Journal of Immunological Methods 391 (2013) 14–21

Contents lists available at SciVerse ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Research paper

Estimating the ratio of CD4 + to CD8 + T cells using high-throughput sequence data

Ryan Emerson [a], Anna Sherwood [a], Cindy Desmarais [a], Sachin Malhotra [b], Deborah Phippard [b], Harlan Robins [c,*]

[a] Adaptive Biotechnologies, 1551 Eastlake Ave E, Seattle, WA 98109, United States
[b] Immune Tolerance Network, 3 Bethesda Metro Center, Suite 400, Bethesda, MD 20814, United States
[c] Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109, United States

ARTICLE INFO

Article history:
Received 17 October 2012
Accepted 10 February 2013
Available online 18 February 2013

Keywords:
T cells
TCR
High-throughput sequencing
TIL

ABSTRACT

Mature T cells express either CD8 or CD4, defining two physiologically distinct populations of T cells. CD8 + T cells, or killer T-cells, and CD4 + T cells, or helper T cells, effect different aspects of T cell mediated adaptive immunity. Currently, determining the ratio of CD4 + to CD8 + T cells requires flow cytometry or immunohistochemistry. The genomic T cell receptor locus is rearranged during T cell maturation, generating a highly variable T cell receptor locus in each mature T cell. As part of thymic maturation, T cells that will become CD4 + versus CD8 + are subjected to different selective pressures. In this study, we apply high-throughput next-generation sequencing to T cells from both a healthy cohort and a cohort with an autoimmune disease (multiple sclerosis) to identify sequence features in the variable CDR3 region of the rearranged T cell receptor gene that distinguish CD4 + from CD8 + T cells. We identify sequence features that differ between CD4 + and CD8 + T cells, including Variable gene usage and CDR3 region length. We implement a likelihood model to estimate relative proportions of CD4 + and CD8 + T cells using these features. Our model accurately estimates the proportion of CD4 + and CD8 + T cell sequences in samples from healthy and diseased immune systems, and simulations indicate that it can be applied to as few as 1000 T cell receptor sequences; we validate this model using in vitro mixtures of T cell sequences, and by comparing the results of our method to flow cytometry using peripheral blood samples. We believe our computational method for determining the CD4:CD8 ratio in T cell samples from sequence data will provide additional useful information for any samples on which high-throughput TCR sequencing is performed, potentially including some solid tumors.

1. Introduction

The human cellular adaptive immune system is mediated by two primary types of T cells, killer T cells and helper T cells. Killer T cells, marked by the surface expression of CD8, recognize short peptides (~8–10 amino acids) presented on the surface of cells by human leukocyte antigen (HLA) Class I molecules (Pamer and Cresswell, 1998). Helper T cells, marked by the surface expression of CD4, recognize longer peptides (~12–16 nucleotides) presented on the surface of cells by HLA Class II molecules (Leddon and Sant, 2010). Both of these T cell types are derived from a common progenitor cell type.

During the development of T cells in the thymus, the DNA loci coding for the alpha and beta chains of the Y-like T cell receptor (TCR) rearrange in a pseudo-random process to form an enormous variety of TCRs (Davis and Bjorkman, 1988). TCR sequence diversity is primarily contained in the complementarity determining region 3 (CDR3) loops of the α and β chains which bind to the peptide antigen, conferring specificity. The nucleotide sequences that encode the CDR3

* Corresponding author.
E-mail address: hrobins@fhcrc.og (H. Robins).

0022-1759/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.jim.2013.02.002

loops are generated by V(D)J recombination: variable (V_β), diversity (D_β) and joining (J_β) genes in the genome rearrange to form a β chain, while V_α and J_α genes rearrange to form an α chain (Davis and Bjorkman, 1988).

After the alpha and beta chains rearrange, while still in the thymus, T cells are both positively and negatively selected against self peptides displayed by Class I and Class II HLA molecules (Jameson et al., 1995; Vukmanovic, 1996). If a TCR binds strongly to a self peptide:HLA complex, the T cell usually dies. Additionally, a T cell is positively selected, requiring some minimal threshold of binding to either a Class I or Class II presented peptide (Jameson et al., 1995; Vukmanovic, 1996). Prior to selection, T cells express both CD4 and CD8 on their surface, and are referred to as double positive T cells. Upon positive selection the T cell halts expression of one of these two surface proteins, leaving a single positive T cell committed as either a helper or killer T cell (Jameson et al., 1995). These two T cell types serve very different functional roles.

In this manuscript we apply the ImmunoSEQ assay for sequencing TCR beta chains at very high throughput to sorted CD8+ and CD4+ T cell populations to assess whether the binding restriction of TCRs to Class I or Class II presented peptides selects different sequence properties of the TCR. Both the differences between Class I and II molecules themselves as well as the significant differences in the length of the peptides they display suggest that CD8+ and CD4+ T cells might be selected for different structures (Pamer and Cresswell, 1998; Leddon and Sant, 2010). However, these differences are sufficiently subtle that a very high-throughput study is required to elucidate the differences.

