RESEARCH

Capturing the Difference in Humoral Immunity between Normal and Tumor Environments from RNA Sequences of B-Cell Receptors Using Supervised Machine Learning

Hiroki Konishi[1,2], Daisuke Komura[1], Hiroto Katoh[1], Shinichiro Atsumi[1,5], Hirotomo Koda[1,4], Asami Yamamoto[1], Yasuyuki Seto[5], Masashi Fukayama[4], Rui Yamaguchi[3], Seiya Imoto[2] and Shumpei Ishikawa[1*]

Abstract

Motivation: The success of immunotherapy in treating tumors has been attracting increasing interest in the adaptive immune system in the tumor microenvironment in recent years. Progress in next-generation sequencing technology now enables sequencing of whole T-cell receptors (TCRs) and B-cell receptors (BCRs)/immunoglobulin (Igs) in the tumor microenvironment. Since BCRs/Igs in tumor tissue have high affinity for tumor-specific antigens, the patterns of amino acid sequences of BCRs/Igs might differ in the normal and tumor environment. However, given the high diversity of BCRs/Igs and the rare recurrent sequences among individuals, it is far more difficult to capture such differences in BCRs/Igs sequences compared to obtaining similar information for TCRs. The aim of this study was to capture this difference by applying supervised machine learning methods to RNA sequences of BCRs/Igs obtained from matched normal and tumor specimens of 90 patients with gastric cancer.

Results: We adopted two approaches to evaluate the possibility of accurately differentiating BCRs/Igs amino acid sequences of the tumor environment. First, we used two kinds of discriminative models for sequence classification between normal and tumor tissues. One takes into account only V and J frame usages and the other deals with the physicochemical properties of amino acid sequences. The results demonstrated clear differences in the physicochemical properties of amino acid sequences between BCRs/Igs derived from normal and tumor tissues, which could not be captured by applying V and J frame usage alone. In the second approach, we assumed a diagnostic situation in which a set of BCR/Ig-sequences is available for each patient, and we classified normal and tumor tissues based on the BCR sequences. Combining BCR/Ig-level classifiers and the Shannon entropy of the BCRs/Igs repertoire, we achieved an area under the curve value of 0.834, suggesting that BCRs/Igs repertoires have sufficient environment-specific features, which can be used to determine the tissue type.

Conclusion: To the best of our knowledge, this is the first study to show that BCRs/Igs sequences derived from tumor and normal tissues have globally distinct patterns, indicating that these tissues can be effectively differentiated using BCRs/Igs repertoires. Our research therefore provides comprehensive benefits in advancing the field of cancer immunity.

Keywords: BCR/Ig; cancer; machine learning

1 Introduction

Recent insight into cancer immunity has provided the possibility of new treatment strategies based on immunotherapy against tumors. Since tumor cells have unique proteins or other molecules or contain proteins in greater abundance than found in normal cells, known as tumor-specific antigens (TSA), evaluation of antigen receptors against TSAs expressed in tumor-infiltrating lymphocytes is important for elucidating cancer immunity. There are two main types of im-

*Correspondence:
[1]Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, 113-8510 Tokyo, Japan
Full list of author information is available at the end of the article

munity conferred by lymphocytes: cellular immunity, which is largely realized by T-cell receptors (TCRs), and humoral immunity realized by antibodies secreted by B-cells.

Recently, advances in next-generation sequencing technology have provided the opportunity to sequence TCRs and B-cell receptors (BCRs) or antibodies at an unprecedented scale [1, 2]. However, most previous studies analyzing the global patterns of antigen receptor sequences have focused on T cells [3–6], which successfully identified specific features of the amino acid motifs in TCRs. In contrast, characterizing humoral immunity is more complex than cellular immunity because B-cells show higher diversity than T-cells due to the high amounts of somatic hypermutations and junctional random sequence insertion/deletion. Therefore, most previous analyses of BCRs have focused on a small number of known antigens or epitopes [7]. Moreover, the conventional approach used for TCR analysis based on analysis of sequence motifs or identical sequences cannot be applied to understanding BCRs/immunoglobulins (Igs) in tumors, because of the greater diversity of the antibody amino acid sequences and the high amount of unique sequences in individual tumors [8]. Nevertheless, given the important role of humoral immunity in cancer, global sequence analysis of BCRs/Igs in tumors is also essential in understanding tumor immunity [9]. Moreover, unlike TCRs that recognize only peptide antigens, antibodies also recognize other types of antigens such as nucleotides, carbohydrates, and lipids; therefore, novel TSAs such as sulfated glycosaminoglycans would only be identified through BCR analysis [8].

Accordingly, we hypothesized that because of the presence of TSAs, BCR/Ig-sequences in the tumor environment might exhibit some different characteristics from those of the normal tissue environment even if no common sequence motif exists. However, to the best of our knowledge, there is currently no available method for capturing such differences at the sequence level.

Therefore, in this study, we tackled this problem by constructing classifiers of BCRs/Igs obtained from RNA-seq data from 90 paired tissue specimens obtained from patients with gastric cancer. These classifiers were based on three supervised machine learning techniques to differentiate the repertoires between the normal and tumor environment. Our analysis is based on the physicochemical properties of amino acid sequences rather than the amino acid sequences themselves as the key features for machine learning. This approach allowed for identifying any distinct physicochemical properties of BCRs/Igs in tumor tissues. We also classified normal and tumor tissues based on the set of BCR/Ig-sequences, assuming a diagnostic situation in which these sets are known for the patient. This approach based on combining the BCR/Ig-clone-level classifiers and the Shannon entropy of the BCR/Ig repertoire was subjected to receiver operating characteristic (ROC) curve analysis to determine the performance of the classifier to sufficiently identify tissue environment-specific features and correctly classify the tissue type. We expect that this approach will help to advance knowledge in the field of cancer research and improve immunotherapy toward better personalized medicine in cancer treatment.

2 Methods

2.1 Clinical samples

Ninety frozen gastric cancer specimens that were surgically resected from patients between 2009 to 2016 at the University of Tokyo Hospital were analyzed in this study after receiving written informed consent. This study was approved by the institutional review boards of the University of Tokyo and Tokyo Medical and Dental University.

