Published in final edited form as:

Cancer Res. 2019 April 01; 79(7): 1671–1680. doi:10.1158/0008-5472.CAN-18-2292.

Biophysicochemical Motifs in T cell Receptor Sequences Distinguish Repertoires from Tumor Infiltrating Lymphocyte and Adjacent Healthy Tissue

Jared Ostmeyer¹, Scott Christley², Inimary T. Toby³, and Lindsay G. Cowell⁴

¹Department of Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9066, jared.ostmeyer@utsouthwestern.edu

²Department of Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9066, scott.christley@utsouthwestern.edu

³Department of Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9066, inimary.toby@utsouthwestern.edu

⁴Department of Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9066

Abstract

Immune repertoire deep sequencing allows comprehensive characterization of antigen receptor-encoding genes in a lymphocyte population. We hypothesized that this method could enable a novel approach to diagnose disease by identifying antigen receptor sequence patterns associated with clinical phenotypes. In this study, we developed statistical classifiers of T cell receptor (TCR) repertoires that distinguish tumor tissue from patient-matched healthy tissue of the same organ. The basis of both classifiers was a biophysicochemical motif in the complementarity determining region 3 (CDR3) of TCR β chains. To develop each classifier, we extracted 4-mers from every TCR β CDR3 and represented each 4-mer using biophysicochemical features of its amino acid sequence combined with quantification of 4-mer (or receptor) abundance. This representation was scored using a logistic regression model. Unlike typical logistic regression, the classifier is fitted and validated under the requirement that greater than or equal to 1 positively labeled 4-mer appears in every tumor repertoire and no positively labeled 4-mers appear in healthy tissue repertoires. We applied our method to publicly available data in which tumor and adjacent healthy tissue were collected from each patient. Using a patient-holdout cross-validation, our method achieved classification accuracy of 93% and 94% for colorectal and breast cancer, respectively. The parameter values for each classifier revealed distinct biophysicochemical properties for tumor-associated 4-mers within each cancer type. We propose that such motifs might be used to develop novel immune-based cancer screening assays.

Correspondence to: Lindsay G. Cowell.

Lindsay Grey Cowell, PhD, Department of Clinical Sciences, UT Southwestern Medical Center, 5323 Harry Hines Boulevard, Mail Code 9066, Dallas, TX 75390-9066, phone 214-648-2289, fax 214-648-3934, lindsay.cowell@utsouthwestern.edu.

Conflict of Interest Statement: A patent has been filed by the University of Texas Southwestern Medical Center covering the techniques described in our manuscript along with its application in diagnosing multiple sclerosis, colorectal cancer, breast cancer, and diseases in general. The authors declare no other potential conflicts of interest.

Significance

This study presents a novel computational approach to identify T cell repertoire differences between normal and tumor tissue.

Keywords

tumor-infiltrating lymphocytes; cancer biomarker; multiple instance learning; immune repertoire

Introduction

The immune system actively responds to solid tumors, resulting in tumor-infiltrating lymphocytes (TILs). Natural immune control is often unsuccessful, however, because the tumor microenvironment contains a mix of immune-activating and immune-suppressing signals [1]. But given the right environmental cues, cytotoxic T lymphocytes in the tumor have the capacity to mediate tumor cell killing in virtue of bearing antigen receptors, T-cell receptors (TCRs), with specificity for tumor-associated antigens [1, 2]. While there is tremendous heterogeneity between patients' antigen landscapes due to patient-specific tumor neoantigens, there is also overlap [2, 3]. Therefore, we reasoned that patients with the same cancer type or subtype may have cytotoxic T-cell responses against a common set of antigens. Indeed, there is evidence for shared immunoreactivity, as well as for shared TCR sequences [4–10]. We further reasoned that if these T-cell responses could be detected, particularly early in the disease course, they could serve as an important addition to the suite of methods under development for the early detection of cancer. As a first step in this direction, we designed this study to determine whether anti-tumor T-cell responses have a cancer-specific signature that can reliably distinguish cancer-associated repertoires from those associated with healthy tissue of the same organ.

We leveraged publicly available TCR deep sequencing data and the Multiple Instance Learning (MIL) machine learning framework. The genes encoding TCRs are somatically generated through a process that creates essentially unique gene sequences at the relevant loci [11]. This results in a tremendously diverse TCR repertoire, in which each TCR has its own distinct profile of antigens it can bind. Immune repertoire deep sequencing has made it possible to comprehensively profile the TCRs of a lymphocyte population and has been widely applied to TILs [12]. The technology has enabled novel approaches for diagnosing and prognosticating diseases with a driving immune component by identifying repertoire patterns associated with clinical phenotypes. Most studies have been purely descriptive and looked for shared amino acid sequences among patients with a common phenotype [7, 8], looked for clusters of sequences overrepresented in one phenotype relative to another [13], or compared repertoire-level summary statistics, such as diversity, between phenotypes (reviewed in [12, 14, 15]). In the latter case, these features have prognostic value for some cancers and therapies [16–18]. We are aware of only a handful of studies developing predictive models [19–23]. With the exception of our study in multiple sclerosis [19], the studies were in the context of distinguishing infected from uninfected individuals or immunized from unimmunized or adjuvant-only immunized individuals where large, mono- or oligoclonal lymphocyte expansion is expected.

Our approach relies on MIL, which provides a rigorous, established framework for relating immune repertoires to phenotypes. Most of the individual receptors in a person are unrelated to any specific phenotype and exist to maintain a diverse set of specificities as a contingency against any possible antigen. Only a small number of receptors are relevant to any specific phenotype. Mapping pertinent receptors in a repertoire to a single phenotype label can formally be described as MIL, which treats problems as bags of instances where the bags are labelled but the instances are not [24]. The receptors from a single repertoire can be thought of as the instances, the repertoire as the bag, and the phenotype as the label. The goal is to predict the phenotype from the receptors.

In this study, we applied MIL to publicly available TCR deep sequencing data from tumor and healthy tissue from patients with colorectal or breast cancer [25, 26]. The locus encoding the TCR β chain (TCR β) was sequenced in all samples. In order to capture features of the TCR β s' antigen binding capabilities, we represented each sequence using biophysicochemical features. We represented the somatically generated portion of each gene, which is the primary determinant of the antigen binding specificity encoded by the gene. We then developed a statistical classifier for each cancer type and obtained classification accuracies by leave-one-out-cross-validation of 93% and 94% for colorectal and breast cancer, respectively. Permutation analyses resulted in classification accuracies of 49% for both data sets. These results demonstrate distinct, biophysicochemical motifs in the TCR β sequences of TILs that are specific to the cancer type and reliably distinguish cancer-associated repertoires from those associated with healthy tissue of the same organ.