We determine a set of sequence features that can unambiguously distinguish CD4+ and CD8+ T cell populations strictly based on TCRβ sequence data. Moreover, we create a simple likelihood model that is able to determine the proportion of CD8+ and CD4+ T cells within a mixed population of cells. Sequencing of TCRs combined with the likelihood model acts in the manner of a virtual flow cytometer on CD4 and CD8. This method may find an application in providing additional information in the case of biological samples for which high-throughput sequencing is commonly performed without accompanying flow cytometry or immunohistochemistry, including T cells infiltrating solid tumors.

2. Materials and methods

2.1. Sample set

For each of the 17 healthy individuals used for the training set, PBMCs were sorted into four populations, CD4+/CD45RA+/CD62L+, CD4+/CD45RA−/CD45RO+, CD8+/CD45RA+/CD62L+, and CD8+/CD45RA−/CD45RO+, using a FacsARIA, (BD Biosciences). In addition, PBMCs from 25 subjects with multiple sclerosis were separated into two fractions. Cells were labeled with either anti-CD4+ or anti-CD8+ antibodies conjugated with phycoerythrin (PE). These labeled cells were treated with anti-PE antibodies coupled to microbeads (Miltenyi Biotech, Auburn, CA). Micro-bead labeled cells were positively selected using Miltenyi cell separator columns. DNA was extracted from all populations using Qiagen DNA isolation kits (QIAGEN Inc., Valencia, CA). For all samples, TCRβ CDR3 chains were amplified and sequenced using the immunoSEQ assay (Adaptive Biotechnologies, Seattle, WA).

2.2. Sequencing

TCRβ CDR3 regions were amplified and sequenced using protocols described by (Robins et al., 2009). Briefly, a multiplexed PCR method was employed to amplify all possible rearranged genomic TCRβ sequences using 52 forward primers, each specific to a TCR Vβ segment, and 13 reverse primers, each specific to a TCR Jβ segment. Reads of length 60 bp were obtained using the Illumina HiSeq system. Raw HiSeq sequence data were preprocessed to remove errors in the primary sequence of each read, and to compress the data. A nearest neighbor algorithm was used to collapse the data into unique sequences by merging closely related sequences, to remove both PCR and sequencing errors. The TCR CDR3 region was defined according to the IMGT collaboration (Yousfi Monod et al., 2004), beginning with the second conserved cysteine encoded by the 3′ portion of the Vβ gene segment and ending with the conserved phenylalanine encoded by the 5′ portion of the Jβ gene segment. The number of nucleotides between these codons determines the length and therefore the frame of the CDR3 region.

2.3. CD4/CD8 estimation

A CD4/CD8 estimation model was trained using 118 samples: 68 sorted samples generated by sorting PBMCs from 17 healthy subjects on CD4/CD8 and CD45RA/CD45RO to produce 4 samples per subject, and 50 sorted samples generated by sorting PBMCs from 25 patients with autoimmune disease on CD4/CD8 to produce 2 samples per subject. The frequency with which each V segment was used in TCRs from each of the 59 CD4 training datasets was determined. The mean of this value across the 59 training datasets was taken to be the proportional usage of that V segment in CD4+ T cells, P(V|CD4). In similar fashion we determined the proportional usage of each J segment, P(J|CD4); the proportion of each CDR3 length given each V segment, P(CDR3|V, CD4); and the proportion of each CDR3 length given each J segment, P(CDR3|J,CD4). These values were likewise calculated for CD8+ T cells using the 59 CD8+ training datasets. The model was then modified to include one observation for each CDR3 length, for each V and J segment, so that no observable value was completely absent from the training data.

To determine the likelihood of new data under this model, a proportion of CD4+ T cells, p, is assumed. The likelihood of a single sequence (with features V, J and CDR3 length) is calculated as:

$$[p*P(V|CD4)*P(J|CD4)*P(CDR3|V,CD4)*P(CDR3|J,CD4)] + [(1-p)*P(V|CD8)*P(J|CD8)*P(CDR3|V,CD8)*P(CDR3|J,CD8)].$$

The likelihood of a dataset is calculated as the product of the likelihoods of its constituent sequences. To determine the proportion of CD4+ T cells in new data, the likelihood of the new data is calculated at each p from 0 to 1 with a granularity of 0.01, and the value of p leading to the highest likelihood of the observed data is chosen as our estimate of the proportion of CD4+ T cells in the sample. Cross-validation was performed by

removing each of the 118 training samples in turn, retraining the model using the other 117 samples, then estimating the proportion of CD4+ T cells in the holdout sample.

2.4. Dataset generation for sample size comparison

To determine the effect of sample size on our model's accuracy, we generated new datasets by treating our model parameters as generative probabilities and running the model forward to create T cell receptor sequences. For each dataset, we determined the number of TCR sequences to be generated and a proportion CD4+ TCR sequences, which was drawn from the uniform distribution. We then generated TCR sequences one at a time, assigned each as being CD4+ or CD8+ based on the proportion determined above, and randomly generated a V segment, J segment and CDR3 length. This process allowed us to generate thousands of datasets with varying numbers of TCR sequences and varying proportions CD4+ TCR sequences but no systematic error with respect to our model parameters.

