Analysis Pipeline

This analysis pipeline script was written by Tsion Minas, Brittany Lord, and Julián Candia to ensure the reproducibility of results reported in "*Association of circulating fatty acids with socio-demographics, diet, FADS1/2 locus, and prostate cancer among Ghanaian, African-American, and European-American men*" by T. Minas, B. Lord et al. For questions and/or comments, please contact Tsion Minas (tsionzminas@gmail.com), Brittany Lord (brittany.lord@nih.gov), Julián Candia (julian.candia@nih.gov), or Stefan Ambs (ambss@mail.nih.gov).

DATA FOLDER:

a) original_data.xlsx: sheet 1 of this dataset contains socio-demographic and clinical information, genotypes for three SNPs in the *FADS1/2* locus, fatty acid measurements, and immune-oncology related pathway scores for 2934 subjects (cases and controls from the Ghana and NCI-MD studies). Each of the variables and their values are defined in sheet 2.

b) clinico-demographic_fattyacid.xlsx: sheet 1 of this dataset contains socio-demographic and clinical information and fatty acid measurements for 2934 subjects. The variables and their values are defined in sheet 2. Sheets 3 of the dataset lists each of the 24 fatty acids and the respective fatty acid classes that they belong to.

c) gleason_ghana_ncimd.xlsx: this dataset contains the Gleason scores for the prostate tumors of cases from the Ghana and NCI-MD studies.

Note: by editorial request, age is only provided for controls. For cases, we only provide age quartile information.

FIGURE 1:

a) Unsupervised hierarchical clustering plots were generated for 24 individual fatty acids in cases and controls separately with JMP v.14.0 statistical software using the following steps:
   1. Upload original_data.xlsx into the JMP statistical software
   2. Hierarchical clustering analysis
      a. Select 24 individual fatty acids for Y, columns
      b. Ordering: none
      c. Label: none
      d. By: case
      e. Method – Ward
      f. Option: Two way clustering
      g. Option: Color map, diverging green to black to red
      h. Option: More color map columns, race

i. Export plots

b) fa_class.do: was used to create a dataset with the six fatty acid classes. The script takes original_data.xlsx and generates fa_class.xlsx as output.

Clustered bar graphs of fatty acid group ratios by case status and population group were generated using Excel v.16.0 using the following steps:

1. Upload fa_class.xlsx into Excel v.16.0
2. Determine average concentrations of 6 fatty acid groups by race and case using =AVERAGE function in Excel
3. Calculate ratios by population group using the =DIVIDE function in Excel
4. To generate clustered bar graphs:
    a. Highlight fatty acid group names and each accompanying ratio value
    b. Insert a clustered bar graph for each ratio comparison by population group
    c. Format data series, fill color

FIGURE 2:

a) multivar_lin_regr.R: was used to perform the multivariate linear regression of each analyte against age, bmi, education, smoking, diabetes and aspirin. The script takes clinico-demographic_fatty acid.xlsx as input and generates Afr_case_pval.txt, Afr_case_res.txt, Afr_case_coef.txt, Afr_ctrl_pval.txt, Afr_ctrl_res.txt, Afr_ctrl_coef.txt, AA_case_pval.txt, AA_case_res.txt, AA_case_coef.txt, AA_ctrl_pval.txt, AA_ctrl_res.txt, AA_ctrl_coef.txt, EA_case_pval.txt, EA_case_res.txt, EA_case_coef.txt, EA_ctrl_pval.txt, EA_ctrl_res.txt, EA_ctrl_coef.txt as output, which is reported as Supplementary Data 1. Whereas multivar_lin_regr_w_gleason.R: was used to perform the multivariate linear regression of each analyte against Gleason score in addition to age, bmi, education, smoking, diabetes and aspirin. The script takes clinico-demographic_fattyacid.xlsx and gleason_ghana_ncimd.xlsx as input and generates Afrcase_gleasoncoef, Afrcase_gleasonFstat, Afrcase_gleasonpval, AAcase_gleasoncoef, AAcase_gleasonFstat, AAcase_gleasonpval, EAcase_gleasoncoef, EAcase_gleasonFstat, and EAcase_gleasonpval as output, which is reported as Supplementary Data 2.

