Subject: Thesis Defense: Daniel Shupp, MS Bioinformatics & Computational Biology
Date: Tuesday, April 7, 2026 at 9:37:15 AM Eastern Daylight Time
From: SSB Faculty List on behalf of Diane St. Germain
To: SSB-FACULTY-LIST-L@LISTSERV.GMU.EDU

Thesis Defense Announcement
To: The George Mason University Community

Candidate: Daniel Shupp

Program: M.S. in Bioinformatics & Computational Biology

Date: April 21, 2026

Time: 10:00 AM Eastern Time (US and Canada)

Location:

Join Zoom Meeting
https://gmu.zoom.us/j/93786298464?pwd=jIru59VWmjetC4L982bzGLkrIbc5U4.1

Meeting ID: 937 8629 8464
Passcode: 758805
One tap mobile
+12678310333,,93786298464#,,,,*758805# US (Philadelphia)
+13017158592,,93786298464#,,,,*758805# US (Washington DC)

Dial by your location
+1 267 831 0333 US (Philadelphia)
+1 301 715 8592 US (Washington DC)
Meeting ID: 937 8629 8464
Passcode: 758805
Find your local number: https://gmu.zoom.us/u/adqryfvEis

Committee Chair: Dr. M. Saleet Jafri

Committee members: Dr. Aman Ullah, Dr. Christopher Lockhart

Title: Machine Learning Classification of Naive and Th1 CD4+ T Cells using RNA-Seq Gene Expression Profiles

Abstract: CD4+ T helper cells are crucial parts of the adaptive immune system. Among their subtypes, Th1 T helper cells are responsible for mediating responses against bacterial, viral, and parasitic infections. RNA-Seq can be used to quantify the relative gene expression in a cell under a certain state, under a condition, and over time. This differential expression across samples helps characterize cell differences, identify disease markers and other biomarkers, and more. Machine learning classification algorithms can also be used to study these differences. In training these classifiers, gene features are identified through feature selection algorithms like LASSO and Random Forest. The difference in these algorithms adjusts which gene features are selected as having high-predicative capabilities, and there is often only minor overlap between these algorithms and traditional differential expression analysis results in gene selection. Raw data was gathered from GEO and relevant literature (total samples: 59), and sent through quality-control, alignment, quantification, and differential expression steps. LASSO and Random Forest feature selection algorithms were used on a normalized gene expression counts matrix to train three classification algorithms (Logistic Regression, Random Forest, and Support Vector Machine) The top differentially expressed genes were saved in addition to the top genes selected by LASSO and Random Forest.

Of the most statistically significant differentially expressed genes, CCR2 and CXCR3 were found to be significantly upregulated. CD38, CFH, and PLXND1 were selected through LASSO feature selection. SMAD7, DMN1L, and MRPL44 were selected through Random Forest feature selection. All identified genes were either of known importance in Th1 biology or have plausible connections to the success of naïve CD4+ differentiation processes. Of the trained models the LASSO-selected Logistic Regression classifier had perfect 1.0 scores across all evaluated metrics (training size n = 42 | test size n = 17). All other models showed perfect or near-perfect scores with the second-best models being both the LASSO-selected Support Vector Machine and Random Forest-selected Random Forest with a mean test accuracy of 0.9278. The lowest scoring model was Random Forest-selected Support Vector Machine with a mean test accuracy and accuracy of 0.7059.

###