Identifying the Performance of Ancient DNA Sequencing Methods.
Final Report

Yukta Sanjiv Chavan

a1873167

August 09, 2024

Report submitted for **Data Science Research Project Part B** at the School of Mathematical Sciences, University of Adelaide

Project Area: **Statistical Data Analysis**

Project Supervisor: Dr. Indu Bala

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

Abstract

The project final report includes the study on Identifying the Performance of Ancient DNA sequencing methodologies, specifically comparing the Twist Bioscience and 1240k capture assays. The main objective is to determine the best effective method for testing tainted and deteriorated ancient DNA material. With the EAGER dataset, which includes variables like single nucleotide polymorphisms (SNPs), input reads, mapped reads, and percentages of endogenous DNA, we employ advanced data processing techniques. These include logistic and linear regression, mixed effects models, and beta regression. The paper evaluates the performance of different sequencing methods and provides recommendations for improving the accuracy and consistency of ancient DNA analysis.

1 Introduction

The study of ancient DNA (aDNA), offers vital information on the origins and development of human societies. It makes it possible for scientists to investigate genetic variety, migration trends, and how prehistoric individuals adapted to their surroundings. Because aDNA is prone to contamination and degradation over time, analysis of this material can be particularly difficult. Reliable data extraction from these old samples requires the use of advanced sequencing techniques [1].

The quality and accuracy of data obtained in genetic research are greatly impacted by the method of DNA sequencing that is selected [2]. The Twist Bioscience capture assay and the 1240k capture assay are two popular techniques for analyzing ancient DNA. The Twist test is a more recent technique that promises to increase capture efficiency and lower contamination, whereas the 1240k assay has been extensively utilized to capture a high number of single nucleotide polymorphisms (SNPs) [3].

The purpose of this experiment is to assess how well these two sequencing techniques function while investigating ancient DNA. We apply multiple statistical models to the EAGER dataset, which contains variables like endogenous DNA percentages, mapped reads, SNPs, and input reads, to evaluate the efficacy of each approach. The main goals are to:

- Compare the effectiveness and precision of the Twist and 1240k tests in identifying ancient DNA.
- Use mixed effects models, beta regression, linear and logistic regression,

and data analysis to examine the influence of various factors on sequencing performance.
- Make suggestions for the best sequencing technique to be used in next ancient DNA studies.

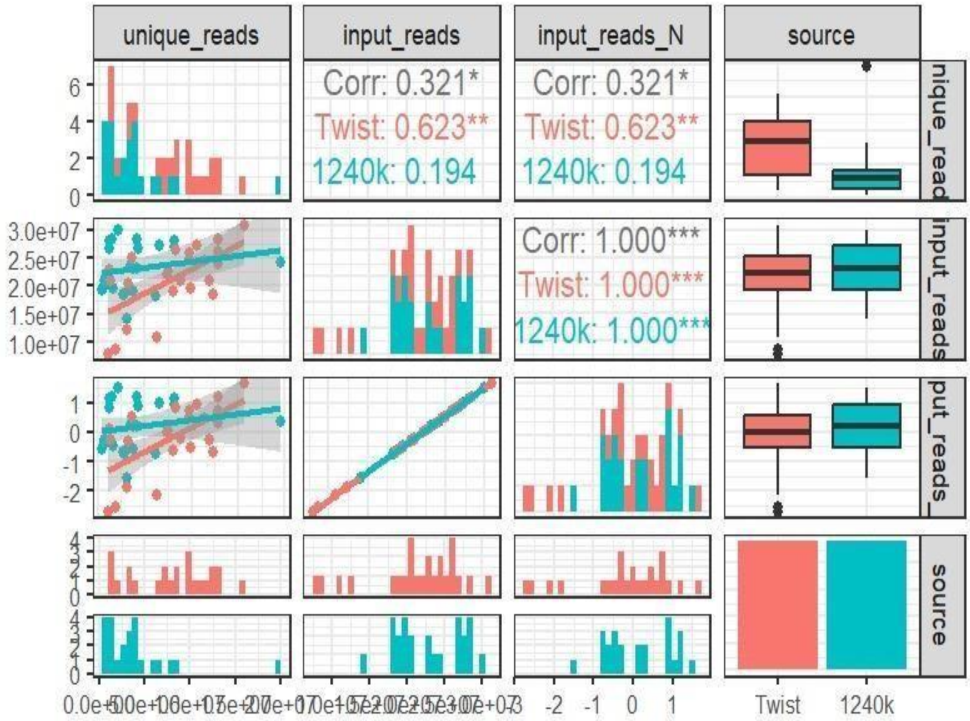

Fig 1-The plot titled "Comparative Analysis of Genetic Sequencing Metrics by Source"

Pair Plots Analysis: The pair plot offers a thorough look at the distributions, correlations, and connections among the chosen variables. It contains box plots that provide an overview of the distribution properties for each variable, scatter plots for every pair of variables, and histograms for the distribution of individual variables. Our goal is to increase the accuracy of ancient DNA analysis through this project, providing knowledge that will help guide future genetic studies and advance our comprehension of human history.