Materials and Methods

Datasets:

We used publicly available TCR β deep sequencing data from 14 colorectal cancer patients (Table 1) [26] and 16 breast cancer patients (Table 2) [25]. In both original studies, tumor and adjacent healthy tissue was biopsied from each patient, and genomic DNA was extracted and sent to Adaptive Biotechnologies for sequencing using a proprietary technology that provides accurate measurements of each receptor sequence's abundance [27]. The sequences can be downloaded from the immuneACCESS database (<https://doi.org/10.21417/B7PP46>; <https://doi.org/10.21417/B7NW5B>) [28].

Representing TCRs:

We utilized a representation of TCR β sequence that captures features relevant to its antigen binding capabilities. We focused on complementarity determining region 3 (CDR3), because it is the somatically generated portion of the gene and the primary determinant of antigen binding specificity. CDR3 residues that directly contact peptide in a peptide-MHC complex are expected to make the largest contribution to a TCR's antigen binding specificity. To determine which TCR β CDR3 residues contact peptide, we analyzed X-ray crystallographic structures of human TCRs bound to peptide-MHC complex (Figure 1a) obtained from the Protein Data Bank (PDB) [29]. We extracted the TCR β CDR3 sequence from each structure. After removing duplicates, 55 were left for analysis (Supplementary Table 1 and Supplementary Figure 1). We annotated each TCR β CDR3 residue as being in contact with

peptide or not being in contact with peptide. Being in contact was defined as being $5\text{\AA}$ from a peptide residue. We used the annotations to perform a multiple sequence alignment, aligning contact positions using clustalw (<http://www.genome.jp/tools-bin/clustalw>) (Figure 1b). The alignment revealed that TCR β CDR3 residues in contact with peptide tend to lie adjacent to each other, forming a contiguous strip (Figures 1a–1b). The size and relative location of this strip varied, but the average length was four, and it rarely included any of the first or last three TCR β CDR3 residues (Figure 1b). For 23 of the structures, the strip was longer than four, and for 33, the strip was shorter than four. In $\sim 1/3$ of the structures, an additional one or two residues were also in contact. To represent each TCR β CDR3, we excluded the first and last three residues and partitioned the remaining sequence into every possible contiguous strip of four amino acids (4-mer) (Figure 1c). Thus, every CDR3 was represented by multiple 4-mers. Based on the alignment, we expect 4-mers to include most of the contact residues for the vast majority of TCR β CDR3 sequences. Our expectation is that, for each TCR β CDR3, at least one of its 4-mers contacts the peptide component of the receptor's cognate antigen.

To identify 4-mers with different amino acid sequences but similar antigen binding capabilities, we represented each 4-mer using numerical values for the biophysicochemical properties of its component amino acids. There are currently at least 566 amino acid indices one could choose from (<https://www.genome.jp/aaindex/>). Many are highly correlated and contain redundant information. At least two efforts have applied dimensionality reduction to large numbers of amino acid indices to derive small numbers of orthogonal properties (factors) that maintain most of the information contained in the original set. Kiddera et al. derived 10 factors from 188 amino acid indices [30], and Atchley et al. derived five factors from 494 amino acid indices [31]. We used Atchley factors, as they were derived from the largest number of indices and would require half as many model parameters as the Kiddera factors. The five Atchley factors correspond loosely to polarity, secondary structure, molecular size/volume, codon diversity, and electrostatic charge. For input into our model, each amino acid in a 4-mer is represented by a vector of its five Atchley factor values (Figure 1d).

T-cells undergo clonal expansion in response to antigen stimulation. Thus, receptor quantity is an important feature for our statistical classifier. We used the logarithm of the 4-mers' relative abundances as a feature in the model. We considered two approaches. The first we refer to as calculating “the 4-mer relative abundance”. First, we identify every TCR β sequence containing the 4-mer in its CDR3 and sum over its template counts, $C^{\text{TCR}\beta}$.¹ This provides the 4-mer count, $C^{4\text{mer}}$, for the sample. We then divide by the total count of all 4-mers in the sample, $T^{4\text{mer}}$, to get the 4-mer relative abundance, RA .

$$C^{4\text{mer}} = \sum_{\text{TCR}\beta \text{ with } 4\text{mer}} C^{\text{TCR}\beta} ; T^{4\text{mer}} = \sum_{4\text{mer}} C^{4\text{mer}} ; RA = \frac{C^{4\text{mer}}}{T^{4\text{mer}}} \quad \#(1)$$

¹We treat TCR β sequences with identical CDR3 sequences as being the same TCR β sequence, ignoring differences upstream of CDR3.

The second approach is to consider only the most abundant TCR β sequence containing the 4-mer in its CDR3. This we refer to as calculating “the TCR β relative abundance”. First, we sum over the TCR β template counts, $C^{\text{TCR}\beta}$, of every TCR β in a sample to get the total count, $T^{\text{TCR}\beta}$. We then divide $C^{\text{TCR}\beta}$ for the most abundant TCR β by $T^{\text{TCR}\beta}$ to get the relative abundance, $RA^{\text{TCR}\beta}$. We use the most abundant TCR β , ignoring all less abundant TCR β s containing the 4-mer.

$$T^{\text{TCR}\beta} = \sum_{\text{TCR}\beta} C^{\text{TCR}\beta}; \quad RA^{\text{TCR}\beta} = \frac{C^{\text{TCR}\beta}}{T^{\text{TCR}\beta}}; \quad RA = \underset{\substack{\text{TCR}\beta \\ \text{with 4mer}}}{\text{Maximum}}\{RA^{\text{TCR}\beta}\} \quad \#(2)$$

It is unclear a priori which approach is better, so we assessed the performance of our classifiers using both.

It is important to normalize the features of a classifier to be on the same scale. We normalized the Atchley factor values so that each has zero mean and unit variance. It was unclear a priori whether it would be appropriate to normalize the 4-mer abundance term, because its values are potentially unbounded. Therefore, we assessed classifier performance with and without normalizing this term.