2.5. In vitro and in vivo validation

In vitro validation was performed by choosing 5 of the subjects used for training, whose PBMCs were sorted on CD4/CD8. For each of these samples, pure sequencing libraries were mixed in three ratios: 3:1, 1:1 and 3:1 CD4:CD8. In each of these mixed samples different sequencing barcodes were used for the CD4 and CD8 components of the sample so that we could directly identify whether each observed sequence in the mixed sample originated from the CD4 or CD8 sorted sample. This process allows us to overcome inaccuracies inherent in quantifying and accurately mixing very small amounts of DNA and exactly determine the CD4:CD8 ratio in each mixed sample, which can then be compared to the estimation of CD4:CD8 ratio from our likelihood model run on the merged data.

For in vivo validation, we obtained PBMC samples from seven healthy subjects not used to train the model, and estimated the CD4:CD8 ratio in these seven samples with flow cytometry (using a FacsARIA, BD Biosciences), and with high-throughput sequencing followed by application of our likelihood method.

3. Results

3.1. Identification of sequence features distinguishing CD4+ and CD8+ T cells

The differences between Class I and II HLA molecules as well as the significant differences in the length of the peptides they display suggest that CD4+ and CD8+ T cells would be subjected to different selective pressures during maturation. In order to assess whether the binding restriction of TCRs to Class I or Class II presented peptides selects different sequence properties of the TCR, we performed high-throughput TCR sequencing of the CDR3 region of TCRs from 42 adults, 17 healthy subjects and 25 multiple sclerosis patients. PBMC samples from each subject were sorted to produce T cell samples separated on CD4+/CD8+. DNA was extracted from these samples and multiplex PCR followed by high-throughput sequencing was performed to generate a total of 602 million TCR sequencing reads spread across 13 million unique TCR sequences.

In order to investigate the sequence features that may differentiate CD4+ and CD8+ T cells, we examined the usage of Vβ segments, usage of Jβ segments, and length of the CDR3 region. First, the frequency with which productively rearranged TCR sequences in each of the CD4+ samples used each Vβ segment was calculated, and the mean of these frequencies was taken to be the population mean usage for that Vβ segment. This was compared to the usage of each segment in CD8+ T cells (Fig. 1). Many of the Vβ segments are used more frequently in either CD4+ or CD8+ cells. To assess statistical significance, we performed a two-tailed unpaired *t*-test for difference of means. 21 of 48 measured Vβ segments have differential usage between CD4+ and CD8+ samples, indicating that the binding requirements imposed on each cell type do influence the frequency with which rearrangements bearing each gene segment survive the selection process. We performed the same analysis for Jβ segments (Fig. 2), and in that case there was significant evidence for differential usage in 3 of the 13 human Jβ segments. Since some of the subjects' cells were also sorted on CD45RA/CD45RO, we compared Vβ and Jβ segment usage in these two sorts. No Vβ or Jβ segments had significantly different usage in memory vs. naïve T cells by this test, with or without accounting for CD4+/CD8+ status first, and for the remainder of our analyses samples sorted on CD4/CD8 and CD45RA/CD45RO were treated simply as CD4+ or CD8+ samples as appropriate.

We also investigated whether the TCR CDR3 regions of CD4+ and CD8+ T cells have different lengths. When taken as a whole, the distribution of CD4+ and CD8+ CDR3 lengths is quite similar (Fig. 3A). However, given any particular Vβ segment, CDR3 length frequently does differ between CD4+ and CD8+ T cells (Fig. 3B). Additionally, we calculated the deviation of each CDR3 sequence's length from the mode length given the Vβ segment it uses and its CD4+/CD8+ status, and this analysis yields a substantial difference between cell types: when comparing TCRs that use the same Vβ segment, CD8+ cells have a considerably more restricted range of CDR3 lengths than CD4+ cells: about 40% of CD8+ cells have a CDR3 region of the mode length, whereas only 30% of CD4+ cells share the most common CDR3 length (Fig. 3C).

3.2. Computational estimation of CD4:CD8 ratio

Having established the existence of sequence features that distinguish CD4+ from CD8+ T cells, we developed a computational method to estimate the proportion of T cells that are CD4+ in an unknown sample using sequence data alone. Briefly, usage frequency for each Vβ segment, each Jβ segment and each CDR3 length was calculated for CD4+ and CD8+ T cells using flow-sorted samples from 42 subjects (see Section 2.1). These values were used to train a likelihood model which treats each observed TCR sequence as independent and uses the observed means as generative probabilities. Given a dataset with an unknown proportion of CD4+ T cells, the model evaluates the likelihood of generating the dataset for each proportion of CD4+ cells and outputs the proportion leading to the highest likelihood of observing the data.

