PROBABILISTIC MIXTURE MODELS IMPROVE CALIBRATION OF PANEL-DERIVED TUMOR MUTATIONAL BURDEN IN THE CONTEXT OF BOTH TUMOR-NORMAL AND TUMOR-ONLY SEQUENCING

1Jordan Anaya, 2John-William Sidhom, 3Alexander Baras*. 1Johns Hopkins, Baltimore, MD, USA; 2Mount Sinai, Baltimore, MD, USA

Background Tumor mutational burden (TMB) has been investigated as a biomarker for immune checkpoint blockade (ICB) therapy. Increasingly, TMB is being estimated with gene panel-based assays (as opposed to full exome sequencing) and different gene panels cover overlapping but distinct genomic coordinates, making comparisons across panels difficult. Previous studies have suggested that standardization and calibration to exome-derived TMB be done for each panel to ensure comparability. However, these studies often propose to use The Cancer Genome Atlas (TCGA) for calibration despite this data having a matched normal to remove germline variants while many labs perform tumor-only sequencing. We suggest this approach has the potential to bias TMB estimates and we propose the use of an alternative dataset for calibration of tumor-only sequencing. Current approaches also propose linear models for TMB calibration. We demonstrate why linear approaches are inappropriate for this data and provide an alternative model.

Methods Our approach to calibration of panel-derived TMB to exomic TMB proposes the use of deep learning to model the probability distribution of the label. We found that a mixture of lognormal distributions can model the nonlinear relationship between panel inputs and TMB as well as the complex error distributions. Using our model we examined the effect of different panel inputs, including nonsynonymous, synonymous, and hotspot counts along with genetic ancestry. To generate a tumor-only version of the TCGA data we reintroduced the germline variants from the matched-normal samples and filtered the data using gnomAD.

Results We were able to model accurately the distribution of both tumor-normal and tumor-only data with our mixture model while a linear model could not (figure 1). Applying a model trained on tumor-normal data to tumor-only input produced biased TMB predictions (figure 2). Including synonymous mutations resulted in better regression metrics across both data types, but ultimately a model able to dynamically weight the various input mutation types exhibited optimal performance. Including genetic ancestry improved model performance only in the context of tumor-only data. Using the probability distribution can help achieve a given positive predictive or negative predictive value.

Conclusions A probabilistic mixture model approach capable of accurately characterizing the distribution of expected exomic TMB from panel inputs better models the nonlinearity and heteroscedasticity of the data. Synthetic tumor-only panel data can allow for the calibration of tumor-only panels to exomic TMB. Leveraging the confidence of the point estimates better informs cohort stratification in terms of TMB.

Acknowledgements The results here are in whole or part based upon data generated by the TCGA Research Network. NGS gene panel coordinates were used from the AACR GENIE consortium. This research was supported by the Mark Foundation for Cancer Research (19-035-ASP), and the philanthropy of Susan Wojcicki and Dennis Troper in support of Computational Pathology at Johns Hopkins.

Abstract 1272 Figure 1 Different modeling strategies for tumor-normal and tumor-only data. Model Fits: Input vs model output scatter plots with overlaid model distribution probabilities (green) along with model uncertainly (95% confidence) of estimate being different from exomic TMB of 10 (grey); TMB Distributions: Histogram of observed distribution of exomic TMB with overlaid model output distribution at the midpoint of the designated range; Model Residuals: conventional residuals plot. (A) tumor-normal data with linear model, (B) tumor-normal data with proposed mixture model, (C) stringent tumor-only data with linear model, (D) stringent tumor-only data with proposed mixture model.

Abstract 1272 Figure 2 Effect of tumor-only sequencing. Prediction versus true plots for a model trained on tumor-normal data and applied to (A) tumor-normal data, (B) tumor-only stringently filtered data, and (C) tumor-only permissively filtered data.