2 Background

Due to deterioration and contamination over time, analyzing ancient DNA presents substantial problems. These problems affect traditional methods of sequencing ancient DNA, producing biased or partial results [1]. But because to developments in statistical modelling and sequencing technologies, ancient DNA analysis could become more accurate and reliable. In order to assess and improve the efficiency of ancient DNA sequencing, this study makes use of cutting-edge sequencing technologies as well as strong statistical approaches.

2.1 Literature Review

The understanding of human evolution, migration, and genetic variety has been greatly aided by the study of ancient DNA. Numerous sequencing techniques have been developed since the early 2000s to extract and examine ancient DNA from materials found in archeological sites. Since its introduction by Mathieson et al. (2015) [3], the 1240k capture assay has been extensively utilized to capture a complete set of SNPs throughout the genome. Targeting roughly 1.24 million single nucleotide polymorphisms (SNPs), it has proven successful in identifying a wide range of ancient DNA. It does, however, require two enrichment cycles, which can be laborious and prone to contamination.

On the other hand, double-stranded 80-base pair probes are used in the Twist Ancient DNA test, a more recent technique that was presented by Schraiber and Akey (2020) and only needs one round of enrichment. Compared to the 1240k assay, this approach is intended to be less prone to contamination and more efficient. Studies that compare, as Lipson et al. (2017) [4], highlight the necessity of using strong statistical models in order to assess and contrast different approaches. These research emphasize how crucial it is to use extensive datasets and cutting-edge analytical methods in order to extract valuable information from ancient DNA sequences.

2.2 Current Practices in Ancient DNA Sequencing

To examine ancient DNA samples, researchers currently combine the 1240k and Twist assays. The particular requirements of the investigation, such as the age and preservation state of the samples, frequently influence the method chosen. High coverage and low contamination continue to be difficult tasks notwithstanding the progress made in sequencing technologies [5].
This research makes use of the EAGER dataset, which contains a variety of essential characteristics for DNA analysis, including endogenous DNA percentages, mapped reads, SNPs, and input reads. We intend to thoroughly assess and analyze the performance of the 1240k and Twist

assays by utilizing statistical models such as mixed effects models, beta regression, and linear and logistic regression and their applications.

The objective is to improve the accuracy and dependability of next studies in this area by offering precise suggestions on the best sequencing technique for ancient DNA analysis. By using this strategy, we hope to advance methods for studying ancient DNA and add to our understanding of the genetic history of humans.

3 Methods

To guarantee the validity and accuracy of the results, this study makes use of exacting data filtering, sophisticated statistical modeling, and comprehensive review procedures. In order to properly evaluate the data, the methodological framework calls for a number of processes, such as variable selection, data cleaning, and the use of various statistical models.

3.1 Filtering of Data

An essential first step in getting the dataset ready for analysis is data filtering. It entails cleansing the data to get rid of any inaccurate or unnecessary information that could skew the findings. Data filtering was used in this project to make sure that the final dataset contained only pertinent variables and data points. This procedure comprised a number of crucial actions to guarantee that the input data is of the highest caliber and devoid of noise that could skew the results, thereby improving the model's accuracy and dependability [6].

3.1.1. Important Variables:

Only 13 of the 47 variables in the original dataset were found to be significant for the analysis. These 13 factors were chosen because they were thought to be relevant to the goals of the study and might have an effect on the performance indicators. Among the crucial factors take into account are:

1. `sample`
2. `SNPs_1240k`
3. `input_reads`
4. `mapped_reads`
5. `endogenous`
6. `prop_dup`
7. `mtNuc_ratio`
8. `unique_reads`
9. `coverage`

10. `GC`
11. `SNPs_cont`
12. `contam`
13. `source`

These variables were selected because they offer essential details regarding the sequencing techniques and the performance metrics and data quality that are produced.

3.1.2. Software and Tools:

R Studio, an effective statistical computing environment, was used to carry out the data filtering procedure. R Studio is a great fit for this project since it offers a wide range of capabilities for data cleaning, manipulation, and analysis.