Logistic Regression Model:

The extracted 4-mers were scored using a logistic regression function that predicts whether a 4-mer is tumor-associated. We used this function because of its widespread use and simplicity, and because it models a binary dependent variable. First, a biased, weighted sum of the 4-mer features (the logit) is computed.

$$\text{logit} = b_0 + W_1 \cdot f_1 + W_2 \cdot f_2 + \dots + W_{20} \cdot f_{20} + W_{21} \cdot \ln RA \quad \#(3)$$

f_1 through f_{20} represent the five Atchley factor values for the four 4-mer residues. RA represents the 4-mer's relative abundance calculated using either equation (1) or (2). The bias term b_0 and weights W_1 through W_{21} are the model parameters and are fit by maximum likelihood using gradient optimization techniques (described below). The same weights W_1 through W_{21} and bias term b_0 are used for all 4-mers. Once the logit is computed, the sigmoid function is applied to obtain a value between 0 and 1.

$$\text{score} = \frac{1}{1 + e^{-\text{logit}}} \quad \#(4)$$

The score represents the probability that the 4-mer is tumor-associated.

Multiple Instance Learning:

The problem of predicting repertoire-level labels from the 4-mers in each repertoire can be formally described as MIL in which the 4-mers are instances, the repertoires are bags, and the bag label is the tissue source of the repertoire (i.e., tumor or healthy) [24]. MIL relies on aggregating instance-level scores to assign a bag-level label. Thus, we need to aggregate the scores from all 4-mers in a repertoire into a single value that predicts whether the repertoire came from tumor or healthy tissue. Only a small number of 4-mers are expected to interact with relevant antigens. Accordingly, under the standard MIL assumption, at least one 4-mer per tumor repertoire must have a high score, while none from healthy tissue repertoires should have high scores. This was implemented by taking the maximum 4-mer score as the repertoire score. Thus, the probability that a repertoire came from tumor tissue given the 4-mer scores is defined as:

$$P(\text{tumor} \mid 4\text{mer}_1, 4\text{mer}_2, 4\text{mer}_3, \dots) = \text{Maximum}\{\text{score}_1, \text{score}_2, \text{score}_3, \dots\} \quad \#(5)$$

The predicted label is tumor when one 4-mer scores ≥ 0.5 , while the predicted label is healthy when every 4-mer scores < 0.5 . The model's parameter values were fit to maximize the assignment of correct labels.

Gradient Optimization:

Specific values for W_1 through W_{21} and b_0 were determined using repertoires with known labels (i.e., tumor vs healthy). The values were selected to maximize the likelihood that each prediction from equation (5) is correct. To search for optimal values, gradient optimization was used, as in [19]. The initial values for b_0 and W_1 through W_{20} were selected as in [19]. Two different protocols for initializing W_{21} were tried (Table 3). We ran gradient optimization from 100,000–375,000 different initial weight values (Table 3). To save compute cycles, only models with good or better performance were run a large number of times.

Overfitting is a concern with any statistical classifier, especially when using small amounts of labeled data. Because our approach uses the same weights for every 4-mer in each sample, our approach has fewer parameters than labeled data points (Supplementary Tables 2–3), which helps alleviate the concern of overfitting. Still, we applied early stopping to regularize the model, and we assessed model generalization using leave-one-out cross-validation (see below). We found that the best performing models generalize best to the holdout data on the last training step (Table 3), indicating that they had not begun to overfit the data. Previously, we applied L1/L2 regularization and dropout to the same models on a different disease and found that both worsened model performance [19]. Therefore, we do not apply them here.

Model Development and Validation:

We applied this approach to the colorectal and breast cancer datasets. Each was treated separately, resulting in one model for each dataset. To assess model performance, we performed patient-holdout cross-validation, where the tumor and patient-matched healthy

sample for a single patient were simultaneously excluded during parameter fitting and then scored after selection of the best model (Figure 2). The two samples were scored independently; the model had no knowledge that one was from a tumor and the other was from healthy tissue.

Several variations of the model were considered, including different methods for calculating the relative abundance term, for normalizing the relative abundance term, and for initializing its weight W_{21} (Table 3). Results for the best performing models are described below.

Results

Colorectal Cancer:

The number of 4-mers per sample in the colorectal cancer dataset ranged from 186 to 7,112, with an average of 3,789, giving ~ 79,566 features per sample (Supplementary Table 2). The best model used 4-mer (rather than TCR β) relative abundance with this term un-normalized and its weight (W_{21}) initialized to 0 (Table 3). This model correctly categorized 93% (26/28) of held-out samples with an average log-likelihood of -0.316 bits. The model always scored the tumor sample above the healthy sample despite having no knowledge that the two samples were from the same patient (Figure 3a). To estimate the probability of correctly classifying 26 out of 28 samples by chance, we performed a permutation analysis. For each permutation, a patient-holdout cross validation was performed where the labels on the training data were permuted but those on the holdouts were not (Supplementary Table 4). The classification accuracies of all 20 permutations were <93%, allowing us to assign $p < 0.05$ to the observed accuracy. The average accuracy over all permutations is 49%, and the average log-likelihood fit is -2.66 bits.

To discern the 4-mer biophysicochemical features that increase the probability of a tumor categorization, we examined the model weights with parameters fit on all 14 patients. The weights reveal how each Atchley factor contributes to the score and the relative importance of each 4-mer position (Figure 3b). We observe negative weights for almost every 4-mer position for Atchley factors II and IV, indicating an increased probability that a sample is tumor-derived if it contains 4-mers comprising residues with a propensity to participate in α -helical segments and that appear infrequently among the space of all protein sequences. We also observe only positive weights for Atchley factor V, indicating an increased probability that a sample is tumor-derived if it contains 4-mers enriched with positively charged residues. The weight on the abundance term favors 4-mers with a large relative abundance. Analysis of the weights for each out hold-out model reveals that the model weights are highly consistent across hold-out patients and with the weights fit on all 28 samples (Supplementary Figure 2, Supplementary Table 5).

The high scoring 4-mers from each holdout-patient also scored high with the model fit to all 14 patients. We aligned all 4-mers that scored highly enough to categorize a sample as being tumor-associated and found that the amino acids vary considerably at each 4-mer position (Figure 3c). These 4-mers would not have been found by looking for shared amino acid sequences. The 4-mers are detected by our method, because they share similar biophysicochemical properties at key positions, as selected by the weights of the model.

