Calculating a summary metric for complex DNA samples

Howard Hu (Lead consultant)

August 13, 2022

1 Exclusive Summary

This report tries to find a single metric that will summarize the variation exhibited by one replicate set to demonstrate how they differ from each other. Since the standard deviation can measure the variation in a group of samples, we choose log standard deviation as our single metric for this data set. For determining the primary contributors to the log standard deviation, we build a linear regression by considering the log standard deviation of total log-likelihood ratios as our response variables and total DNA and dose ratio as our predictors. However, we do not find any statistically significant predictors except the interception. The "Ratio=0.66" has the smallest p-value, which is 0.066. It indicates the "Ratio=0.66" is the most influential factor for the log standard deviation of total log-likelihood ratios

2 Introduction

2.1 General Background

This report aims to analyze Short tandem repeat (STR) length polymorphisms of subjects' DNA differentiate between two people. A sample may contain a minimal amount of DNA and DNA from multiple people, which makes identification difficult. We use likelihood ratios to examine the likelihood that a subject's DNA is present in a sample based on mathematical models. As we discussed with our client, Dr. Keith Inman, the numerator of likelihood ratios is the probability of evidence given subject is present. The denominator of likelihood ratios is the probability of evidence given subject is not present. In the given data set, there is another method called European Forensic Mixtures (Euroformix/EFM) to calculate the probability of each locus. However, we focus on the results of likelihood ratios in this report due to the discussion with our client.

2.2 Objectives

The main objective of our client is to determine a single metric that will summarize the variation exhibited by one replicate set to demonstrate how they differ from each other. The secondary objective is to determine the primary contributors to the single metric. To be more specific, we are trying to find a mathematical statistic to compare the variation in different locus.

2.3 Exploratory Data Analysis

Our data set is given by Dr. Keith Inman from California State University East Bay. We received two data sets called AllResultsTbl and AllReults Table. After discussing with our client, we decided to focus on the AllResultsTbl data set. There are 240 samples in the AllResultsTbl data set, and every five samples form a replicate set. It means we have 48 different replicate sets. For the replicate set, we have 15 different pre-locus likelihood ratios. By adding those 15 different pre-locus likelihood ratios, we have a sum of each locus log-likelihood ratio called "TotlogLR." Besides, we record the number of contributors in the mixture as "NumCount", and the total ng of DNA between both contributors as "TotDNA." Since we are only concerned with 2 person mixtures, we use C1 to represent contributor 1 and C2 to represent contributor 2. In addition, we use dose 1 and dose 2 to describe the C1 to

C2 mixture ratio of DNA. For visualizing the relationship between the mixture ratio of DNA with TotDNA, we create a new variable called "Ratio" by the following equations.

$$Ratio_i = \frac{\text{Target Contributor's Dose Ratio}_i}{\text{Target Contributor's Dose Ratio}_i + \text{Other Contributor's Dose Ratio}_i}$$

$Ratio_i$: the ratio of mixture DNA for ith replicate set

For sample 13, we have missing values for 15 loci. Therefore, we use the mean log-likelihood ratio from the same replicate set to replace those missing values. Moreover, we removed any observations with "StFilt=on" and "Progrm=EFM." Then we have 80 observations for "TCon=A" (TCon means the true contributor). Then we can use box plots to see the variation of those 80 observations in 15 loci.

