Factors Influencing Variation in Likelihood-Ratios of DNA Tests

Executive Summary: The likelihood-ratio (LR) is a numerical metric that describes the likelihood that a DNA sample comes from an individual. A high likelihood-ratio implies that the target contributor is more likely to be responsible for the sample. This report attempts to discern the significant factors which effect the likelihood-ratio (LR) of DNA tests. A multivariate linear regression (MLR) was fitted, and its conclusions were analysed. According to the MLR model, three factors statistically significantly affect the likelihood-ratio: the totalDNA, the log amount of DNA from target contributor, and the ID of the target contributor.

Introduction

DNA evidence is important for crime scene investigations, for they allow the positive identification of suspects. Most DNA evidence is generated through the analysis of items which were on the crime scene. In analysing these items, an investigator usually tries to find biological traces of DNA. Once these biological traces of DNA are found, the investigator then visualizes the DNA traces. A list of suspects and their DNA profiles are then provided. The investigator then matches the visualized DNA traces with the DNA profiles of the suspect and computes a likelihood-ratio (LR) score based on a mathematical model. The higher the likelihood ratio score, the more likely the DNA sample comes from the suspect.

Given the importance of the likelihood score for a criminal investigation, it is somewhat surprising that the parameters that determine the likelihood-ratio have not been studied in any systematic fashion. These variations are important: if there is high variance in the likelihood scores, then there is a higher probability that an important piece of evidence will be misidentified. Misidentification of evidence can be very damaging for the integrity of the criminal investigation and may lead to unsupported convictions. In this report, data from a forensics consulting firm is used to investigate the possible factors leading to variation in the likelihood-ratio. A direct analysis of the variance of the likelihood-ratio is conducted using a multivariate linear regression (MLR). This report will be divided into three parts. The first part (***Data Cleaning and Exploratory Data Analysis***) discusses steps taken to prepare the data for analysis. The second part (***The Multivariate Linear Regression Model***) discusses issues relating to the modelling of the data. The third part (***Conclusion and Discussion***) discusses some potential implications and limitations of our model.

Data Cleaning and Exploratory Data Analysis

Raw Data Description

The raw data used in this study consisted of 33 columns[1] and 7933 rows. The 33 columns can be categorized into roughly seven subsets of columns:

- Sample and Replicate Columns: The sample (column 1) and replicate (column 2) columns describe the sampling methodology. The sample column represents the unique ID of each DNA sample. The replicate column represents trial replicate responsible for the likelihood-ratio measurement.
- Likelihood-Ratio Columns: These columns consist of the likelihood ratio column (column 3), as well as the list of loci components of the ILR (column 18-33). The sum of columns 18-33, or the loci components, should equal to the likelihood ratio column.
- Procedure Column: The Procedure column (column 5) is a factor and gives us a description of what mathematical model was used. There were two mathematical models used in our current analysis: EFM and likeLTD
- Stutter Column: In many cases, the DNA evidence received may have some noise in them. Most mathematical models have a way in which the noise is automatically filtered out. Stutter (column 6) is a binary column, consisting of the values 'Present' or 'Absent.' If the value of the Stutter column is 'Present,' then the calculations for Likelihood-Ratio included automated filter for stutter. The inverse holds true if the value of Stutter column is 'Absent.'
- Contributor Columns: The contributor columns include the number of contributors (column 6) and the target contributor (column 7). The number of contributors (NC) is a discrete variable that describes how many people contributed to the DNA sample. The target contributor (TC) is a factor that indicates the individual for whom the likelihood ratio is calculated. In other words, TC is the suspected 'owner' of the DNA sample. The individual is identified by a letter ranging from A to H.
- DNA Column: All observations come with two columns of DNA data. (Columns 8-9) The first column, TotDNA, describes the total DNA (column 8) in pg. The second column, amtDNA, describes the amount of DNA from the target contributor (column 9).
- Contributor Detail Columns: The contributor detail columns (10-18) give details as to which individual contributed to the DNA sample, and in what proportion. The first four columns (C1-C4) identify the person responsible for the sample. The latter four columns (D1-D4) represent the ratio of the DNA mixture. For example, a DNA sample might be the result of a one-to-one mixture of the DNA of two individuals A and B. If this were so, then only C1 and C2 would be filled with the identifying letter of the individuals responsible for the DNA mixture. C3 and C4 would be filled NANs. Similarly, since the mixture is one-to-one, then D1 and D2 would be filled with one, and D3 and D4 would be NANs. An example of the contributor detail column can be seen in Fig. 1.

	Sample	Rep	lLR	Eff	Pr	St	NC	TC	TotDNA	amtDNA	C1	C2	C3	C4	D1	D2	D3
1	1	1	10.665196	0.607	EFM	Absent	2	C	500	250	C	E	<NA>	<NA>	1	1	NA

Figure 1: Example of Contributor Detail Columns for a One-One Mixture between Individuals C and E

[1] We exclude Eff (column 4) from our description, as we do not use column 4 in any of our preparation or analysis.

Data Cleaning Steps

Four main steps were taken to process the raw data:

1. NA values in VWA were dropped
2. Experiment ID were generated using dense rank
3. Outliers within the loci values were smoothed using median polish, and new likelihood-ratios were calculated from these smoothed loci values
4. The median of each experimental ID likelihood-ratios was calculated

Dropping NA Values in VWA

One possible source of noise in the dataset is the NAN VWA values. The VWA is one of the 15 loci components that sum up to make the likelihood-ratio score. In the raw dataset, around 225 observations had NAN VWA values. There are two ways in which we could have dealt with these NAN values. We could have replaced such NAN values with 0. Nevertheless, this would have been illogical, as we noticed that there were no other observations in which any of the loci values were 0. Thus, we decided to delete the 225 observations where the VWA loci values were NANs. This reduced the number of usable observations from 7933 to 7708.

Experimental ID Generation

The aim of this of study is to study what factors influence intra-/inter-experimental variation. To do this, we must know which experiment each likelihood-ratio experiment was taken from. Nevertheless, the raw dataset did not have a unique value to identify each experiment. We fixed this issue by applying dense rank to the Procedure, Stutter, Contributor, DNA, and Contributor Detail columns. The dense rank algorithm which partitions the dataset by matching the specified column values. Performing dense rank on the above-mentioned columns yielded an ID for each unique experiment. This vector was added as a new column to the dataset. We called this new column experimental ID, or e_id for short.

Smoothing Outliers

Another source of noise in the dataset is outliers. Outliers are usually unintended anomalies that result from the data generation process. As such, outliers may dilute the strength of whatever trends might be discovered in the analysis. Thus, it is important to correct for outliers. We decided to correct for outliers in the loci values instead of the processed likelihood-ratio values. We did this because there might be a scenario in which the processed likelihood-ratio could have been a result of outlier observations within the loci values.