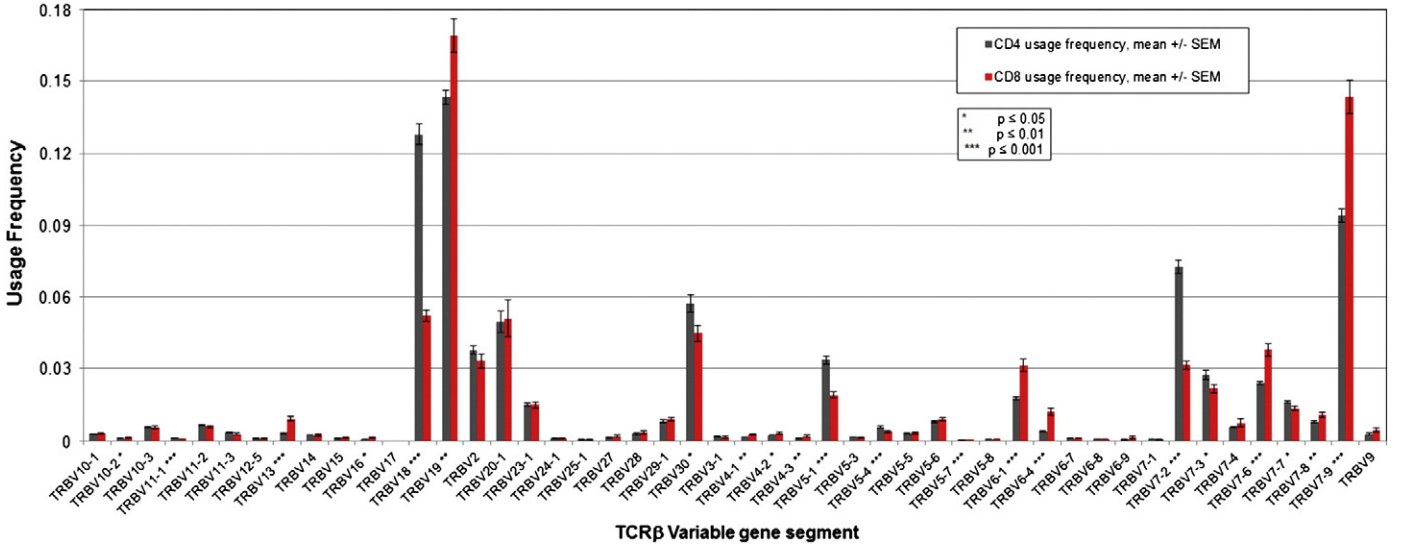

Fig. 1. V gene usage in CD4+ and CD8+ T cells. Pictured is the mean usage frequency in 42 subjects of each Variable gene segment in CD4+ and CD8+ T cell receptor sequences as determined by high-throughput TCRβ sequencing. Error bars represent the standard error of the mean. Each V gene segment was analyzed using a two-tailed unpaired *T*-test for difference of means, and V segment names are marked accordingly: *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. A considerable difference in usage frequency exists between CD4+ and CD8+ T cells for several V gene segments.

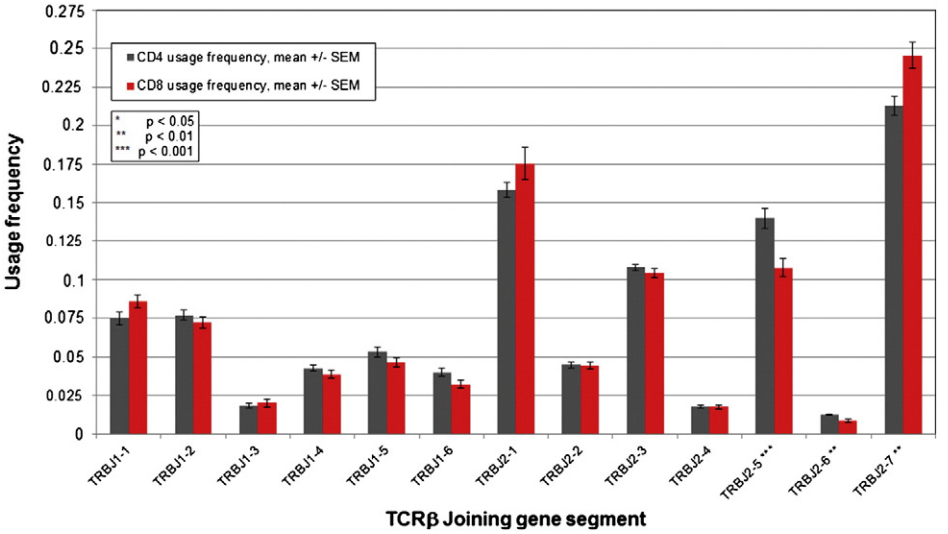

Fig. 2. J gene usage in CD4+ and CD8+ T cells. Pictured is the mean usage frequency in 42 subjects of each joining gene segment in CD4+ and CD8+ T cell receptor sequences as determined by high-throughput TCR sequencing. Error bars represent the standard error of the mean. Each J gene segment was analyzed using a two-tailed unpaired *T*-test for difference of means, and J segment names are marked accordingly: *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. A considerable difference in usage frequency exists between CD4+ and CD8+ T cells for several J gene segments.

In order to test the effectiveness of our model, a cross-validation experiment was performed in which each of the training samples was left out one at a time, the model's parameters were re-estimated using the remainder of the training samples and the proportion of CD4+ T cells in the holdout sample was estimated. The results of this test are shown in Fig. 4, which shows results for all samples ($n=118$) as well as the cross-validation results from only the healthy subjects and only the subjects with autoimmune disease. The results for both groups are very similar, indicating that our model performs as well in diseased subjects as in healthy subjects. In all cases, a large majority of the CD4+ training datasets are estimated to have ≥90% CD4+ T cells, and likewise most training CD8+ datasets are estimated to have ≤20% CD4+ T cells. For both CD4+ and CD8+ samples, the median result is that the model estimates a nearly-pure sample, as expected for a model with good accuracy but imperfect precision. These results indicate that even when accounting for over-fitting, the model is generally accurate in its estimates of the proportion of CD4+ cells in unknown samples.