b) multivar_lin_regr_heatmap.R: was used to generate the heatmap presented in Figure 2. The script takes Afr_case_pval.txt, Afr_case_res.txt, Afr_case_coef.txt, Afr_ctrl_pval.txt, Afr_ctrl_res.txt, Afr_ctrl_coef.txt, AA_case_pval.txt, AA_case_res.txt, AA_case_coef.txt, AA_ctrl_pval.txt, AA_ctrl_res.txt, AA_ctrl_coef.txt, EA_case_pval.txt, EA_case_res.txt, EA_case_coef.txt, EA_ctrl_pval.txt, EA_ctrl_res.txt, and EA_ctrl_coef.txt as input and generates multivar_lin_regr_heatmap_by_caco.pdf and multivar_lin_regr_heatmap_by_race.pdf as output, which is displayed as Figure 2 in the manuscript.

FIGURE 3:

FADS_snps.do: was used to create separate files for each SNP separated by case status and population group. The script takes original_data.txt as input and generates rs174556_Af_controls.txt, rs174556_AA_controls.txt , rs174556_EA_controls.txt , rs174556_Af_cases.txt , rs174556_AA_cases.txt , rs174556_EA_cases.txt , rs174583_Af_controls.txt , rs174583_AA_controls.txt , rs174583_EA_controls.txt , rs174583_Af_cases.txt , rs174583_AA_cases.txt , rs174583_EA_cases.txt as output.

a. rs174556_Af_controls.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174556) in *FADS1* gene in Ghanaian controls. The script takes rs174556_Af_controls.txt as input and generates rs174556_Af_controls.pdf as output.
b. rs174556_Af_cases.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174556) in *FADS1* gene in Ghanaian cases. The script takes rs174556_Af_cases.txt as input and generates rs174556_Af_cases.pdf as output.
c. rs174556_AA_controls.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174556) in *FADS1* gene in African-American controls. The script takes rs174556_AA_controls.txt as input and generates rs174556_AA_controls.pdf as output.
d. rs174556_AA_cases.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174556) in *FADS1* gene in African-American cases. The script takes rs174556_AA_cases.txt as input and generates rs174556_AA_cases.pdf as output.
e. rs174556_EA_controls.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174556) in *FADS1* gene in European-American controls. The script takes rs174556_EA_controls.txt as input and generates rs174556_EA_controls.pdf as output.
f. rs174556_EA_cases.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174556) in *FADS1* gene in European-American cases. The script takes rs174556_EA_cases.txt as input and generates rs174556_EA_cases.pdf as output.
g. rs174583_Af_controls.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174583) in *FADS2* gene in Ghanaian controls. The script takes rs174583_Af_controls.txt as input and generates rs174583_Af_controls.pdf as output.
h. rs174583_Af_cases.R: was used to simultaneously assess the variance analysis for the levels of each of the 24 fatty acids as a function of a SNP (rs174583) in *FADS2* gene in Ghanaian cases. The script takes rs174583_Af_cases.txt as input and generates rs174583_Af_cases.pdf as output.
i. rs174583_AA_controls.R: was used to simultaneously assess the variance analysis for the levels of each of the 24 fatty acids as a function of a SNP (rs174583) in *FADS2* gene in

African-American controls. The script takes rs174583_AA_controls.txt as input and generates rs174583_AA_controls.pdf as output.

j. rs174583_AA_cases.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174583) in *FADS2* gene in African-American cases. The script takes rs174583_AA_cases.txt as input and generates rs174583_AA_cases.pdf as output.

k. rs174583_EA_controls.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174583) in *FADS2* gene in European-American controls. The script takes rs174583_EA_controls.txt as input and generates rs174583_EA_controls.pdf as output.

l. rs174583_EA_cases.R: was used to simultaneously assess the variance for the levels of each of the 24 fatty acids as a function of a SNP (rs174583) in *FADS2* gene in European-American cases. The script takes rs174583_EA_cases.txt as input and generates rs174583_EA_cases.pdf as output.