There are many methods in which we could use to correct outliers. The most common method is to use the 1.5 IQR rule. The 1.5 IQR rule excludes data points which are below the lower and

upper IQR bound (defined as lower_bound = Median – 1.5 IQR and upper_bound = Median + 1.5 IQR respectively). Nevertheless, we did not follow this method for reasons of preserving data observations. We reasoned that, under the 1.5 IQR rule, we would have to drop observations where only one loci value was an outlier. Thus, this may mean that we drop a significant amount of observation because of only one error.

We instead decided to use median polish to correct for the outliers. Our decision to use median polish is since, at this point in the investigation, we do not know the underlying distribution of the likelihood-ratios. Median polish is advantageous in this instance, for it does not impose many assumptions on the data. The algorithm for the median polish outlier algorithm is described in Figure 2.

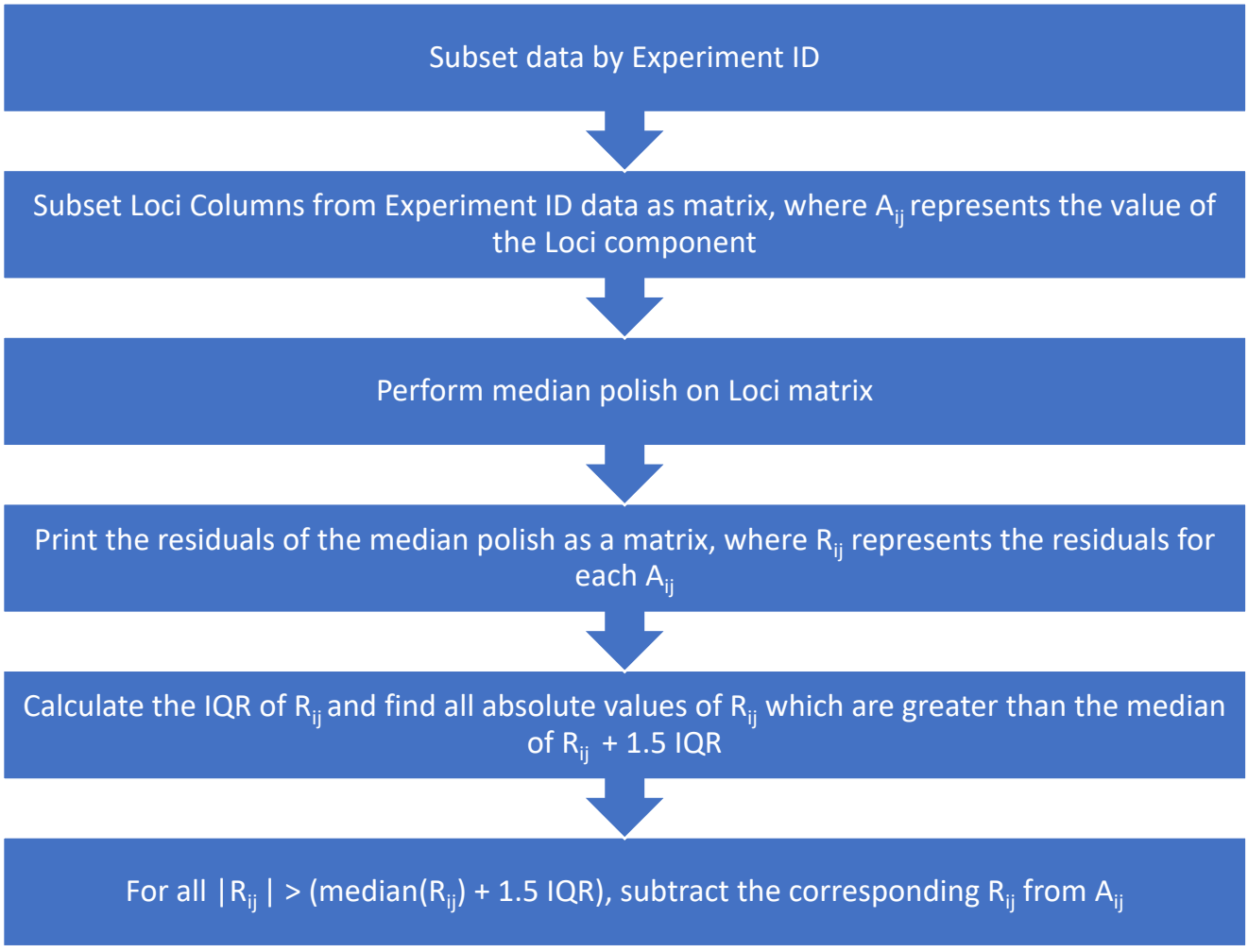

Figure 2: Median Polish Outlier Algorithm

We repeated the Median Polish Outlier Algorithm for all experimental IDs, and then summed the new loci values to create a new variable called smoothed ILR. A diagram of the smoothed and unsmoothed LR values can be found in Figure 3. The diagram shows that the median polish outlier algorithm preserved the shape of the distribution of the original LR values. The changes occurred most in the higher and lower bins of the distribution. It seems that the original

distribution had a higher frequency of likelihood-ratio scores in the 0-5 bins, whilst the smoothed likelihood-ratios had higher frequencies in the 10-20 bins.

