensive T-cell epitope mapping technique suitable for human infectious and immune diseases when large numbers of potential epitopes derived from hundreds of potentially related proteins need to be tested in a large cohort of subjects. Our long-term goal is a tolerogenic therapy for CD based on gluten T-cell epitopes. However, strategies developed for CD may also be applicable to T-cell epitope mapping for viral diseases, such as HIV or HCV, where hundreds of related proteins include potential T-cell epitopes.

Here we focus on the algorithmic and data analysis component of the method. We propose a method to comprehensively screen all peptide sequences of a protein family for potential epitopes. We describe an algorithm to select short 20mer sequences that contain all-unique 12mer sequences in the protein family. We further present a statistical model and method to subsequently analyze the data resulting from a large number of T-cell assays. We show the performance of our suggested methods on large datasets generated for the study of CD. We have designed peptide libraries to screen for possible T-cell epitopes in gluten proteins. We have subsequently developed, tested and applied our method to statistically analyze the 140 000 T-cell assays, currently in our database, resulting from our studies.

2 METHODS

The following steps were performed: (1) Select families of gluten proteins from GENBANK. (2) Select 20mer peptides to screen for epitopes. (3) Synthesize screening grade peptides for all selected 20mers. (4) Recruit CD patients and perform gluten challenge. (5) Perform ELISPOT T-cell assays using the blood from CD patients and the synthesized peptides in 96-well plates. (6) Analyze data from T-cell assays. (7) Based on previously found positive 20mer peptides design

shorter peptides, synthesize them in screening grade, perform T-cell assays and analyze the data. (8) Based on previously confirmed positive peptides, group similar peptides into families of epitopes, synthesize high-grade peptides and use them for further tests. Here, we describe steps (1)–(6), focusing especially on steps (2) and (6).

We have previously shown that gluten peptide-specific T-cells are undetectable in CD blood, but CD subjects who normally follow a gluten-free diet, have gluten-specific T-cells in their blood 6 days after eating gluten. Such T-cells are readily detected using a quantitative assay (ELISPOT) that enumerates antigen-specific T-cells stimulated to secrete interferon-gamma. ELISPOT is now the most sensitive assay for detection of antigen/peptide specific T-cells. The readout of the ELISPOT assay is numerical, i.e. the number of T-cells specific for a particular candidate epitope that secrete the cytokine of interest (typically gamma-interferon) present in a sample of blood. Since gluten contains many thousands of potential epitopes and epitope mapping using T-cell clones is limited by technical and theoretical uncertainties, no study has attempted or claimed to have comprehensively identified all gluten epitopes. However, this is a critical step for development of tolerogenic therapies for CD.

T-cell assays are used to measure how many T-cells react against a certain peptide epitope. Blood is taken from a patient, and the peripheral blood mononuclear cells are isolated from the whole blood. 200 000–1 000 000 blood cells are then incubated with the peptide epitope that is to be tested. Figure 1 illustrates the process. The assay is carried out in a 96-well plate format and the resulting spots, which indicate the responding T-cells, are photographed and counted using an ELISPOT reader.

Here we use this assay to determine against which parts of the gluten proteins in wheat, rye, barley or oats people having CD react, i.e. which peptides of these proteins are T-cell stimulating epitopes. CD patients who had followed a gluten-free diet for at least 6 months then ate defined amounts of wheat or rye flour, barley grains, or oat flakes for 3 days. Three days after finishing the gluten 'challenge', blood was collected from subjects and used to perform between 500 and 2000 T-cells assay with different peptides.

We screened 20mer peptides containing all-unique 12mer sequences occurring within all families of gluten proteins, i.e. gliadins, glutenins, avenins, hordeins and secalins. T-cell epitopes are expected to have a length of 9–12 amino acids. The number of all-unique 12mer peptides would be too large, i.e. >15 000 and impractical for testing. Therefore, we use a set of 20mer peptides that incorporates all 12mer peptides (and therefore all 9mers as well) in glutens, exploiting the opportunity that each 20mer peptide can cover up to nine different 12mers.

We use an iterative algorithm to compute such a set of peptides (Fig. 2). The algorithm starts with the protein sequences of a group of unaligned proteins from a protein family and

