Hindawi
BioMed Research International
Volume 2021, Article ID 6639698, 10 pages
https://doi.org/10.1155/2021/6639698

Research Article

Identifying the Immunological Gene Signatures of Immune Cell Subtypes

Yu-Hang Zhang,[1,2] **Zhandong Li**,[3] **Tao Zeng**,[4] **WenCong Lu**,[5] **Tao Huang**,[6] **and Yu-Dong Cai**[1]

[1]*School of Life Sciences, Shanghai University, Shanghai 200444, China*
[2]*Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[3]*College of Food Engineering, Jilin Engineering Normal University, Changchun, China*
[4]*Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China*
[5]*Department of Chemistry, College of Sciences, Shanghai University, Shanghai 200444, China*
[6]*Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Received 24 December 2020; Revised 25 January 2021; Accepted 10 February 2021; Published 19 February 2021

Academic Editor: Luis Alberto Morales Quintana

The immune system is a complicated defensive system that comprises multiple functional cells and molecules acting against endogenous and exogenous pathogenic factors. Identifying immune cell subtypes and recognizing their unique immunological functions are difficult because of the complicated cellular components and immunological functions of the immune system. With the development of transcriptomics and high-throughput sequencing, the gene expression profiling of immune cells can provide a new strategy to explore the immune cell subtyping. On the basis of the new profiling data of mouse immune cell gene expression from the Immunological Genome Project (ImmGen), a novel computational pipeline was applied to identify different immune cell subtypes, including $\alpha\beta$ T cells, B cells, $\gamma\delta$ T cells, and innate lymphocytes. First, the profiling data was analyzed by a powerful feature selection method, Monte-Carlo Feature Selection, resulting in a feature list and some informative features. For the list, the two-stage incremental feature selection method, incorporating random forest as the classification algorithm, was applied to extract essential gene signatures and build an efficient classifier. On the other hand, a rule learning scheme was applied on the informative features to construct quantitative expression rules. A group of gene signatures was found as qualitatively related to the biological processes of four immune cell subtypes. The quantitative expression rules can efficiently cluster immune cells. This work provides a novel computational tool for immune cell quantitative subtyping and biomarker recognition.

1. Introduction

The immune system is a complicated defensive system that comprises multiple functional cells and molecules acting against endogenous and exogenous pathogenic factors [1–3]. With organism evolution, the immune system gradually becomes complicated and finally forms layered defensive mechanisms, including innate and adaptive immune systems, in advanced creatures such as mammals [4, 5]. Both these immune systems are complicated and constitute cellular (immune cells) and noncellular components (immunoregulatory molecules) [6]. The cellular component is diverse on structural and functional levels [6]. In particular, each single immune response could involve multiple subtypes of immune cells, and each immune cell subtype may play various important roles in multiple immune responses, thereby constituting a complicated regulatory network on the cellular level [6].

Identifying the subtypes of immune cells and recognizing their unique immunological functions are difficult due to the complicated cellular components and immunological functions of the immune system. The only standards are the

typical molecular markers recognized by cytobiology [7–9]. However, such biomarkers cannot accurately reflect the components of immune cells and reveal their immunoregulatory mechanisms in vivo. With the development of transcriptomics and high-throughput sequencing [10, 11], the gene expression profiling of immune cells can provide a new strategy to explore the complicated immunoregulatory mechanisms, e.g., detailed immune cell subtyping. Gene expression profiling with transcriptomic analysis can reflect the typical gene expression pattern of each cell subgroup. In accordance with the central dogma of molecular biology, cells with different gene expression patterns may have varying proteomic features and biological functions and therefore must be clustered into different cell subtypes [12, 13]. Differentially expressed genes or transcripts may be potential biomarkers for the identification of a given cell subgroup/subtype. Therefore, transcriptomic sequencing is a novel technique for immune cell subtyping through the recognition of cell subgroups and their respective biomarkers and functions.

The Immunological Genome Project (ImmGen) [14, 15] is aimed at establishing a systematic panorama of gene expression and regulatory networks of all immune cells by using a mouse model. Initiated in 2008, this collaborative study [15] has analyzed the differentiation, maturation, active responses, effector stages, tissue localizations, and genetic variations of more than 250 subtypes of immune cells in mouse models. A systematic immunoregulatory network in mice, which encompasses the innate and adaptive immune systems, has been established by this project through systematic quality check control and standardized analyzed conditions. The analyzed data and constructed networks can be accessed by using a dedicated data browser, and the raw sequencing and microarray data can be accessed in a public database [16]. This project provides reliable resources for immune cell gene expression profiling for further exploration and research.

As reported by various previous publications, the immune system is composed by multiple immune cell subtypes which are impossible for us to study one by one. For further analyses on the mouse immune cell gene expression profiling, we focused on four basic subtypes of immune cells: $\alpha\beta$ T cells [17], B cells [18], $\gamma\delta$ T cells [19], and innate lymphocytes [20]. Among them, $\alpha\beta$ T cells are the majority of T cells with $\alpha\beta$ TCR, contributing to immune-mediated cell death as adaptive immune responders [21]. As for $\gamma\delta$ T cells, as quite a functioning minority of T cells, which are different from $\alpha\beta$ T cells mainly functioning for adaptive immune responses, there are three major functions for $\gamma\delta$ T cells: (1) regulating other immune cells for central and peripheral immune responses [22, 23], (2) contributing to thermogenesis to maintain body temperature [23], and (3) regulating autoimmune responses [24]. B cells are the main participator for antibody medicated adaptive humoral immune responses with complicated activation processes relying either on T cells or not [25]. Innate lymphocytes are a group of immune cells that participate in the innate immune responses with cytotoxic natural killer (NK) cells in the circulating system and innate lymphoid cells (ILCs) in the tissue-resident microenvironment [26]. Therefore, as introduced above, the four major subgroups of immune cells have quite different biological functions, controlling the basic biological functions of the immune system: innate and adaptive immune responses. Therefore, considering the significance of such four subgroups of cells in the immune system, we selected such immune cell subtypes for detailed classification studies on the murine transcriptomic level in this study.

A new batch of profiling data for mouse immune cell gene expression has been released by ImmGen on the Gene Expression Omnibus (GEO) database provided by NCBI [16]. On the basis of these data, a novel computational pipeline was applied to distinguish different immune cell subtypes including $\alpha\beta$ T cells, B cells, $\gamma\delta$ T cells, and innate lymphocytes. These four subtypes of immune cells are the major effective immune cells in mouse immune systems, contribute to different immune responses, and constitute a complicated immunoregulatory system by playing unique roles and interacting with each other. The powerful feature selection method, the Monte-Carlo Feature Selection (MCFS) [27], was first applied to the profiling data. We obtained a feature list and some informative features. Of the feature list, the two-stage incremental feature selection (IFS) [28] method with random forest (RF) [29] as the classification algorithm was executed to extract essential gene signatures and build an efficient RF classifier. Furthermore, some quantitative expression rules were constructed on the informative features via a rule learning scheme. Extensive analysis on gene signatures and rules were performed by literature review. All in all, we recognized the typical gene signatures and rules of each key immune cell subtype, which were helpful to explore the complicated regulatory mechanisms of immune systems.