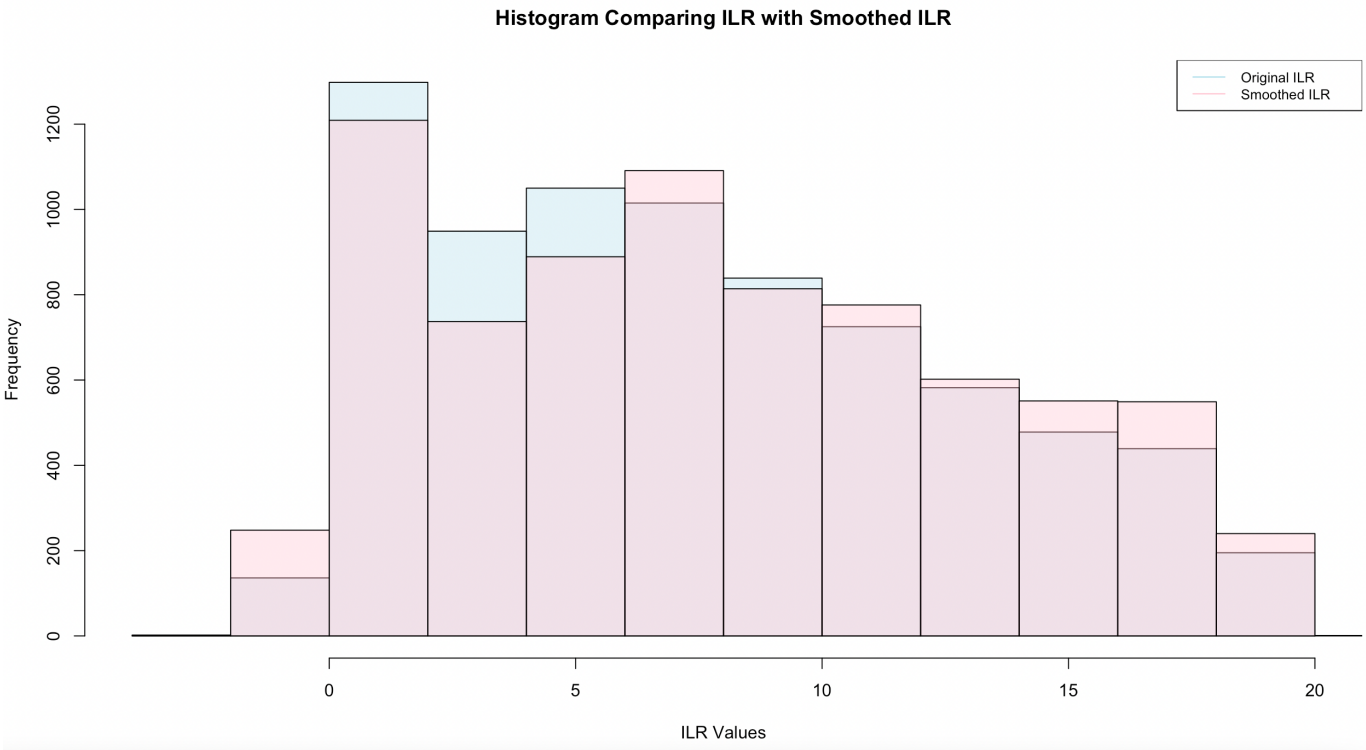

Figure 3: Histogram Comparison between Smoothed and Original LR

Calculating Median for Experimental ID

For reasons which will eventually become clear later, we aim to fit a multivariate linear regression model to this data. A multivariate linear regression model imposes strict assumptions onto the data. One of these assumptions is that the prediction errors must be independent and normally distributed. This raises problems for data with replication. Theoretically, say we have a response variable vector $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)$ that represents the likelihood-ratio for an experiment with *n* replications. Say we fit a linear regression model which produces the estimation vector $\widehat{\boldsymbol{Y}} = (\hat{Y}_1, \hat{Y}_2, \dots \hat{Y}_n)$. We then define the error term as $\boldsymbol{E} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$**.** If the errors were truly independent, then we should expect that the correlation of two partitions of the error terms to be 0 i.e. $\rho(E_i E_j) = 0$. Nevertheless, in cases of replicates, $\rho(E_i E_j)$ cannot theoretically be 0. This is because each replicate data point is produced using the same exact conditions. Because of this, we would expect the sources of errors in each replicate to be the same. The errors would then be correlated with each other.

To pre-emptively correct for the issues of correlated errors, we decided to find the median outlier-correct likelihood ratio of each experimental unit. We used the median, and not the mean, because the median imposes fewer restrictions on the distribution of the data. After completing this, the number of observations were reduced to 1,548 observations from an original of 7,708.

Exploratory Data Analysis (EDA)

The first step that we performed in doing EDA was to plot a histogram of the response variable i.e. the smoothed median ILR. The first four moments of this distribution (mean, variance, skewness, kurtosis) and the median of the distribution were also calculated. The results of this computation can be seen in Figure 4.

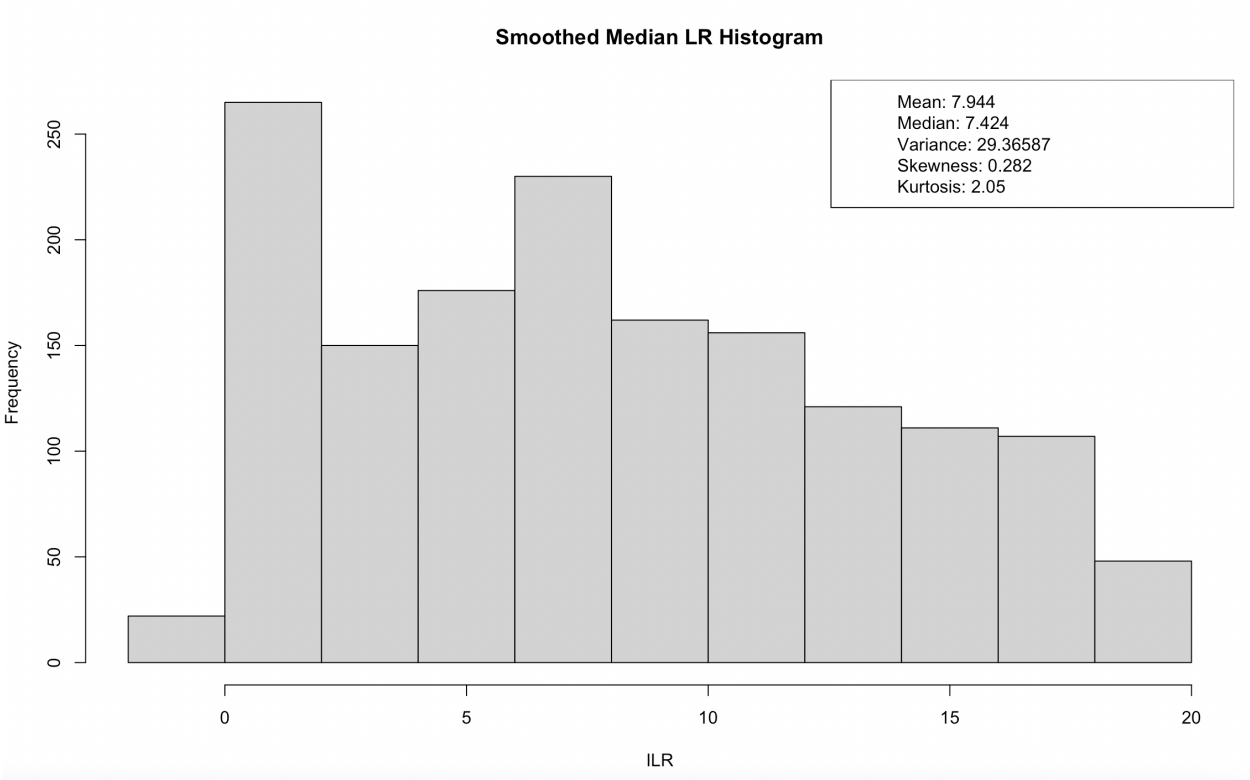

Figure 4: Smoothed Median LR

Visually, the distribution of the smoothed median ILR seems to have two peaks, from 0-2.5pg and 5-7pg. The fact that the mean is more than the median, and that the skewness is more than 0, also indicates that the overall distribution itself is slightly right skewed. The kurtosis of 2.05 indicates that the outliers of the distribution of the smoothed ILR are less influential when compared to the kurtosis of a normal distribution of 3.