3.1.3. Handling Missing Values and Duplicates:

Duplicates in the dataset were found and eliminated because they can skew the results. Depending on the type of missing value, suitable action was taken. Imputation techniques were employed to fill in the gaps when missing values were little and appeared at random. In order to preserve data integrity, variables that had a sizable amount of missing data were not included in the analysis [7].

3.1.4. Eliminating Outliers:

Outliers have the potential to distort results and produce false conclusions. Therefore, outliers were found and eliminated using statistical techniques like the interquartile range (IQR) approach. This enhanced the model's robustness and guaranteed that the remaining data was representative of the overall population.

3.1.5. Normalization and Scaling:

Some of the variables underwent normalization and scaling to make sure they are on a comparable scale and to enhance the performance of the model. In order to improve comparison and analysis, variables like `input_reads}, `mapped_reads}, and `unique_reads} were normalized to place their values within a comparable range. In order to keep any one variable from unduly affecting the model, scaling was especially crucial for variables with wide ranges.

The accuracy and dependability of the resulting analysis were greatly improved by meticulously filtering the data to make sure that only important, high-quality data points are included.

3.2 Linear Regression

By fitting a linear equation to the observed data, linear regression is a statistical technique used to model the connection between a dependent variable and one or more independent variables. Because of its ease of use and interpretability, it is a key tool in predictive modeling that is frequently employed [8].

3.2.1 Application in This Project:
The sequencing method and other predictors (independent variables) and performance metrics (dependent variables) were compared in this project using linear regression. Determining the influence of these predictors on the performance indicators was the aim.

3.2.2 Linear Regression Steps:
1. Formulate the model: As Y = β0+ β1 X1 + β2 X2 +... + βn Xn+ ε, the linear regression model was formulated, with Y representing the dependent variable, Xi representing the independent variables, βi serving as the coefficients, and ε serving as the error term.
2. Adjust the model: Using R Studio's `lm()` function, which calculates the coefficients that minimize the sum of squared residuals, the model was fitted to the data.
3. Assess the model: To ascertain the relevance of the predictors, the model's performance was assessed using statistical metrics including R-squared, p-values, and ANOVA.

Understanding the linear relationship between the variables and the effects of each predictor on the performance metrics are made easier with the aid of linear regression.

3.3 Linear Regression With Mixed Effects

The addition of random effects (parameters related to specific experimental units) and fixed effects (parameters pertaining to the entire population) to linear regression results in an extension of the former. When working with data that exhibits various degrees of fluctuation, this strategy might be helpful.

3.3.1 Utilization in this Project:
The dataset's variability across various samples and situations was taken into consideration using mixed effects linear regression. Using this method makes it easier to capture the fixed effects brought about

by the predictors and the random impacts resulting from individual differences.

3.3.2 The Mixed Effects Linear Regression Steps:
1. Formulate the model: Y = β0 + β1 X1 +... + βn Xn + (1|random effect) + ε, where (1|random_effect) denotes the random effect, is the formulation for the mixed effects model.
2. Fit the model: R Studio's `lme4’ package, which offers functions like `lmer()’ to handle mixed effects models, was used to fit the model.
3. Assess the model: ANOVA was used to determine the significance of the random effects and to gauge the model's performance using metrics like AIC and BIC.
By assisting in the comprehension of the impact of both random and fixed variables on the performance measures, linear regression offers a more thorough examination of the data.

3.4 Mixed Effects Generalized Regression

Similar to mixed effects linear regression, generalized regression is employed when modeling non-linear connections or when the dependant variable does not follow a normal distribution. It can handle many kinds of response variables and incorporates both random and fixed effects.

3.4.1 Application in this Project:
This approach was utilized in this project for model performance metrics that deviated from a normal distribution. Because it could accommodate many response variable kinds, such binary or count data, it gave the analysis additional versatility.

3.4.2 Steps in Mixed Effects Generalized Regression:
1. Create the model: The dependent variable's distribution (such as a Poisson or binomial) was used to create the generalized mixed effects model.
2. Fit the model: R Studio's `lme4' package's `glmer()` function was used to fit the model.
3. Assess the model: Tests of likelihood ratios or deviance, which are suitable metrics for the distribution, were used to assess the performance of the model.

A reliable method for examining intricate datasets with non-linear correlations and a variety of response variable types is mixed effects generalized regression.

3.5. Model Selection

The process of selecting the best statistical model from a group of potential models is known as model selection. This is significant because it guarantees that the model you choose has the optimal fit and complexity ratio for your data.