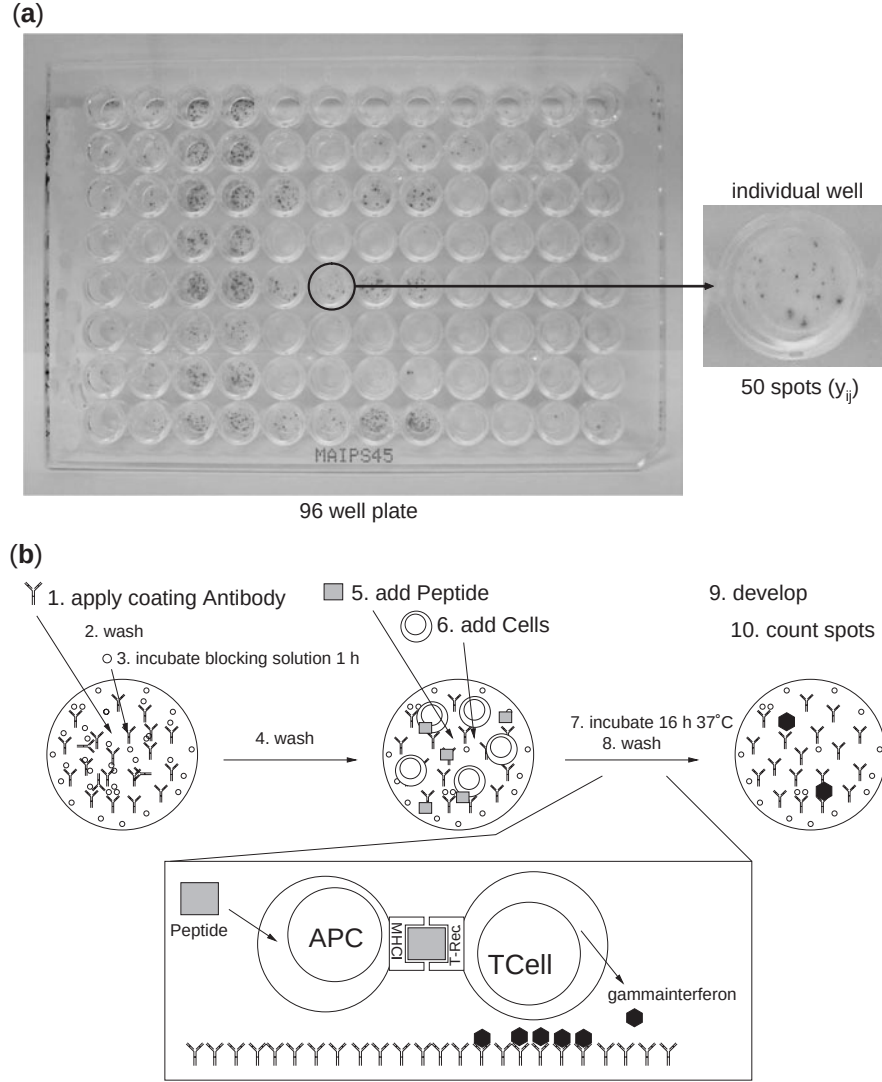


Fig. 1. T-cell assay scheme. (a) T-cell assays are performed in 96-well plate format and spots are counted using an ELISPOT reader. (b) shows a schematic representation of the assay. In each well blood cells of a patient are incubated with a different peptide and the T-cell response is measured.

generates all unique overlapping 20mer and 12mer peptides. It selects 20mers until a set is selected that covers all unique 12mers. The lengths of the peptides can be adjusted in the algorithm and the results can be checked in a sequence alignment viewer.

3 STATISTICS

The data generated as a result of T-cell epitope screening are counts of spots, representing the number of T-cells of a certain patient responding to a given peptide. We do not know, however, whether the patient is responding to the given antigen or whether a spot count could be attributed to a non-specific background response. Ideally, we want to find out how many people respond to each peptide and what is the expected response rate

(i.e. count of spots) for the patients who do respond to this peptide. We also have to include some normalization between patients, as we observe that different patients respond with quite different overall response rates.

We model this as an incomplete data problem. Our observed data y_{ij} are the counts of spots for the different patients i and peptides j , which we get from the ELISPOT reader as a result of our T-cell assays. Whether a patient responds or not is unknown at this stage. We assume a response indicator z_{ij} that is either 1, indicating that patient i responds to peptide j , or 0.

We want to estimate the following parameters (θ):

- α_i : factor indicating the overall responsiveness of patient i .

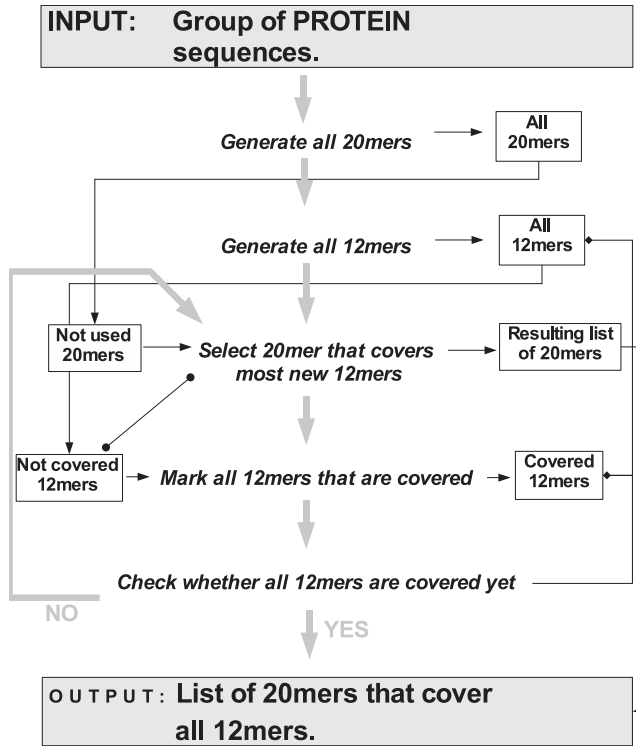


Fig. 2. Diagrammatic representation of the algorithm used to select a set of 20mer peptides that incorporates all existing 12mer peptide variants within a group of proteins.