The second step in the EDA process was to find the variables which account for some of the variation seen in the LR data. We grouped our variables into two sets: the first set of variables consist of continuous variables. In our case, we could only find one continuous variable, which is amtDNA. The correlation plot, as well as the fitted line, of amtDNA and LR is reproduced in Figure 5.

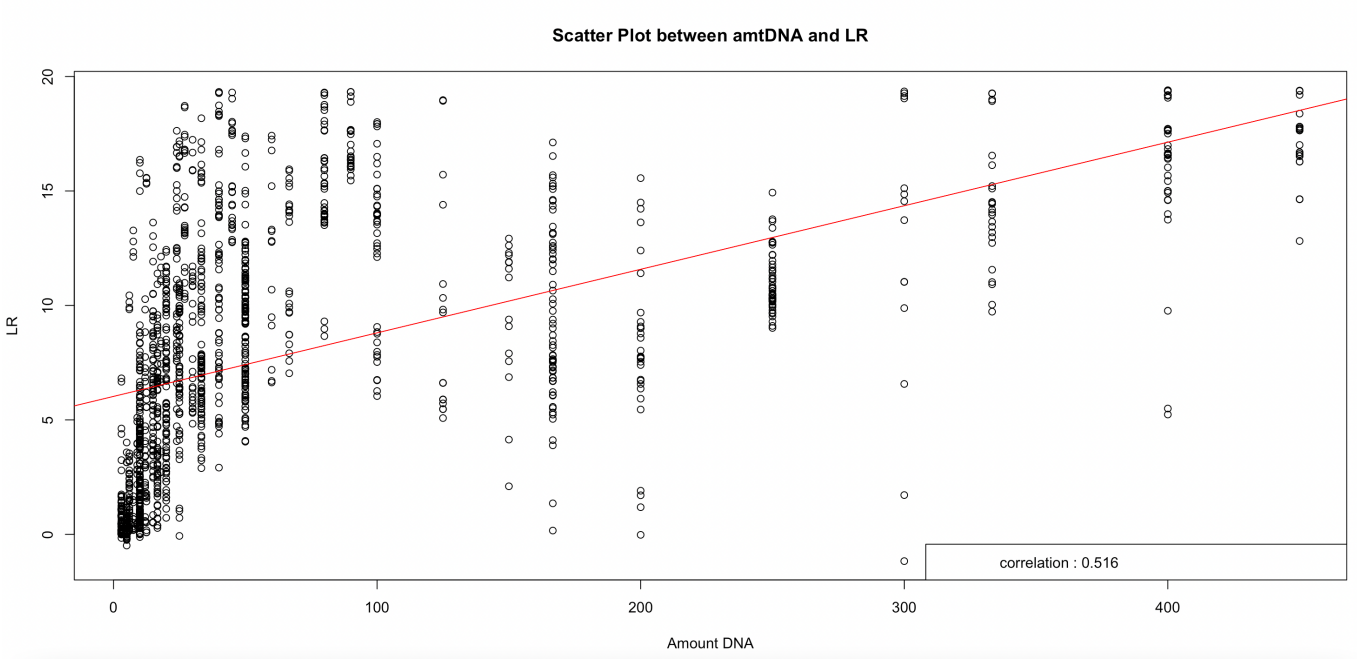

Figure 5: Scatterplot between amtDNA and LR

In general, the LR seems to be increasing as the amtDNA increases. This is confirmed by the simple regression line in red, as well as the positive correlation. Nevertheless, the positive correlation seems to not tell us the whole story, for we see many exceptions to this rule. For example, the regression line seems to be quite unable to predict the LR of 350 microgram DNA samples. The regression line, in fact, seems to be better able to predict LRs for samples with low amtDNA as opposed to high amtDNA.

We also produced boxplots to investigate group means of the various factor variables we have. The boxplot for each factor variable is reproduced in Figure 6, 7 and 8.

