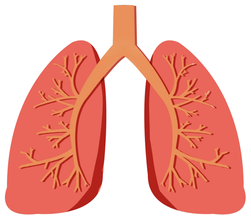
The TBSignatureProfiler: A Novel R Package for Comparing Tuberculosis Gene Expression Signatures

Aubrey Odom[1,2], David Jenkins[3], Yue Zhao[3], W. Evan Johnson[3,4]

[1]Boston University Bioinformatics BRITE REU Program, Summer 2019, Boston, MA
[2]Department of Statistics, Brigham Young University, Provo, UT
[3]Program in Bioinformatics, Boston University, Boston, MA
[4]The Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA

BACKGROUND & MOTIVATION

- **Tuberculosis (TB)** is the leading cause of infectious disease mortality worldwide
- ~1.4 million deaths in 2017[12]
- Traditional **phlegm/sputum tests not always accurate** in TB diagnosis
    - e.g. pediatric, slow-growth
- Published **gene expression signatures can be used instead** as blood-based disease biomarkers[6]
- Most of 30+ **published signatures lack cross-condition validation testing**
    - e.g., testing TB in samples from diverse geographic and comorbidity backgrounds of the signatures performed
- **Aim: Formally aggregate these signatures as a single, unified resource, and develop open source software for their visual & quantitative comparison**
- Developed the **"TBSignatureProfiler"** R package to characterize gene signatures' diagnostic ability in multiple comorbidity settings.



1.4 MILLION DEATHS IN 2017

METHODS

- **Collected set of 34 published signatures**
    - These differentiate patients with multiple disease backgrounds
        - e.g., patients with TB, at risk of TB treatment failure, or have latent TB infection (LTBI) that will likely progress to active TB disease
- **Profiling methods analyze expression levels for gene groups**
    - Known as "gene expression profiles", which may then be scored and correlated to TB outcomes
        - (e.g., active TB vs. LTBI)
    - 6 main methods**:** *GSVA[10], ssGSEA[8], ASSIGN[9], singscore[7], PLAGE[5],* and *comparing Z-scores[11]*
- **TBSignatureProfiler features**
    - Signature strength estimation: bootstrapping estimates' AUC and leave-one-out cross-validation (LOOCV) of logistic regression
    - Visualization: sample-signature score heatmaps, bootstrap area under the curve (AUC, a statistical evaluation metric) & LOOCV boxplots, and results tables
    - Heatmaps for individual signature composition and between-signature comparison
- **Analysis of malnutrition comorbidity data**
    - Cohort from Chennai and Bengaluru, India
    - Study focused on identifying active TB from LTBI in severely malnourished individuals

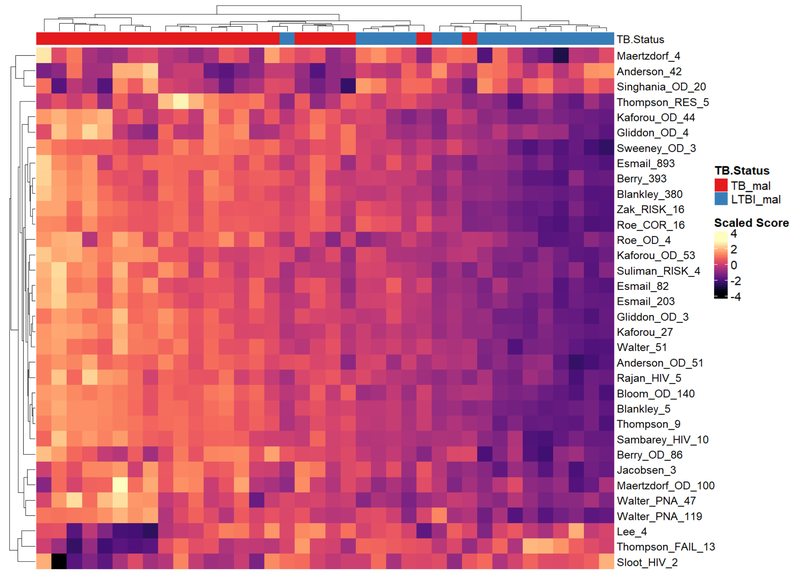


Figure 1. **Heatmap of scaled ssGSEA scores** for all 34 signatures (rows) for the malnourished/HIV infected TB and LTBI individuals (columns). The topmost color bar designates whether the sample is from an LTBI individual (red) or an individual with active TB (blue). Collectively, these signatures accurately separate most of the TB samples from the LTBI samples, and the scores are largely concordant. This heatmap was generated using the SignatureHeatmap() function from the TBSignatureProfiler.