Several of the TCR β CDR3 sequences correspond to large clones, but many of them do not (Figure 3c). These CDR3 sequences would not have been found by examining only the most abundant clones.

Breast Cancer:

For the breast cancer dataset, the number of 4-mers ranged from 2,518 to 39,354 with an average of 20,261, giving ~425,487 features per sample (Supplementary Table 3). The best performing model uses the TCR β (rather than 4-mer) relative abundance (Table 3). It is otherwise like that for colorectal cancer (Table 3). The model correctly categorized 94% (30/32) of held-out samples with an average log-likelihood error of -0.283 bits (Figure 4a). As with colorectal cancer, the model always scores the tumor sample above the patient-matched healthy sample (Figure 4a). Permutation analysis gave a classification accuracy of 49% and an average log-likelihood fit of -2.71 bits (Supplementary Table 6). The classification accuracies of all 20 permutations were $<94\%$, allowing us to assign $p < 0.05$ to the observed accuracy.

We examined the model weights with parameters fit on all 16 patients (Figure 4b). The direction and magnitude of the weights differed considerably from those obtained on the colorectal samples, indicating that the model is specific to cancer type. For all Atchley factors, the weights are position-dependent. For example, 4-mer scores are increased for 4-mers with hydrophobic residues at the first two positions and hydrophilic residues at the last two positions (Figure 4b). The one similarity with the colorectal results is that the model assigns a high score to 4-mers with a high relative abundance. The weights of all models (holdout models and the model fit on all samples) cluster more tightly than for the colorectal cancer models (Supplementary Figure 3 and Supplementary Table 7), possibly due to the much larger number of features per sample.

The high scoring 4-mers from each holdout tumor sample also scored high with the model fit to all 16 patients except for the 4-mer GSYN, which is one of two high-scoring 4-mers for patient 3 during cross-validation. We aligned all 4-mers that scored high enough to categorize a sample as tumor and found, as with the colorectal cancer model, that the amino acids vary considerably at each 4-mer position (Figure 4c). In contrast to what was observed for colorectal cancer, however, we observe that almost every TCR β CDR3 sequence containing a high-scoring 4-mer corresponds to a top clone (Figure 4c).

Discussion:

Improved methods for cancer early detection are urgently needed. For the vast majority of cancers, there is currently no test that is both sensitive enough to detect early stage disease and specific enough to mitigate overdiagnosis and overtreatment. For those cancers, screening of average-risk populations is not recommended, and cancer is typically detected only after it has progressed enough to cause symptoms. For the small number of cancers for which screening of average-risk populations is advised, there are still significant downsides. Current, guideline-endorsed approaches have sensitivities and specificities lower than ideal and therefore require frequent re-screening and have the potential for overdiagnosis and overtreatment. Furthermore, they primarily involve detecting abnormal tissue changes by

imaging or cytology and require follow-up by invasive tissue collection, which has associated risks.

The holy grail of cancer detection is a highly specific, highly sensitive test that detects early stage disease and does not require invasive tissue collection. Many potential blood-borne biomarkers are under investigation, including protein markers and circulating cell-free tumor DNA [32]. Some of them have also been found in other tissues accessible by minimally invasive procedures, such as cervical cytology samples [33]. The most promising results have been obtained by assaying for combinations of markers [34]. While the results are promising, the sensitivities are highly variable, depend strongly on organ site, and are less promising for early stage disease [33–38]. Thus, complementary biomarkers are needed.

Given the specificity of adaptive immune responses, lymphocyte recirculation, and the sensitivity of tests for detecting rare lymphocyte clones [39, 40], we find it plausible that anti-tumor lymphocyte responses could provide one such complementary biomarker. Success of the approach would require that anti-tumor lymphocyte responses have cancer-specific signatures, that the signatures appear early in the disease, and that they be detectable in readily accessible tissue. We designed the current study to address the first requirement by determining whether tumor-associated T-cell repertoires have a cancer-specific signature that can reliably distinguish cancer-associated TCR repertoires from those associated with healthy tissue of the same organ.

With the results presented here, we have successfully demonstrated that TIL repertoires include TCRs with cancer-specific, biophysicochemical motifs. Specifically, we detected distinct biophysicochemical motifs for colorectal cancer and breast cancer tumor-infiltrating T lymphocytes and demonstrated that these motifs distinguish tumor repertoires from healthy tissue repertoires with classification accuracies of 93% and 94%, respectively, by leave-one-out cross-validation. We further show by permutation analysis that the probability of obtaining these accuracies by chance is <0.05 . These results suggest that the first requirement given above, that anti-tumor lymphocyte responses have cancer-specific signatures, could be met, at least for some cancer types. A definitive answer will require follow up studies on larger patient cohorts and in additional cancer types.

As in our original approach, we represented each TCR β 's CDR3 by partitioning it into every possible contiguous strip of four amino acids (4-mer) and representing each 4-mer using the Atchley factor values for its component residues. We improved the approach, adding a feature that quantifies the relative abundances of the 4-mers. This feature is critical to model performance, increasing accuracy on the colorectal and breast cancer datasets from 36% to 93% and 67% to 94%, respectively (Table 3). Our approach is not merely identifying highly expanded T-cell clones, however. In the colorectal cancer dataset, the TCR β CDR3 sequences containing high scoring 4-mers correspond to a top-ten most abundant clone for only four of the fourteen patients.

The biophysicochemical motifs detected by our method are specific to a cancer type. Four-mers with the following properties are classified as tumor-associated by the colorectal cancer model: hydrophilic residues in the 2nd and 3rd 4-mer positions with hydrophobic

residues in the 1st and 4th positions; amino acids that tend to form α -helices at all four 4-mer positions; small residues in the 1st, 2nd, and 4th positions with large residues in the 3rd position; and positively charged residues at all four 4-mer positions. In contrast, 4-mers with the following properties are classified as tumor-associated by the breast cancer model: hydrophilic residues in the 3rd and 4th positions with hydrophobic residues in the 1st and 2nd positions; amino acids that tend to form α -helices in the 1st 4-mer position with amino acids that tend to form bends, coils, or turns in the remaining positions; large residues in the 1st, 2nd, and 4th 4-mer positions with small residues in the 3rd position; and negatively charged amino acids in the 2nd and 4th 4-mer positions with positively charged residues in the 1st and 3rd positions. Furthermore, both cancer motifs are quite different from the one reported for multiple sclerosis [19]. Note that the Atchley factors have some moderate, inter-factor correlation. Thus, two factors could have large weights at a particular position but only one is important to antigen binding.