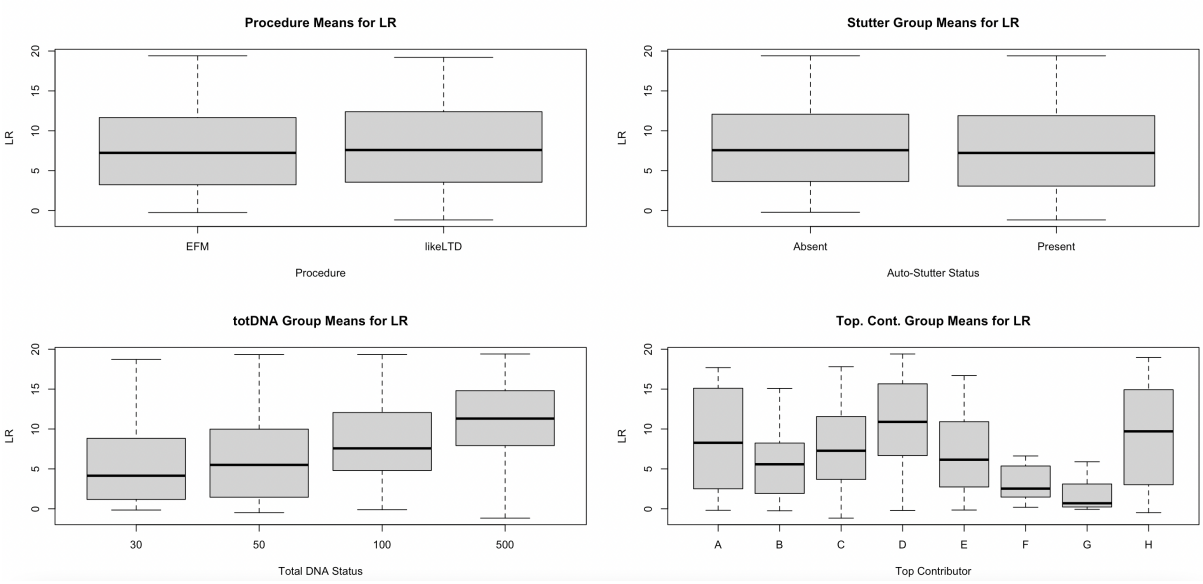

Figure 6: Boxplots for Factor Variable Group Means I

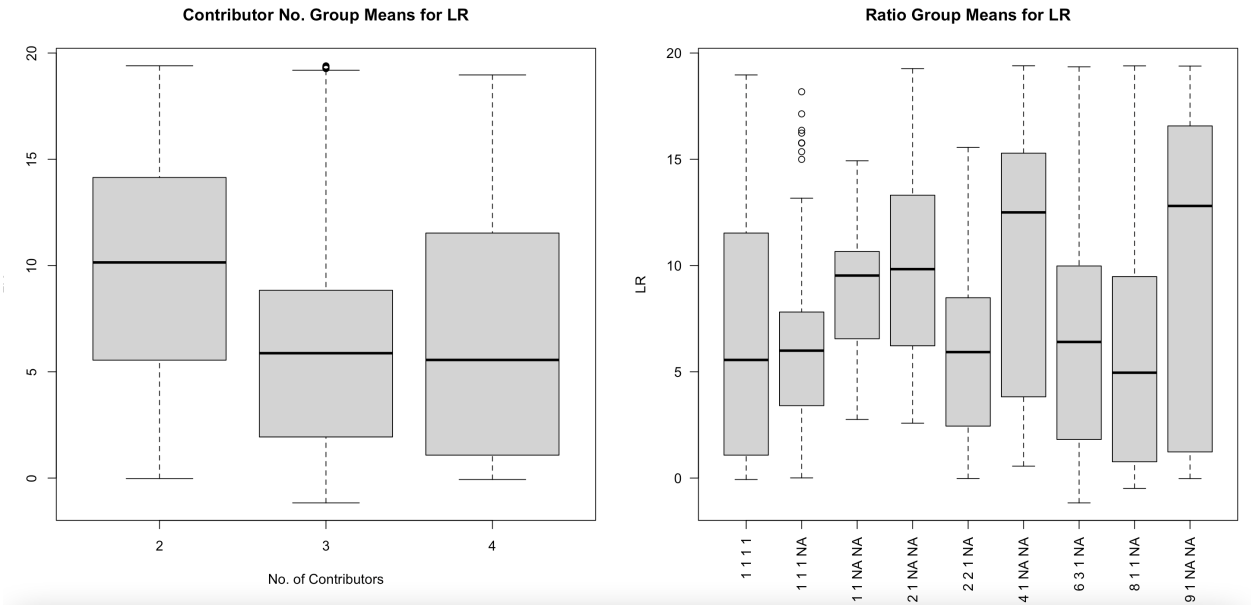

Figure 7: Boxplot for Factor Means, II

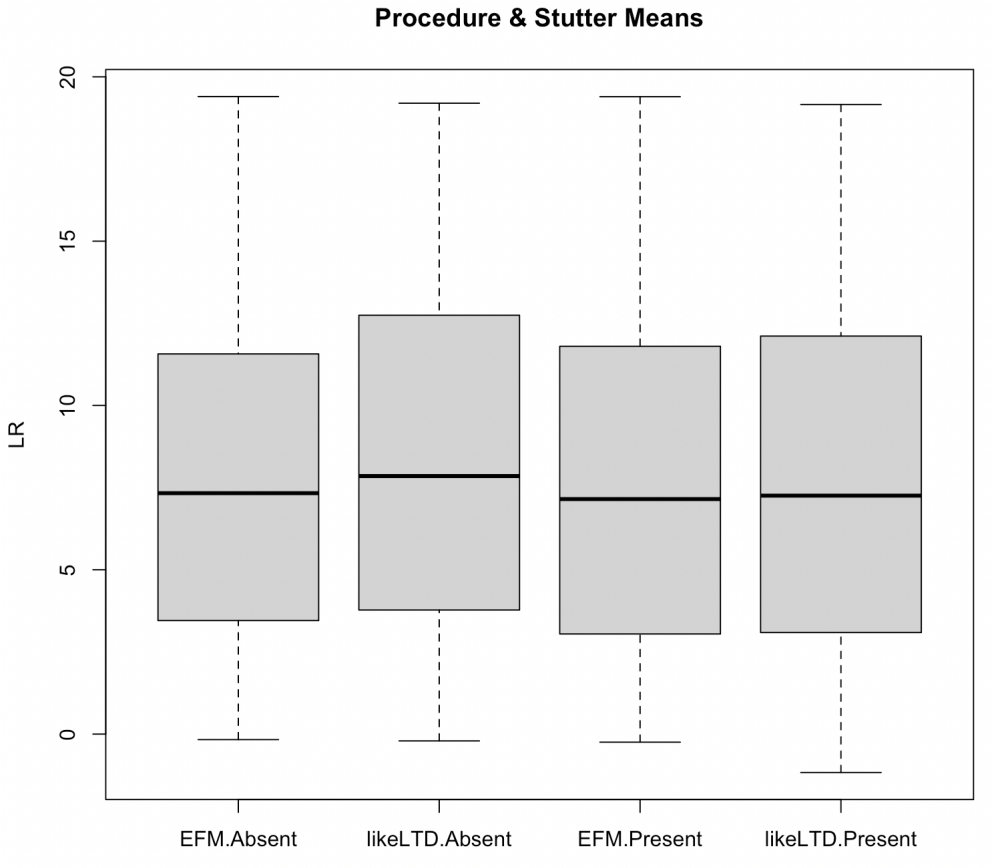

Figure 8: Boxplot for Factor Means, III

From Fig. 6, it seems that *Procedure* and *Stutter* did not produce any noticeable differences in LR means. However, total DNA and top contributor seemed to produce noticeable differences in LR. In the total DNA case, it seems that higher total DNA led to a higher LR. The intuition for this seems to be unclear, for total DNA does not always assume that the target contributor's DNA is more represented. There also seems to be noticeable differences in the mean LR of the top contributor whisker plot. From the plot, we can surmise that individual H has the highest LR rate of all the individuals. The reason for this, again, is unclear: it might be a result of unbalanced sampling. It might also be that the procedures which were used to compute the LR score responded more sensitively to individual H's DNA make-up.

From Fig. 7, it seems that the number of contributors and ratio all have noticeable differences in their group LR means. In terms of the number of contributors, it seems that mixtures with two contributors have a higher LR than mixtures with three or four contributors. This is expected, as mixtures with two people are more homogenous. It is interesting to note, however, that the interquartile range of mixtures with three people is larger than the interquartile range of mixtures with two people. This is unexpected, since we would expect more dilution in a mixture of four people than with three. Another factor which had significant differences in group means is ratio. It seems that mixtures with a 9:1:0:0 and 4:1:0:0 mixture ratio have the highest LR means. There is, again, no *a priori* reason for why this might be so.

In Fig. 8, we tested a possible interaction between procedure and stutter. The reasoning for testing such an interaction is that some procedures might have differing ways of removing stutter. Nevertheless, from Fig. 8, we see that the differences between group LR means for the interaction between stutter and procedure is not high enough to be noticed.

