An evaluation of machine learning and traditional statistical methods for discovery in large-scale translational data

Megan Hollister

Department of Biostatistics, Vanderbilt University

1. Background

- A recent paper in the journal *Nature Methods* examined multiple approaches for high-dimensional discovery in large-scale translational data, such as identifying dysregulated genes in genomic data
- They concluded that random forests (a machine learning approach) outperformed traditional methods rooted in p-value adjustments
- However, the implementation of machine learning methods used prior knowledge that was not available to the traditional methods
- We repeated this investigation using a modified approach that did not favor any particular method and considered the addition of within person correlation

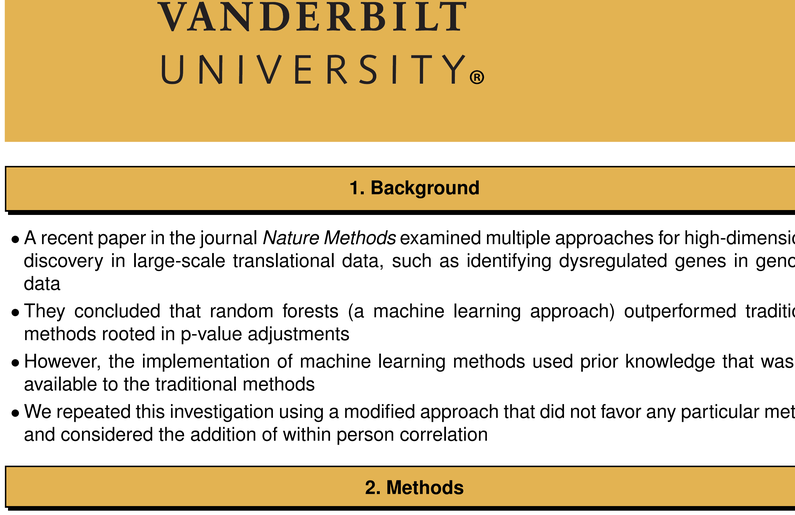
2. Methods

- **Objectives:**
  - To examine the *Nature Methods* claims under a broad set of conditions that is unbiased and fair for all methods
  - To estimate the accuracy of typical machine learning methods and traditional statistical methods for high-dimensional discovery
  - To identify the methods with the best performance characteristics

2.1 Simulated Study Population

- Gene expression data of 40 genes from 20 people
- 10 people are phenotype positive and 10 are phenotype negative (e.g., blue or brown eyes)
- 25% of the genes (10) were set to be ”dysregulated” across phenotype
- Allowed for within person correlation across genes
- Used pseudo-counts to smooth data, which allowed us to easily distinguish gene differences

Individual Gene Expression Data



2.2 Discovery Methods Examined

- A variety of methods were used to identify dysregulated genes
- For the traditional methods we used a pre-specified significance level of 0.05
- To appropriately compare all methods the top 10 ranked p-values/importance levels were taken

P-value adjustments

- Raw p-values
- Benjamini-Hochberg p-values
- Second-generation p-values
- Bonferroni p-values