- p_j : proportion of people responding to peptide j .
- λ_j : rate of response for patients responding to peptide j .
- λ_0 : rate of response for patients not responding.

We model the observed data using independent Poisson distributions: $y_{ij} \sim \text{Poisson}(\alpha_i \lambda_j)$, if patient i is responding to peptide j , i.e. $z_{ij} = 1$ or $y_{ij} \sim \text{Poisson}(\alpha_i \lambda_0)$, if the patient is not responding, $z_{ij} = 0$. The complete data likelihood is thus:

$$L_c(\theta) = \prod_i \left\{ \prod_j p_j^{z_{ij}} (1 - p_j)^{1-z_{ij}} \frac{\lambda_j^{y_{ij}} e^{-\lambda_j}}{y_{ij}!} \right\}$$

with

$$\lambda_{ij} = z_{ij} \alpha_i \lambda_j + (1 - z_{ij}) \alpha_i \lambda_0$$

Given a set of parameters, we can then estimate expected values for z_{ij} using the probability (pr) distributions:

$$\begin{aligned} \hat{z}_{ij} &= \text{pr}(z_{ij} = 1 | y_{ij}) = \frac{\text{pr}(z_{ij} = 1) \text{pr}(y_{ij} | z_{ij} = 1)}{\text{pr}(y_{ij})} \\ &= \frac{p_j e^{-\alpha_i \lambda_j} (\alpha_i \lambda_j)^{y_{ij}}}{p_j e^{-\alpha_i \lambda_j} (\alpha_i \lambda_j)^{y_{ij}} + (1 - p_j) e^{-\alpha_i \lambda_0} (\alpha_i \lambda_0)^{y_{ij}}} \end{aligned}$$

We are using the Expectation Maximization (EM) framework (Dempster *et al.*, 1977) to estimate the parameters and unobserved values. The EM algorithm cycles between assuming

current parameters and computing expected values $\hat{z}_{ij}$, and maximizing the complete data likelihood L_c using the expected values $\hat{z}_{ij}$ and recomputing new parameters. The algorithm starts with initial guesses for the parameters and cycles until convergence (it usually converges very fast) or for a fixed number of cycles.

Maximizing L_c results in the following new $(c+1)$ estimates for the parameters, given the current (c) estimates of $\hat{z}_{ij}$:

$$\begin{aligned} p_j^{(c+1)} &= \frac{\sum_i \hat{z}_{ij}^{(c)}}{\text{number_of_patients}}, \\ \alpha_i^{(c+1)} &= \frac{\sum_j [\hat{z}_{ij}^{(c)} y_{ij} + (1 - \hat{z}_{ij}^{(c)}) y_{ij}]}{\sum_j [\lambda_j^{(c)} \hat{z}_{ij}^{(c)} + \lambda_0^{(c)} (1 - \hat{z}_{ij}^{(c)})]}, \\ \lambda_j^{(c+1)} &= \frac{\sum_i \hat{z}_{ij}^{(c)} y_{ij}}{\sum_i \alpha_i^{(c)} \hat{z}_{ij}^{(c)}}, \\ \lambda_0^{(c+1)} &= \frac{\sum_{ij} (1 - \hat{z}_{ij}^{(c)}) y_{ij}}{\sum_{ij} \alpha_i^{(c)} (1 - \hat{z}_{ij}^{(c)})}. \end{aligned}$$

When maximizing the likelihood we observe two problems. First, it is not possible to estimate the α s and λ s independently, as α_i appears in both the background distribution as well as in the response distribution. To fix this we adjust the α s after each iteration to have a mean of 1. Second, the algorithm tends to fit the majority of the data, which in our case is usually the background, very well, and therefore uses λ_j for the background rates and λ_0 as a global response rate. To prevent this from happening, we need to put in the constraint that $\lambda_j \geq \lambda_0$. We can show that this can be done iteratively in each EM cycle, i.e. we compute initial estimates for $\lambda_0^{(c+1)}$ and $\lambda_j^{(c+1)}$ as above and then iteratively (we used a fixed number of three iterations) recompute new values for λ_0 as follows:

$$\lambda_0^{(c+1)} = \text{iterate}_{k=0, \dots, 3} \left\{ \frac{\sum_{ij} [(1 - \hat{z}_{ij}^{(c)}) y_{ij}] + \sum_{\lambda_j^{(c+1)} \leq \lambda_0^{(c+1)}} \sum_i [\hat{z}_{ij}^{(c)} y_{ij}]}{\sum_{ij} \alpha_i^{(c)} (1 - \hat{z}_{ij}^{(c)}) + \sum_{\lambda_j^{(c+1)} \leq \lambda_0^{(c+1)}} \sum_i \alpha_i^{(c)} \hat{z}_{ij}^{(c)}} \right\}.$$

Afterwards λ_j is updated:

$$\lambda_j^{(c+1)} = \begin{cases} \lambda_j^{(c+1)}, & \text{if } \lambda_j^{(c+1)} > \lambda_0^{(c+1)} \\ \lambda_0^{(c+1)}, & \text{if } \lambda_j^{(c+1)} \leq \lambda_0^{(c+1)}. \end{cases}$$