2.2 RNA-seq data of BCRs

Total RNA was extracted from approximately 10 sequential frozen sections (10 µm thick) of gastric cancer tissues using the RNeasy Mini kit (Qiagen, Hilden, Germany). Quantity and quality assessments of the extracted total RNAs were determined on the Agilent Bioanalyzer (Agilent Technologies, Foster City, CA, USA). Multiplex PCR primers targeting BCRs/Igs genes (iRepertoire, Inc., AL, USA) were used to amplify the BCRs/Igs repertoires in each sample of tumor and normal tissues according to the manufacturer's protocol. Each repertoire library was then sequenced on an Illumina MiSeq instrument (Illumina) with 2x 300-bp paired-end sequencing according to the manufacturer's protocol. The results of paired-end nucleotide sequencing were obtained from two FASTQ files, read from the 5'-end to the 3'-end and from the 3'-end to 5'-end, respectively.

2.3 Pre-processing of BCR sequences

An overview of the workflow for the analysis and classification of BCR sequences is provided in Figure 1. First, the BCR/Ig-sequences obtained for each patient were summarized and transformed to a feature matrix that is suitable for machine learning using the following steps.

2.3.1 Alignment of BCR sequences

A heavy chain of BCR/Ig is composed of variable(V), diversity (D), and joining (J) gene segments. To align paired-end nucleotide sequences of BCRs/Igs and estimate the appropriate assignment of V, D, and J gene

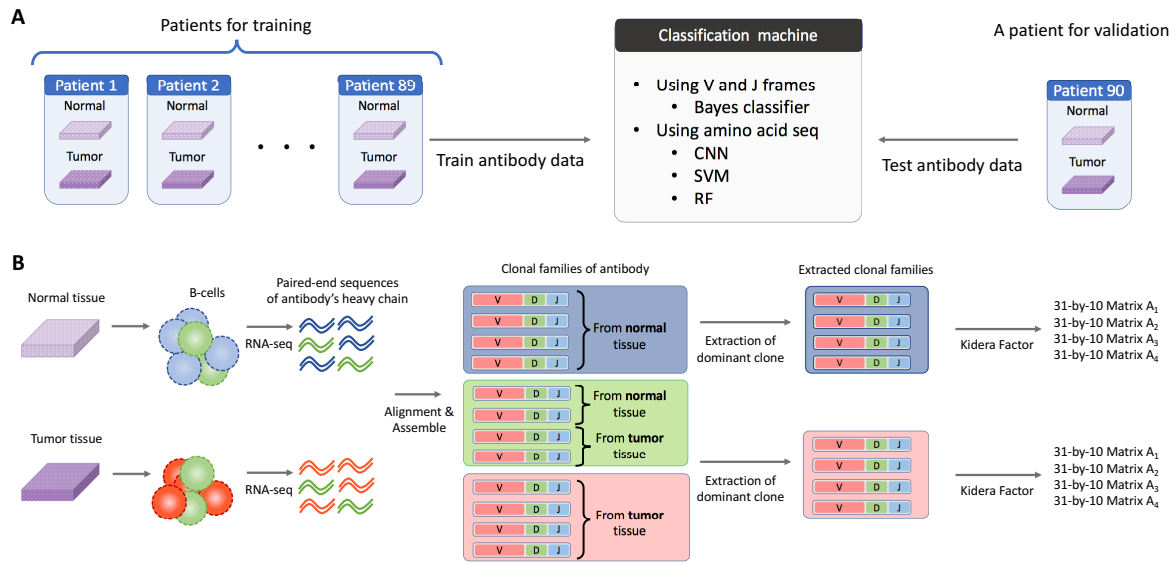

Figure 1 Overview of the machine-learning approach for determination of BCR sequences. **(A)** Example of leave-one-out validation. The classification machine is trained using samples from 89 patients and classification performance is determined based on the sample of the other patient. **(B)** Method for obtaining the BCRs/Igs data that is the query for the classification machine from a normal or tumor tissue.

segments, we used MiXCR, which is a tool for the universal framework processing of large immunome data from raw sequences for quantitating clonotypes [10]. Since the D frame is very short and surrounded by random nucleotides, the accuracy of D frame assignment is generally very low compared to that for the V and J frames. Thus, we did not apply D frame usage for further analysis.

2.3.2 Defining clonotypes

Estimation of the clonal lineage of BCRs was also conducted using MiXCR. The following clustering criteria were applied as described by Uduman et al [11]:

- The same combination of V- and J-frames
- The same length of CDR3
- A maximum of three nucleotide mismatches in CDR3

BCRs that satisfy the properties above are defined as a clonal family.

2.3.3 Extraction of dominant BCR clones

After constructing the clonal families, BCR clones used for training of the classification machine were selected. Each clone includes BCRs/Igs derived from tumor tissues as well as those derived from normal tissues. Thus, a clone with a tumor-derived BCRs/Igs content of 90% or more and the number of nucleotide-reads greater than 50 was defined as a tumor-specific clone. The same threshold was applied to defining a normal clone. The BCRs/Igs of these extracted clones can be used in training, but it is necessary to eliminate the influence of biased clonal size. Thus, three BCRs/Igs sequences were sampled randomly from each extracted clone.

The pipeline for the pre-processing of BCR sequences is outlined in Figure 1B.

