Correct Reference Analysis: Visualization Scripts

Overview

The files listed in the Correct Reference Analysis folder create visuals of the correctly specified reference group results across two to five contributors in a mixture. Each script calculates the false positive rate (POI-) for each group involved in the analysis across a number of contributors when the threshold is greater than 0 and stores the results in a data frame. The file **FNR_FPR_Varying_Thresholds.R** calculates both the false positive rate and the false negative rate across different thresholds and stores the results in separate data frames. Each group is named according to their genetic diversity (AEH), but the **informed consent** file contains both the group name and genetic diversity. Lastly, each file creates a visual of the distribution of false positive rates across each possible number of contributors per mixture.

Understanding The Scripts

FBI_Plot.R

This script creates a linear plot of the distribution of FPRs across varying numbers of contributors for the FBI reference groups.

To run this script you will need the following files:

- **Informed_Consent_AEHs.csv:** this file contains the AEH and names of the 83 groups involved in the analysis
- **Direct_not_in_mix_3dresults_051724.RDS:** this file is a 3D matrix of the likelihood ratios of each group per contributor in a mixture per number of simulations (matrix is 83 x 5x 100,000).

Lines 1-23: Sets the working directory, loads in all of the library packages and the 3D matrix file

Lines 25-37: This section calculates the false positive rate and stores the results in a matrix.

- The 3D matrix file contains some -INF values, which cannot be graphed unless set to a different value and in this case we set all -INFs = -10000.

- A nested for loop was used to fill the empty matrix created in line 29. The for loop calculates the percentage of log(LR) that are greater than 0, which can be referred to as the false positive rate.
- The **Informed_Consent_AEHs.csv** file is loaded in and reordered alphabetically in lines 37 and 39 to ensure proper labeling.

Lines 40 - 75: This section subsets the data to only include the FBI populations.

- The 8 FBI populations are subsetted from the FPR_matrix and combined into one array using the rbind() function. The data is now stored in an array labeled **FBI_groups_combined**
- **FBI_groups_combined's** columns are labeled according to the possible number of contributors per mixture (2-6). The **informed_consent_AEH.csv** file is used to label the rows of the array according to the group's respective genetic diversity (AEH).
- Line 74 adds a column labeled "contribs" to the data frame and line 76 creates a key ID name for this column for graphing purposes.

Lines 79-98: This section builds the FBI plot using ggplot() package

- The color brewers package is called for a specific set of colors to represent each FBI group that will be plotted
- The FBI plot is built using the ggplot() and geom_point() function. The legend included in this plot uses the specific names of the FBI groups involved in this analysis rather than their genetic diversity (AEH), but AEH is used to order the legend and FBI group data points.
- On line 95 scale_y_continuous(trans = "log10") function is used to create a log scaled y-axis

FNR_FPR_Varying_Thresholds.R

This script creates a panel plot consisting of 4 rows and 2 columns. Each row will display a violin plot of the distribution of false positive rates and false negative rates under varying thresholds. Column 1 will display the false positive rate and column 2 will display the false negative rate.

To run this script you will need the following files:

- **Informed_Consent_AEHs.csv:** this file contains the AEH and names of the 83 groups involved in the analysis
- **Direct_in_mix_3dresults_051924.RDS:** this file is a 3D matrix of the likelihood ratios of each group per number of contributors in a mixture per number of simulations when the person of interest is included in the mixture (POI+). (The matrix is 83 x 5 x 100,000).
- **Direct_not_in_mix_3dresults_051724.RDS:** this file is a 3D matrix of the likelihood ratios of each group per number of contributors in a mixture per number of simulations when the person of interest is not included in the mixture (POI-). (The matrix is 83 x 5x 100,000).

Lines 1-27: Sets the working directory and loads in the necessary library packages, POI+ result file, and the informed consent file. Line 27 reorders the file alphabetically.

Line 30-46: In this section of the script 4 different thresholds will be used to calculate the false negative rate and the results of each threshold will be stored in a matrix. In total this section builds 4 matrices.

- Line 30: the threshold values that will be used to calculate the false negative rate are stored in a vector (**thresholds**)
- Lines 32-43: this nested for loop will create a new matrix (83x5) to store the false negative rate of each group per number of contributors in a mixture. The for loop will iterate through the empty matrix to fill each row and column with the false negative rate that is calculated in line 39.
- Line 45: names the newly created matrices based on the threshold value used to calculate the false negative rates and saves the matrices in the global environment

Lines 48-52: This section turns the newly created matrices into data frames using the as.data.frame() function

Lines 53-62: This section assigns column names to each data frame using the possible number of contributors per mixture from each simulation. The row names are also assigned using the AEH values stored in the **informed_consent_AEH.csv** file

Lines 63-77: This section of the script restructures the data frames for ease of graphing.