4 RESULTS

We have applied and tested these methods in a comprehensive study to determine the T-cell epitopes involved in CD. Patients with CD react against proteins in wheat (gliadins or glutenins), rye (secalins), barley (hordeins) or oats (avenins). The database entries from GENBANK for each of these protein families were collected and 20mer peptides were designed for

Table 1. Summary of peptides designed and tested contributing distinct potential epitopes within different families of proteins

Protein family	No. of sequence	No. of amino acids	No. of unique kmers		Generated 20mers
			20	12	
Gliadins (wheat)	107	28 124	5805	4465	721
LMW Glutenins	77	21 554	5489	3945	645
HMW Glutenins	52	26 059	6979	4799	786
Avenins (oats)	18	3413	1366	1279	199
Hordeins (barley)	59	13 792	3022	2672	416
Secalins (rye)	14	2033	1080	957	155

Generated 20mer peptides cover all-unique 12mers.

testing as described in Section 2. As can be seen in Table 1, there is considerable redundancy in 12mer sequences within the set of unique 20mers. The selection of 20mers by our algorithm reduces this redundancy and produces a list of peptides that is practical in cost and lab-work to be assayed. The resulting peptides were subsequently synthesized and tested in our study.

In the following text, some of our experimental findings are summarized to illustrate the feasibility of the model and to show that our methods and algorithms are able to produce useful results. Currently, our database stores about 140 000 T-cell assays performed on more than 200 patients and almost 5000 different peptide antigens. The full data of this study and the biological findings and applications will be submitted in a biomedical journal.

To test how well our Poisson assumption fits the data and how well we fit the parameters, we have designed several control experiments. Among them we have included repeated measurements for several peptides. The deamidated wheat gliadin peptide P04722 77-93 Q → E85 ('P04722E', QLQFPQPPELPYPQPQP) has previously been shown (Anderson *et al.*, 2000) to be a very potent peptide causing strong T-cell responses, and has been utilized as a positive control in our study. Furthermore, all 20mers tested in our study containing the motif PFPQP[QE]LPY show consistently high responses and are predicted by our algorithms to be among the most highly reactive epitopes. P04722E has an estimated response rate of 90.3 in our database with an estimate of 78% of all patients responding. We performed 19 replicate measurements of P04722E in each of 12 patients and the results are shown in Figure 3.

We observe that our model and estimates fit the data well, although some over-dispersion is apparent. Some degree of difference will be expected, as the parameters are fitted on all measurements in the database, i.e. the patient effects $\alpha_1, \dots, \alpha_{12}$ are derived from 721 measurements for each patient and the peptide effect $\lambda_{P04722E}$ from 364 measurements for the peptide. The estimated response indicator z_{ij} distinguishes between measurements that seem to follow background rate and those which are true responses, the

estimated value of z_{ij} is usually 1.0 or very close for responses. It gets more difficult to decide for weak responders, especially when there are no observed responses for a peptide, i.e. $\lambda_j = \lambda_0$ in these cases the estimated z_{ij} is usually between 0.2 and 0.5. In the example above, the estimated response rate for the peptide ($\lambda_{P04722E} = 90.3$) is the same for all patients, we see that patient effect α_i is able to normalize between the different strengths of response in the different people.

Figure 4 shows a normal Q-Q plot of the residuals, for the measurements described above. These residuals were computed based on the deviances calculated as described in McCullagh and Nelder (1989). We define the deviance residuals as:

$$r_{ij} = \text{sign}[y_{ij} - \hat{E}(y_{ij})] \sqrt{y_{ij} \log \frac{y_{ij}}{\hat{E}(y_{ij})} - [y_{ij} - \hat{E}(y_{ij})]}$$

$$\hat{E}(y_{ij}) = \hat{\alpha}_i [\hat{\lambda}_j \hat{z}_{ij} + \hat{\lambda}_0 (1 - \hat{z}_{ij})]$$

If the model fits well the residuals r_{ij} should be normally distributed. The mean of the deviances (r_{ij}^2) is a measure of how well the observed variances match to the Poisson variance. Its value should be ideally 1. We observe a mean deviance of 4.8 for the data shown in Figure 4, and 2.1 for all measurements in the database, meaning that the distribution is somewhat over-dispersed. We note that the variance of measurements with very high spot counts is often higher than the theoretical Poisson variance—this is consistent with mean versus variance plots we did with other repeatedly measured peptides as well. This is not unexpected, as with very high spot counts the ELISPOT reader is unable to determine the number of spots accurately. It is not overly concerning, though, as for high spot counts the background and response distribution are easily distinguished. Also the model should still work satisfactorily, even if the variances are a multiple of the estimated Poisson variances.

Figure 5 shows the distributions of parameters we have estimated for the measurements performed on the 20mers summarized in Table 1. Responsive peptides were selected for further analysis. In order to select only peptides sufficiently different from non-responsive peptides we chose cutoffs based on the estimated response rate λ_j and the estimated number of people responding. We further analyzed peptides with a mean response rate >10 , which is well above background level, and we required that at least 3 people respond (with $z_{ij} > 0.8$) in order to reduce outliers. The product of the estimated response rate and proportion of responders $\lambda_j p_j$ is a good indicator of how potent a peptide is as an epitope.