2.3.4 Classification using V- and J-frame assignment

Given a BCR/Ig sequence whose V frame and J frame are v and j, respectively, the probability that the sequence is derived from tumor tissues or not can be determined by comparing $P_{\mathscr{C}|V,J}(\mathscr{C}_N|v,j)$ and $P_{\mathscr{C}|V,J}(\mathscr{C}_T|v,j)$, in which the subscripts of the conditional probability functions P indicate the types of the arguments, and $\mathscr{C}_N$ and $\mathscr{C}_T$ represent the classes of BCRs/Igs corresponding to normal tissues and tumor tissues, respectively. Using a maximum a posteriori (MAP) decision rule, classification is then conducted so that the posterior probability is maximized. In other words, if the former probability is higher, the BCR/Ig will be classified into $\mathscr{C}_N$, and if it is lower it will be classified into $\mathscr{C}_T$. Therefore, it is necessary to calculate these posterior probability mass functions in the training phase. By applying the Bayes theorem, the two probability functions can be transformed as $P_{\mathscr{C}|V,J}(\mathscr{C}_N|v,j) = \frac{P_{V,J|\mathscr{C}}(v,j|\mathscr{C}_N)P_{\mathscr{C}}(\mathscr{C}_N)}{P_{V,J}(v,j)}$ and $P_{\mathscr{C}|V,J}(\mathscr{C}_T|v,j) = \frac{P_{V,J|\mathscr{C}}(v,j|\mathscr{C}_T)P_{\mathscr{C}}(\mathscr{C}_T)}{P_{V,J}(v,j)}$, respectively, where $P_{V,J}(v,j)$ is irrelevant to the magnitude of the relationship between the two posterior probabilities. Thus, the determination will be done by comparing

$P_{V,J|\mathscr{C}}(v,j|\mathscr{C}_N)P_{\mathscr{C}}(\mathscr{C}_N)$ and $P_{V,J|\mathscr{C}}(v,j|\mathscr{C}_T)P_{\mathscr{C}}(\mathscr{C}_T)$. Calculation of the prior probabilities $P_{\mathscr{C}}(\mathscr{C}_N)$ and $P_{\mathscr{C}}(\mathscr{C}_T)$ is very easy because they are immediately obtained from the ratio of the class sizes included in the training data. The conditional probabilities $P_{V,J|\mathscr{C}}(v,j|\mathscr{C}_N)$ and $P_{V,J|\mathscr{C}}(v,j|\mathscr{C}_T)$ are then obtained by simply calculating the distribution of the V- and J-frames of BCRs/Igs sequences used in training data for each class.

Since this classification machine does not contain any hyperparameter, the performance of the machine is simply measured by conducting a single routine leave-one-out cross validation (LOOCV). The sketch of the LOOCV is shown as Algorithm 1 and an overview of an example training and validation process for a single sample (leave-one-out of patient 90) is provided in Figure 1A.

Algorithm 1 LOOCV of the Bayes classifier

```
procedure
    let 𝒜 be a whole set of BCRs/Igs
    split 𝒜 into sub-samples 𝒜^(1), ..., 𝒜^(90) depending on which of the 90 patients the BCRs/Igs belong to.
    𝒮 ← {𝒜^(1), ..., 𝒜^(90)}
    for 𝒜_test in 𝒮 do
        Calculate P_𝒞(𝒞_N), P_𝒞(𝒞_T), P_{V,J|𝒞}(v,j|𝒞_N) and P_{V,J|𝒞}(v,j|𝒞_T) using training data 𝒮 \ {𝒜_test}.
        Calculate test accuracy for 𝒜_test.
    Calculate averaged accuracy.
```

Figure 2 shows the distributions of the V and J germline gene assignment over all BCRs/Igs, and the classification strategy deducted by the joint distribution using the MAP rule.

2.4 Classification using amino acid feature vectors

As an alternative classification approach, we also used the amino acid sequences of BCRs/Igs. To take into account the physicochemical properties of BCRs/Igs, we adopted Kidera factors, which define the transformation of an amino acid into a 10-dimensional vector. The 10-dimensional vector was originally derived by applying multivariate analysis to 188 physicochemical properties of the 20 amino acids [12].

We focused on the complementarity determining regions (CDRs) of BCRs/Igs, which are supposed to be located at the center of the antigen-binding site. There are three CDRs for each BCR/Ig: CDR1, CDR2, and CDR3. Only CDRs are used in this classification, and their individual lengths are decisively fixed to be the same among all BCRs/Igs. The fixed lengths of CDRs were determined by the length distribution as 8, 8, and 15 for CDR1, CDR2, and CDR3, respectively. The following methodology was applied for fixing the sequence length:

- If the original length is the same as the target length, there is no modification to the sequence.
- If the original length is larger than the target length, the center of the sequence is trimmed so that the consequent length of the sequence is equivalent to the target length.
- If the original length is smaller than the target length, a pseudo amino acid whose feature vector is a zero vector is inserted into the center of the sequence so that the consequent length of the sequence is equivalent to the target length.

In this way, a 31(8 + 8 + 15)-by-10 matrix is obtained for each BCR/Ig. Therefore, the objective is to construct a classification machine denoted by $\mathscr{M}_{\boldsymbol{\theta}}$, such that $\mathscr{M}_{\boldsymbol{\theta}} : A_{\text{in}} \mapsto \mathscr{C}_N$ if a BCR/Ig is derived from normal tissue and $\mathscr{M}_{\boldsymbol{\theta}} : A_{\text{in}} \mapsto \mathscr{C}_T$ if a BCR/Ig is derived from tumor tissue, where $\boldsymbol{\theta}$ is a vector of free parameters of a classification machine and A_{in} is a 31-by-10 matrix obtained by applying Kidera factors transformation to the amino acid sequence of an input BCR/Ig.

2.5 Implementation details

We tackled the classification problem of BCRs/Igs using three machine learning methods: convolutional neural networks (CNN), support vector machine (SVM), and random forest (RF). The performances of the three methods were then compared to determine the most efficient approach for classification of BCRs/Igs into normal and tumor tissues.

2.5.1 CNN

CNN was implemented using Python-API of Tensorflow [13]. The properties of the neural network architecture were as follows:

1 Convolution layers
2 Max-pooling layers
3 Fully connected layers
4 Dropout
5 Rectified linear units (ReLU) as an activation function of hidden layers
6 Softmax function as an activation function of the output layer
7 Adam optimizer

The possible models of the CNN infinitely spread in the hyperparameter space. Therefore, a random search is used for hyperparameter optimization [14].