- Lines 63-72: transposes the data frames and adds a new column labeled **contribs** for specification of the number of contributors in a mixture for each false negative rate per group.
- Lines 73-77: Sets a key name to the contribs column so that the false negative rate distributions can be graphed against the number of contributors in a mixture.

Lines 80-214: This section of the script builds a separate data frame for the 8 groups that have been selected to represent the range of genetic diversity of the groups involved in this analysis.

- Lines 80-113: Creates a new matrix of the 8 selected groups when the threshold is >0
- Lines 114-148: Creates a new matrix of the 8 selected groups when the threshold is > 2
- Lines 148-181: Creates a new matrix of the 8 selected groups when the threshold is > 4
- Lines 182-214: Creates a new matrix of the 8 selected groups when the threshold is > 6

Lines 215 - 316: This section of the script builds the individual threshold violin plots displaying the distribution of false negative rates and combines each plot into one panel plot that has 1 column and 4 rows.

- Lines 217-219: This is a function that changes the way the y-axis tick mark labels will be written. Rather than writing the values in scientific notation (1e-5) the labels will be written as (1.0 x 10^-5).
- Lines 221-24: Builds a violin distribution plot of the false negative rate where the selected 8 populations are graphed linearly across the number of contributors in a mixture when the threshold >0 . This graph has a legend that will be removed later on
- Lines 243-256: This function extracts a legend from a plot and is used to extract the legend from **fn_0_plot**
- Lines 260-264: This section makes the y-axis log scaled and removes the legend.
- Lines 266-280: Builds violin plot when threshold is >2
- Lines 283-296: Builds violin plot when threshold is >4
- Lines 299-311: Builds violin plot when threshold is >6
- Lines 313-316: Combines all 4 plots into one panel plot and adds y-axis title “False Negative Rate”

Lines 320-322: This section loads in the POI- file (**Direct_not_in_mix_3dresults_051724.RDS**) and turns all -INFs values into another value for graphing purposes.

Line 324-337: In this section of the script 4 different thresholds will be used to calculate the false positive rate and the results of each threshold will be stored in a matrix. In total this section builds 4 matrices.

- Lines 327-333: this nested for loop will create a new matrix (83x5) to store the false positive rate of each group per number of contributors in a mixture. The for loop will iterate through the empty matrix to fill each row and column with the false positive rate that is calculated in line 332.
- Line 336: names the newly created matrices based on the threshold value used to calculate the false positive rates and saves the matrices in the global environment

Lines 338-342: This section turns the newly created matrices into data frames using the as.data.frame() function

Lines 343-352: This section assigns column names to each data frame using the possible number of contributors per mixture from each simulation. The row names are also assigned using the AEH values stored in the **informed_consent_AEH.csv** file

Lines 353-367: This section of the script restructures the data frames for ease of graphing.

- Lines 354 - 357: transposes the data frames and adds a new column labeled **contribs** for specification of the number of contributors in a mixture for each false positive rate per group.
- Lines 359-362: Sets a key name to the contribs column so that the false positive rate distributions can be graphed against the number of contributors in a mixture.

Lines 369-522: This section of the script builds a separate data frame for the 8 groups that have been selected to represent the range of genetic diversity of the groups involved in this analysis.

- Lines 370-406: Creates a new matrix of the 8 selected groups when the threshold is >0
- Lines 407-445: Creates a new matrix of the 8 selected groups when the threshold is > 2
- Lines 446-483: Creates a new matrix of the 8 selected groups when the threshold is > 4
- Lines 485-522: Creates a new matrix of the 8 selected groups when the threshold is > 6

Lines 525 - 611: This section of the script builds the individual threshold violin plots displaying the distribution of false positive rates and combines each plot into one panel plot that has 1 column and 4 rows.

- Lines 525-544: Builds a violin distribution plot of the false negative rate where the selected 8 populations are graphed linearly across the number of contributors in a mixture when the threshold >0 .
- Lines 546-562: Builds violin plot when threshold is >2

- Lines 564-580: Builds violin plot when threshold is >4
- Lines 582-598: Builds violin plot when threshold is >6
- Lines 600-611: Combines all 4 plots into one panel plot and adds y-axis title “False Positive Rate”

FPR_AEH_Distribution_Plot.R

This script creates 5 scatter plots of the distribution of false positive rates across the genetic diversity of the 83 groups involved in this analysis when the person of interest is not included in the mixture (POI-). Each plot represents the distribution of FPRs per number of contributors in a mixture. All 5 plots will be combined into one panel plot that consists of 1 column and 5 rows.