2. Materials and Methods

2.1. Dataset. The system-wide mouse RNA-Seq data released by the Immunological Genome Project (ImmGen) were downloaded from GEO (Gene Expression Omnibus) under accession number of GSE109125 [30]. The reads were mapped to the mouse reference genome (mm10), and then the uniquely mapped reads were assigned to genes according to GENCODE annotation (vM12). The genes were quantified as counts per million (CPM) using edgeR [31]. A total of 49,480 genes and 112 samples were obtained from the four types of cells: 46 $\alpha\beta$ T cells, 33 B cells, 13 $\gamma\delta$ T cells, and 20 innate lymphocytes. The original dataset included more samples and cell types (46 $\alpha\beta$T cells, 33 B cells, 20 innate lymphocytes, 13 $\gamma\delta$ T cells, 12 stromal cells, 11 stem cells, 8 dendritic cells, 7 macrophages, 6 granulocytes, and 1 mast cell). Since the sample sizes of many cell types was extremely small, we only kept the top four cell types with enough samples (46 $\alpha\beta$ T cells, 33 B cells, 20 innate lymphocytes, and 13 $\gamma\delta$ T cells). The gene expression signatures were identified for such four major types of immunological cells.

2.2. Feature Selection. The MCFS [27] was first used to identify interpretable information about gene discrimination among the different groups of immune cells. Then, the two-stage IFS [28] method was applied to obtain genes with

strong classification ability to improve the component recognition for the immune system.

2.2.1. Monte-Carlo Feature Selection. MCFS is a classic feature selection method for distinguishable features and ranks the features through guided sampling. For specific steps, multiple feature subsets with m features were arbitrarily chosen from the original M features ($m << M$), the bootstrap dataset was trained for each specific feature subset, and the generated p decision trees were evaluated. The $p \times t$ decision trees were obtained by repeating the above steps t times. The accuracy weights of generated multiple decision trees provide a relative importance (RI) score for each feature, which was calculated as below.

$$\mathrm{RI}_f = \sum_{\tau=1}^{pt} (\mathrm{wAcc})^u \mathrm{IG}\left(n_f(\tau)\right) \left(\frac{\mathrm{no.in}\, n_f(\tau)}{\mathrm{no.in}\, \tau}\right)^v, \quad (1)$$

where wAcc is the weighted accuracy and $n_f(\tau)$ is a node of feature f in decision tree τ. The information gain of $n_f(\tau)$ is expressed as $\mathrm{IG}(n_f(\tau))$, no.in $n_f(\tau)$ is the number of training samples in $n_f(\tau)$, and u and v are the two weighting factors. Here, we adopted the MCFS program obtained from http://www.ipipan.eu/staff/m.draminski/mcfs.html. For convenience, default parameters were used.

After obtaining the RI scores of all features, we ranked all features in a list with the decreasing order of their RI scores. In addition to the feature list, the MCFS method also yields some most important features, called informative features in this study, which are some top features in the list. These features are accessed by a permutation test on class labels and one-sided Student's t-test.

2.2.2. Two-Stage Incremental Feature Selection. IFS is a feature selection method that accurately distinguishes samples from different classes by screening a set of optimal features. The features in the ranked list from MCFS can be sorted in a descending order according to their RI scores as mentioned above. Such feature list can help the classification algorithm in producing optimal performance. The original IFS must test all possible feature subsets, which are constructed from the feature list, to filter out the optimal feature subset that can identify samples' classes with best performance. Here, due to the large number of features (~50000), inducing lots of time to test all feature subsets, we designed a two-stage IFS method. In the first stage, we constructed candidate feature subsets with a large step size. Taking 10-step size as an example, in N feature subsets $F = [F_1^1, F_2^1, \cdots, F_N^1]$, the i-th feature subset contains $i \times 10$ high-ranked features, denoted as $F_i^1 = [f_1, f_2, \cdots, f_{i\times 10}]$. Classifiers on samples with each feature subset were learned, which were further tested with 10-fold cross-validation [32–37]. Then, the feature interval containing the feature subset with the highest performance was determined and denoted as [min, max]. In the second stage, a series of feature subsets which contained top min, min + 1,…, max-1, max features were constructed. Likewise, a classifier for each feature subset was built and evaluated by 10-fold cross-validation. Accordingly, the classifier with the best performance can be found and termed as the optimal classifier. The corresponding feature subset is defined as the optimal feature subset. In this study, we selected RF to construct classifiers.

2.3. Random Forest. RF [29] is a classic machine learning algorithm, which contains a large number of decision tree classifiers, that is RF is an assemble classification algorithm. It is widely used in computational biology as one of the most common machine learning methods [38–43]. The output sample class/category of RF is determined by these tree classifiers (i.e., decision trees) in an aggregating vote manner. A RF consists of multiple decision trees with subtle differences. Thus, the mean of the predictions of all decision trees is usually taken as the final consensus results. Although this approach can lead to interpretability loss and slight increase in the model bias, it can avoid overfitting and improve the performance robustness. To quickly implement RF, the tool "RandomForest" in Weka [44, 45] was employed. Such a tool was executed with its default parameters.

2.4. Rule Learning Scheme. The IFS method with RF is helpful to construct a powerful classifier. However, such a classifier is absolutely a black-box classifier. It is very hard to capture the classification principle from such a classifier. Thus, we further employed a rule learning scheme to extract classification rules from the cell expression data.

To save time, we directly used the informative features yielded by the MCFS method. These features were first processed by the Johnson reducer algorithm [46, 47]. Some nonessential features were discarded, and the remaining features had the similar classification ability to the original informative features. Then, the repeated incremental pruning to produce the error reduction (RIPPER) algorithm [48] was applied on the remaining features to construct rules. The RIPPER algorithm is a specific method for constructing rule-based classifiers. The main frame of the RIPPER algorithm is based on IF-ELSE rules and consists of two parts: rule generation and rule optimization. The rule generation is a two-layer loop: the outer loop generates a rule each time after pruning and adds it to the rule pool, and the inner loop adds a predecessor to the rule each time. The rule optimization constructs alternatives based on the rules in the pool and finally selects the optimal rule to update the rule pool.

The above procedures are implemented and integrated in the MCFS program downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html. We directly used it to produce rules.