In the subsequent analysis fine mapping was performed using shorter peptides that were derived from responsive 20mers. These peptides were synthesized as high-grade peptides for further testing. T-cell assays performed with high-grade peptides were analyzed in the same manner to estimate the response rates of the peptides and proportions of responders.

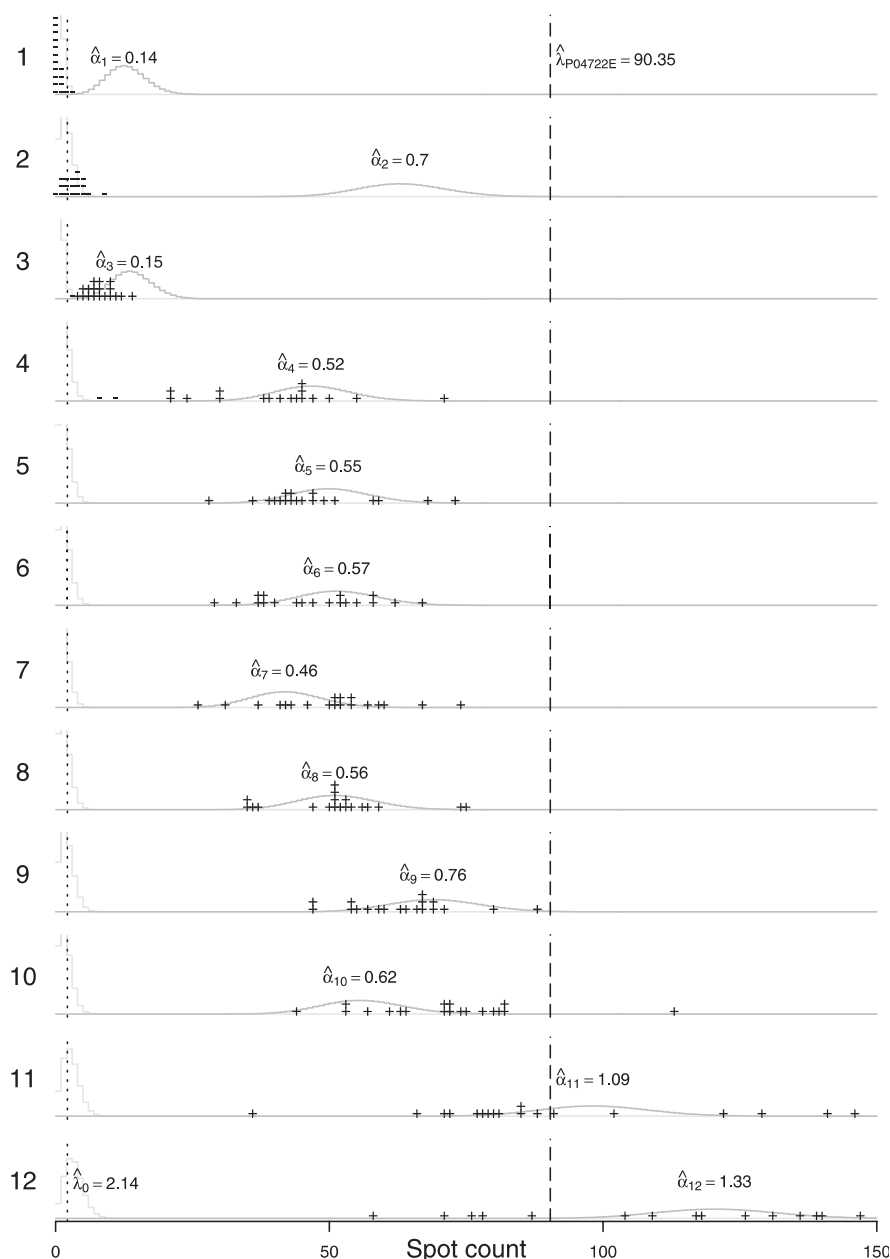


Fig. 3. Peptide P04722E was tested in 12 patients, each having 19 replicates. The light gray line (at the left of each row) shows the estimated background distribution [$\text{Poisson}(\alpha_i \lambda_0)$], the dark gray line show the estimated response distribution [$\text{Poisson}(\alpha_i \lambda_j)$] for each patient. The estimated response rate for the peptide λ_{P04722E} is indicated by a dashed vertical line and the estimated background response λ_0 by a dotted line. Measurements where the estimated indicator [$z_{ij} > 0.8$] indicates a response are shown as '+', other measurements are shown as '-'.

As an example, we compare the results of our screen on the gliadin data with results from literature. Gliadins are the group of gluten proteins best studied for CD epitopes. We screened 721 20mer peptides designed from gliadin proteins of each of the 26 patients. Of the 721 20mer peptides we selected 217 14mer peptides for second-round testing. These 14mers were designed from responsive 20mers either based

on the knowledge of the DQ2 binding motif or as overlapping peptides. Table 2 summarizes how our methods performed in selecting previously published epitopes suggested to play a role in CD (Arentz-Hansen *et al.*, 2000; Vader *et al.*, 2002a,b, 2003). All the known epitopes were found using our method and our cutoff criteria. In addition many new epitopes were found.