To run this script you will need the following files:

- **Informed_Consent_AEHs.csv:** this file contains the AEH and names of the 83 groups involved in the analysis
- **Direct_not_in_mix_3dresults_051724.RDS:** this file is a 3D matrix of the likelihood ratios of each group per number of contributors in a mixture per number of simulations when the person of interest is not included in the mixture (POI-). (The matrix is 83 x 5x 100,000).

Lines 1-26: Sets the working directory, loads the necessary library packages, and reads in the POI- file.

- Line 26: -INF values are changed into another value (-10000) to make these values graphable

Line 30-35: A nested for loop will create a new matrix (83x5) to store the false negative rate of each group per number of contributors in a mixture. The for loop will iterate through the empty matrix to fill each row and column with the false positive rate that is calculated in line 33.

Line 38: **informed_consent_AEH** file is loaded in

Line 42-47: The FPR matrix get restructured into a data frame

- Line 44-46: columns are labeled using the possible numbers of contributors per mixture and the AEH of each group is added as a new column in the dataframe

Line 49-108: builds the scatter distribution plot for each number of contributors

- Lines 49-58: builds the scatter distribution plot for 2 contributor mixtures
- Lines 60-70: builds the scatter distribution plot for 3 contributor mixtures
- Lines 73-83: builds the scatter distribution plot for 4 contributor mixtures
- Lines 86-96: builds the scatter distribution plot for 5 contributor mixtures
- Lines 99-108: builds the scatter distribution plot for 6 contributor mixtures

Line 110-117: combines all plots into one panel plot and labels the x and y-axes

FPR_linear_violin_plot.R

This script creates a violin distribution plot of the false positive rates across all numbers of contributors when the person of interest is not included in the mixtures (POI-). The y-axis of this plot will be linear.

To run this script you will need the following files:

- **Informed_Consent_AEHs.csv:** this file contains the AEH and names of the 83 groups involved in the analysis
- **Direct_not_in_mix_3dresults_051724.RDS:** this file is a 3D matrix of the likelihood ratios of each group per number of contributors in a mixture per number of simulations when the person of interest is not included in the mixture (POI-). (The matrix is 83 x 5x 100,000).

Lines 1-29: Sets the working directory, loads the necessary library packages, and reads in the POI- file and the informed consent file.

- Line 25: -INF values are changed into another value (-10000) to make these values graphable

Line 33-38: A nested for loop will create a new matrix (83x5) to store the false negative rate of each group per number of contributors in a mixture. The for loop will iterate through the empty matrix to fill each row and column with the false positive rate that is calculated in line 33.

Line 41-54: The FPR matrix get restructured into a data frame

- Line 43-46: columns are labeled using the possible numbers of contributors per mixture and
- Line 51-54: the AEH of each group are added as a new column in the dataframe

Lines 57-92: This section of the script builds a separate data frame for the 8 groups that have been selected to represent the range of genetic diversity of the groups involved in this analysis.

- The **fpr_combined_df** is restructured in the same manner as the **FPR_DF**

Line 94-111: This section builds the violin distribution plot of false positive rates (y-axis) across the number of contributors in a mixture (x-axis). The 8 groups selected to represent the range of genetic diversity of all groups involved in this analysis are displayed as line graphs overlaying the violin plot. The y-axis scale of this graph is linear.

FPR_log_violin_plot.R

This script creates a violin distribution plot of the false positive rates across all numbers of contributors when the person of interest is not included in the mixtures (POI-). The y-axis of this plot will be log-scale.

To run this script you will need the following files:

- **Informed_Consent_AEHs.csv:** this file contains the AEH and names of the 83 groups involved in the analysis
- **Direct_not_in_mix_3dresults_051724.RDS:** this file is a 3D matrix of the likelihood ratios of each group per number of contributors in a mixture per number of simulations when the person of interest is not included in the mixture (POI-). (The matrix is 83 x 5x 100,000).

Lines 1-29: Sets the working directory, loads the necessary library packages, and reads in the POI- file and the informed consent file.

- Line 25: -INF values are changed into another value (-10000) to make these values graphable

Line 33-38: A nested for loop will create a new matrix (83x5) to store the false negative rate of each group per number of contributors in a mixture. The for loop will iterate through the empty matrix to fill each row and column with the false positive rate that is calculated in line 33.

Line 41-54: The FPR matrix get restructured into a data frame

- Line 43-46: columns are labeled using the possible numbers of contributors per mixture and
- Line 51-54: the AEH of each group are added as a new column in the dataframe

Lines 57-92: This section of the script builds a separate data frame for the 8 groups that have been selected to represent the range of genetic diversity of the groups involved in this analysis.

