Using High Dimensional Flow Cytometry and Machine Learning to Evaluate T Cell Function in Hematopoietic Stem Cell Transplant Patients Undergoing Graft-versus-Host Disease

L. Blaha3, M. Perez1, C. David1, Y. Shaw3, A. Patel3, Z. McIver2, and J. Grayson1

1 Department of Microbiology and Immunology, Wake Forest School of Medicine, Winston-Salem, NC
2 Department of Hematology and Oncology, Wake Forest School of Medicine, Winston-Salem, NC
3 Wake Forest School of Medicine, Winston-Salem, NC

Abstract

Hematopoietic stem cell transplantation (HSCT) seeks to reconstitute the hematopoietic function of patients with certain hematologic diseases (primarily leukemia) by transplantation of immunologically normal hematopoietic stem cells from a donor. HSCT can be curative for leukemia, but graft-versus-host-disease (GVHD) is a potentially serious complication which leads to significant morbidity and mortality. Currently, there is a lack of accurate predictive algorithms for a patient's post-HSCT course. The ability to predict GVHD outcome and severity would allow clinicians better prophylaxis, reduce morbidity and mortality, greatly improve the efficacy of HSCT, and allow expansion of its use. The overarching goal of this study and its counterparts is to produce accurate predictive algorithms for a patient's post-HSCT course. Future directions include building an appropriate model to predict GVHD development and severity in future studies.

Methods

Healthy Controls

Gender and racial distribution representative of the USA

Patient Samples

5-15 days post-HSCT

PBMC Isolation

Lymphoprep

Stimulation

Phytohemagglutinin

Acquisition

Fortessa X-20

Staining

17 color panel: Zombie Green viability dye, CD3, CD4, CD8a, CD127, HLA-DR, IFN-γ, TNF-α, IL-2, IL-4, IL-9, IL-10, IL-17, Fox-P3, T-bet, GZB, CD19, CD25, CD69, CD107, CD62L

Initial Analysis

BD FACS Diva

Initial Gating

FlowJit

Import to R

Application of machine learning algorithms

K-Means Clustering

Unsupervised

Self-Organizing Maps

Supervised

Logistic Regression

Random Forest

Results

Table 1. Results of machine learning algorithms, divided by CD4+ and CD8+ T cells. Optimum number of clusters was selected based on the "elbow method" and accounting for spikes on the CD8+ plot.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>CD4+ Error Rate</th>
<th>CD8+ Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>22%</td>
<td>33%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>22%</td>
<td>33%</td>
</tr>
<tr>
<td>K-means Clustering</td>
<td>30 Clusters</td>
<td>40 Clusters</td>
</tr>
</tbody>
</table>

Figure 1. WSS plots demonstrating the number of significant cell clusters generated by K-means clustering analysis. Left (a) shows clusters of CD4+ T cells. Right (b) shows clusters of CD8+ T cells. Optimum number of clusters was selected based on the "elbow method" and accounting for spikes on the CD8+ plot.

Figure 2. Self-organizing maps of significant CD8+ T cell clusters. Left (a) demonstrates cluster prevalence in healthy control samples. Right (b) demonstrates cluster prevalence in post-HSCT transplant patients.

Figure 3. IFN-γ-heavy clusters, indicating mature effector cells (Fig. 2a); patient samples favor CD45RA-heavy clusters, indicating naïve cells (Fig. 2b), as expected due to clinical circumstances.

Conclusions

We conclude that it is possible to use machine learning to classify post-HSCT patients and healthy controls based on T cell surface markers and cytokine production. Logistic regression and random forest may be used with 22-33% accuracy. CD8+ SOMs are dramatic in visual representation of disparate clusters between patients and controls in a predictable pattern, confirming the utility of high dimensional flow and the SOM algorithm (Fig. 2). Control samples favor IFN-γ-heavy clusters, indicating mature effector cells (Fig. 2a); patient samples favor CD45RA-heavy clusters, indicating naïve cells (Fig. 2b), as expected due to clinical circumstances.

Future Directions

The best fit of 30 CD4+ clusters and 40 CD8+ clusters (Fig. 1) is especially important in identifying the unexpected immunological heterogeneity of HSCT patients, warranting further investigation of these clusters as possible predictors of patient outcome. In the future, these machine learning algorithms will be applied to a larger sample size with patients broken out into separate outcome groups, with the goal of creating an accurate algorithm for early prediction and prevention of HSCT patient morbidity and mortality.

References

Acknowledgements: this project was supported by the Wake Forest Clinical and Translational Science Institute and the Laura Scas Cancer Fund.