FIGURE 4:

a) fa_caco.do: was used to create a dataset with the 24 fatty acids for each population group. The script takes original_data.xlsx and generates Af_caco_fa.xlsx, AA_caco_fa.xlsx, and EA_caco_fa.xlsx as output.

b) The differences in the level of each fatty acids in cases vs. controls for each population groups were computed in the form of log2FoldChange and -LOG10(p-value) in Excel v.16.0 using the following steps:

   1. Upload Af_caco_fa.xlsx into Excel v.16.0
   2. Transpose entire dataset
   3. Determine average concentrations for each of the 24 fatty acids in cases and controls using =AVERAGE function in Excel
   4. Fold Change (FC) is calculated by dividing the average fatty acid concentrations for cases by the average concentration of fatty acids for controls in Excel
   5. To get log2 fold change value from the FC the =IMLOG2 function in Excel was used
   6. To get the p-value the =TTEST function in Excel was used
   7. -LOG10(p-value) was generated using the =LOG10 function in Excel
   8. The transposed data with the calculations is saved as Af_caco_fa_transposed.xlsx
   9. Steps 1-7 were repeated for AA_caco_fa.xlsx and EA_caco_fa.xlsx

c) The volcano plots were generated using Graphpad Prism using the following steps

   1. Select ‘XY’ Table and graph
   2. For Data table, select ‘Enter or import data into a new table’
   3. For X values, select ‘Numbers’
   4. For Y values, select ‘Enter and plot a single Y value for each point’
   5. Click the ‘Create’ button
   6. Paste in the log2(FC) and -Log10(p-values) calculated using steps 1-7 above
   7. For the graph select the points only option

8. Shift the Y-axis to the left by clicking the 'Format Axes' button and under 'Frame and Origin' tab change the 'Set origin' option to 'Lower left'
9. Add significant threshold line (p-value of 0.002 or –log10(p-value) of 2.70) onto the plot using the following steps
    a. Click on 'Format Axes' button
    b. Select 'Left Y axis' tab
    c. Under 'Additional ticks and grid lines' enter a y=2.70
    d. Select the 'line' box
10. Add a FC=1 or log2FC=0 threshold line using the following steps
    a. Click on 'Format Axes' button
    b. Select 'X axis' tab
    c. Under 'Additional ticks and grid lines' enter a x=0
    d. Select the 'line' box

FIGURE 5:

a) correlation_heatmap_data.do: was used to create a dataset containing the six fatty acid classes and seven immune-oncology related pathway scores. The script takes original_data.xlsx and generates correlation_heatmap_data.xlsx as output.

b) Correlation heatmaps were generated by case status and population groups using JMP v.14.0 statistical software with the following steps:
    1. Upload correlation_heatmap_data.xlsx into the JMP statistical software
    2. Highlight all rows with control participants and subset a new table in JMP called correlation_heatmap_data_controls.jmp (n=1,503).
    3. Fit Y by X analysis – Determine correlation using Bivariate Fit
        a. Y, response: apoptosis, autophagy, chemotaxis, inflammation, promoteTI, suppressTI, vasculature
        b. X, factor: saturated, cis_mono, omega3, omega6, omega6_3, trans
        c. Block: None
        d. Weight: None
        e. Freq: None
        f. By: race
        g. Option: Density Ellipse = 0.95
        h. Correlation – make combined data table
        i. Option: Select matching cells to clear blank rows
        j. Option: Delete rows
    4. Use Graph Builder in JMP To generate the correlation heatmaps:
        a. Graph choice: Heatmap
        b. Group X: race
        c. Y: Y (immune oncological pathways)
        d. X: X (fatty acid groups)
        e. Color: Correlation
        f. Option: legend gradient, color theme, diverging green to black to red
        g. Export plots

TABLE 1:

table1.do: was used (1) to generate z-score values for each of the 24 fatty acids and (2) to estimate prostate cancer risk per one standard deviation increase for each fatty acid using multivariable logistic regression analysis. The script inputs original_data.xlsx and generates adjusted odds ratios (OR) and 95% confidence intervals (CI) for all study participants and for Ghanaians, African American, and European American men separately as displayed in Table 1.

TABLE 2:

table2.do: was used (1) to divided each trans fatty acid blood concentration (elaidic acid, palmitelaidic acid, and linoelaidic acid) into tertiles (i.e. low, intermediate, and high) for the NCI-MD and Ghana studies separately and (2) to assess the association of high trans fatty acid levels with prostate cancer in a dose-dependent manner across the three population groups using multivariable logistic regression analyses. The script inputs original_data.xlsx and generates adjusted odds ratios (OR) and 95% confidence intervals (CI) for Ghanaian, African American, and European American men displayed in Table 2.