2.5. Performance Measurement. The Matthew Correlation Coefficient (MCC) [49–57] is a common method used to evaluate the performance of dissimilar classifiers. This variable correlation coefficient calculates the correlation between the target and prediction classes with return value between -1 and +1. MCC considers true and false positives and negatives and is generally considered as a balanced measurement, even when the sample categories have different sizes. In this study, MCC within 10-fold cross-validation was used to evaluate classification performance.

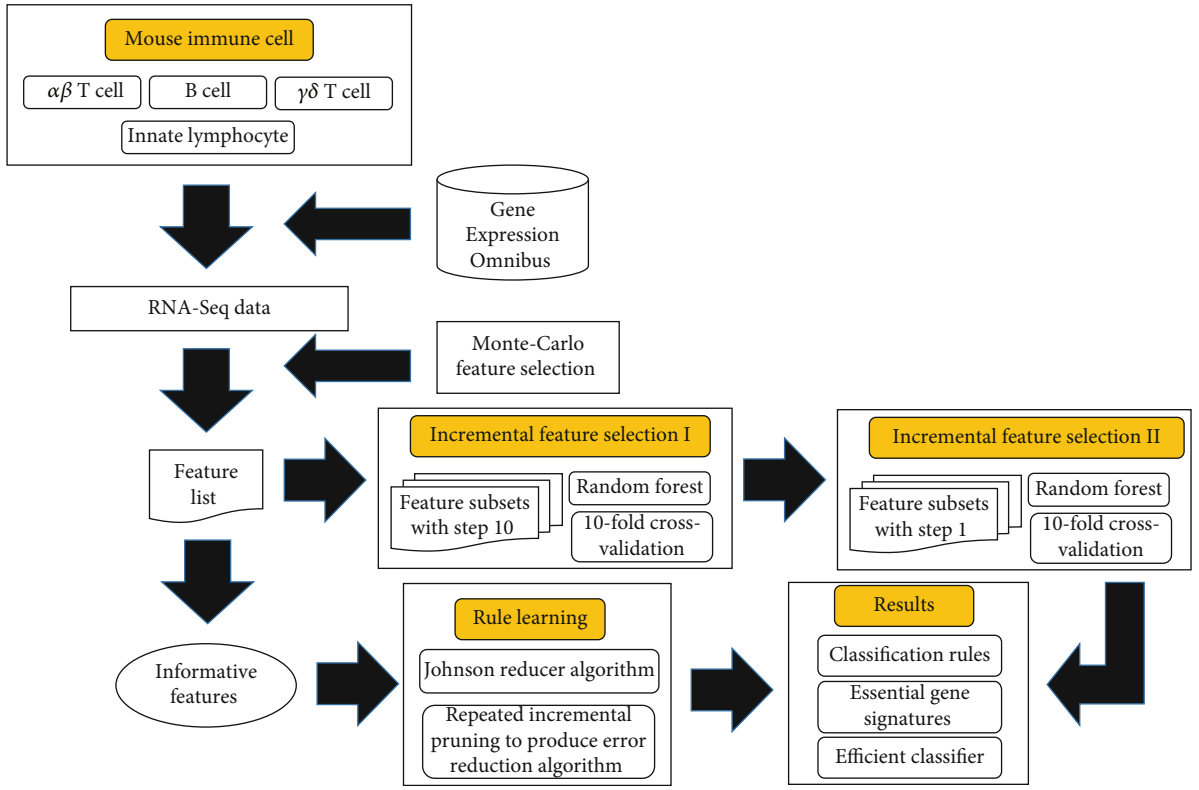

Figure 1: Entire procedures of the computational analysis on RNA-Seq data of mouse immunological cells. The data is retrieved from the Gene Expression Omnibus and is analyzed by the Monte-Carlo Feature Selection method. One feature list and some informative features are produced. A two-stage incremental feature selection method with random forest as the classification algorithm was applied on the feature list to extract essential gene signatures and one efficient classifier. Furthermore, a rule learning scheme is executed on the informative features for constructing classification rules.

Besides, we also employed accuracy on each cell type and overall accuracy (ACC) to fully evaluate the performance of different classifiers.

3. Results

In this study, we adopted several computational methods to analyze the RNA-Seq data of mouse immunological cells. The entire procedures are illustrated in Figure 1. The purpose was to extract essential gene signatures and rules for different immunological cell types. This section gives detailed results of each step of the procedures.

3.1. Results of the MCFS Method. The RNA-Seq data was first analyzed by the MCFS method. Accordingly, each feature was assigned a RI score. Then, a feature list was constructed with the decreasing order of their RI scores, which are provided in Table S1. Moreover, some informative features were also yielded by the MCFS method, which were the top 84 features in the list provided in Table S1.

3.2. Results of the IFS with Random Forest. A two-stage IFS method, incorporating RF as the classification algorithm, was applied to the feature list. In the first stage, we ran IFS with a step size of 10 on the feature list from MCFS. A RF classifier was built based on each constructed feature subset. Then, all classifiers were assessed by 10-fold cross-validation. The predicted results were counted as MCCs, accuracies on four types, and ACCs, which are available in Table S2. For an easy observation, we plotted the obtained MCCs on a coordinate system with the number of used features as the x-axis, as shown in Figure 2(a). It can be seen that when the top ten features were adopted, the RF classifier gave a perfection prediction with MCC = 1. Thus, we determined the min = 1 and max = 50 to do the second IFS stage. Feature subsets containing the top 1-50 features were built, on each of which a RF classifier was set up. Each classifier was evaluated by 10-fold cross-validation. The predicted results are provided in Table S3. Figure 2(b) listed the performance of the RF classifier based on the top ten feature subsets. The RF with the top six features yielded the perfect classification. Thus, such RF classifier was called the optimal classifier, and the corresponding feature subset was termed as the optimal feature subset.

3.3. Results of the Rule Learning. Besides the RF black-box classifier, we also used a rule learning scheme to give a clearer description on the classification procedure, thereby evidently elaborating the differences on four immunological cell types.