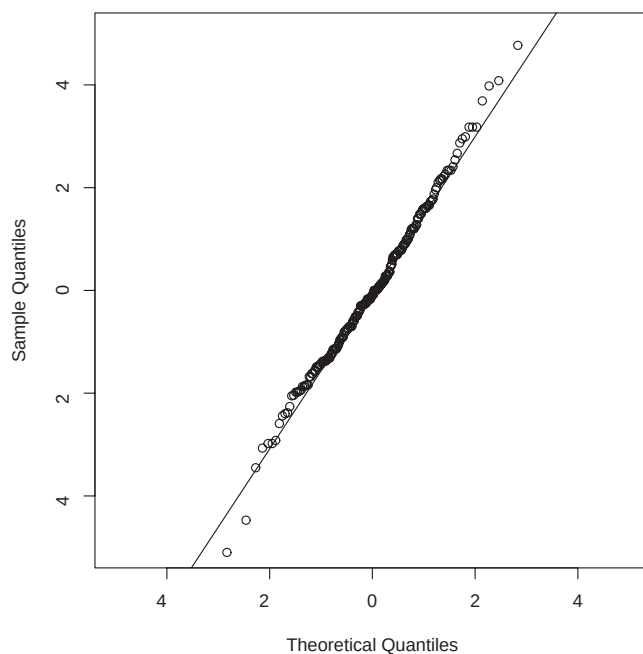


Fig. 4. Normal Q–Q plot of estimated residuals. For the 364 measurements on P04722E the deviance residuals, i.e. amount of deviation between observed and expected values, are computed, and the quantiles are compared with the theoretical quantiles from a normal distribution.

Table 2. Summary of epitopes suggested to play a role in CD and results in our screening

Known epitope	Tested 20mer peptides		Avg. of estimated parameters	
	No. tested	No. selected	$\bar{\lambda}$	$\bar{p}$
PFPQPQLPY	6	6	49.1	0.40
PQPQLPYQP	4	4	83.5	0.44
PYPQPQLPY	1	1	91.6	0.46
PQSFPPQQ	6	4	11.3	0.32
PFSQQQPPV	3	1	47.9	0.12
IIQPQQPAQ	5	1	26.7	0.18

20mer peptides were selected for additional testing if $\lambda_j \geq 10$ or at least three people having $z_{ij} > 0.8$ ($\approx p \geq 0.15$).

5 DISCUSSION

We have proposed a novel methodology to study immune responses, including a comprehensive screen for T-cell epitopes and subsequent statistical data analysis. This amount of peptide- and patient-specific data has not been available in previous T-cell epitope mapping studies of human immune-mediated diseases. These methods were designed and tested in CD. We believe, however, that similar methodology can be applied as well in other diseases where fairly large-scale ELISPOT studies have already been performed,

e.g. HIV (Keating *et al.*, 2002), hepatitis (Anthony *et al.*, 2002), diabetes (Meierhoff *et al.*, 2002) and cancer (Hobeika *et al.*, 2005).

Methods for T-cell epitope mapping are tailored to the disease of interest according to what is known about the pathogenic antigen/s and the availability of relevant T-cells. ‘Antigen challenge’ occurs whenever new pathogens are encountered or reencountered, food antigens consumed or organs grafted after autoimmune destruction (e.g. pancreatic islet transplantation for type-1 diabetes). Hence, T-cells may be available in peripheral blood in a number of clinical situations that would be amenable to large-scale systematic analysis by highly sensitive assays such as ELISPOT.

Availability of large databases, such as GENBANK, offers the opportunity for systematic rather than serendipitous selection of candidate protein sequences for epitope mapping. In T-cell epitope mapping, few investigators have adopted methods not at philosophical extremes, either hypothesis driven (one polymorphism of one candidate protein) or non-hypothesis driven (combinatorial peptide libraries). Often researchers have relied upon T-cell clones because of the rarity of relevant T-cells, and have consequently generated data that are categorical—a clone specific for a peptide is either present or absent in an individual with or without the disease of interest.

Gluten challenge in CD generates substantial numbers of gluten-specific T-cells in peripheral blood that can be assayed in a quantitative fashion by the interferon-gamma ELISPOT assay. Systematic design of 20mer peptides including all-unique 12mers encoded by gluten genes entered in GENBANK produced a peptide library that could be practically assessed in cohorts of 20 or more volunteers after eating the physiologically relevant grain (wheat, barley, rye or oats). During one year, a single laboratory worker was able to perform all the assays required from over 150 such gluten-challenge experiments and generate over 140 000 data points. Putative epitopes which were found in our study are undergoing further testing, and are being used to design diagnostics, vaccines and non-toxic wheat.

There may be several factors which make our methods especially suitable for CD research. Since the families of gluten protein show high homology and significant repeats, we could limit the number of peptides to be tested. CD being correlated strongly with the MHC Class II variants DQ2 and DQ8, we expect a relatively limited number of different responses with high consistency among different patients.

Certain issues downstream in the analysis have not been addressed here, e.g. clustering structurally related responding peptides into families and determining the non-redundant 9mer epitopes from the tested 20mer sequences. Occasionally, responding 20mers included several overlapping epitopes, and we observed that their effects are generally additive. These problems will be addressed elsewhere.