1. Convolution layers. In order to capture the local connectivity of the amino acid patterns of BCRs/Igs, the neural network model was composed of convolution layers with convolution filters. The shape of the convolution filter is of particular significance, because it defines the range pertaining to the local connectivity

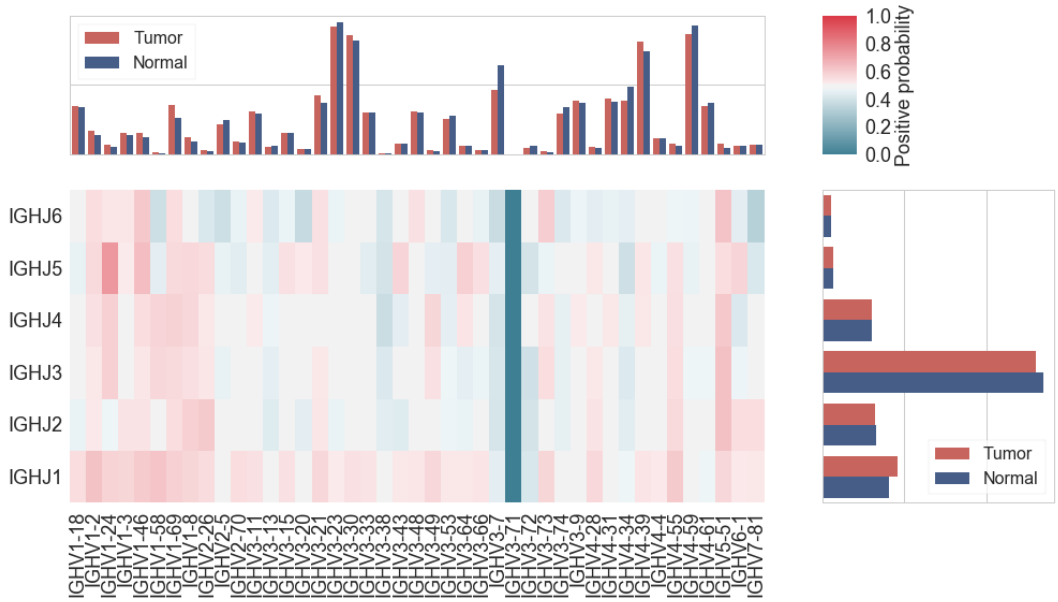

Figure 2 V and J germline gene assignment histogram for the normal and tumor environment, and heatmap of the positive probability of a BCR/Ig sequence whose V and J germline gene assignment is detected.

among amino acid sequences. The number of convolution layers and the number of filters to use in each layer also have a significant influence on the behavior of the networks. Therefore, these are all considered to be hyperparameters.

2. Max-pooling layers. Each convolution layer is followed by a max-pooling layer with max-pooling filters. The size of the filter is a hyperparameter.

3. Fully connected layers. A convolution and pooling block, which is a repetitive structure of convolution layers and pooling layers, is followed by some fully connected layers, in which the number of the layers and the number of units of each layer are hyperparameters.

4. Dropout. A dropout technique is used to prevent overfitting in the fully connected layers [15]. A dropout rate defines the rate of units in the fully connected layers that are inactivated in a minibatch iteration. This rate is a hyperparameter.

5. ReLU. To realize the non-linearly separatable classification, ReLU is used as an activation function. ReLU has relatively low computational complexity and is supposed to solve the gradient elimination problem.

6. Softmax function. The Softmax function, which is a generalized form of the logistic function, is used as an activation function of the output layer. Because our problem is the classification of BCRs/Igs, it is desirable to obtain the confidence of the estimation from the classification machine. Toward this end, the output of the Softmax function is normalized and is in the range $[0, 1]$; thus, use of this function is reasonable in this situation.

7. Adam optimizer. The Adam optimizer is used as the optimization method of the networks [16].

2.5.2 SVM

SVM was implemented with scikit-learn using the radial basis function (RBF) kernel as the kernel. The values C and γ of the SVM hyperparameters were tuned using a grid search, where the parameter C trades off any misclassification of training examples against the simplicity of the decision surface [17], and γ parameter defines the extent of the influence of a single training example. The possible values of the hyperparameters C and γ are defined as $[10^0, 10^1, 10^2, 10^3]$ and $[10^{-2}, 10^{-3}, 10^{-4}, 10^{-5},]$, respectively.

2.5.3 RF

RF was also implemented with scikit-learn [17]. the number of features to consider when looking for

the best split is tuned as a hyperparameter of the RF model and its possible values are $\sqrt{|\text{features}|}$, $\log_2 |\text{features}|$, and $|\text{features}|$, where $|\text{features}|$ is the number of features of an input BCR/Ig.

2.6 Model selection of machine learning

To determine the hyperparameters of the classification machines, double cross-validation was conducted [18]. Since the diversity of BCRs/Igs differs among individuals, for this study, it was guaranteed that BCRs/Igs from a given patient would not be included in both the training and test datasets. An overview of the double cross-validation step is shown as Algorithm 2. Double cross-validation was conducted for each machine learning method independently so that the best model was selected for CNN, SVM, and RF, respectively. Once the best models were selected, the test accuracy of the classification for one patient was calculated 10 times by iterating the training and test phases because the results showed some variance. The performance of CNN depends on how the network is initialized and the order of data fed to it. The performance of SVM and RF varies depending on which data is used (these batch-learning algorithms cannot deal with whole BCRs/Igs datasets).

Algorithm 2 Double cross-validation

procedure
  let $\boldsymbol{\mathcal{A}}$ be a whole set of BCRs/Igs
  split $\boldsymbol{\mathcal{A}}$ into sub-samples $\boldsymbol{\mathcal{A}}^{(1)}$, ..., $\boldsymbol{\mathcal{A}}^{(90)}$ depending on which of the 90 patients the BCRs/Igs belong to.
  $\boldsymbol{\mathcal{S}} \leftarrow \{\boldsymbol{\mathcal{A}}^{(1)}, \ldots, \boldsymbol{\mathcal{A}}^{(90)}\}$
  $\boldsymbol{\mathcal{S}}^{(1)} \leftarrow \{\boldsymbol{\mathcal{A}}^{(1)}, \ldots, \boldsymbol{\mathcal{A}}^{(45)}\}$
  $\boldsymbol{\mathcal{S}}^{(2)} \leftarrow \{\boldsymbol{\mathcal{A}}^{(46)}, \ldots, \boldsymbol{\mathcal{A}}^{(90)}\}$
  for $i = 1, 2$ **do**
    For each candidate model $m \in \boldsymbol{\mathcal{M}}$, calculate the LOOCV averaged accuracy and select the best model $\hat{m}$.
    $\hat{m} \leftarrow \arg\max_{m \in \boldsymbol{\mathcal{M}}} LOOCV(m, \boldsymbol{\mathcal{S}}^{(i)})$
    for $\boldsymbol{\mathcal{A}}_{test}$ in $\boldsymbol{\mathcal{S}} \setminus \boldsymbol{\mathcal{S}}^{(i)}$ **do**
      for $j = 1, \ldots, 10$ **do**
        Train the classification machine with $\boldsymbol{\mathcal{A}} \setminus \boldsymbol{\mathcal{A}}_{test}$.
        Test $\boldsymbol{\mathcal{A}}_{test}$ with the trained machine.
      Calculate the averaged test accuracy and variance. of $\boldsymbol{\mathcal{A}}$ using model $\hat{m}$