According to the MCFS results, 84 informative features were obtained (see the first 84 features in Table S1). Then, the Johnson reducer algorithm was applied on these features to further select the most essential features. The RIPPER algorithm followed to extract rules with the

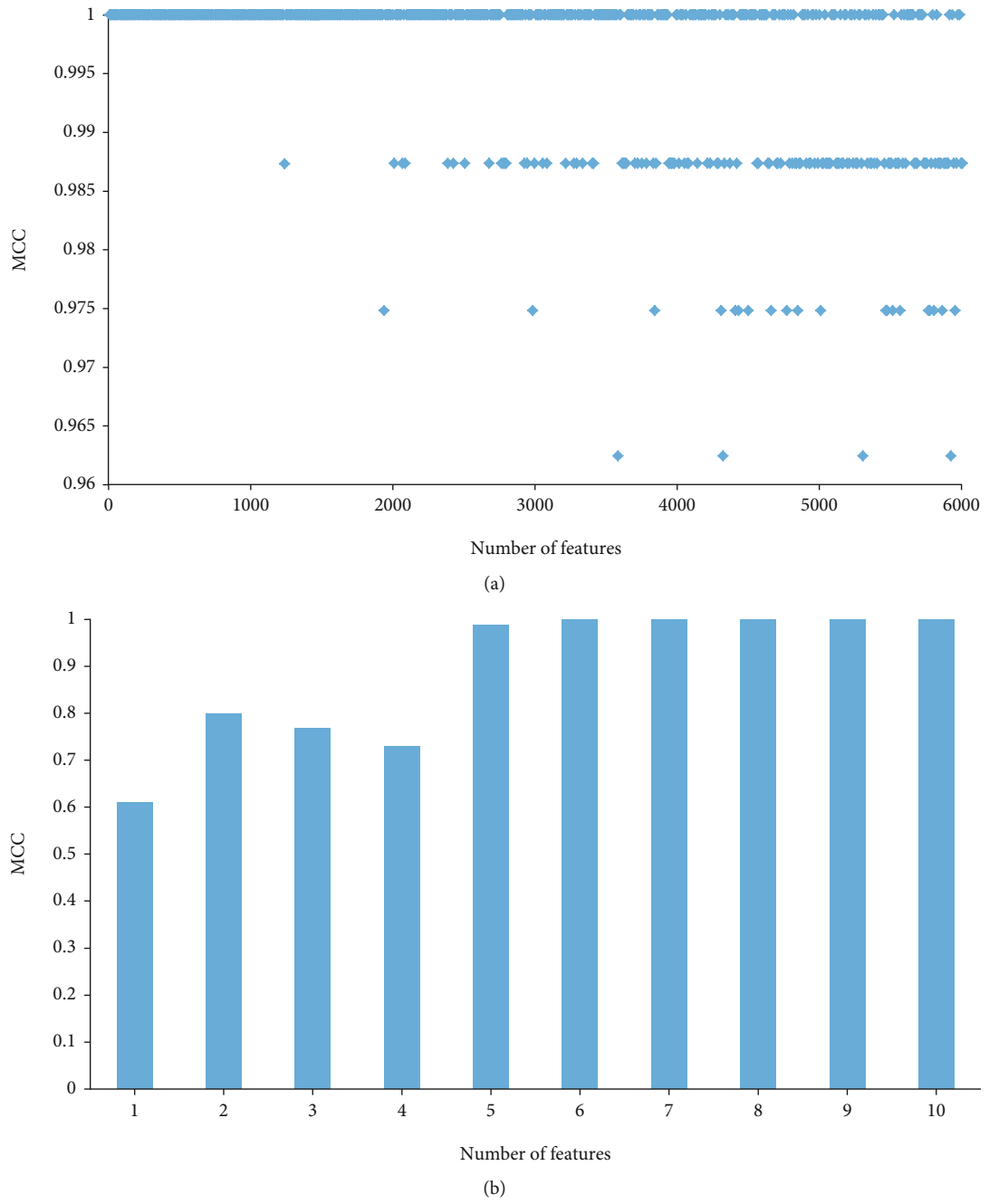

Figure 2: Performance curve of incremental feature selection (IFS) with random forest (RF): (a) performance of RF classifiers with different numbers of features in the first stage of IFS method; (b) performance of RF classifiers with different numbers of features in the second stage of IFS method.

remaining features, resulting in four rules, which are listed in Table 1. To indicate the utility of these rules, two measurements, support and accuracy, were calculated for each rule, which are also listed in Table 1. It can be seen that each rule can cover several immunological cells, and the efficiency of each rule was quite high.

Furthermore, to elaborate the utility of the procedures for constructing the rules, we did the 10-fold cross-validation three times. The accuracies on four cell types are shown in Figure 3. Except the accuracy on the $\gamma\delta$ T cell (82.05%), other accuracies were all no less than 90%. The ACC was 93.15% and MCC was 0.903. All these indicated that such a rule learning scheme was quite effective to extract efficient rules, also indicating the reliability of the rules in Table 1.

4. Discussion

We analyzed the following four typical cell subtypes in the mouse immune system: $\alpha\beta$ T cells, B cells, $\gamma\delta$ T cells, and innate lymphocytes, to screen detailed immune genes and establish standards for cell subgrouping. Basing on the gene expression profiling of individual cells, we performed

Table 1: Classification rules from RIPPER.

Rules	Criteria	Patients	Support[a]	Accuracy[b]
Rule 1	Tcrg-V4 ≥ 62.7560	$\gamma\delta$ T cells	10.71%	100%
Rule 2	Aifm2 ≥ 14.9503	Innate lymphocytes	17.86%	95.00%
Rule 3	Abcb9 ≤ 10.6151	B cells	29.46%	96.97%
Rule 4	Others	$\alpha\beta$ T cells	43.75%	93.88%

[a]Support is defined as the proportion of immunological cells satisfying the rule. [b]Accuracy is defined as the proportion of correctly predicted immunological cells among the cells satisfying the rules.

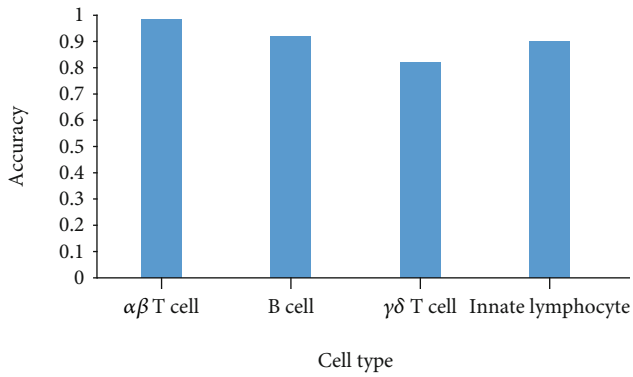

Figure 3: Performance of the rule learning scheme with 10-fold cross-validation three times. The accuracy on each cell type is quite high.

qualitative prediction on cell subtypes, identification of candidate immune cell-associated genes (noted as ImmGen-associated genes), and quantitative screening for the detailed recognition criteria of each cell subtype in a rule manner. According to recent publications, all identified ImmGen-associated genes and quantitative rules can be supported and confirmed by existing experiments and analysis, thus validating the efficacy and accuracy of our prediction. The detailed analysis of high-ranked ImmGen-associated genes and corresponding quantitative rules can be seen below.

4.1. Cell Type-Specific Function of ImmGen-Associated Genes. With the MCFS method, we ranked features (genes) in a list (Table S1). Here, we selected the top ten genes, which are listed in Table 2, for detailed analysis.

