Analysis Pipeline

This analysis pipeline script was written by Tsion Minas and Julián Candia to ensure the reproducibility of results reported in "Serum proteomics links suppression of tumor immunity to ancestry and lethal prostate cancer" by T. Minas, J. Candia et al. For questions and/or comments, please contact Tsion Minas (tsion.minas@nih.gov), Julián Candia (julian.candia@nih.gov), or Stefan Ambs (ambss@mail.nih.gov).

DATA FOLDER:

a) original_data.xlsx: dataset that contains identifiers, socio-demographic and clinical information, and protein expression measurements for 2790 subjects (cases and controls from the Ghana and NCI-MD studies).

b) WestAfrAncestry_NCIMDcontrols.txt: dataset that contains the West-African ancestry index for 795 control subjects from the NCI-MD study.

c) Olink markers grouped by biological processes.xlsx: dataset that provides a classification of Olink proteins by biological process.

d) biologicalprocess_clinicaldemo_survival_cases.xlsx: dataset with survival information for 819 case subjects from the NCI-MD study.

e) immune markers with survival data.xlsx: dataset with merged survival, socio-demographic, clinical and proteomic information used for elastic net regression.

f) Table 1 statistics.xlsx: dataset with Table 1 results used to analyze statistical differences across groups.

FIGURE 1:

a) correlation_matrix.do: was used to create separate files of the 82 immune-oncological proteins for African American cases and controls. The script takes original_data.xlsx and generates AA_cases.xlsx and AA_controls.xlsx as output.

b) Correlation matrix plots were generated using Broad Institute's web-based matrix visualization and analysis platform - Morpheus (https://software.broadinstitute.org/morpheus/) using the following steps:

i. Upload AA_cases.xlsx or AA_controls.xlsx onto the Morpheus software
ii. Remove row annotation
iii. Hierarchical clustering
   1. Metric- one minus Pearson correlation
   2. Linkage method- average
   3. Cluster rows and columns
   4. Group columns by – none selected
   5. Group rows by – none selected

iv. Similarity matrix
   6. Metric- Pearson correlation
   7. Compute matrix for- columns

v. Display options
   8. Link rows and columns
   9. Change row and column size to 16
   10. Show grid
   11. Show values

vi. Export plots

FIGURE 2:

a) multivar_lin_regr.R: was used to perform the multivariate linear regression of each analyte against age, bmi, education, aspirin, smoking, diabetes and PSA, whose results are reported in Supplementary Data 1.

b) multivar_lin_regr_heatmap.R: was used to generate the heatmap presented in Figure 2. Also, this script analyzes differences across cohorts and generates diff_controls.txt as output, which is reported as an additional tab in Supplementary Data 1.

FIGURE 3:

a) remove_outliers.R: was used to set each protein's range of abundance values to saturate at the 1st and 99th percentiles in order to avoid spurious effects from outliers in heatmap plots. The script takes original_data.xlsx as input and generates original_data_outliers_removed.txt as an output for each subject group.

b) Heatmap plots were generated using Broad Institute's web-based matrix visualization and analysis platform - Morpheus (https://software.broadinstitute.org/morpheus/) using the following steps:

i. Upload only the values for the controls (case=0) from the original_data_outliers_removed.txt file onto the Morpheus software

ii. Transpose

iii. Use race as the column annotation

iv. Group markers by race

v. Unsupervised hierarchical clustering of all individuals
   1. Metric- one minus Pearson correlation
   2. Linkage method- average
   3. Cluster rows and columns
   4. Group columns by – none selected
   5. Group rows by – none selected

vi. Export heatmap

c) clustering_enrichment.R: was used to generate dendrogram cuts to assess the association between clusters and cohorts, whose results are reported in Supplementary Figure 5.

FIGURE 4:

analysis_of_variance.R: was used to simultaneously assess the variance analysis for the levels of each of the 82 immune-oncological protein as a function of the genetic estimate of West African admixture among men without prostate cancer from the NCI-Maryland study. The script takes WestAfrAncestry_NCIMDcontrols.txt as input and generates analysis_of_variance.pdf as output.

FIGURE 5:

a) biological_processes.do: was used to create separate files for each biological process. The script takes original_data_outliers_removed.txt as input and generates apoptosis_controls.xlsx, autophagy_controls.xlsx, chemotaxis_controls.xlsx, promoteTI_controls.xlsx, suppressTI_controls.xlsx, and vasculature_controls.xlsx as output.

b) Heat map plots were generated using Broad Institute's web-based matrix visualization and analysis platform - Morpheus (https://software.broadinstitute.org/morpheus/) using the following steps:

i. Upload each of the exported files for each biological process (apoptosis_controls.xlsx, autophagy_controls.xlsx, chemotaxis_controls.xlsx, promoteTI_controls.xlsx, suppressTI_controls.xlsx, and vasculature_controls.xlsx) onto the Morpheus software
ii. Transpose
iii. Use race as the column annotation
iv. Group markers by race
v. Unsupervised hierarchical clustering of all individuals
    1. Metric- one minus Pearson correlation
    2. Linkage method- average
    3. Cluster rows and columns
    4. Group columns by – race
    5. Group rows by – none selected
vi. Export heatmaps

c) biological_processes_averagezscores.do: was used to generate average Z-scores for each biological process (i.e. apoptosis, autophagy, chemotaxis, suppression of tumor immunity, promotion of tumor immunity, and vasculature). The script takes original_data.xlsx as input and generates original_data_BPaveragezscores.xlsx and biological_processes_averagezscores_controls.txt as output.

d) violin_plots.R: was used to generate violin plots for each biological process' average scores grouped by race. The script takes biological_processes_averagezscores_controls.txt as input and generates violin plots for each of the biological processes scores by race (apoptosis_averagezscores_controls.tiff, autophagy_averagezscores_controls.tiff, chemotaxis_averagezscores_controls.tiff, promoteTI_averagezscores_controls.tiff, suppressTI_averagezscores_controls.tiff, and vasculature_averagezscores_controls.tiff).

FIGURE 6:

survival_analysis_cases.do: was used to estimate the unadjusted and adjusted hazard ratios (HR) and 95% confidence intervals (CI) for all-cause mortality, cancer-related mortality, and prostate cancer-specific mortality of cases using Cox regression model. The script takes biologicalprocess_clinicaldemo_survival_cases.xlsx as input and generates the output results in Supplementary Tables 11, 13, and 14. The adjusted Hazard Ratios and the confidence intervals were used to generate the panels in Figure 6 using Graphpad Prism.

FIGURE 7:

eNetXplorer.R: was used for the regularized Cox regression models performed on the African American cohort, whose results are reported in Figure 7, Supplementary Figures 6-7, and Supplementary Data 4-5.

TABLE 1:

a) suppressTI_NCCN.do: was used (1) to group suppression of tumor immunity scores into low (≤median) and high (>median) with cutoffs determined using the distribution of the score among population controls of the NCI-Maryland study and (2) to evaluate the association of high suppression of tumor immunity score with aggressive prostate cancer (NCCN risk score) using multivariable logistic regression analysis. The script inputs original_data_BPaveragezscores.xlsx and generates adjusted odds ratios (OR) and 95% confidence intervals (CI) for all cases, African American cases, and European American cases displayed in Table 1.

b) table_1_differences.R: was used to assess the statistical significance of the differences between AA and EA cohorts reported in Table 1.