3.3. Effect of sample size on CD4+/CD8+ estimation

In order to assess the ability of our model to accurately predict the proportion of CD4+ T cells in mixed samples of varying sizes, we performed an experiment in which the parameters of our model were used to directly generate datasets of varying sizes. We then applied our model to predict the proportion of CD4+ T cell sequences in these datasets. Since these datasets were generated in perfect accordance with all features of our model, the error in these estimates reflects fundamental constraints on our model's accuracy with respect to the number of TCR sequences available and on the amount of information contained in the differences in Vβ and Jβ usage frequency and CDR3 length between CD4+ and CD8+ T cells. The results of this experiment are shown in Fig. 5. Panels A through D represent 1000 replicates each using datasets containing 10; 100; 1000; or 10,000 TCR sequences respectively. As expected, our model has no predictive capacity when examining 10 TCR sequences, since the differences in VJ usage and CDR3 length are subtle. When using 100 TCR sequences, our model performs poorly but better than would be expected by chance. Our model performs quite well when given 1000 TCR sequences, and at 10,000 TCR sequences or more sample size is not a substantial constraint on our model's accuracy in these simulations.

3.4. Validation

In order to further validate our computational method for estimation of CD4+/CD8+ T cells, we performed an experiment in which DNA from sorted CD4+ or CD8+ samples was mixed in known proportion, and then estimated the proportion CD4+ in these samples using our method. CD4+ or CD8+ sequencing libraries from 5 of the subjects used in our initial exploration were mixed at three ratios; mixing very small amounts of DNA in exact proportions was difficult, but we utilized a scheme which allowed us to determine the realized CD4+/CD8+ ratio in each of these mixed samples post-facto (see Section 2.5). The results of this experiment are summarized in Fig. 6, which compares the known proportion CD4+ in these mixed samples to the estimates obtained for each sample from our likelihood model. Our method resulted in very good estimates of the proportion of CD4+ sequences in each of the fifteen mixed samples, with a mean difference of 0.05 (standard deviation 0.03) between our estimate and the true value.

To validate our method in vivo, we sequenced PBMC samples from seven additional healthy subjects that were not used in the training set. We estimated the CD4:CD8 ratio in

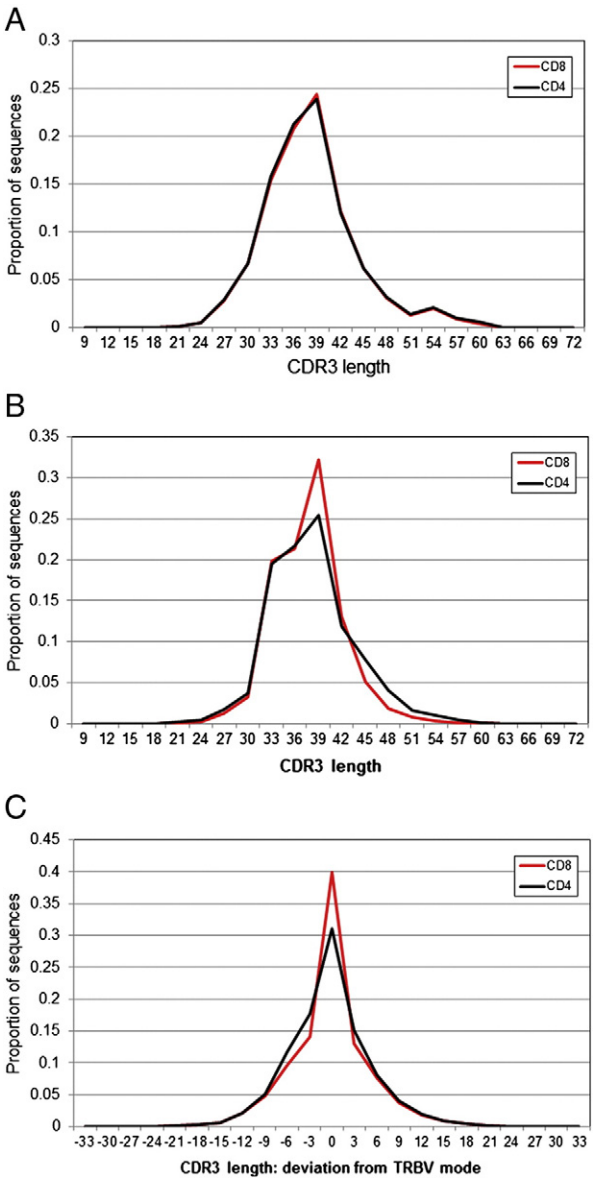

Fig. 3. CDR3 length distribution in CD4+ and CD8+ T Cells. A: Distribution of CDR3 length among all CD4+ (black) and CD8+ (red) T cells from 42 subjects. Only productive rearrangements are shown, so all CDR3 lengths are multiples of 3. The proportion of total TCR sequences with CDR3 regions of each length is shown. B: Distribution of CDR3 length among CD4+ (black) and CD8+ (red) T cells using TRBV6-4. Despite similar profiles overall, CDR3 lengths do differ between CD4+ and CD8+ T cells when considering only TCR sequences using the same V segment usage. C: Distribution of CDR3 length among all CD4+ (black) and CD8+ (red) T cells, presented as deviation from mode CDR3 length given V segment usage. CD8+ T cells have a more restricted range of CDR3 lengths than CD4+ T cells when controlling for V segment usage.