The top gene in the ranked feature list is *Ighv1-72*, which encodes the variable region in the heavy chain of immunoglobulin [58]. According to recent publications, this gene participates in antigen-responding antibody synthesis [58]. All biological processes involving antibody synthesis mostly occur in one of our candidate cell subtypes, i.e., B cells, but not in the other cell subtypes [59–61]. Therefore, the expression pattern of *Ighv1-72* in B cells may be different from those in the other three cell subtypes. This finding validates the potential distinguishing role of *Ighv1-72*.

The next high-ranked gene is *Cd5*, which encodes a famous cluster of differentiation. In the mouse immune system, Cd5 is a T-cell surface glycoprotein that regulates T cell inhibition [76, 77]. According to recent publications, *Cd5* is expressed in $\alpha\beta$ T cells and $\gamma\delta$ T cells and is a potential biomarker for T cell subgroups [62, 63]. Although *Cd5* has a specific role of encoding protein T cell surface glycoprotein, its protein products are found on the surface of a specific subgroup of B cells [64, 65]. This finding implies that this biomarker may distinguish innate lymphocytes from the other three immune cell subtypes.

The next predicted gene is *Klrb1b*, which encodes a specific lectin-like receptor on the surface of natural killer cells. Our predicted gene *Nlrb1b* (*Klrb1b*) encodes a functional subunit of a receptor-ligand system in NK cells and may regulate an MHC-independent immune surveillance mechanism [66]. For its cell subtype specific expression pattern, *Nlrb1b* plays an irreplaceable role in natural killer cells and T cells [66], i.e., a key subtype of innate lymphocyte. Hence, Nlrb1b may be a potential marker for innate lymphocytes.

Phka1 exhibits a differential expression pattern among the different cell subtypes. Encoding an alpha chain of the phosphorylase kinase, this gene has a differential expression pattern in different T cells, B cells, and innate lymphocyte subtypes and even under different cell activation status [67–69]. Therefore, *Phka1* is definitely a potential biomarker for the distinction of four immune cell subtypes due to its substantially biological functions and alternative expression patterns in different immune cell subtypes under various immune conditions.

Trdv5, *Trbj1-2*, *Trbj1-3*, *Tcrg-V4*, and *Trbj1-7* are the feature genes encoding different regions of the T cell receptor (TCR) [78, 79]. The differential expression pattern of these genes in four cell subtypes indicates or reflects that of T cell receptors in different cell subtypes. $\alpha\beta$ T and $\gamma\delta$ T cells have a high expression of T cell receptors [80, 81]. These two cell subtypes can be further distinguished according to feature genes due to the differential expression pattern of Tcrg-V4, which encodes a unique region of the gamma chain [82]. A high expression pattern of Tcrg-V4 can be found in $\gamma\delta$ T cells but not in $\alpha\beta$ T cells, thus confirming the distinguishing capacity of our predicted gene signatures [72, 73]. For the remaining cell subtypes, B cells and innate lymphocytes, the former does not have the expression of all TCR-associated genes. A specific subtype of innate lymphocyte, namely, natural killer T cells, has a unique expression pattern of T cell receptors [70, 71]. All identified natural killer T cells with T cell receptor expression would also have the specific $\alpha\beta$ T cell receptors but not the $\gamma\delta$ T cell receptors [71]. Therefore, some subgroups of innate lymphocytes may also have alternative expression patterns of Trdv5, Trbj1-2, Trbj1-3, and Trbj1-7 but not Tcrg-V4. This finding reflects the distinguishing effects of our predicted ImmGen-associated genes involved in T cell receptors.

Table 2: Information of top ten genes selected by Monte-Carlo Feature Selection method.

Rank	Gene	RI score	Cell types	Reference
1	*Ighv1-72*	0.7660	B cells	[58–61]
2	*Cd5*	0.5902	$\alpha\beta$ T cells/$\gamma\delta$ T cells and B cells (with another specific pattern)	[62–65] (for B cells)
3	*Klrb1b*	0.5720	Innate lymphocytes	[66]
4	*Phka1*	0.5535	$\alpha\beta$ T cells, $\gamma\delta$ T cells, B cells, and innate lymphocytes (with different expression level)	[67–69]
5	*Trdv5*	0.5223	$\alpha\beta$ T cells/$\gamma\delta$ T cells and innate lymphocytes (with another specific pattern)	[70, 71]
6	*Trbj1-2*	0.4789		
7	*Trbj1-3*	0.4752		
9	*Trbj1-7*	0.4454		
8	*Tcrg-V4*	0.4738	$\gamma\delta$ T cells	[72, 73]
10	*EBF1*	0.4403	B cells	[74, 75]

Various T cell receptor-associated genes can still be found in the top-ranked genes of our feature gene list, thereby implying the unique differential capacity of the T cell receptor expression pattern and validating the efficacy and accuracy of our prediction approach. In addition to T cell receptor coding genes, we also identified a unique B cell recognizing gene named *EBF1*. Acting as a transcription factor, this gene contributes to the maintenance of B cell identity and prevention of alternative fates in committed cells, such as transferring to the T cell lineage [74, 75]. Therefore, the high expression pattern of EBF1 can only be identified in B cells, implying its potential as a biomarker for this cell type.

Owing to the limitation of the article length, we cannot individually analyze the discriminative genes. However, the above-mentioned high-ranked genes have cell type-specific expression patterns in immune cell subtypes, thus validating the efficacy and accuracy of our prediction and analysis.

4.2. Cell Type-Specific Expression Pattern of ImmGen-Associated Rules. Besides the gene signatures, we also obtained some classification rules via a rule learning scheme (Table 1). They were analyzed as follows. The analysis was based on the expression level measured by Fragments Per Kilobase of transcript per Million mapped reads (FPKM).

The first identified quantitative parameter is *Tcrg-V4*. As a specific encoding gene for the $\gamma\delta$ T cell receptors, this gene may have high expression in $\gamma\delta$ T cells. In accordance with our predicted expression rules, the expression abundance of *Tcrg-V4* is higher than 62.755978 (FPKM) [72, 73]. According to the Mouse Genome Informatics database [83], the expression level and relative expression quantity of *Tcrg-V4* in the T cell subgroup, $\gamma\delta$ T cells, basically conforms to our predicted rules [84, 85], thereby validating the efficacy and accuracy of our prediction.

The next classification rule of quantitative parameter involves a specific gene named *Aifm2*, which contributes to the identification of innate lymphocytes. According to the Mouse Genome Informatics database [83], this gene may have a unique higher expression pattern in mucosal tissues, which are full of innate immune cells, than in the blood system and lymphoid node, which are full of T cells and B cells [86, 87]. Therefore, *Aifm2* may be highly expressed in innate lymphocytes with a threshold of approximately 15 FPKM.