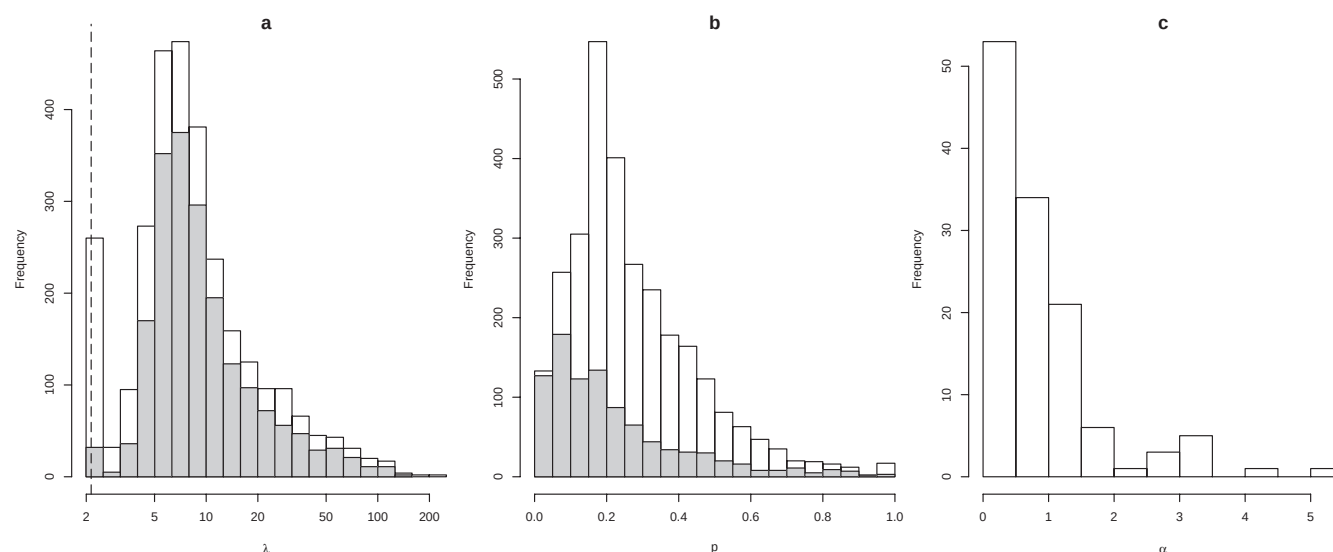


Fig. 5. Distributions of estimated parameters from T-cell assays in our database. **(a)** Distribution of estimated peptide responses λ_j from 2922 peptides. The gray part of the bars indicates the number of peptides where responses ($z_{ij} > 0.8$) were observed in at least three different patients. The dashed line indicates λ_0 . **(b)** Distribution of proportions p_j of responding people to each peptide. The gray part of the bars indicates the number of peptides with a response rate > 10 . **(c)** Distribution of responsiveness α_l from 125 patient samples.

6 CONCLUSION

In our study we have presented methods to comprehensively screen for potential T-cell epitopes in groups or families of related proteins. We have introduced an algorithm that assists in selecting a comprehensive, non-redundant set of peptides to be tested. We have further introduced a statistical model to analyze ELISPOT T-cell assay data. Using this model we were able to distinguish what responses are likely to be significantly different from background. Furthermore, we were able to estimate parameters, such as the expected response rate for each tested peptide and the proportion of people showing positive responses to each peptide. We have shown that these methods were useful in CD research, and we believe that they will be useful in other areas of disease-related research, as well.

ACKNOWLEDGEMENTS

We thank Pratyaksha Wirapati and Matthew Ritchie for help and suggestions, the WEHI IT department for IT support and Mark Stewart for technical assistance. This work was supported by BTG International Plc. R.P.A. holds the DW Keir Fellowship (Melbourne Health) and Lions Cancer Council of Victoria Fellowship. J.A.T.-D. was supported by an NHMRC Postgraduate Scholarship. T.B. was supported by a fellowship of the Deutsche Forschungsgemeinschaft and the WEHI NHMRC Transitional Institute Grant 215499.

REFERENCES

- Anderson,R., Degano,P., Godkin,A., Jewell,D. and Hill,A. (2000) *In vivo* antigen challenge in celiac disease identifies a single

transglutaminase-modified peptide as the dominant A-gliadin T-cell epitope. *Nat. Med.*, **6**, 337–342.

Anthony,D.D., Valdez,H., Post,A.B., Carlson,N.L., Heeger,P.S. and Lehmann,P.V. (2002) Comprehensive determinant mapping of the hepatitis C-specific CD8 cell repertoire reveals unpredicted immune hierarchy. *Clin. Immunol.*, **103**, 264–276.

Arentz-Hansen,H., Körner,R., Molberg,O., Quarsten,H., Vader,W., Kooy,Y., Lundin,K., Koning,F., Roepstorff,P., Sollid,L. and McAdam,S. (2000) The intestinal T cell response to alpha-gliadin in adult celiac disease is focused on a single deamidated glutamine targeted by tissue transglutaminase. *J. Exp. Med.*, **191**, 603–612.