From our EDA, we assume that the variables that will be most responsible for any variation within the results will be totDNA, target contributor, number of contributors, amtDNA, and ratio. Some other factors that may be significant, but do not exhibit noticeable group mean differences include procedure, stutter and the interaction between procedure and stutter.

The Multiple Linear Regression (MLR) Model

Model Justification

The aim of this study is to ascertain which parameters best predict the LR of a DNA sample. As such, the task at hand is not merely a question of creating a model which is able to *predict* future LR from a given set of independent variables. Rather, the task at hand is twofold. Firstly, we must create a model which is able to predict future LR from an unseen set of independent variables. Secondly, and perhaps more importantly for this study, we must create a model which would be readily interpretable. In statistical literature, there exists many models which could be implemented to predict future LRs from new data. Nevertheless, the MLR is a model which balances interpretability and prediction well. As such, we chose to implement an MLR model to investigate the parameters which influence the LR.

MLR Data Constraints

The aim of MLR is to predict a dependent variable given a series of independent variables. In doing so, it imposes a set of strict considerations on the data, as stated before. In this section, we present two constraints with the dataset for performing MLR. The first constraint is the issue of multicollinearity. In a MLR model, all independent variables should contain new information that is not reflected in another variable. There are some cases in our dataset where this assumption may be violated. For example, say we put both the number of contributors and ratio as explanatory variables for the MLR model. This is a case of multicollinearity, for the ratio already has information about the number of contributors embedded into it. Such relationships exist between many of the variables. As such, building an MLR model in this situation requires a deliberate parameter choice. The second constraint is a more basic issue of model selection. For our model to *say* something, we must assume that the way in which all variables were generated are not dependent on each other in some way. This then excludes the usage of loci values as explanatory variables, as the sum of loci values results in the LR value. Say we were to put loci values as explanatory variables. Then the conclusions of our model would be trivial, for it would simply say that there is a strong relationship between the sum of numbers and the individual numbers themselves. This second constraint also precludes the inclusion of the eight Contributor Detail columns.

These two constraints imply that the amount of potential explanatory variables is reduced by at least 24 columns (15 loci value columns, eight contributor detail columns, and either ratio or number of contributors). This issue is then exacerbated by the fact that one cannot use replicate, or sample as dependent variables. From the original 33 columns, it follows that only six columns (excluding the dependent variable, LR) can be used as explanatory variables. These columns are Procedure, Stutter, Number of Contributors *OR* Ratio, Total DNA, and amount DNA.

Model Estimation

We estimated three models, whose form is described below:

1. $\hat{y} = \beta_0 + \widehat{\beta_1}\,\text{Procedure} + \widehat{\beta_2}\,Stutter + \widehat{\beta_3}TotDNA + \widehat{\beta_4}\log(amtDNA) + \widehat{\beta_5}totDNA + \widehat{\beta_6}TopContributor + \varepsilon$
2. $\hat{y} = \beta_0 + \widehat{\beta_1}\,\text{Procedure} + \widehat{\beta_2}\,Stutter + \widehat{\beta_3}TotDNA + \widehat{\beta_4}\log(amtDNA) + \widehat{\beta_5}totDNA + \widehat{\beta_6}Ratio + \varepsilon$
3. $\hat{y} = \beta_0 + \widehat{\beta_1}\,\text{Procedure} + \widehat{\beta_2}\,Stutter + \widehat{\beta_3}TotDNA + \widehat{\beta_4}\log(amtDNA) + \widehat{\beta_5}totDNA + \widehat{\beta_6}NumberOfContributor + \varepsilon$

The models were evaluated on a two-step basis. Firstly, we evaluate whether the model passed the diagnostic tests. The two main diagnostic tests used were the residuals [2] vs. fitted plot and the theoretical quantile plot. The residual vs. fitted plot tests the quality of the prediction. The results of each estimation can be seen in Table 1, 2 and 3.

[2] The differences between the actual values of the LR and values predicted by the model

```
lm(formula = reg_aggr$median_ILR ~ reg_aggr$Pr + reg_aggr$St + 
    reg_aggr$TotDNA + log(reg_aggr$amtDNA) + reg_aggr$TC)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.0424  -2.0311  -0.0966   1.7315   9.7410 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -5.6316     0.4108 -13.709  < 2e-16 ***
reg_aggr$PrlikeLTD    0.1017     0.1516   0.671  0.50240    
reg_aggr$StPresent   -0.2105     0.1516  -1.388  0.16521    
reg_aggr$TotDNA50    -1.9162     0.2217  -8.645  < 2e-16 ***
reg_aggr$TotDNA100   -3.2985     0.2549 -12.939  < 2e-16 ***
reg_aggr$TotDNA500   -8.7275     0.3879 -22.498  < 2e-16 ***
log(reg_aggr$amtDNA)  5.2501     0.1155  45.456  < 2e-16 ***
reg_aggr$TCB         -1.0390     0.3382  -3.073  0.00216 ** 
reg_aggr$TCC         -2.6045     0.3107  -8.382  < 2e-16 ***
reg_aggr$TCD          1.7356     0.3118   5.566 3.07e-08 ***
reg_aggr$TCE         -1.9220     0.2983  -6.444 1.55e-10 ***
reg_aggr$TCF         -4.2421     0.7898  -5.371 9.04e-08 ***
reg_aggr$TCG         -5.5807     0.7898  -7.066 2.41e-12 ***
reg_aggr$TCH          3.8621     0.4381   8.815  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.976 on 1534 degrees of freedom
Multiple R-squared:  0.701,     Adjusted R-squared:  0.6984 
F-statistic: 276.6 on 13 and 1534 DF,  p-value: < 2.2e-16
```

Table 1: Results of the Estimation 1.)

```
Residuals:
     Min       1Q   Median       3Q      Max 
-15.3975  -2.2094  -0.3652   1.8606  12.8755 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)            -5.81563    0.51520 -11.288  < 2e-16 ***
reg_aggr$PrlikeLTD      0.16540    0.16510   1.002 0.316601    
reg_aggr$StPresent     -0.27420    0.16510  -1.661 0.096959 .  
reg_aggr$TotDNA50      -1.78742    0.24198  -7.387 2.46e-13 ***
reg_aggr$TotDNA100     -2.99491    0.27988 -10.701  < 2e-16 ***
reg_aggr$TotDNA500     -8.12663    0.43165 -18.827  < 2e-16 ***
log(reg_aggr$amtDNA)    4.99794    0.12901  38.741  < 2e-16 ***
reg_aggr$ratio1 1 1 NA -2.06233    0.47379  -4.353 1.43e-05 ***
reg_aggr$ratio1 1 NA NA -1.24620    0.47777  -2.608 0.009186 ** 
reg_aggr$ratio2 1 NA NA  0.01974    0.47518   0.042 0.966869    
reg_aggr$ratio2 2 1 NA  -1.80988    0.47256  -3.830 0.000133 ***
reg_aggr$ratio4 1 NA NA  1.50410    0.47199   3.187 0.001468 ** 
reg_aggr$ratio6 3 1 NA  -0.22769    0.47203  -0.482 0.629621    
reg_aggr$ratio8 1 1 NA   0.92850    0.47332   1.962 0.049982 *  
reg_aggr$ratio9 1 NA NA  2.38014    0.46867   5.079 4.27e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.243 on 1533 degrees of freedom
Multiple R-squared:  0.6451,    Adjusted R-squared:  0.6419 
F-statistic: 199.1 on 14 and 1533 DF,  p-value: < 2.2e-16
```

Table 2: Result of Estimation 2.)