Another quantitative parameter is *Abcb9*, whose low expression (lower than 10 FPKM) is indicative of B cells rather than T cells or innate lymphocytes. Mucosal tissues are full of innate immune cells, and thymus tissues are full of T cells. By contrast, the anatomic area, the spleen, is full of mature B cells and thus has low expression level of *Abcb9* (<5 FPKM) [83].

The cells that do not follow the three rules mentioned above may be $\alpha\beta$ T cells. Thus, all typical expression patterns can be set up for corresponding immune cell subtypes and have been confirmed by recent studies, thereby validating the efficacy and accuracy of our analysis.

5. Conclusions

By using our newly presented computational approach, we identified a group of signature genes that are qualitatively related to the biological processes of four immune cell subtypes. We also set up a set of quantitative expression rules for the detailed clustering of immune cells based on the absolute expression levels measured by FPKM. This work provides a novel computational tool for the quantitative subtyping of immune cells and biomarker recognition.

Data Availability

The data used to support the findings of this study have been deposited in the Gene Expression Omnibus repository (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109125).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Yu-Hang Zhang and Zhandong Li contributed equally to this work.

Acknowledgments

This work was supported by the Shanghai Municipal Science and Technology Major Project [2017SHZDZX01], the National Key R&D Program of China [2018YFC0910403, 2017YFC1201200], the National Natural Science Foundation of China [31701151], the Shanghai Sailing Program [16YF1413800], the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) [2016245], the Strategic Leading Science & Technology Programme of the Chinese Academy of Sciences (XDB38050200), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002].

Supplementary Materials

Supplementary 1. Table S1: ranked features by MCFS

Supplementary 2. Table S2: performance of IFS with RF when using different numbers of features ranked by MCFS with step size of 10

Supplementary 3. Table S3: performance of IFS with RF when using different numbers of features ranked by MCFS with step size of 1.

References

[1] M. Hasegawa and N. Inohara, "Regulation of the gut microbiota by the mucosal immune system in mice," *International Immunology*, vol. 26, no. 9, pp. 481–487, 2014.

[2] S. C. Latet, V. Y. Hoymans, P. L. van Herck, and C. J. Vrints, "The cellular immune system in the post-myocardial infarction repair process," *International Journal of Cardiology*, vol. 179, pp. 240–247, 2015.

[3] J. Parkin and B. Cohen, "An overview of the immune system," *Lancet*, vol. 357, no. 9270, pp. 1777–1789, 2001.

[4] N. Labrecque and N. Cermakian, "Circadian clocks in the immune system," *Journal of Biological Rhythms*, vol. 30, no. 4, pp. 277–290, 2015.

[5] S. F. Martin, "Adaptation in the innate immune system and heterologous innate immunity," *Cellular and Molecular Life Sciences*, vol. 71, no. 21, pp. 4115–4130, 2014.

[6] O. Osborn and J. M. Olefsky, "The cellular and signaling networks linking the immune system and metabolism in disease," *Nature Medicine*, vol. 18, no. 3, pp. 363–374, 2012.

[7] J. Goverman, T. Hunkapiller, and L. Hood, "A speculative view of the multicomponent nature of T cell antigen recognition," *Cell*, vol. 45, no. 4, pp. 475–484, 1986.

[8] T. A. Springer, "Adhesion receptors of the immune system," *Nature*, vol. 346, no. 6283, pp. 425–434, 1990.

[9] T. A. Springer, M. L. Dustin, T. K. Kishimoto, and S. D. Marlin, "The lymphocyte function-associated LFA-1, CD2, and LFA-3 molecules: cell adhesion receptors of the immune system," *Annual Review of Immunology*, vol. 5, no. 1, pp. 223–252, 1987.

[10] M. Guerau-de-Arellano, H. Alder, H. G. Ozer, A. Lovett-Racke, and M. K. Racke, "miRNA profiling for biomarker discovery in multiple sclerosis: from microarray to deep sequencing," *Journal of Neuroimmunology*, vol. 248, no. 1-2, pp. 32–39, 2012.

[11] A. A. Alizadeh and L. M. Staudt, "Genomic-scale gene expression profiling of normal and malignant immune cells," *Current Opinion in Immunology*, vol. 12, no. 2, pp. 219–225, 2000.

[12] I. San Segundo-Val and C. S. Sanz-Lozano, "Introduction to the gene expression analysis," *Methods in Molecular Biology*, vol. 1434, pp. 29–43, 2016.

[13] C. Pilarsky, L. K. Nanduri, and J. Roy, "Gene expression analysis in the age of mass sequencing: an introduction," *Methods in Molecular Biology*, vol. 1381, pp. 67–73, 2016.

[14] C. C. Kim and L. L. Lanier, "Beyond the transcriptome: completion of act one of the Immunological Genome Project," *Current Opinion in Immunology*, vol. 25, no. 5, pp. 593–597, 2013.

[15] The Immunological Genome Project Consortium, T. S. P. Heng, M. W. Painter et al., "The Immunological Genome Project: networks of gene expression in immune cells," *Nature Immunology*, vol. 9, no. 10, pp. 1091–1094, 2008.

[16] E. Clough and T. Barrett, "The Gene Expression Omnibus database," *Methods in Molecular Biology*, vol. 1418, pp. 93–110, 2016.

[17] C. D. Castro, C. T. Boughter, A. E. Broughton, A. Ramesh, and E. J. Adams, "Diversity in recognition and function of human $\gamma\delta$ T cells," *Immunological Reviews*, vol. 298, no. 1, pp. 134–152, 2020.

[18] D. Nemazee, "Mechanisms of central tolerance for B cells," *Nature Reviews Immunology*, vol. 17, no. 5, pp. 281–294, 2017.

[19] R. M. Rezende, A. J. Lanser, S. Rubino et al., "$\gamma\delta$ T cells control humoral immune response by inducing T follicular helper cell differentiation," *Nature Communications*, vol. 9, no. 1, pp. 1–13, 2018.

[20] K. Neumann, K. Karimi, J. Meiners et al., "A proinflammatory role of type 2 innate lymphoid cells in murine immune-mediated hepatitis," *The Journal of Immunology*, vol. 198, no. 1, pp. 128–137, 2017.

[21] M. S. Cruz, A. Diamond, A. Russell, and J. M. Jameson, "Human $\alpha\beta$ and $\gamma\delta$ T cells in skin immunity and disease," *Frontiers in Immunology*, vol. 9, p. 1304, 2018.

[22] B. Silva-Santos, S. Mensurado, and S. B. Coffelt, "$\gamma\delta$ T cells: pleiotropic immune effectors with therapeutic potential in cancer," *Nature Reviews Cancer*, vol. 19, no. 7, pp. 392–404, 2019.

[23] A. C. Kohlgruber, S. T. Gal-Oz, N. M. LaMarche et al., "$\gamma\delta$ T cells producing interleukin-17A regulate adipose regulatory T cell homeostasis and thermogenesis," *Nature Immunology*, vol. 19, no. 5, pp. 464–474, 2018.