2.7 Tissue-level classification

Consider a classification machine $\mathscr{N}$ such that $\mathscr{N} : \boldsymbol{\mathcal{A}}_{\text{in}} \mapsto \mathscr{C}_N$ if the BCRs/Igs in $\boldsymbol{\mathcal{A}}_{\text{in}}$ are derived from a normal tissue, and $\mathscr{N} : \boldsymbol{\mathcal{A}}_{\text{in}} \mapsto \mathscr{C}_T$ if the BCRs/Igs in $\boldsymbol{\mathcal{A}}_{\text{in}}$ are derived from a tumor tissue, where $\boldsymbol{\mathcal{A}}_{\text{in}}$ is the input set of BCRs/Igs that corresponds to a dataset obtained from a tissue of a patient. In order to simulate this diagnostic simulation, whole BCRs/Igs $\boldsymbol{\mathcal{A}}$ were split into sub-samples $\boldsymbol{\mathcal{A}}_N^{(1)}$, ..., $\boldsymbol{\mathcal{A}}_N^{(90)}$, $\boldsymbol{\mathcal{A}}_T^{(1)}$, ..., $\boldsymbol{\mathcal{A}}_T^{(90)}$ where the subscripts N and T indicate that the set of BCRs/Igs belongs to normal tissues and tumor tissues, respectively. Thus, BCRs/Igs that are derived from the same patient and the same tissue type (normal or tumor) are in one sub-sample.

Although only environment-specific clonal families were used when constructing BCR/Ig-clone-level classification machine, all of the RNA sequence data of BCRs/Igs were used when estimating the tissue-types(normal or tumor), taking into account actual diagnostic situations.

2.7.1 Classification using the clonal entropy of BCRs/Igs

Suppose that when conducting clonal clustering to a set of BCRs/Igs $\boldsymbol{\mathcal{A}}$, n clustered sets of BCRs/Igs $\{\boldsymbol{\mathcal{A}}_1, \ldots, \boldsymbol{\mathcal{A}}_n\}$ can be obtained. The frequency f_i for $\boldsymbol{\mathcal{A}}_i$ is then calculated as $f_i = \frac{|\boldsymbol{\mathcal{A}}_i|}{|\boldsymbol{\mathcal{A}}|}$ for all $i = 1, \ldots, n$. Let $\boldsymbol{f}$ be $\{f_1, \ldots, f_n\}$, and the Shannon entropy for $\boldsymbol{f}$ is defined as

$$H(\boldsymbol{f}) = -\sum_{f \in \boldsymbol{f}} (f \log f) \tag{1}$$

The Shannon entropy was calculated using Alakazam, which is part of the adaptive immune receptor repertoire sequencing framework Change-O [19].

The value of the Shannon entropy can be used to solve the classification problem, because it is known that BCRs/Igs in tumor tissues are likely to comprise relatively less diverse clonal families than those found in normal tissues [8]. Herein, we defined the tissue-level classification machine $\mathscr{N}_\xi^{\text{Shannon}}$ that uses the Shannon entropy as

$$\mathscr{N}_\xi^{\text{Shannon}} : \boldsymbol{\mathcal{A}} \mapsto \begin{cases} \mathscr{C}_N & (H(\boldsymbol{f}) > \xi) \\ \mathscr{C}_T & (\text{otherwise}) \end{cases} \tag{2}$$

where $\boldsymbol{f}$ is the frequency obtained by $\boldsymbol{\mathcal{A}}$, and ξ is a decision threshold of the Shannon entropy. By changing the threshold ξ, the ROC curve can be drawn.

2.7.2 Classification using the averaged score of BCR/Ig-clone-level classification

Since the CNN model achieved the best performance in BCR/Ig-clone-level classification from the trained classification machine, $\mathscr{M}_{\boldsymbol{\theta}}$ derived using CNN was used for classifying a BCR/Ig into $\mathscr{C}_N$ or $\mathscr{C}_T$; we denote the trained machine as $\mathscr{M}_{\boldsymbol{\theta}}^{(\text{CNN})}$. CNN provides the probability that an input BCR/Ig is derived from tumor tissues because CNN uses the softmax function as the activation function of the output layer. Therefore, we can define $\mathscr{M}_{\boldsymbol{\theta}}^{(\text{CNN})} : A_{\text{in}} \mapsto P_t$, where P_t is the

probability. Here, classification of a set of n BCRs/Igs $\boldsymbol{\mathcal{A}}(n)$ by $\mathscr{N}_{\boldsymbol{\theta},\xi}^{(\mathrm{CNN})}$ is defined as follows:

$$\mathscr{N}_{\boldsymbol{\theta},\xi}^{(\mathrm{CNN})} : \boldsymbol{\mathcal{A}} \mapsto \begin{cases} \mathscr{C}_N & \left(\frac{1}{n} \sum_{A \in \boldsymbol{\mathcal{A}}(n)} \mathscr{M}_{\boldsymbol{\theta}}^{(\mathrm{CNN})}(A) \ \leq \ \xi \right) \\ \mathscr{C}_T & (\text{otherwise}) \end{cases} \tag{3}$$

whose performance was also measured based on the ROC curve.