Finding biophysicochemical TCR β motifs that, within a cancer type, are shared across the TILs repertoires of all patients but are largely absent from healthy tissue repertoires is consistent with the hypothesis that we are detecting TCR with specificity for a tumor-associated antigen that is shared between patients. We are not the first to hypothesize that patients may have shared TCRs responding to a common antigen, nor are we the first to look for the corresponding TCRs in colorectal and breast cancer TILs [7, 10, 25, 26, 41, 42]. To address this hypothesis, other investigators have searched for CDR3 amino acid sequences that were shared between multiple patient TILs repertoires [7, 10, 25, 26, 41, 42]. None of the studies found sequences shared across all patients in a study, and the degree of sharing varied tremendously. In the breast cancer studies, it was concluded that the shared TCR β CDR3 sequences most likely correspond to public clones rather than receptors responding to common cancer antigens, because the same TCR β CDR3 sequences were frequently found in databases of TCR β repertoires from presumed healthy individuals [7, 25, 41, 42]. Additionally, Beausang et al. analyzed the sequence properties of the shared sequences and found that they had many features in common with the TCR β CDR3 sequences of public clones, such as being shorter in length and having few insertions [25]. The Munson et al. paper stands out, because they sequenced transcripts from both TCR chains [7]. They identified 14 TCR α and β pairs that were present in 7 of 20 patient TILs repertoires (including one present in 15 TILs repertoires) but not present in peripheral blood repertoires from six presumed healthy individuals [7].

Public clones are typically defined as TCRs with identical amino acid sequences observed across multiple individuals of a given species [43]. In our case, the TCR β CDR3s bearing high-scoring 4-mers have different amino acid sequences across patients, and therefore are not formally considered part of public TCRs. We conducted our analysis using biophysicochemical representations of amino acid sequence in order to find TCR β CDR3s that could be expected to have similar antigen binding capabilities, even in the absence of having identical amino acid sequences. Despite their similar biophysicochemical features, however, we cannot say that the TCR β CDR3s we have identified are capable of binding to the same antigen. This needs to be tested experimentally.

The motifs we found are tumor-associated, suggesting that, if TCRs bearing the motif can bind the same antigen, the antigen is tumor-associated. The permutation analysis indicates that our method cannot find a shared biophysicochemical motif that uniquely distinguishes any random grouping of TCR β repertoires. Additionally, the motif we have identified in each case is unique to the TILs repertoires. This suggests that the motif is related to the fact that the T-cells are found in the tumor. This may mean that they have specificity for a cancer antigen shared across the patients. There are alternative explanations, however. The T-cells may be responding to tissue damage in the tumor, or they may be T regulatory cells contributing to immunosuppression. There are likely other interpretations. Again, experimental follow-up studies are needed.

It may seem surprising that our statistical classifier generalizes across patients with presumably different HLA genes. We expect this is because TCR:MHC interaction primarily happens via contacts in CDRs 1 and 2, while peptide contacts are primarily via CDR3 [44–47]. We further expect that using a 4-mer (rather than a longer k-mer) allows us to isolate the residues of the TCR β CDR3 that are responsible for peptide interaction. Additionally, we speculate that having patient-matched healthy tissue is important, because it provides HLA-matched controls, enabling the model to achieve good performance in the absence of HLA information. It is quite possible, however, that the models' performances would improve if HLA type were included for each patient. This is something we will determine in future studies.

For both colorectal and breast cancer, we observed multiple TCR β CDR3 sequences that contain multiple high-scoring 4-mers. If these 4-mers are the residues that contact peptide, then this suggests that the corresponding receptors may interact with cognate antigen via multiple 4-mers (Supplementary Figure 4). The same receptor may bind the same peptide:MHC complex in multiple ways. The same receptor may bind the same peptide in the context of different MHC molecules using different 4-mers. Or the receptor may bind different peptides with the different 4-mers. In this case, we would expect the different peptides to be highly similar in terms of their biophysicochemical properties. In either case, the TCR β CDR3 loop must exhibit considerable conformational flexibility. This kind of loop flexibility has been demonstrated [48, 49].

Our study has several limitations. First, our approach is designed to detect biophysicochemical motifs in TCRs with specificity for shared antigens. It cannot identify TCRs with specificity for patient-specific tumor neoantigens. Secondly, the number of crystal structures examined is quite small relative to the space of possible TCR-peptide-MHC interactions. Since we examined all PDB structures from which we could extract the CDR3 amino acid sequence, increasing the number will require that additional crystal structures become available. Thirdly, we used TCR β sequences and thus have only part of the TCR antigen binding site. Studies incorporating both TCR β and TCR α are needed. It is not clear whether a model using both would find the same TCR β biophysicochemical motifs together with a TCR α motif, or whether a novel motif would be detected. Fourthly, we do not know the T lymphocyte subset to which the T-cells bearing the TCRs with high-scoring 4-mers belong. Thus, we do not know their functional dispositions (e.g., cytotoxic, regulatory) or whether they could participate in tumor cell killing. Fifthly, our patient sets

are relatively small. Thus, follow up studies on larger patient cohorts are needed. Finally, although we have successfully used our method on three different diseases and with both B-cell and T-cell receptors, our approach is not yet a turnkey solution for identifying biomarkers in immune repertoires. In each case, we considered several variations of the model to determine which one worked best for each disease. Each time we identify a new biomarker, however, we move one step closer to developing automated approaches that relate immune repertoires to labelled clinical phenotypes.

Our approach also has benefits. It requires only collecting tissue from a small number of patients, sequencing the immune receptors, and fitting the model. No prior knowledge of the antigens or receptor specificities is needed, and no additional experiments are required to enrich disease-specific receptors. The model's predictions are easily interpretable and may elucidate individual immune receptors associated with diseases such as cancer. Our future work will include improving our methodology and applying it to other diseases and tissue types.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

This work was supported by a training grant from the Cancer Prevention and Research Institute of Texas (RP160157) and an R01 from the National Institute of Allergy and Infectious Diseases (R01AI097403). We are grateful that the colorectal and breast cancer datasets have been made available online, and we appreciate Adaptive Biotechnologies for hosting this data on immuneACCESS (<https://clients.adaptivebiotech.com/immuneaccess>). Computing time on the UT Southwestern BioHPC computing cluster was made available through the Harold C. Simmons Comprehensive Cancer Center.