[24] D. Liang, H. Shao, W. K. Born, R. L. O'Brien, H. J. Kaplan, and D. Sun, "High level expression of A2ARs is required for the enhancing function, but not for the inhibiting function, of $\gamma\delta$ T cells in the autoimmune responses of EAU," *PLoS One*, vol. 13, no. 6, article e0199601, 2018.

[25] S. Garaud, L. Buisseret, C. Solinas et al., "Tumor-infiltrating B cells signal functional humoral immune responses in breast cancer," *JCI insight*, vol. 5, no. 18, article e129641, 2019.

[26] E. R. Kansler and M. O. Li, "Innate lymphocytes—lineage, localization and timing of differentiation," *Cellular & Molecular Immunology*, vol. 16, no. 7, pp. 627–633, 2019.

[27] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics*, vol. 24, no. 1, pp. 110–117, 2008.

[28] H. A. Liu and R. Setiono, "Incremental feature selection," *Applied Intelligence*, vol. 9, no. 3, pp. 217–230, 1998.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] H. Yoshida, C. A. Lareau, R. N. Ramirez et al., "The cis-regulatory atlas of the mouse immune system," *Cell*, vol. 176, no. 4, pp. 897–912.e20, 2019, e20.

[31] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139-140, 2010.

[32] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International joint Conference on artificial intelligence*, Lawrence Erlbaum Associates Ltd, San Francisco, CA, USA, 1995.

[33] J.-P. Zhou, L. Chen, and Z.-H. Guo, "iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs," *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.

[34] J. Che, L. Chen, Z. H. Guo, S. Wang, and Aorigele, "Drug target group prediction with multiple drug networks," *Combinatorial Chemistry & High Throughput Screening*, vol. 23, no. 4, pp. 274–284, 2020.

[35] Y. Zhu, B. Hu, L. Chen, and Q. Dai, "iMPTCE-Hnetwork: A Multilabel Classifier for Identifying Metabolic Pathway Types of Chemicals and Enzymes with a Heterogeneous Network," *Computational and Mathematical Methods in Medicine*, vol. 2021, Article ID 6683051, 12 pages, 2021.

[36] H. Liu, B. Hu, L. Chen, and L. Lu, "Identifying protein subcellular location with embedding features learned from networks," *Current Proteomics*, vol. 17, 2020.

[37] S. Wang, Q. Zhang, J. Lu, and Y. D. Cai, "Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm," *Current Bioinformatics*, vol. 13, no. 1, pp. 3–13, 2018.

[38] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[39] X. Zhao, L. Chen, Z. H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.

[40] H. Liang, L. Chen, X. Zhao, and X. Zhang, "Prediction of drug side effects with a refined negative sample selection strategy," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1573543, 16 pages, 2020.

[41] X. Zhang, L. Chen, Z. H. Guo, and H. Liang, "Identification of human membrane protein types by incorporating network embedding methods," *IEEE Access*, vol. 7, pp. 140794–140805, 2019.

[42] Y. Jia, R. Zhao, and L. Chen, "Similarity-based machine learning model for predicting the metabolic pathways of compounds," *IEEE Access*, vol. 8, pp. 130687–130696, 2020.

[43] J. Li, L. Lu, Y. H. Zhang et al., "Identification of synthetic lethality based on a functional network by using machine learning algorithms," *Journal of Cellular Biochemistry*, vol. 120, no. 1, pp. 405–416, 2019.

[44] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[45] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*, Kaufmann, San Francisco, Morgan, 2nd edition, 2005.

[46] D. S. Johnson, "Approximation algorithms for combinatorial problems," *Journal of Computer and System Sciences*, vol. 9, no. 3, pp. 256–278, 1974.

[47] A. Ohrn, *Discernibility and rough sets in medicine: tools and applications, in Department of Computer and Information Science*, Norwegian University of Science and Technology, Trondheim, 1999.

[48] W. W. Cohen, "Fast effective rule induction," in *The Proceeding of Proceeding of the Twelfth International. Conference of Machine Learning*, pp. 115–123, Tahoe City, CA, USA, July 9–12, 1995.

[49] B. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[50] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 381–389, 2019.

[51] L. Chen, C. Chu, Y. H. Zhang et al., "Identification of drug-drug interactions using chemical interactions," *Current Bioinformatics*, vol. 12, no. 6, pp. 526–534, 2017.

[52] L. Chen, S. Wang, Y. H. Zhang et al., "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[53] Y.-H. Zhang, H. Li, T. Zeng et al., "Identifying transcriptomic signatures and rules for SARS-CoV-2 infection," *Frontiers in Cell and Developmental Biology*, vol. 8, p. 627302, 2021.

[54] Y.-H. Zhang, Z. Li, T. Zeng et al., "Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles," *Frontiers in Genetics*, vol. 11, p. 599970, 2021.

[55] X. Pan, H. Li, T. Zeng et al., "Identification of protein subcellular localization with network and functional embeddings," *Frontiers in Genetics*, vol. 11, p. 626500, 2021.

[56] Y.-H. Zhang, Z. Li, T. Zeng et al., "Distinguishing glioblastoma subtypes by methylation signatures," *Frontiers in Genetics*, vol. 11, p. 604336, 2020.

[57] L. Chen, Z. Li, T. Zeng et al., "Identifying robust microbiota signatures and interpretable rules to distinguish cancer subtypes," *Frontiers in Molecular Biosciences*, vol. 7, p. 604794, 2020.

[58] N. Kono, L. Sun, H. Toh et al., "Deciphering antigen-responding antibody repertoires by using next-generation sequencing and confirming them through antibody-gene synthesis," *Biochemical and Biophysical Research Communications*, vol. 487, no. 2, pp. 300–306, 2017.

[59] P. M. Glassman, L. Abuqayyas, and J. P. Balthasar, "Assessments of antibody biodistribution," *Journal of Clinical Pharmacology*, vol. 55, Suppl 3, pp. S29–S38, 2015.

[60] J. W. Larrick, P. W. H. I. Parren, J. S. Huston et al., "Antibody engineering and therapeutics conference. The annual meeting of the antibody society, Huntington Beach, CA, December 7-11, 2014," *MAbs*, vol. 6, no. 5, pp. 1115–1123, 2014.

[61] L. Presta, "Antibody engineering for therapeutics," *Current Opinion in Structural Biology*, vol. 13, no. 4, pp. 519–525, 2003.

[62] H. Yokozeki, K. Watanabe, K. Igawa, Y. Miyazaki, I. Katayama, and K. Nishioka, "Gammadelta T cells assist alphabeta T cells in the adoptive transfer of contact hypersensitivity to para-phenylenediamine," *Clinical and Experimental Immunology*, vol. 125, no. 3, pp. 351–359, 2001.