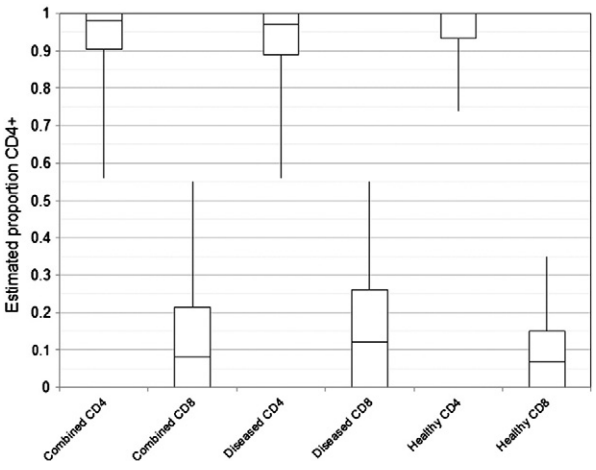

Fig. 4. In silico CD4:CD8 estimation, cross-validation results. Above are shown the results of in silico estimation of the proportion of CD4+ T cells on flow-sorted training data using 118 CD4+ or CD8+ samples from 17 healthy subjects and 25 subjects with multiple sclerosis. The results for each sample were calculated by training the model on the other 117 samples and then estimating the proportion CD4+ in the missing sample. Results are presented as a box-and-whisker plot: the open box for each sample represents the second and third quartiles of estimated CD4+ proportion. Top whisker represents the maximum and bottom whisker represents the minimum. Data are presented for all 118 samples, and separately for samples derived from diseased and healthy subjects.

these samples and compared it to the CD4:CD8 ratio obtained by flow cytometry. The results are presented in Fig. 7. Our method generates results similar to those from flow cytometry; the mean error for these 7 samples was 0.02 (mean proportion CD4+ was 0.60 by flow cytometry and 0.62 by our method) and the standard deviation of the error was 0.08. In addition to placing these samples near the correct proportion CD4+, the ranking of these samples by our method is similar to the ranking calculated from flow cytometry (Spearman's $\rho = 0.82$, $p < 0.02$ by permutation). Together, these experiments demonstrate that our method can correctly estimate the proportion of CD4+ T cell sequences in in vitro and in vivo mixed samples across a wide range of values.

4. Discussion

The CDR3 region of the T cell receptor β locus, which plays a paramount role in determining the binding affinity and specificity of the final T cell receptor, is produced by rearrangement of the genomic TCRβ locus during the course of T cell development in the thymus. In this study, we have leveraged the extremely high-throughput sequencing of TCRβ rearrangements made possible by the ImmunoSEQ platform to identify multiple sequence features in the CDR3 region of rearranged TCRβ loci that distinguish CD4+ T cells from CD8+ T cells, consistent with the hypothesis that positive selection to bind to Class I or Class II MHC molecules selects for slightly different subsets among the many possible TCRβ rearrangements generated during the T cell maturation process. This selective process generates subtle differences in both the usage frequency of the various Vβ and Jβ segments as well as the distribution of lengths observed for the CDR3 region in mature T cells. Utilizing the depth of data available using modern sequencing methods, however, these subtle effects can be informative nonetheless.

In addition to identifying distinguishing sequence features, we have taken advantage of these features to develop a simple but effective method to computationally determine