Financial support: This project was supported by an NIAID-funded R01 (AI097403) to LGC, a training grant to the Simmons Comprehensive Cancer Center at UT Southwestern from the Cancer Prevention and Research Institute of Texas (RP160157), and funding to LGC from UT Southwestern and the Simmons Comprehensive Cancer Center.

References

1. Chen DS and Mellman I, Elements of cancer immunity and the cancer-immune set point. *Nature*, 2017 541(7637): p. 321–330. [PubMed: 28102259]
2. Kvistborg P, van Buuren MM, and Schumacher TN, Human cancer regression antigens. *Curr Opin Immunol*, 2013 25(2): p. 284–90. [PubMed: 23566921]
3. Dhodapkar K and Dhodapkar M, Harnessing shared antigens and T-cell receptors in cancer: Opportunities and challenges. *Proc Natl Acad Sci U S A*, 2016 113(29): p. 7944–5. [PubMed: 27410048]
4. Romero P, et al., Ex vivo staining of metastatic lymph nodes by class I major histocompatibility complex tetramers reveals high numbers of antigen-experienced tumor-specific cytolytic T lymphocytes. *J Exp Med*, 1998 188(9): p. 1641–50. [PubMed: 9802976]
5. Dhodapkar KM, et al., SOX2-specific adaptive immunity and response to immunotherapy in non-small cell lung cancer. *Oncoimmunology*, 2013 2(7): p. e25205. [PubMed: 24073380]
6. Dhodapkar MV, et al., Prospective analysis of antigen-specific immunity, stem-cell antigens, and immune checkpoints in monoclonal gammopathy. *Blood*, 2015 126(22): p. 2475–8. [PubMed: 26468228]
7. Munson DJ, et al., Identification of shared TCR sequences from T cells in human breast cancer using emulsion RT-PCR. *Proc Natl Acad Sci U S A*, 2016 113(29): p. 8272–7. [PubMed: 27307436]

8. Massa C, et al., Identification of patient-specific and tumor-shared T cell receptor sequences in renal cell carcinoma patients. *Oncotarget*, 2017 8(13): p. 21212–21228. [PubMed: 28177902]
9. Bai X, et al., Characteristics of Tumor Infiltrating Lymphocyte and Circulating Lymphocyte Repertoires in Pancreatic Cancer by the Sequencing of T Cell Receptors. *Sci Rep*, 2015 5: p. 13664. [PubMed: 26329277]
10. Nakanishi K, et al., Characterization of the T-cell receptor beta chain repertoire in tumor-infiltrating lymphocytes. *Cancer Med*, 2016 5(9): p. 2513–21. [PubMed: 27465739]
11. Fugmann SD, et al., The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annu Rev Immunol*, 2000 18: p. 495–527. [PubMed: 10837067]
12. Kirsch I, Vignali M, and Robins H, T-cell receptor profiling in cancer. *Mol Oncol*, 2015 9(10): p. 2063–70. [PubMed: 26404496]
13. Galson JD, et al., Analysis of B Cell Repertoire Dynamics Following Hepatitis B Vaccination in Humans, and Enrichment of Vaccine-specific Antibody Sequences. *EBioMedicine*, 2015 2(12): p. 2070–9. [PubMed: 26844287]
14. Miho E, et al., Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. *Front Immunol*, 2018 9: p. 224. [PubMed: 29515569]
15. Chaudhary N and Wesemann DR, Analyzing Immunoglobulin Repertoires. *Front Immunol*, 2018 9: p. 462. [PubMed: 29593723]
16. Jia Q, et al., Diversity index of mucosal resident T lymphocyte repertoire predicts clinical prognosis in gastric cancer. *Oncoimmunology*, 2015 4(4): p. e1001230. [PubMed: 26137399]
17. Postow MA, et al., Peripheral T cell receptor diversity is associated with clinical outcomes following ipilimumab treatment in metastatic melanoma. *J Immunother Cancer*, 2015 3: p. 23. [PubMed: 26085931]
18. Hosoi A, et al., Increased diversity with reduced “diversity evenness” of tumor infiltrating T-cells for the successful cancer immunotherapy. *Sci Rep*, 2018 8(1): p. 1058. [PubMed: 29348598]
19. Ostmeyer J, et al., Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics*, 2017 18(1): p. 401. [PubMed: 28882107]
20. Emerson RO, et al., Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet*, 2017 49(5): p. 659–665. [PubMed: 28369038]
21. Sun Y, et al., Specificity, Privacy, and Degeneracy in the CD4 T Cell Receptor Repertoire Following Immunization. *Front Immunol*, 2017 8: p. 430. [PubMed: 28450864]
22. Cinelli M, et al., Feature selection using a one dimensional naive Bayes’ classifier increases the accuracy of support vector machine classification of CDR3 repertoires. *Bioinformatics*, 2017 33(7): p. 951–955. [PubMed: 28073756]
23. Thomas N, et al., Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics*, 2014 30(22): p. 3181–8. [PubMed: 25095879]
24. Carbonneau M-A, et al., Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 2018 77: p. 329–353.
25. Beausang JF, et al., T cell receptor sequencing of early-stage breast cancer tumors identifies altered clonal structure of the T cell repertoire. *Proc Natl Acad Sci U S A*, 2017 114(48): p. E10409–E10417. [PubMed: 29138313]
26. Sherwood AM, et al., Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. *Cancer Immunol Immunother*, 2013 62(9): p. 1453–61. [PubMed: 23771160]
27. Carlson CS, et al., Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun*, 2013 4: p. 2680. [PubMed: 24157944]
28. DeWitt WS, et al., A Public Database of Memory and Naive B-Cell Receptor Sequences. *PLoS One*, 2016 11(8): p. e0160853. [PubMed: 27513338]
29. Rose PW, et al., The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res*, 2015 43(Database issue): p. D345–56. [PubMed: 25428375]