[63] T. Sugie, H. Kubota, M. Sato, E. Nakamura, M. Imamura, and N. Minato, "NK 1+ CD4- CD8- alphabeta T cells in the peritoneal cavity: specific T cell receptor-mediated cytotoxicity and selective IFN-gamma production against B cell leukemia and myeloma cells," *Journal of Immunology*, vol. 157, no. 9, pp. 3925–3935, 1996.

[64] V. L. Palmer, V. K. Nganga, M. E. Rothermund, G. A. Perry, and P. C. Swanson, "Cd1d regulates B cell development but not B cell accumulation and IL10 production in mice with pathologic CD5(+) B cell expansion," *BMC Immunology*, vol. 16, no. 1, p. 66, 2015.

[65] R. R. Hardy and K. Hayakawa, "Perspectives on fetal derived CD5+ B1 B cells," *European Journal of Immunology*, vol. 45, no. 11, pp. 2978–2984, 2015.

[66] Q. Zhang, M. M. A. Rahim, D. S. J. Allan et al., "Mouse Nkrp1-Clr gene cluster sequence and expression analyses reveal conservation of tissue-specific MHC-independent immunosurveillance," *PLoS One*, vol. 7, no. 12, article e50561, 2012.

[67] W. R. Osborne and C. R. Scott, "The metabolism of deoxyguanosine and guanosine in human B and T lymphoblasts. A role for deoxyguanosine kinase activity in the selective T-cell defect associated with purine nucleoside phosphorylase deficiency," *The Biochemical Journal*, vol. 214, no. 3, pp. 711–718, 1983.

[68] R. M. Goldblum, F. C. Schmalstieg, J. A. Nelson, and G. C. Mills, "Adenosine deaminase (ADA) and other enzyme abnormalities in immune deficiency states," *Birth Defects Original Article Series*, vol. 14, no. 6A, pp. 73–84, 1978.

[69] R. B. Trelease, R. A. Henderson, and J. B. Park, "A qualitative process system for modeling NF-kappaB and AP-1 gene regulation in immune cell biology research," *Artificial Intelligence in Medicine*, vol. 17, no. 3, pp. 303–321, 1999.

[70] A. Rodríguez-Caballero, A. C. García-Montero, P. Bárcena et al., "Expanded cells in monoclonal TCR-alphabeta+/CD4+/NKa+/CD8-/+dim T-LGL lymphocytosis recognize hCMV antigens," *Blood*, vol. 112, no. 12, pp. 4609–4616, 2008.

[71] M. Verykokakis, M. D. Boos, A. Bendelac, E. J. Adams, P. Pereira, and B. L. Kee, "Inhibitor of DNA binding 3 limits development of murine slam-associated adaptor protein-dependent "innate" gammadelta T cells," *PLoS One*, vol. 5, no. 2, article e9303, 2010.

[72] T. Washburn, E. Schweighoffer, T. Gridley et al., "Notch activity influences the $\alpha\beta$ versus $\gamma\delta$ T cell lineage decision," *Cell*, vol. 88, no. 6, pp. 833–843, 1997.

[73] L. Riera-Sans and A. Behrens, "Regulation of alphabeta/gammadelta T cell development by the activator protein 1 transcription factor c-Jun," *Journal of Immunology*, vol. 178, no. 9, pp. 5690–5700, 2007.

[74] R. Nechanitzky, D. Akbas, S. Scherer et al., "Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells," *Nature Immunology*, vol. 14, no. 8, pp. 867–875, 2013.

[75] I. Gyory, S. Boller, R. Nechanitzky et al., "Transcription factor Ebf1 regulates differentiation stage-specific signaling, proliferation, and survival of B cells," *Genes & Development*, vol. 26, no. 7, pp. 668–682, 2012.

[76] M. Bamberger, A. M. Santos, C. M. Gonçalves et al., "A new pathway of CD5 glycoprotein-mediated T cell inhibition dependent on inhibitory phosphorylation of Fyn kinase," *The Journal of Biological Chemistry*, vol. 286, no. 35, pp. 30324–30336, 2011.

[77] W. Luo, H. Van de Velde, I. von Hoegen, J. R. Parnes, and K. Thielemans, "Ly-1 (CD5), a membrane glycoprotein of mouse T lymphocytes and a subset of B cells, is a natural ligand of the B cell surface protein Lyb-2 (CD72)," *Journal of Immunology*, vol. 148, no. 6, pp. 1630–1634, 1992.

[78] E. Ruggiero, J. P. Nicolay, R. Fronza et al., "High-resolution analysis of the human T-cell receptor repertoire," *Nature Communications*, vol. 6, no. 1, p. 8081, 2015.

[79] D. R. Thapa, R. Tonikian, C. Sun et al., "Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing," *Arthritis Research & Therapy*, vol. 17, no. 1, p. 132, 2015.

[80] K. Ohshima, K. Karube, R. Kawano et al., "Classification of distinct subtypes of peripheral T-cell lymphoma unspecified, identified by chemokine and chemokine receptor expression: analysis of prognosis," *International Journal of Oncology*, vol. 25, no. 3, pp. 605–613, 2004.

[81] B. M. Hall, "T cells: soldiers and spies–the surveillance and control of effector T cells by regulatory T cells," *Clinical Journal of the American Society of Nephrology*, vol. 10, no. 11, pp. 2050–2064, 2015.

[82] S. Huck, P. Dariavach, and M. P. Lefranc, "Variable region genes in the human T-cell rearranging gamma (TRG) locus: V-J junction and homology with the mouse genes," *The EMBO Journal*, vol. 7, no. 3, pp. 719–726, 1988.

[83] C. J. Bult, J. T. Eppig, J. A. Blake, J. A. Kadin, J. E. Richardson, and the Mouse Genome Database Group, "Mouse genome database 2016," *Nucleic Acids Research*, vol. 44, no. D1, pp. D840–D847, 2016.

[84] J. S. Heilig and S. Tonegawa, "Diversity of murine gamma genes and expression in fetal and adult T lymphocytes," *Nature*, vol. 322, no. 6082, pp. 836–840, 1986.

[85] C. Hetzer-Egger, M. Schorpp, A. Haas-Assenbaum, R. Balling, H. Peters, and T. Boehm, "Thymopoiesis requires Pax9 function in thymic epithelial cells," *European Journal of Immunology*, vol. 32, no. 4, pp. 1175–1181, 2002.

[86] G. Diez-Roux, S. Banfi, M. Sultan et al., "A high-resolution anatomical atlas of the transcriptome in the mouse embryo," *PLoS Biology*, vol. 9, no. 1, article e1000582, 2011.

[87] S. Magdaleno, P. Jensen, C. L. Brumwell et al., "BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system," *PLoS Biology*, vol. 4, no. 4, article e86, 2006.