2.7.3 Ensemble of the Shannon entropy and averaged score of BCR/Ig-clone-level classification

Next, we consider a classification machine that uses both the Shannon entropy and the averaged score of $\mathscr{M}_{\boldsymbol{\theta}}^{(\mathrm{CNN})}$. In other words, this is classification of $\boldsymbol{\mathcal{A}}$ using the following two features:

- $H(\boldsymbol{f})$
- $\frac{1}{n} \sum_{A \in \boldsymbol{\mathcal{A}}(n)} \mathscr{M}_{\boldsymbol{\theta}}^{(\mathrm{CNN})}(A)$

To estimate the class of input from these two heterogeneous values, the classification was conducted using the linear SVM implemented with scikit-learn [17]. The performance was determined by iterating one LOOCV and by drawing the ROC curve.

3 Results

3.1 BCR/Ig-clone-level classification

As shown in Figure 3, we demonstrated that CNN is the best classifier among the three machine learning methods tested for classifying an input into $\mathscr{C}_N$ and $\mathscr{C}_T$ using the amino acid information.

Although CNN emerged as the best model, all three methods showed a significantly better ability to classify the BCRs/Igs from normal and tumor tissues than random assignment (the accuracy was greater than 0.5 in all cases).

Figure 4 shows the ROC curves of the performance of the CNN model compared to that obtained using a Bayes classifier based on V and J germline gene assignment of BCR/Ig, which was constructed by averaging 10 iterations of the test estimations. As mentioned above, since the classification performance of CNN is dependent on the initialization of the network and the order of data fed to the networks to some extent, we have also drawn the 1-standard deviation error area in the curve to determine the overall performance. By contrast, the Bayes classifier is a deterministic model that simply calculates the joint distributions of V and J frames and classifies test data according to the MAP rule. Therefore, there is no error area corresponding to the ROC curve of Bayes classifier. Comparison of the ROC curves clearly demonstrates that CNN has greater classification performance than the Bayes clas-

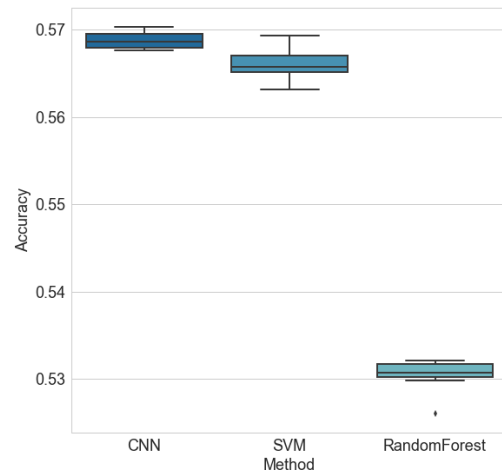

Figure 3 Boxplots of the performance for three different machine learning methods: CNN, SVM, and RF. One box represents information of 10 iterations of test estimations for each method.

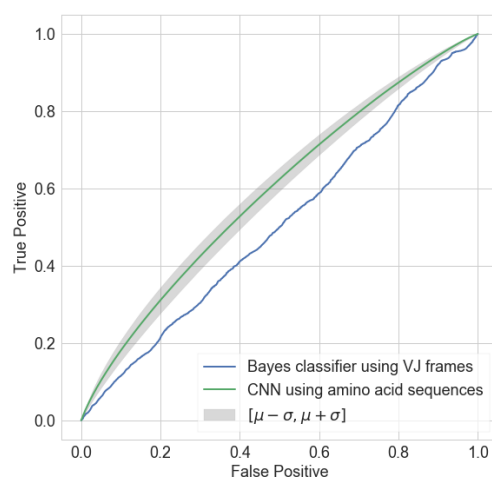

Figure 4 ROC curves of two different methods using different features of BCRs/Igs. The blue curve is the ROC curve of the Bayes classifier using the V and J germline gene assignment, and the green curve is that of the CNN classifier using amino acids feature-vectors, or Kidera factors. The grey shaded area shows the width of 1 standard deviation of the ROC curve of the CNN obtained by calculating the ROC curves of 10 iterations of the test estimation of the CNN.

sifier. According to this result, by using the amino acid sequences of BCRs/Igs and the feature vectors of amino acids, the classification machine is able to detect some environment-specific features of BCRs/Igs.

3.2 Tissue-level classification

One of the main goals of this work is to precisely estimate the type of tissue, normal or tumor, using BCRs/Igs data obtained from the tissue sample. This problem is considered as the classification of a set of BCRs/Igs into two classes, which differs from the BCR/Ig-clone-level classification described above in that we can use many BCRs/Igs to determine the type of tissue.

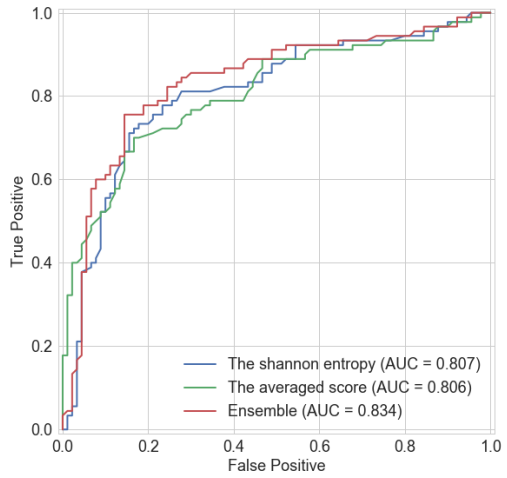

Figure 5 ROC curves of three different strategies used to estimate the tissue environment. The blue line is the ROC curve of a classifier based on the Shannon entropy of BCRs/Igs obtained from the target tissue. The green line is the ROC curve of a classifier based on the averaged positive probability of the target tissue calculated by the trained BCR/Ig-clone-level classification machine. The red line is the ROC curve of a linear support vector classifier using both of the features indicated above.

Figure 5 shows three ROC curves corresponding to the following classifiers:

- Classifier setting threshold of the Shannon entropy.
- Classifier setting threshold of the averaged score.
- Support vector classifier that makes a decision line following a two-variables plane for both the Shannon entropy and the averaged score (ensemble).

Although the three ROC curves showed similar results, the ensemble method showed better performance with a higher area under the curve (AUC) value.