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.3968  -2.4444  -0.3778   2.0512  11.7164

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -3.7584     0.4028  -9.332  < 2e-16 ***
reg_aggr$PrlikeLTD     0.1453     0.1762   0.825   0.4097
reg_aggr$StPresent    -0.2541     0.1762  -1.442   0.1495
reg_aggr$TotDNA50     -1.5049     0.2574  -5.846 6.13e-09 ***
reg_aggr$TotDNA100    -2.3290     0.2945  -7.909 4.93e-15 ***
reg_aggr$TotDNA500    -6.5506     0.4452 -14.713  < 2e-16 ***
log(reg_aggr$amtDNA)   4.4448     0.1312  33.878  < 2e-16 ***
reg_aggr$NC3          -1.7068     0.1887  -9.044  < 2e-16 ***
reg_aggr$NC4          -0.9477     0.4551  -2.082   0.0375 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.462 on 1539 degrees of freedom
Multiple R-squared:  0.594,     Adjusted R-squared:  0.5919
F-statistic: 281.5 on 8 and 1539 DF,  p-value: < 2.2e-16
```

Table 3: Result of Estimation 3.)

Ideally, the residuals v. fitted plot should be distributed equally around a mean of 0. This implies that our MLR model has adequately ‘explained away’ any trends present in the dataset, and that any errors are due to noise. The theoretical quantile plot (QQ plot) measures the distribution of the residuals against the theoretical quantile of a normal distribution. If both plot to a 45 degree line, then the residuals are normally distributed. A normally distributed residual ensures, through the Central Limit Theorem, that the model will predict the true values in the long run. The residuals v. fitted and QQ plots of each model are given in Fig. 8, Fig. 9, and Fig. 10.

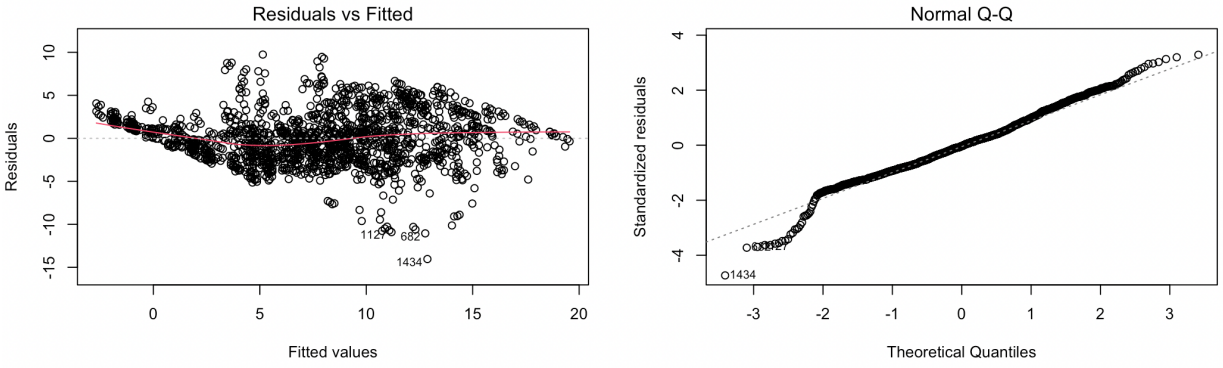

Figure 8: Diagnostic Plot for Model 1.)

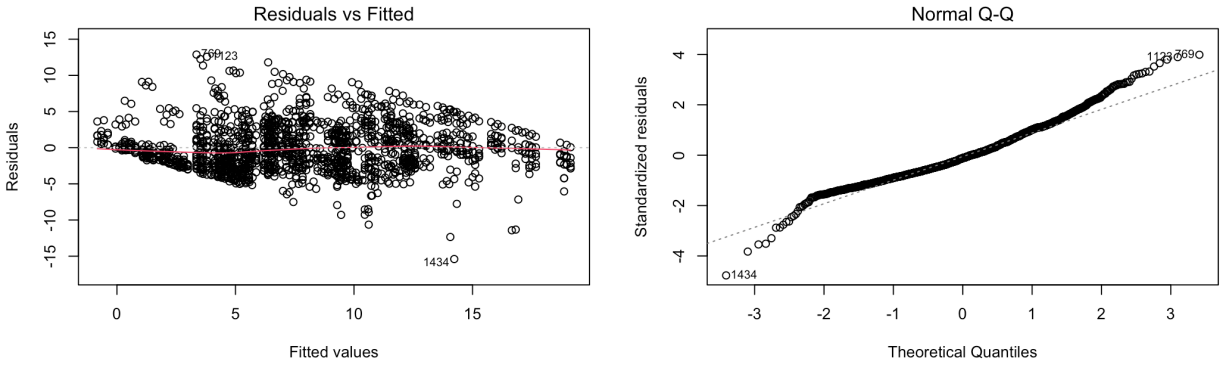

Figure 9: Diagnostic Plots for Model 2.)

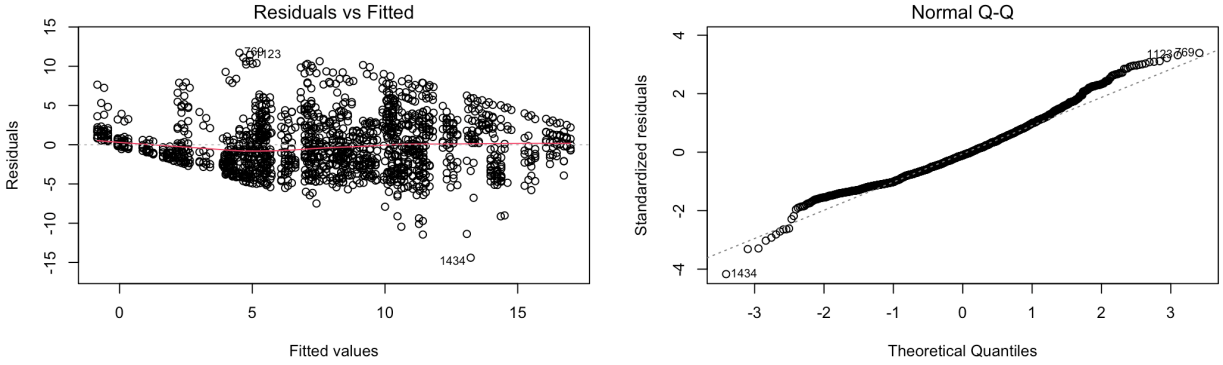

Figure 10: Diagnostic Plots for Model 3.)

The three diagnostic plots are not too noticeably different. In terms of each model's residuals v. fitted plots, all three models seem to have some sort of systematic error in predicting lower and higher values. Nevertheless, such systematic error does not detract from the fact that the errors of the residuals for all three models are basically normally distributed around a mean of 0. The same goes for the qq-plot: all three models have similar qq-plots, although it could be argued that the qq-plot for 1.) seems to be the most closely aligned to the 45-degree line. Nevertheless, there seems to be no noticeable difference, with the residuals of all three models having long tail ends. Because of the similarity in diagnostic plots, we decided to choose the model based on only the adjusted-R squared value. As such, 1.) seems to be the best choice, as it has the highest adjusted R-squared of all three models.

According to model 1.), the statistically significant parameters which effect the variation in the LR score include the totalDNA, the log(amtDNA), and the identity of the Target Contributor. In terms of totalDNA, it seems that DNA samples with 30 units has the highest LR. It is interesting to note that this number conflicts with the whisker plot in Fig. 6, which shows the group LR mean of totDNA samples with 500 units is higher than all other LR means. The unexpected sign in the presence of other controls indicates that the figure we saw in Fig. 6 was probably the

result of another confounding factor. We suspect that the confounding factor is amtDNA, as an increased in totalDNA will also lead to an increase in maximum amtDNA possible. Recall that amtDNA has a positive relationship with the LR in model 1.). Thus, we suspect that the positive LR that we see from the Fig. 6 is explained in the positive coefficient of LR. [3]However, this does not mean that the totalDNA has no explanatory power, as we can see from the ANOVA breakdown of model 1.).