30. Kidera A, et al., Statistical-Analysis of the Physical-Properties of the 20 Naturally-Occurring Amino-Acids. *Journal of Protein Chemistry*, 1985 4(1): p. 23–55.
31. Atchley WR, et al., Solving the protein sequence metric problem. *Proc Natl Acad Sci U S A*, 2005 102(18): p. 6395–400. [PubMed: 15851683]
32. Babayan A and Pantel K, Advances in liquid biopsy approaches for early detection and monitoring of cancer. *Genome Med*, 2018 10(1): p. 21. [PubMed: 29558971]
33. Kinde I, et al., Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci Transl Med*, 2013 5(167): p. 167ra4.
34. Cohen JD, et al., Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*, 2018 359(6378): p. 926–930. [PubMed: 29348365]
35. Krimmel JD, et al., Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc Natl Acad Sci U S A*, 2016 113(21): p. 6005–10. [PubMed: 27152024]
36. Fernandez-Cuesta L, et al., Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. *EBioMedicine*, 2016 10: p. 117–23. [PubMed: 27377626]
37. Bettegowda C, et al., Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med*, 2014 6(224): p. 224ra24.
38. Newman AM, et al., An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat Med*, 2014 20(5): p. 548–54. [PubMed: 24705333]
39. Korde N, et al., Treatment With Carfilzomib-Lenalidomide-Dexamethasone With Lenalidomide Extension in Patients With Smoldering or Newly Diagnosed Multiple Myeloma. *JAMA Oncol*, 2015 1(6): p. 746–54. [PubMed: 26181891]
40. Wu D, et al., Detection of minimal residual disease in B lymphoblastic leukemia by high-throughput sequencing of IGH. *Clin Cancer Res*, 2014 20(17): p. 4540–8. [PubMed: 24970842]
41. Levy E, et al., Immune DNA signature of T-cell infiltration in breast tumor exomes. *Sci Rep*, 2016 6: p. 30064. [PubMed: 27452728]
42. Wang T, et al., The Different T-cell Receptor Repertoires in Breast Cancer Tumors, Draining Lymph Nodes, and Adjacent Tissues. *Cancer Immunol Res*, 2017 5(2): p. 148–156. [PubMed: 28039161]
43. Venturi V, et al., The molecular basis for public T-cell responses? *Nat Rev Immunol*, 2008 8(3): p. 231–8. [PubMed: 18301425]
44. Garcia KC, et al., The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nat Immunol*, 2009 10(2): p. 143–7. [PubMed: 19148199]
45. Rossjohn J, et al., T cell antigen receptor recognition of antigen-presenting molecules. *Annu Rev Immunol*, 2015 33: p. 169–200. [PubMed: 25493333]
46. Rudolph MG, Stanfield RL, and Wilson IA, How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol*, 2006 24: p. 419–66. [PubMed: 16551255]
47. Zhang H, et al., The contribution of major histocompatibility complex contacts to the affinity and kinetics of T cell receptor binding. *Sci Rep*, 2016 6: p. 35326. [PubMed: 27734930]
48. Reiser JB, et al., A T cell receptor CDR3beta loop undergoes conformational changes of unprecedented magnitude upon binding to a peptide/MHC class I complex. *Immunity*, 2002 16(3): p. 345–54. [PubMed: 11911820]
49. Ayres CM, et al., Differential utilization of binding loop flexibility in T cell receptor ligand selection and cross-reactivity. *Sci Rep*, 2016 6: p. 25070. [PubMed: 27118724]

Figure 1:

(a) X-ray crystallographic structure of a human T-cell receptor β -chain (gray) bound to a peptide (blue) in complex with MHC (not shown). The CDR3 is shown in green, and the portion of the CDR3 in direct contact ($\sim 5\text{\AA}$) with the peptide is shown in red. The MHC complex and α -chain are omitted for clarity. (b) CDR3 sequences extracted from 55 X-ray crystallographic structures of human T-cell receptors bound to peptide:MHC. Residues $\sim 5\text{\AA}$ from peptide (red) are used to align the sequences. The alignment was created using clustalw. The bar chart shows the proportion of structures in which the corresponding CDR3 position was in direct contact with peptide. (c) To profile the specificity of a CDR3 sequence, the CDR3 is cut into every possible 4-mer excluding the first and last 3 residues. (d) Each 4-mer is converted into a biophysicochemical representation. For each residue, there are 5 Atchley factor values describing the residues biophysicochemical properties, resulting in a 4-mer representation consisting of 20 numeric values.

Figure 2:

Workflow for model selection and parameter fitting. The diagram shows how the data were used to train and validate each model. The performance of each model was assessed by a patient-holdout cross-validation, where the tumor and healthy samples from the same patient were excluded for validation. Data from the remaining N-1 patients were used to fit the model. For each model, between 100,000 and 375,000 initial sets of weights were generated. To save compute cycles, only models with good or better performance were run a large number of times. Each set of weights was used for exhaustive leave-one-out cross-validation over all N patients. Each run of cross-validation with each set of initial weights was run for 2,500 iterations of gradient optimization. The best fit to the N-1 training samples from among all runs was used to evaluate the excluded validation data.

Figure 3:

Colorectal cancer results. (a) Classification accuracy obtained by patient-holdout cross-validation, where the tumor and healthy tissue from the same patient are excluded for validation. (b) Illustration of the classifier weights after fitting the model to all 14 patients. For each of the five Atchley factors, the weights are shown for the four residue positions. The weight for the log-frequency of the 4-mer is also shown. Positive weight values are shown pointing up, and negative weight values are shown pointing down. The length of the *arrow* corresponds to the weight's magnitude. (c) All 4-mers with a score above 0.5 (middle column) shown for each of the 14 patients (leftmost column). Each 4-mer is shown in the context of its respective CDR3. When the 4-mer appears in multiple CDR3 sequences, the CDR3 with the largest relative abundance is shown. The CDR3 sequences are ranked according to their relative abundance in the sample (rightmost column). A rank of 1 indicates the largest relative abundance in the sample. In patient 6, there are two CDR3s that each have two high-scoring 4-mers. MYRE and YREV are both found in the TCR β CDR3 sequence CASSMYREVEAFF, and the 4-mers ERFY and RERF are both found in the TCR β CDR3 sequence CASSRERFYEQYF.