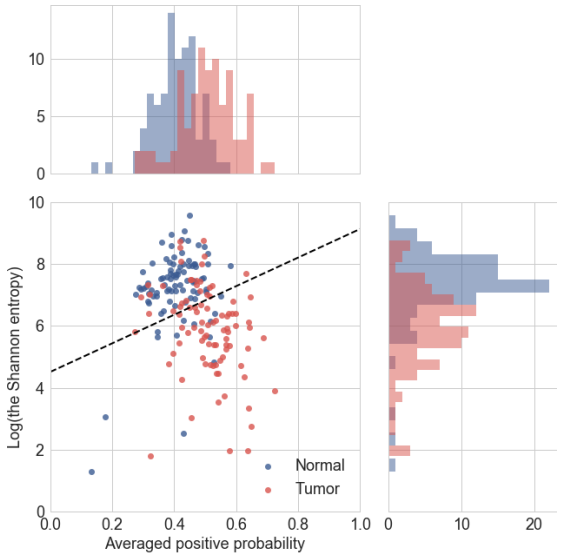

Figure 6 Three types of information to evaluate the performance of the classifier. First, the histogram that shares the x-axis of the scatter plot describes the distributions of the averaged positive probability for normal and tumor samples. Therefore, the total count of the distribution is equivalent to 180, which is the number of samples obtained from the 90 patients. Second, the other histogram shows the distribution of the Shannon entropy of the samples. Finally, the scatter plots represent the joint distribution of these two variables. The line drawn in the scatter plot is the decision line determined by SVM using the linear kernel.

Figure 6 clearly shows the behavior of linear classification in two-dimensional space. The x-axis is the logarithm of the Shannon entropy value and the y-axis is the averaged score. The decision line of the linear classifier is not parallel to x-axis or y-axis.

Therefore, these results demonstrated that the performance of classification is sufficient to accurately estimate the tissue environment and both the Shannon entropy of clonal families and averaged score of the BCR/Ig-clone-level classifier are important for the estimation.

4 Discussion

In the first test of BCR/Ig-clone-level classification, the classification performance was poor when using only the V and J germline frame assignment. However, the performance was improved when incorporating the physicochemical properties of the amino acid sequences of BCRs/Igs. These results indicate that the insertion of random residues during the VDJ recombination process or somatic hypermutations in BCRs/Igs create contribute to the specificities of BCRs/Igs in recognizing normal and tumor tissues. Of note, the classification machine could differentiate BCRs/Igs between normal and tumor microenvironments from a

single patient based on training with the BCRs/Igs from the other patients, which indicates that the amino acid patterns of normal and tumor tissues have shared properties among patients.

Although the BCR/Ig-clone-level classification performance was significantly better than random, it is far from perfect, which leads to difficulty for the in-depth characterization of BCR/Ig-sequences in TSAs (e.g., identifying critical amino acid residues of CDRs in TSAs). This is probably because the classification problem was considered under label-contaminated conditions. Ideally, we would like to classify BCRs/Igs that can and cannot recognize tumor-specific antigens. However, in the present study, we defined the $\mathscr{C}_T$ as the class of BCRs/Igs that are derived from the tumor tissue. Although we extracted dominant clonal families to obtain more refined data, not all of the extracted BCRs/Igs recognize tumor-specific antigens. Insufficient performance is therefore inevitable when solving the classification problem without any assumption or knowledge of antigens or epitopes.

Although this classification problem is difficult to resolve completely, the second experiment based on tissue-level classification achieved an AUC value of 0.834, which indicates that when using BCRs/Igs repertoires alone, cancer could be diagnosed even if the tissues do not contain cancer cells. It is well known that biopsy evaluations are not perfect. Indeed, single biopsies of gastric cancer tissues fail to show malignancy in 20% to 30% of cancer cases [20, 21]. Therefore, our approach helps to rescue some of the false negatives of gastric biopsies to diagnose gastric cancers correctly by examining BCRs/Igs repertoires of B-cells near around the cancer cells. Further experimental validation will be required to determine the clinical application of this concept.

5 Conclusion

We developed classification machines that can estimate whether an input BCR/Ig amino acid sequence is derived from tumor or normal tissue. Using feature vectors that take into account the physicochemical properties of amino acid sequences on CDRs, we revealed that BCRs/Igs in the normal and tumor microenvironment have distinct physicochemical properties. To test the performance of this approach, we conducted a computer experiment to detect the environment of an input tissue from a set of antibodies. In this experiment, we used 90 paired normal and tumor tissues from patients as a dataset and achieved an AUC value of 0.834 from the ROC curve, representing good classification power. Although we focused specifically on BCRs/Igs in gastric cancer in this study, the same approach could be applied to TCRs and BCRs/Igs for other diseases such as other cancers and autoimmune and infectious diseases. From these perspectives, our method can provide new knowledge related to humoral immunity.

Ethics approval and consent to participate
This study has been approved by the institutional review boards of the University of Tokyo and Tokyo Medical and Dental University. Ninety frozen specimens of gastric cancer that were surgically resected between 2009 to 2016 at the University of Tokyo Hospital were analyzed in this study with written informed consent from the patients.

Consent for publication
Not applicable

Competing interests
The authors declare no competing interests.

Funding
This study was supported by the AMED Basic Science and Platform Technology Program for Innovative Biological Medicine (17am0301011h0004), and the AMED Project for Cancer Research and Therapeutic Evolution (P-CREATE) (17cm0106514h0002) to S.I.

Authors' contributions
H. Konishi, D.K., H. Katoh, and S. Ishikawa conceived of the study and design. H. Katoh, A.Y., H.Koda, and S.A. were involved in processing tissue samples and acquisition of repertoire sequencing data. Y.S. and M.F. contributed clinical tissue sample collection. H. Konishi implemented the pipeline for the computer experiments. S. Ishikawa, S. Imoto, and R.Y. are supervisors of this research and provided meaningful suggestions and discussions. H. Konishi, D.K., and S.Ishikawa wrote the manuscript.