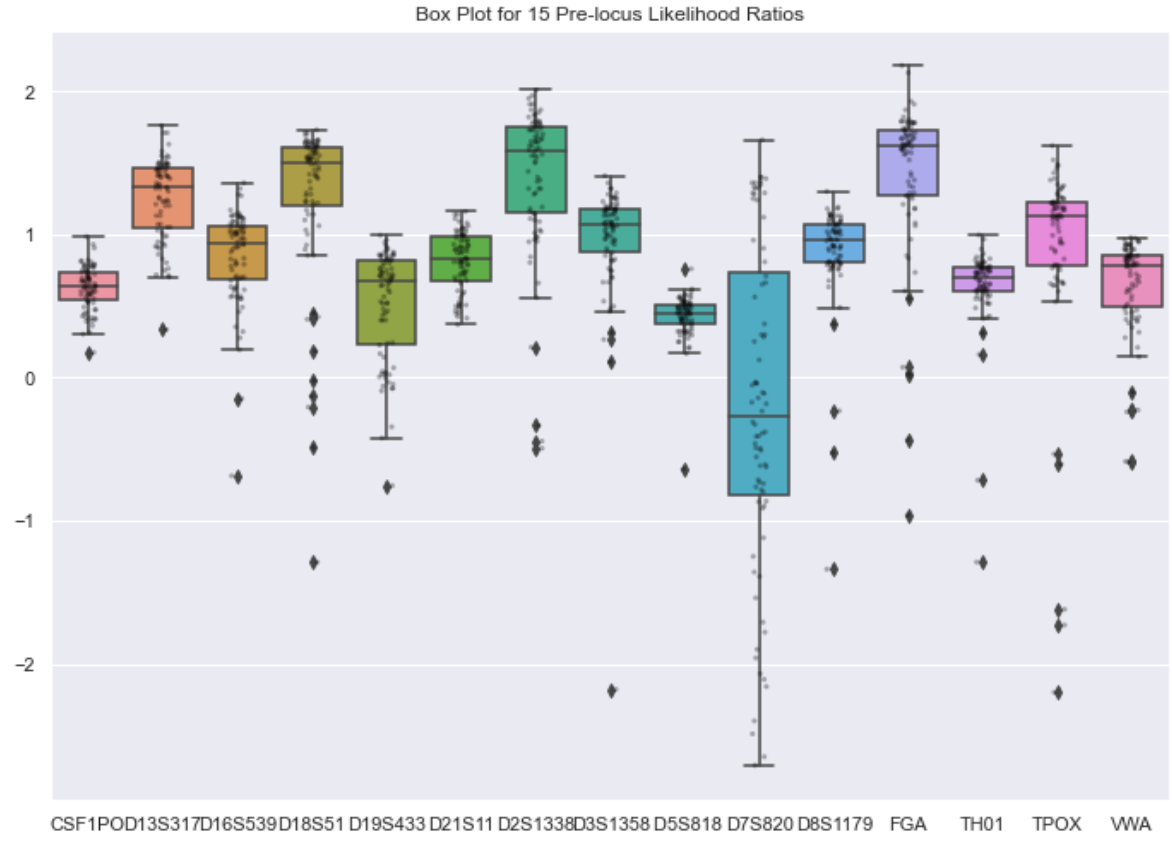

Figure 1: Box Plot for 15 Pre-locus Likelihood Ratios

From figure 1, we find out there are a lot of outliers for locus D18S51. It indicates that there is a considerable variation in locus D18S51. We also find that the maximum of the box plot of locus D7S820 is much larger than the minimum. It implies there is a significant variation in locus D7S820. For quantifying the amount of variation of each locus, we can calculate the log standard deviation of each pre-locus likelihood ratio and the total log-likelihood ratios.

$$\log \sigma_j = \log \sqrt{\frac{\sum (x_i - \mu)}{N}}$$

$\log \sigma$: the log sample standard deviation of jth replicate set
x_i : the likelihood ratio for ith locus in jth replicate set
μ : the sample mean of the likelihood ratio for ith locus
N : the size of the replicate set which is 5

We use the log standard deviation as our single metric to summarize the variation in replicate sets. For each replicate set, we have five samples. Then we have 16 different replicate sets, which create a 16×21 data set. We can use box plots to plot the log standard deviation of those 16 different replicate sets in 15 loci.

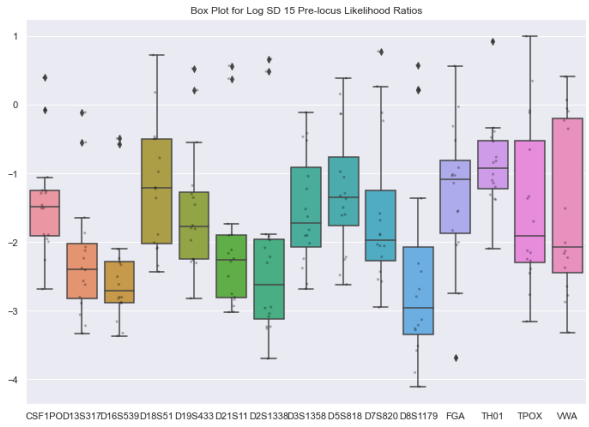

Figure 2: Box Plot for Log SD 15 Pre-locus Likelihood Ratios

Figure 3: Line Plot for Log SD of Pre-locus D18S51 TPOX VWA Likelihood Ratios

From figure 2, we find that the maximum box plot of locus D18S51 is much larger than the minimum. It shows a large variation in the log standard deviation of locus D18S51. Similarly, we assume there is a large variation for TPOX and VWA. Figure 3 shows that locus D18S51, TPOX, and VWA are zigzagging, showing a large variation in those loci. Since we already know the total log-likelihood ratio is a summation of 15 pre-locus likelihood ratios, we want to know how the total DNA and dose ratio affects the total log-likelihood ratio.

