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

Aubrey Odom1,2, David Jenkins3, Yue Zhao3, W. Evan Johnson3,4
1Boston University Bioinformatics BRITE REU Program, Summer 2019, Boston, MA
2Department of Statistics, Brigham Young University, Provo, UT
3Program in Bioinformatics, Boston University, Boston, MA
4The 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
• 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
• Most of 30+ published signatures lack cross-condition validation testing
• e.g., testing TB in samples from diverse geographic and comorbidity backgrounds
• Aim: Formally aggregate these signatures as a single, unified resource, and develop open source software for their visual & quantitative comparison
• Developed the “TBSsignatureProfiler” R package to characterize gene signatures’ diagnostic ability in multiple comorbidity settings.

METHODS
• Collected set of 34 published signatures
• These differentiated 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: GSVA10, ssGSEA4, ASSIGN9, sngscore7, PLAGE3, and comparing Z-scores11
• TBSsignatureProfiler 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

RESULTS

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 TBSsignatureProfiler.

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.

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 605-05, NSF grant 1541127-615NSF, NSF grant DIB-1539829 (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.

Department of Statistics
9Nature 466 (7309) 973–77.