Author details
[1]Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, 113-8510 Tokyo, Japan. [2]Institute of Medical Science, Health Intelligence Center, The University of Tokyo, 108-8639 Tokyo, Japan. [3]Institute of Medical Science, Human Genome Center, The University of Tokyo, 108-8639 Tokyo, Japan. [4]Department of Pathology, Graduate School of Medicine, The University of Tokyo, 113-0033 Tokyo, Japan. [5]Department of Gastrointestinal Surgery, Graduate School of Medicine, The University of Tokyo, 113-0033 Tokyo, Japan.

References

1. Wang, C., Liu, Y., Cavanagh, M.M., Le Saux, S., Qi, Q., Roskin, K.M., Looney, T.J., Lee, J.-Y., Dixit, V., Dekker, C.L., Swan, G.E., Goronzy, J.J., Boyd, S.D.: B-cell repertoire responses to varicella-zoster vaccination in human identical twins. Proc. Natl. Acad. Sci. **112**(2), 500–505 (2015). doi:10.1073/pnas.1415875112. NIHMS150003
2. Rechavi, E., Lev, A., Lee, Y.N., Simon, A.J., Yinon, Y., Lipitz, S., Amariglio, N., Weisz, B., Notarangelo, L.D., Somech, R.: Timely and spatially regulated maturation of B and T cell repertoire during human fetal development. Sci. Transl. Med. **7**(276), 276–25 (2015). doi:10.1126/scitranslmed.aaa0072
3. Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L.E., Rubelt, F., Ji, X., Han, A., Krams, S.M., Pettus, C., Haas, N., Arlehamn, C.S.L., Sette, A., Boyd, S.D., Scriba, T.J., Martinez, O.M., Davis, M.M.: Identifying specificity groups in the T cell receptor repertoire. Nature **547**(7661), 94–98 (2017). doi:10.1038/nature22976
4. Dash, P., Mcclaren, J.L., Iii, T.H.O., Rothwell, W., Todd, B., Morris, M.Y., Becksfort, J., Reynolds, C., Brown, S.A., Doherty, P.C., Thomas, P.G., Wang, G.C., Dash, P., McCullers, J.A., Doherty, P.C., Thomas, P.G., Fiore-Gartland, A.J., Hertz, T., Wang, G.C., Sharma, S., Souquette, A., Crawford, J.C., Clemens, E.B., Nguyen, T.H.O., Kedzierska, K., La Gruta, N.L., Bradley, P., Thomas, P.G.: Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature **121**(1), 128–4212842 (2017). doi:10.1038/nature22383. NIHMS150003
5. Thomas, N., Best, K., Cinelli, M., Reich-Zeliger, S., Gal, H., Shifrut, E., Madi, A., Friedman, N., Shawe-Taylor, J., Chain, B.: 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 **30**(22), 3181–3188 (2014). doi:10.1093/bioinformatics/btu523
6. Epstein, M., Barenco, M., Klein, N., Hubank, M., Callard, R.E.: Revealing individual signatures of human T cell CDR3 sequence repertoires with Kidera Factors. PLoS One **9**(1), 1–10 (2014). doi:10.1371/journal.pone.0086986
7. Trück, J., Ramasamy, M.N., Galson, J.D., Rance, R., Parkhill, J., Lunter, G., Pollard, A.J., Kelly, D.F.: Identification of Antigen-Specific B Cell Receptor Sequences Using Public Repertoire Analysis. J. Immunol. **194**(1), 252–261 (2015). doi:10.4049/jimmunol.1401405
8. Katoh, H., Komura, D., Konishi, H., Suzuki, R., Yamamoto, A., Kakiuchi, M., Sato, R., Ushiku, T., Yamamoto, S., Tatsuno, K., Oshima, T., Nomura, S., Seto, Y., Fukayama, M., Aburatani, H., Ishikawa, S.: Immunogenetic Profiling for Gastric Cancers Identifies Sulfated Glycosaminoglycans as Major and Functional B Cell Antigens in Human Malignancies. Cell Rep. **20**(5), 1073–1087 (2017). doi:10.1016/j.celrep.2017.07.016
9. Fridman, W.H., Pagès, F., Sautès-Fridman, C., Galon, J.: The immune contexture in human tumours: impact on clinical outcome. Nat. Rev. Cancer **12**(4), 298–306 (2012). doi:10.1038/nrc3245. NIHMS150003
10. Bolotin, D.A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I.Z., Putintseva, E.V., Chudakov, D.M.: MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods **12**(5), 380–381 (2015). doi:10.1038/nmeth.3364
11. Uduman, M., Shlomchik, M.J., Vigneault, F., Church, G.M., Kleinstein, S.H.: Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. J. Immunol. **192**(3), 867–74 (2014). doi:10.4049/jimmunol.1301551
12. Kidera, A., Konishi, Y., Oka, M., Ooi, T., Scheraga, H.a.: Statistical analysis of the physical properties of the 20 naturally occurring amino acids. Journal of Protein Chemistry **4**, 23–55 (1985)
13. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org (2015). http://tensorflow.org/
14. Bergstra, J., Bengio, Y.: Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research **13**, 281–305 (2012)
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014). doi:10.1214/12-AOS1000. 1102.4807
16. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG], 1–15 (2014). 1412.6980
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
18. Fran, R.: Double Cross Validation for Model Based Classification (2006). https://www.r-project.org/conferences/useR-2006/Abstracts/Francois+Langrognet.pdf
19. Gupta, N.T., Vander Heiden, J.A., Uduman, M., Gadala-Maria, D., Yaari, G., Kleinstein, S.H.: Change-o: a toolkit for analyzing large-scale b cell immunoglobulin repertoire sequencing data. Bioinformatics, 1–3 (2015). doi:10.1093/bioinformatics/btv359
20. Graham DY, S.J., Cain GD, G.F.: Prospective evaluation of biopsy number in the diagnosis of esophegeal and gastric carcinoma. Gastroenterology, 82–228 (1982)
21. Choi, Y., Choi, H.S., Jeon, W.K., Kim, B.I., Park, D.I., Cho, Y.K., Kim, H.J., Park, J.H., Sohn, C.I.: Optimal number of endoscopic biopsies in diagnosis of advanced gastric and colorectal cancer. J. Korean Med. Sci. **27**(1), 36–39 (2012). doi:10.3346/jkms.2012.27.1.36