Beißbarth,T., Sun,J., Kavathas,P. and Ortmann,B. (2000) Increased efficiency of folding and peptide loading of mutant MHC class I molecules. *Eur. J. Immunol.*, **30**, 1203–1213.

Bian,H., Reidhaar-Olson,J.F. and Hammer,J. (2003) The use of bioinformatics for identifying class II-restricted T-cell epitopes. *Methods*, **29**, 299–309.

Cease,K., Margalit,H., Cornette,J., Putney,S., Robey,W., Ouyang,C., Streicher,H., Fischinger,P., Gallo,R. and DeLisi,C. (1987) Helper T-cell antigenic site identification in the acquired immunodeficiency syndrome virus gp120 envelope protein and induction of immunity in mice to the native protein using a 16-residue synthetic peptide. *Proc. Natl Acad. Sci. USA*, **84**, 4249–4253.

Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data using the EM algorithm. *J. R. Statist. Soc.*, **39**, 1–38.

Dicke,W., Weijers,H. and van de Kamer,J. (1953) Coeliac disease. II. The presence in wheat of a factor having a deleterious effect in cases of coeliac disease. *Acta Paediatr.*, **42**, 34–42.

Elsawa,S.F., Rodeberg,D.A. and Celis,E. (2004) T-cell epitope peptide vaccines. *Expert. Rev. Vaccines*, **3**, 563–575.

- Hobeika,A.C., Morse,M.A., Osada,T., Ghanayem,M., Niedzwiecki,D., Barrier,R., Lyster,H.K. and Clay,T.M. (2005) Enumerating antigen-specific T-Cell Responses in peripheral blood: a comparison of peptide MHC tetramer, ELISpot, and intracellular cytokine analysis. *J. Immunother.*, **28**, 63–72.
- Janetzki,S., Schaed,S., Blachere,N.E.B., Ben-Porat,L., Houghton,A. and Panageas,K. (2004) Evaluation of Elispot assays: influence of method and operator on variability of results. *J. Immunol. Meth.*, **291**, 175–183.
- Jung,G., Fleckenstein,B., von der Mülbe,F., Wessels,J., Niethammer,D. and Wiesmüller,K. (2001) From combinatorial libraries to MHC ligand motifs, T-cell superagonists and antagonists. *Biologicals*, **29**, 179–181.
- Keating,S.M., Bollinger,R.C., Quinn,T.C., Jackson,J.B. and Carruth,L.M. (2002) Cross-clade T lymphocyte-mediated immunity to HIV type 1: implications for vaccine design and immunodetection assays. *AIDS Res. Hum. Retroviruses*, **18**, 1067–1079.
- Lemmel,C. and Stevanovic,S. (2003) The use of HPLC-MS in T-cell epitope identification. *Methods*, **29**, 248–259.
- McCullagh,P. and Nelder,J. (1989) *Generalized Linear Models*, 2nd edn. Chapman and Hall/CRC, London, Chapter Log-linear models, pp. 193–213.
- Meierhoff,G., Ott,P.A., Lehmann,P.V. and Schloot,N.C. (2002) Cytokine detection by ELISPOT: relevance for immunological studies in type 1 diabetes. *Diabetes Metab. Res. Rev.*, **18**, 367–380.
- Sollid,L. (2000) Molecular basis of celiac disease. *Annu. Rev. Immunol.*, **18**, 53–81.
- Sospedra,M., Pinilla,C. and Martin,R. (2003) Use of combinatorial peptide libraries for T-cell epitope mapping. *Methods*, **29**, 236–247.
- Srinivasan,K., Zhang,G., Khan,A., August,J. and Brusic,V. (2004) Prediction of class I T-cell epitopes: evidence of presence of immunological hot spots inside antigens. *Bioinformatics*, **20** (Suppl. 1), I297–I302.
- Townsend,A., Rothbard,J., Gotch,F., Bahadur,G., Wraith,D. and McMichael,A. (1986) The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides. *Cell*, **44**, 959–968.
- Vader,L.W., de Ru,A., van der Wal,Y., Kooy,Y.M.C., Benckhuijsen,W., Mearin,M.L., Drijfhout,J.W., van Veelen,P. and Koning,F. (2002a) Specificity of tissue transglutaminase explains cereal toxicity in celiac disease. *J. Exp. Med.*, **195**, 643–649.
- Vader,W., Kooy,Y., Veelen,P.V., Ru,A.D., Harris,D., Benckhuijsen,W., Peña,S., Mearin,L., Drijfhout,J.W. and Koning,F. (2002b) The gluten response in children with celiac disease is directed toward multiple gliadin and glutenin peptides. *Gastroenterology*, **122**, 1729–1737.
- Vader,W., Stepniak,D., Kooy,Y., Mearin,L., Thompson,A., van Rood,J.J., Spaenij,L. and Koning,F. (2003) The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses. *Proc. Natl Acad. Sci. USA*, **100**, 12390–12395.
- van de Wal,Y., Kooy,Y., van Veelen,P., Peña,S., Mearin,L., Papadopoulos,G. and Koning,F. (1998) Selective deamidation by tissue transglutaminase strongly enhances gliadin-specific T cell reactivity. *J. Immunol.*, **161**, 1585–1588.