Figure 4: Breast cancer results. (a) Classification accuracy obtained by patient-holdout cross-validation, where the tumor and healthy tissue from the same patient are excluded for validation. (b) Illustration of the classifier weights after fitting the model to all 16 patients. For each of the five Atchley factors, the weights are shown for the four residue positions. The weight for the log-frequency of the receptor is also shown. Positive weight values are shown pointing up, and negative weight values are shown pointing down. The length of the *arrow* corresponds to the weight's magnitude. (c) All 4-mers with a score above 0.5 (middle column) shown for each of the 16 patients (leftmost column). Each 4-mer is shown in the context of its respective CDR3. When the 4-mer appears in multiple CDR3 sequences, the CDR3 with the largest relative abundance is shown. The CDR3 sequences are ranked according to their relative abundance in the sample (rightmost column). A rank of 1 indicates the largest relative abundance in the sample. As with colorectal cancer, we observed TCR β CDR3 sequences containing multiple high-scoring 4-mers. In patient 1, LSRS and RSNQ appear in the TCR β CDR3 sequence CASSLSRSNQPHF. In patient 10, SSPH, AYNQ, and AAYN appear in the TCR β CDR3 sequence CASSSPHRAAYNQPHF.

Table 1:

Microsatellite instability status and the number of unique T-cell receptor β (TCR β) complementarity determining region 3 (CDR3) sequences from the tumor and healthy tissue samples are shown for the 14 colorectal cancer patients. MSS: no microsatellite instabilities detected. MSI-H: microsatellite instability detected at two or more markers.

COLORECTAL SAMPLES From [26]			
Patient # (Patient ID)	Tumor		Healthy
	MSI-status	Unique TCR β s	Unique TCR β s
1 (400464)	MSS	1,836	1,466
2 (400480)	MSS	2,432	1,773
3 (400488)	MSI-H	2,090	699
4 (400600)	MSS	862	984
5 (400712)	MSS	203	667
6 (400728)	MSS	41	1,110
7 (401144)	MSS	1,390	1,040
8 (401176)	MSS	723	883
9 (401248)	MSS	391	1,844
10 (401256)	MSS	1,711	1,068
11 (401264)	MSS	3,849	910
12 (401304)	MSS	1,659	1,612
13 (401320)	MSS	2,933	1,667
14 (401336)	MSI-H	988	1,228

Table 2:

Breast cancer type, receptor status, and the number of unique T-cell receptor β (TCR β) complementarity determining region 3 (CDR3) sequences from the tumor and healthy tissue samples are shown for the 16 breast cancer patients. IDC: invasive ductal carcinoma. ILC: invasive lobular carcinoma. IMC: invasive mucinous carcinoma. ER: estrogen receptor. PR: progesterone receptor. HER2: human epidermal growth factor receptor 2.

BREAST SAMPLES From [25]				
Patient # (Patient ID)	Tumor			Healthy
	Type	ER/PR/HER2	Unique TCR β s	Unique TCR β s
1 (BR01)	IDC	+ / + / -	50,667	18,848
2 (BR05)	IDC	+ / + / -	21,559	7,923
3 (BR07)	IDC	+ / + / -	22,345	12,334
4 (BR13)	IDC	+ / + / -	8,276	2,609
5 (BR14)	ILC	+ / + / -	34,203	5,577
6 (BR15)	IDC	+ / + / -	16,341	3,316
7 (BR16)	IDC	+ / + / -	8,237	22,483
8 (BR17)	IDC	+ / + / -	8,686	7,748
9 (BR18)	IDC	+ / + / -	5,324	812
10 (BR19)	ILC	+ / + / -	8,571	8,865
11 (BR20)	ILC	+ / + / -	15,956	13,611
12 (BR21)	IDC	+ / + / -	18,597	10,593
13 (BR22)	IMC	+ / + / -	51,097	22,774
14 (BR24)	IDC	- / - / -	45,953	10,903
15 (BR25)	IDC	- / - / +	16,004	4,276
16 (BR26)	ILC	+ / + / -	6,250	3,397

Table 3:

Model variations considered for each cancer type. (1st column) Cancer type. (2nd column) Strategy for computing 4-mer relative abundance. (3rd column) Approach to normalization of the relative abundance term. (4th column) Different schemes for initializing the weight W_{21} . (5th column) Other variations of the model that were considered. (6th column) The number of initializations that were run. (7th column) The performance of each variation of the model. The performances of the best performing models are shown in bold font.

Cancer Type	Calculation of RA (i.e. relative abundance)	Normalization of "Log RA"	Initial Value W_{21} (the weight term on "Log RA")	Miscellaneous	Number of Initializations	Patient Holdout Cross-Validation
Colorectal Cancer SHERWOOD ET AL., 2013	-	-	-		125,000	10/28 $\approx$ 36%
	Equation (1)	Un-normalized	$W_{21} = 0$		250,000	26/28 $\approx$ 93%
	Equation (1)	Un-normalized	$W_{21} = 0$	Batch Normalization	125,000	16/28 $\approx$ 57%
	Equation (1)	Un-normalized	$W_{21} = 0$	Batch Normalization & Early Stopping	125,000	19/28 $\approx$ 67%
	Equation (1)	$\mu=0, \sigma=1$	$W_{21} \sim N(0, 1/21)$		250,000	18/28 $\approx$ 64%
	Equation (2)	Un-normalized	$W_{21} = 0$		375,000	21/28 $\approx$ 75%
	Equation (2)	Un-normalized	$W_{21} = 0$	Early Stopping	375,000	23/28 $\approx$ 82%
	Equation (2)	Un-normalized	$W_{21} \sim N(0, 1/21)$		125,000	18/28 $\approx$ 64%
Breast Cancer BEAUSANG ET AL., 2017	-	-	-		250,000	21/32 $\approx$ 67%
	Equation (1)	Un-normalized	$W_{21} = 0$		125,000	14/32 $\approx$ 44%
	Equation (1)	Un-normalized	$W_{21} = 0$	Early Stopping	125,000	23/32 $\approx$ 72%
	Equation (1)	Un-normalized	$W_{21} = 0$	Smaller Step Size	100,000	13/32 $\approx$ 41%
	Equation (1)	Un-normalized	$W_{21} = 0$	Smaller Step Size & Early Stopping	100,000	22/32 $\approx$ 69%
	Equation (2)	Un-normalized	$W_{21} = 0$		250,000	30/32 $\approx$ 94%
	Equation (2)	Un-normalized	$W_{21} \sim N(0, 1/21)$		250,000	27/32 $\approx$ 84%
	Equation (2)	Un-normalized	$W_{21} \sim N(0, 1/21)$	Early Stopping	250,000	28/32 $\approx$ 87%