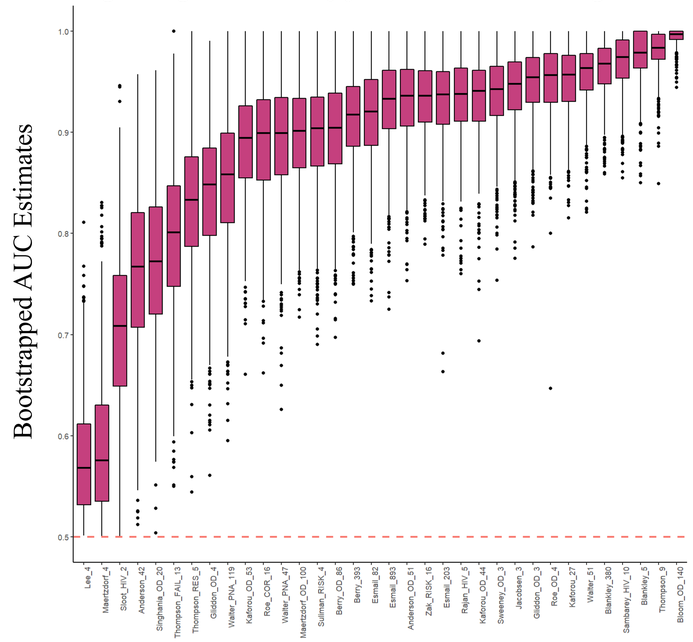


Figure 2. Boxplots of bootstrapped AUC estimates (y-axis) from bootstrapped (n = 1000) samples for each signature (x-axis) using the ssGSEA algorithm. All AUC estimates were above the 0.50 mark.

RESULTS

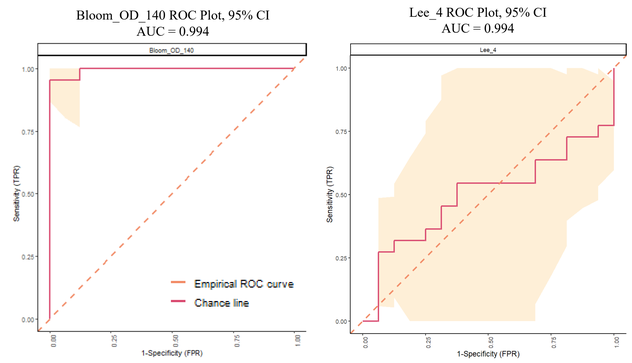


Figure 3. ROC plots for the best and worst performing signatures – Bloom_OD_140 and Lee_4, respectively. We obtain AUC estimates by finding the area under the pink curve. The tan ribbons illustrate 95% CI bands. Across the samples, 82% of signatures perform well with AUCs above 80%. These ROC plots were generated using the signatureROCplot_CI() function from the TBSignatureProfiler.

DISCUSSION & FUTURE WORK

- Existing blood RNA signatures of TB generally work in the undernutrition setting
- Some differences may reflect gene signature size (i.e., smaller signatures may not perform as well) and/or original data training setting
- Lee_4 (AUC = 0.50), Maertzdorf_4 (AUC = 0.54), & Sloot_HIV_2 (AUC = 0.70) signatures do not perform well in the setting of severe undernutrition
- Findings suggest that most TB signatures are robust and could work with many different settings/comorbidities
- Further studies needed to understand impact of additional comorbidities on signature performance
    - (e.g., diabetes, alcoholism, pregnancy)

REFERENCES

Thanks to Evan Johnson for his continual support and guidance, and Gary Benson for inviting me to be a part of the BRITE REU. This work was funded in part by CRDF grant #61407, NIH grant 5U19 AI111276-05NSF, NSF grant DBI-1559829 (awarded to the Boston University Bioinformatics BRITE REU program), and a grant awarded by the Warren Alpert Foundation with co-funding from BU School of Medicine.

[5]Barbie, D.A., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 108-112.

[6]Berry, Matthew P. R., et al. 2010. "An Interferon-Inducible Neutrophil-Driven Blood Transcriptional Signature in Human Tuberculosis." Nature 466 (7309): 973–77.

[7]Foroutan, M. et al. (2018). Single sample scoring of molecular phenotypes. *BMC Bioinformatics*, 19.

[8]Lee, E. et al. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comp Biol*, 4(11):e1000217.

[9]Shen, Y. et al. (2015). ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. *Bioinformatics*, 31, 1745-1753.

[10]Subramanian, A. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102, 15545-15550.

[11]Tomfohr, J. et al. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225.

[12]WHO. "TB Global Health Observatory (GHO) Data." *World Health Organization*, World Health Organization, 27 Dec. 2018, www.who.int/gho/tb/epidemic/cases_deaths/en/.