Machine Learning

- Random Forest importance levels
- Neural Net prediction weights

3. Results

3.1 Comparisons

- Accuracy statistics were computed

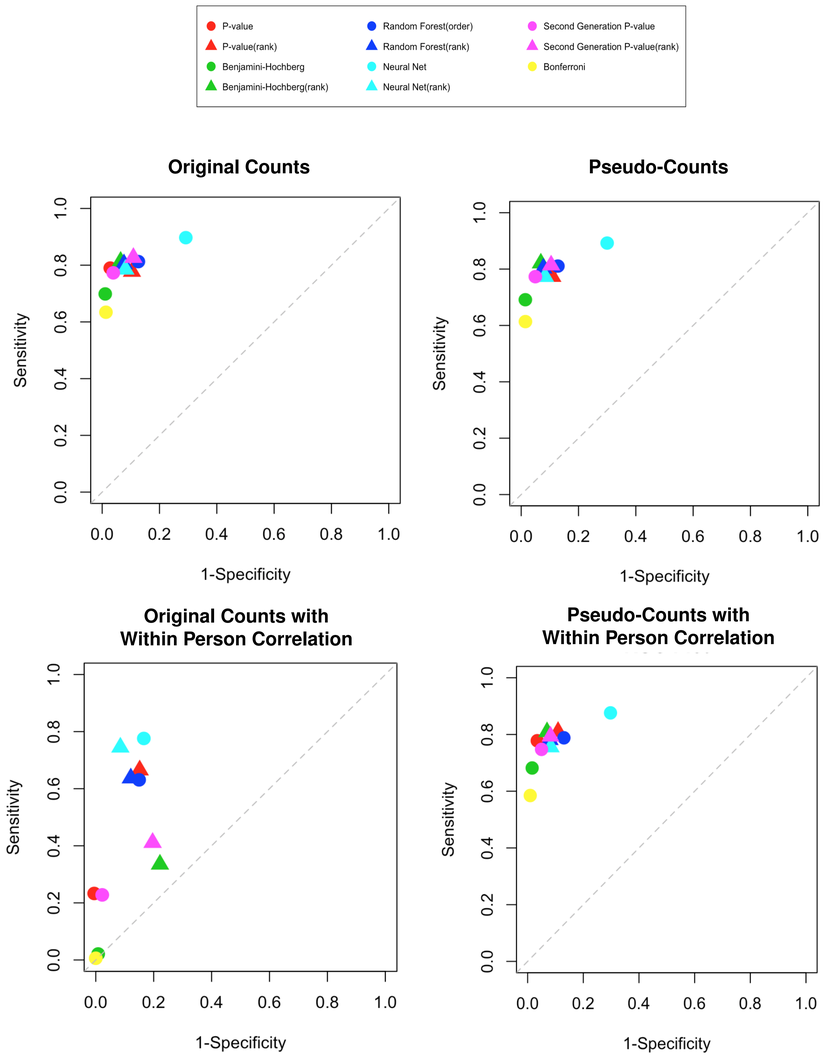
Sensitivity or Power

→ Proportion of ”dysregulated” genes identified as ”dysregulated”

1-Specificity or Type I Error rate

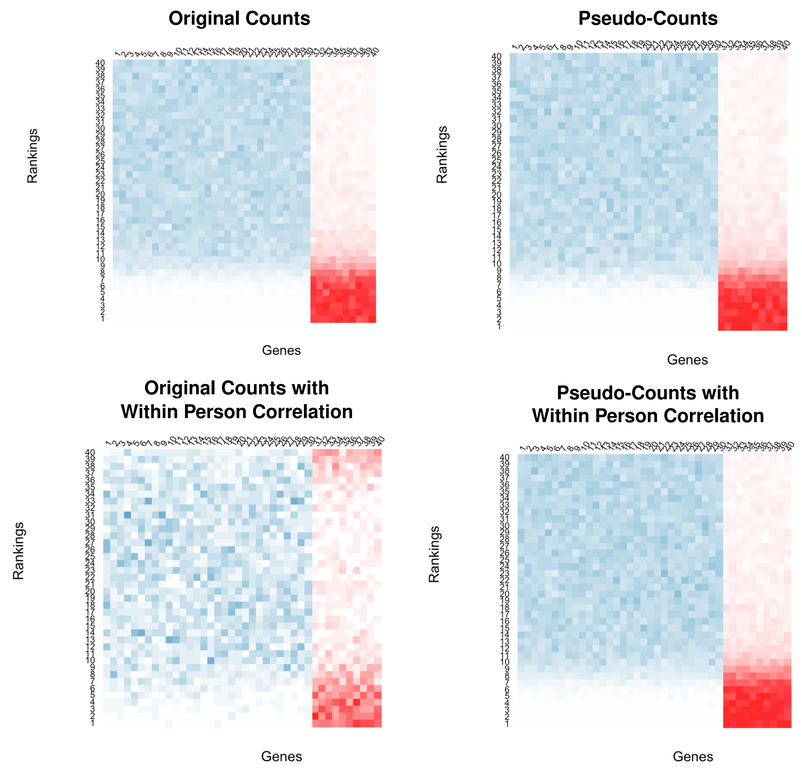
→ Proportion of ”properly regulated” genes identified as ”dysregulated”

ROC Plots



3.2 Rankings

Heat Maps of Rankings by Genes



3.3 Findings

- Performance characteristics depend on which method of comparison is used, either a pre-specified significance level or the top 10 ranked values
- After smoothing, the within person correlation does not change the relative performance, but it slightly changes the accuracy levels of all methods
- Almost uniformly, the machine learning methods did not yield improved accuracy and they depend heavily on the a priori chosen number of ”dysregulated” genes

4. Conclusion

- Machine learning methods only outperform standard methods when they are given extra information
  - Their additional complexity does not lead to improved accuracy in this situation
- The choice of an analysis method for large-scale translation data is critical to the success of any statistical investigation

5. Acknowledgments

NIH Clinical and Translational Science Awards (CTSA) TL1 Training Grant
Vanderbilt University Biostatistics Department:
Dr. Jeffrey Blume, Dr. Thomas Stewart, and Valerie Welty
The Statistical Evidence in Data Science (SEDS) Lab