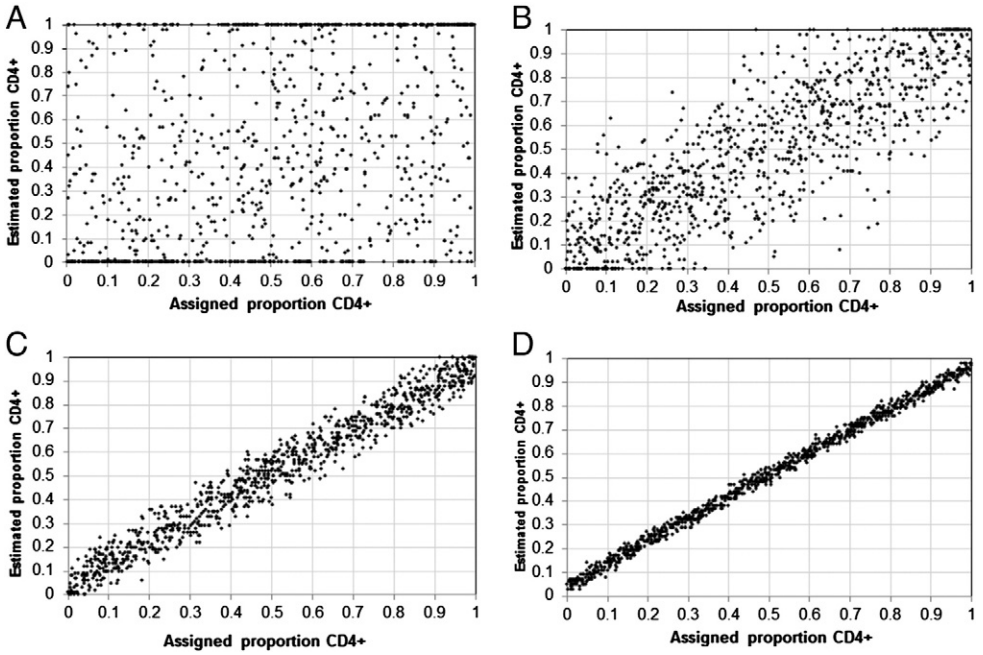

Fig. 5. CD4:CD8 estimation: effect of sample size. In each panel, 1000 experiments are shown in which a proportion CD4+ was chosen at random, and our likelihood model was then run forward to generate a new dataset which was subsequently assessed for proportion CD4+. This experiment elucidates the effect of sample size on our model's ability to estimate the CD4+:CD8+ ratio. A: 1000 datasets with 10 simulated TCRβ sequences each. B: 1000 datasets with 100 simulated TCRβ sequences each. C: 1000 datasets with 1000 simulated TCRβ sequences each. D: 1000 datasets with 10,000 simulated TCRβ sequences each.

the CD4+/CD8+ makeup of an unknown sample. Based on results from both healthy subjects and multiple sclerosis patients, we report that our method can accurately classify sorted samples of CD4+ or CD8+ T cells in healthy and diseased immune systems. We further report that our method can accurately estimate the proportion of CD4+ T cells in a mixed sample as indicated by results obtained from both in vitro mixes of sorted CD4+ and CD8+ T cell sequences and a

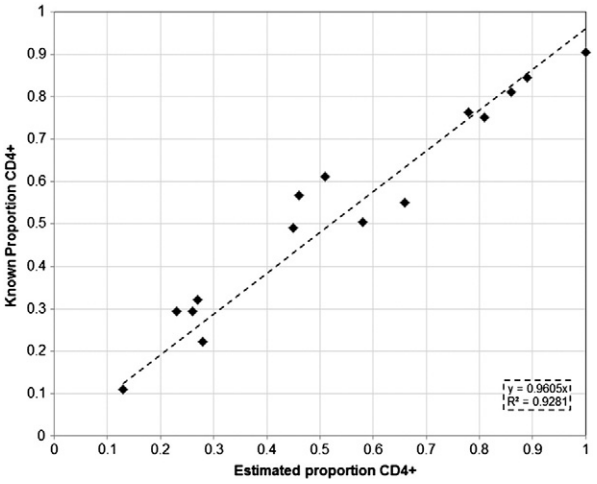

Fig. 6. CD4:CD8 estimation: in vitro validation results. Shown above are the results of 15 experiments in which sorted CD4+ and CD8+ T cell DNA samples from five subjects were mixed to produce samples with nominal proportions of CD4+ sequences of 25%, 50%, or 75%; however, sequencing protocols allowed us to determine the exact proportion CD4+ in each mixed sample. Samples were sequenced, and then assessed computationally to predict the proportion CD4+ T cell sequences in each sample. The black dashes represent a line of best fit. Our model correctly estimates the proportion of CD4+ T cells in samples assembled in vitro from sorted CD4+ and CD8+ TCRβ sequences.

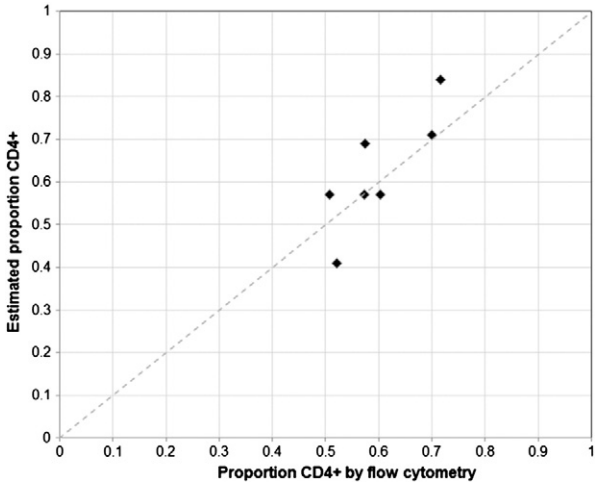

Fig. 7. CD4:CD8 estimation: comparison to flow cytometry. Above, the results generated by our method on PBMC samples from 7 healthy control subjects are compared to data obtained from flow cytometry. Each sample was analyzed by flow cytometry to determine the ratio of CD4+ to CD8+ T cells. Subsequently, the ratio of CD4+ to CD8+ T cells was also estimated by sequencing each sample and applying our computational method. The dashed gray line on the diagonal represents expected results. Our model agrees well with the results of flow cytometry, both in absolute terms (the mean error in proportion CD4+ among these 7 samples was 0.02, with a standard deviation of 0.08) and in relative terms (Spearman's $\rho = 0.82$; $p < 0.02$ by permutation).

comparison of our method to flow cytometry using PBMC samples with various ratios of CD4+:CD8+ T cells.