- The **fpr_combined_df** is restructured in the same manner as the **FPR_DF**

Line 94-112: This section builds the violin distribution plot of false positive rates (y-axis) across the number of contributors in a mixture (x-axis). The 8 groups selected to represent the range of genetic diversity of all groups involved in this analysis are displayed as line graphs overlaying the violin plot. The y-axis scale of this graph is log-scaled.

- Line 108: the **scale_y_continuous(trans = “log10”)** is what causes this change in y-axis

hdr_FPR_linear_violin_plot.R

This script creates a violin distribution plot of the false positive rates across all numbers of contributors when the person of interest is not included in the mixtures (POI-). The LRs used to calculate the false positive rates were calculated under conditions with a higher dropout rate (prDHet = 0.02 and prDHom=0.04). The y-axis of this plot will be linear.

To run this script you will need the following files:

- **Informed_Consent_AEHs.csv:** this file contains the AEH and names of the 83 groups involved in the analysis
- **Direct_not_in_mix_3dresults_041224.RDS:** this file is a 3D matrix of the likelihood ratios of each group per number of contributors in a mixture per number of simulations when the person of interest is not included in the mixture (POI-). The likelihood ratios are calculated under conditions where a higher dropout rate has been implemented (prDHet = 0.02 and prDHom=0.04). The matrix is 83 x 5x 100,000.

Lines 1-29: Sets the working directory, loads the necessary library packages, and reads in the high dropout rate (hdr) POI- file and the informed consent file.

- Line 25: -INF values are changed into another value (-10000) to make these values graphable

Line 33-38: A nested for loop will create a new matrix (83x5) to store the false negative rate of each group per number of contributors in a mixture. The for loop will iterate through the empty matrix to fill each row and column with the false positive rate that is calculated in line 33.

Line 41-54: The hdr_FPR_matrix get restructured into a data frame

- Line 43-46: columns are labeled using the possible numbers of contributors per mixture and
- Line 51-54: the AEH of each group are added as a new column in the dataframe

Lines 57-92: This section of the script builds a separate data frame for the 8 groups that have been selected to represent the range of genetic diversity of the groups involved in this analysis.

- The **hdr_combined_df** is restructured in the same manner as the **hdr_FPR_DF**

Line 94-111: This section builds the violin distribution plot of false positive rates (y-axis) across the number of contributors in a mixture (x-axis). The 8 groups selected to represent the range of genetic diversity of all groups involved in this analysis are displayed as line graphs overlaying the violin plot. The y-axis scale of this graph is linear.

hdr_FPR_violin_plot.R

This script creates a violin distribution plot of the false positive rates across all numbers of contributors when the person of interest is not included in the mixtures (POI-). The LRs used to calculate the false positive rates were calculated under conditions with a higher dropout rate (prDHet = 0.02 and prDHom=0.04). The y-axis of this plot will be log scaled.

To run this script you will need the following files:

- **Informed_Consent_AEHs.csv:** this file contains the AEH and names of the 83 groups involved in the analysis
- **Direct_not_in_mix_3dresults_041224.RDS:** this file is a 3D matrix of the likelihood ratios of each group per number of contributors in a mixture per number of simulations when the person of interest is not included in the mixture (POI-). The likelihood ratios are calculated under conditions where a higher dropout rate has been implemented (prDHet = 0.02 and prDHom=0.04). The matrix is 83 x 5x 100,000.

Lines 1-29: Sets the working directory, loads the necessary library packages, and reads in the high dropout rate (hdr) POI- file and the informed consent file.

- Line 25: -INF values are changed into another value (-10000) to make these values graphable

Line 33-38: A nested for loop will create a new matrix (83x5) to store the false negative rate of each group per number of contributors in a mixture. The for loop will iterate through the empty matrix to fill each row and column with the false positive rate that is calculated in line 33.

Line 41-54: The hdr_FPR_matrix get restructured into a data frame

- Line 43-46: columns are labeled using the possible numbers of contributors per mixture and
- Line 51-54: the AEH of each group are added as a new column in the dataframe

Lines 57-92: This section of the script builds a separate data frame for the 8 groups that have been selected to represent the range of genetic diversity of the groups involved in this analysis.

- The **hdr_combined_df** is restructured in the same manner as the **hdr_FPR_DF**

Line 94-112: This section builds the violin distribution plot of false positive rates (y-axis) across the number of contributors in a mixture (x-axis). The 8 groups selected to represent the range of genetic diversity of all groups involved in this analysis are displayed as line graphs overlaying the violin plot. The y-axis scale of this graph is log scaled.

- Line 108: the function **scale_y_continuous(trans = “log10”)** is what changes the y-axis into log scale