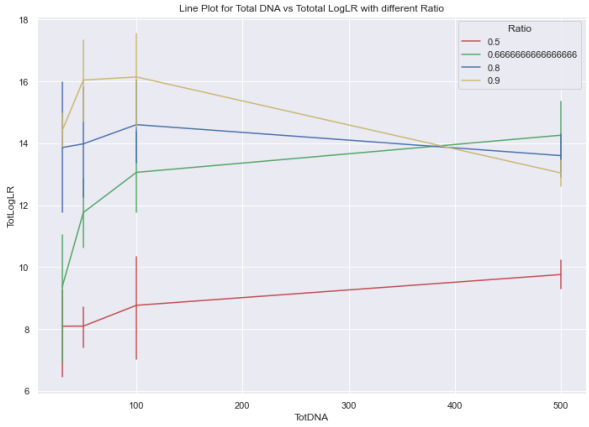

Figure 4: Line Plot for Total DNA vs Tototal LogLR with different Ratio

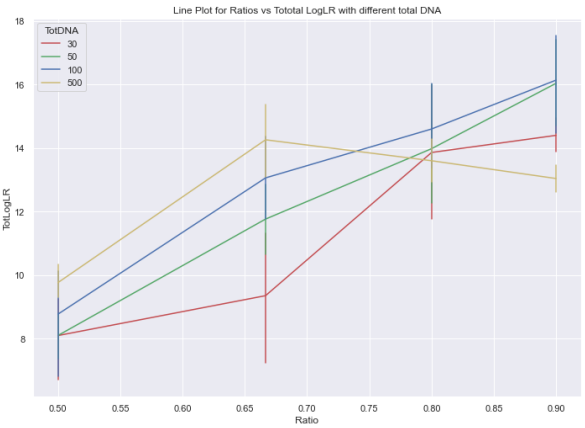

Figure 5: Line Plot for Ratios vs Tototal LogLR with different total DNA

Figure 4 shows that when we increase the total DNA amount, the log standard deviation of total log-likelihood ratios will sightly increase except "Ratio=0.9" and "Ratio=0.8." Figure 5 shows that when we increase the target contributor's dose ratio, the total log-likelihood ratios will vastly increase except "TotDNA=500."

3 Approach

We calculate the log standard deviation as our single metric that will summarize the variation exhibited by one replicate set to demonstrate how they differ from each other. Then we can build our linear

regression model by factorizing total DNA and Ratio.

$$\log \sigma_{TotLogLR} = \text{interception} + \text{Total DNA} + \text{Ratio} + e$$

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-2.2825	0.459	-4.971	0.001	-3.321	-1.244
C(TotDNA)[T.50]	0.6978	0.491	1.421	0.189	-0.413	1.808
C(TotDNA)[T.100]	0.5957	0.491	1.214	0.256	-0.515	1.706
C(TotDNA)[T.500]	0.7993	0.491	1.628	0.138	-0.311	1.910
C(Ratio)[T.0.6666666666666666]	-1.0282	0.491	-2.095	0.066	-2.139	0.082
C(Ratio)[T.0.8]	-0.2268	0.491	-0.462	0.655	-1.337	0.884
C(Ratio)[T.0.9]	-0.6333	0.491	-1.290	0.229	-1.744	0.477

Omnibus:	0.068	**Durbin-Watson:**	2.479
Prob(Omnibus):	0.967	**Jarque-Bera (JB):**	0.152
Skew:	-0.110	**Prob(JB):**	0.927
Kurtosis:	2.576	**Cond. No.**	5.57

From the previous table, we do not find any p-value that is smaller than the statistically significant level of 0.05 except the p-value of interception. However, "Ratio=0.66"(Target Contributor's Dose Ratio/Sum of Dose Ratio= 2/3) has the lowest p-value with 0.066. It means "Ratio=0.66" might be the most influential factor for the log standard deviation of total log-likelihood ratios.

4 Conclusion

For summarizing the variation in different replicate sets, we use log standard deviation as our single metric. After building a linear regression model, we do not find any statistically significant predictors. However, the p-value of "Ratio=0.66" is 0.066, the closest value for a statistically significant level of 0.05. It means "Ratio=0.66" might be the most influential factor for the log standard deviation of total log-likelihood ratios.

5 Appendix