While our method will be of direct relevance in any situation in which the ratio of CD4+ to CD8+ T cells is desired and sequence data are readily available, we expect that this method may be particularly useful in research and clinical contexts in which high-throughput sequence data are available, and in which traditional flow cytometry is difficult and immunohistochemistry is subject to human error. We expect that some solid tumor samples will fit into this category. Presence and abundance of Tumor Infiltrating Lymphocytes (TILs) are emerging as independent and informative prognostic factors for some immunogenic cancers. For colorectal cancers, TIL presence and abundance more accurately predicts prognosis than the currently used prognostic factors (tumor histology, nodal involvement, and metastases), and the immunological characteristics of TILs may play an important role in prognosis (Sato et al., 2005; Mlecnik et al., 2011). In some cases CD4+:CD8+ ratio has been shown a better prognostic indicator than total TIL density (Sato et al., 2005). Already our method could be used to estimate the CD4+:CD8+ ratio of a TIL sample, and could be used in conjunction with an estimate of total TIL density to determine the density of CD8+ TILs within a tumor, and we expect that further investigations of the relationship between the immunological and TCR sequence characteristics of T cells may provide further answers; information about the proportion of regulatory T cells (and therefore the CD8+/Treg ratio) may be particularly useful for prognosis since the CD8+:Treg ratio has been significantly correlated with clinical outcomes in several studies (Barnett et al., 2010; Gooden et al., 2011). In addition, sequencing assays inherently provide information about the diverse v. oligoclonal nature of the TIL population, which is itself important (Stumpf et al., 2009). In cases such as this, the ability to computationally estimate additional characteristics such as the relative proportions of CD4+ and CD8+ T cells after high-throughput sequencing of TCRs from a solid tissue sample has the potential to become very valuable.

We expect that further work will continue to elucidate the mechanisms and effects of thymic selection on the TCR rearrangements present in mature T cells, as well as improve our ability to computationally infer functional characteristics of mature T cells in healthy and diseased immune systems from high-throughput sequence data.

References

Barnett, J.C., Bean, S.M., Whitaker, R.S., Kondoh, E., Baba, T., Fujii, S., Marks, J.R., Dressman, H.K., Murphy, S.K., Berchuck, A., 2010. Ovarian cancer tumor infiltrating T-regulatory (T(reg)) cells are associated with a metastatic phenotype. Gynecol. Oncol. 116, 556.

Davis, M.M., Bjorkman, P.J., 1988. T-cell antigen receptor genes and T-cell recognition. Nature 334, 395.

Gooden, M.J., de Bock, G.H., Leffers, N., Daemen, T., Nijman, H.W., 2011. The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. Br. J. Cancer 105 (1), 93.

Jameson, S.C., Hogquist, K.A., Bevan, M.J., 1995. Positive selection of thymocytes. Annu. Rev. Immunol. 13, 93.

Leddon, S.A., Sant, A.J., 2010. Generation of MHC class II-peptide ligands for CD4 T-cell allorecognition of MHC class II molecules. Curr. Opin. Organ Transplant. 15, 505.

Mlecnik, B., Tosolini, M., Kirilovsky, A., Berchuck, A., Bindea, G., Meatchi, T., Bruneval, P., Trajanoski, Z., Fridman, W.H., Pages, F., Galon, J., 2011. Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction. J. Clin. Oncol. 29, 610.

Pamer, E., Cresswell, P., 1998. Mechanisms of MHC class I-restricted antigen processing. Annu. Rev. Immunol. 16, 323.

Robins, H.S., Campregher, P.V., Srivastava, S.K., Wacher, A., Turtle, C.J., Kahsai, O., Riddell, S.R., Warren, E.H., Carlson, C.S., 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114, 4099.

Sato, E., Olson, S.H., Ahn, J., Bundy, B., Nishikawa, H., Qian, F., Jungbluth, A.A., Frosina, D., Gnjatic, S., Ambrosone, C., Kepner, J., Odunsi, T., Ritter, G., Lele, S., Chen, Y.T., Ohtani, H., Old, L.J., Odunsi, K., 2005. Intraepithelial CD8+ tumor-infiltrating lymphocytes and a high CD8+/regulatory T cell ratio are associated with favorable prognosis in ovarian cancer. Proc. Natl. Acad. Sci. U.S.A. 102, 18538.

Stumpf, M., Hasenburg, A., Riener, M.O., Jutting, U., Wang, C., Shen, Y., Orlowska-Volk, M., Fisch, P., Wang, Z., Gitsch, G., Werner, M., Lassmann, S., 2009. Intraepithelial CD8-positive T lymphocytes predict survival for patients with serous stage III ovarian carcinomas: relevance of clonal selection of T lymphocytes. Br. J. Cancer 101, 1513.

Vukmanovic, S., 1996. The molecular jury: deciding whether immature thymocytes should live or die. J. Exp. Med. 184, 305.

Yousfi Monod, M., Giudicelli, V., Chaume, D., Lefranc, M.P., 2004. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. Bioinformatics 20 (Suppl. 1), i379.