```
Analysis of Variance Table

Response: reg_aggr$median_ILR
                       Df  Sum Sq Mean Sq   F value  Pr(>F)
reg_aggr$Pr             1     9.0     9.0    1.0200 0.31268
reg_aggr$St             1    28.1    28.1    3.1693 0.07523 .
reg_aggr$TotDNA         3  8202.0  2734.0  308.7248 < 2e-16 ***
log(reg_aggr$amtDNA)    1 17767.6 17767.6 2006.3172 < 2e-16 ***
reg_aggr$TC             7  5837.4   833.9   94.1660 < 2e-16 ***
Residuals            1534 13584.8     8.9
```

Table 4: ANOVA Breakdown of 1.)

From table 3 and 4, we can also conclude that the parameters which are not statistically significant include Procedure and Stutter. It is perhaps interesting to note that Procedure and Stutter are not significant in any of our estimations. This perhaps demonstrates that the differences of LR score between each LR calculation procedure is minimal. As such, in explaining the LR, it seems that the log(amtDNA), totDNA, and the Target Contributor (in that order) are the most significant parameters in explaining the variation in the data.

[3] This is confirmed when we re-estimate the model without log(amtDNA) in. The results for such a regression are shown below. As we suspected, the omission of the log(amtDNA) variable led to the totDNA variables to exhibit a different sign.

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.5069     0.4781  13.611  < 2e-16 ***
reg_aggr$PrlikeLTD  0.1336     0.2322   0.575   0.5651
reg_aggr$StPresent -0.2424     0.2322  -1.044   0.2967
reg_aggr$TotDNA50   0.7657     0.3272   2.340   0.0194 *
reg_aggr$TotDNA100  3.0225     0.3272   9.237  < 2e-16 ***
reg_aggr$TotDNA500  5.9725     0.3281  18.202  < 2e-16 ***
reg_aggr$TCB       -3.3497     0.5120  -6.543 8.21e-11 ***
reg_aggr$TCC       -1.0850     0.4731  -2.293   0.0220 *
reg_aggr$TCD        1.9067     0.4775   3.993 6.83e-05 ***
reg_aggr$TCE       -1.7822     0.4568  -3.902 9.96e-05 ***
reg_aggr$TCF       -5.7843     1.2085  -4.786 1.86e-06 ***
reg_aggr$TCG       -7.1230     1.2085  -5.894 4.62e-09 ***
reg_aggr$TCH        0.2834     0.6601   0.429   0.6678
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion and Discussion

Conclusion and Practical Implications

The three factors that explain the variation in the LR data is as follows: the log(amtDNA), the totDNA, and the Target Contributor. Factors which do not significantly explain the variation in the LR include Procedure and the Presence/Absence of Stutter. To expand in more detail: it seems that, *cetera paribus*, an increase in totDNA seems to have an inverse effect on the LR. On the other hand, *cetera paribus,* an increase in the log(amtDNA) seems to increase the LR. In terms of individuals, it seems that individuals H has the highest LR when compared to all other individuals. Out of the previous conclusion, only the conclusion relating to totDNA is inconsistent with our EDA from Fig. 6. Nevertheless, as noted in Footnote 3, the whisker plot in Fig. 6 is a result of not controlling for amtDNA. Once controlled for this, the sign of the totDNA variables is reversed.

There are two practical implications of these conclusions. Firstly, more totalDNA does not always correlate with a higher LR score. In fact, in some instances, more totalDNA may lead to a decreased LR score. Secondly, it seems that the biological make-up of person H is quite agreeable to the algorithms which calculate the LR score. As such, one can be surer of LR scores calculated by individuals with H's biological make-up than others.

Discussion

One major strength of this study is that it provides one of the first, highly interpretable and highly predictive model in which one could study variations in the LR measure. The adjusted R-squared of 0.6984 implies that out-of-sample predictions would be of a relatively high quality.

One major limitation of this study is the lack of explanatory variables. As noted in the MLR Data Considerations section, only 8 of the 33 data columns could be used in a linear regression setting. As such, some of our conclusions are trivial: for example, there was already a strong *a priori* reason to believe in a correlation between the amtDNA given by a target individual and the LR for that target individual. In the future, we would have liked more quantitative variables, such as frequency or height of alleles. Such quantitative variables would have allowed us to analyze the sources of LR variation more accurately. Another major limitation of this study is experimental design. More specifically, we do not if the experiment was designed to maximize the power of each hypothesis test. We did not know which factors we would have considered to be blocked, or what exact hypothesis to test. Because of this, our findings must be merely preliminary, as they explore what possible parameters *may* contribute to variation.