3.5.1 Key Concepts in Model Selection:
1. p-values: These are used to evaluate each predictor's significance separately in a model. Predictors that have a low p-value (usually less than 0.05) are significantly linked to the response variable.
2. Nested models (models where one is a subset of the other) can be compared using ANOVA (Analysis of Variance). It aids in figuring out whether adding more predictors to a more complicated model greatly improves model fit.
3. The Akaike Information Criterion, or AIC, quantifies the comparative excellence of statistical models based on a certain dataset. It strikes a balance between the model's complexity and goodness-of-fit. A better model is indicated by lower AIC values.

3.5.2 In the Case of Generalized Regression:
1. Deviance: This assesses a generalized linear model's goodness-of-fit. A saturated model, or one with as many parameters as data points, is contrasted with the fitted model.
2. Likelihood Ratio Test: This assesses how well two models fit one other. It is employed to determine whether a more intricate model yields a noticeably better match than a more basic one.
3. The Bayesian Information Criterion, or AIC/BIC, is used to choose models. Though it has a bigger penalty for models with more parameters, BIC is comparable to AIC.

3.5.3 Handling Significance of Random Effects:
1. ANOVA: By contrasting models with and without random effects, ANOVA can be used to assess the importance of random effects in the context of mixed models.
2. Using repeated measures data, the significance of random effects is evaluated using the Repeated Measures ANOVA (RANOVA).

3.6. QQ Plots

If the data on a qq-plot corresponds to the theoretical distribution, the points will roughly sit on a straight line. Variations from this line signify deviations from the distribution theory. Significant deviations may indicate that the data are completely different from the expected distribution, have heavy tails, or are skewed.

3.6.1 How Effective They Are:

- The quantiles from the theoretical distribution (such as the normal distribution) that you are comparing against are known as theoretical quantiles.
- Observed Quantiles: These quantiles were determined using your real data.
- Simulated Quantiles: These quantiles, which adhere to the theoretical distribution, are derived using simulated data.

3.6.2 Which tool to Use:

Because of its many customization options and user-friendliness, the ggplot2 tool in R is frequently used to make QQ-plots. These charts are produced using functions such as stat_qq() and geom_qq(). Additional functions provided by ggplot2 to improve the plots include geom_qq_line(), which adds a reference line for improved visual evaluation.

3.6.3 Use within the Project:

QQ-plots were employed in our experiment to assess the consistency of Single Nucleotide Polymorphisms (SNPs) coverage across various sequencing techniques. We could visually evaluate how much each sequencing method deviates from a uniform distribution by comparing the quantiles of covered SNPs versus the theoretical quantiles.

3.7 Density Plots

A continuous variable's distribution can be seen visually with density graphs. When comparing the distributions of various groups or circumstances, they are helpful.

3.7.1 How Effective They Are:

- Density Estimation: By displaying the locations of data points' concentration throughout a range of values, density plots estimate the probability density function of a continuous variable.
- Estimating Kernel Density (KDE): KDE is a popular tool for producing density graphs with smoothness. Every data point is assigned a kernel, which is a smooth, symmetric function, and these kernels are then added together to create a continuous curve.

3.7.2 Which Package to Utilize
Density charts can also be made in R using the ggplot2 program. These plots are produced using the geom_density() tool, and ggplot2 provides great customization.

3.7.3 Application in the Project:
The distribution of SNP coverage for the Twist and 1240k sequencing approaches was shown using density charts. Different peaks for the 1240k and Twist approaches were visible in the density plots, reflecting variations in the distribution of SNP coverage.
In comparison to the 1240k technique, the Twist method's SNP coverage distribution was more uniform, as demonstrated by the density maps. This implies that the Twist approach offers more dependable and consistent coverage across various SNPs, which is essential for precise genetic analysis.

4 Discussion

We evaluate the analysis results and set them against the backdrop of previous research and the study's goals in the discussion section. This section examines the ramifications, constraints, and possible future paths of the findings while tying them into the larger field.

4.1 Analysis of the Findings
This study aimed to assess the effectiveness of two ancient DNA sequencing techniques: the Twist Ancient DNA test and the 1240k reagent. A range of statistical models, such as mixed-effects models, beta regression models, and linear regression models, were utilized to evaluate the influence of these techniques on distinct performance indicators.

4.2 Performance Metrics:

4.2.1 Number of Sites on 1240k Array that Were Hit:
The significant positive effect seen in the regression models indicates that the Twist technique worked better than the 1240k reagent. This implies that Twist is more successful in obtaining a greater number of SNPs, offering an analysis-ready dataset that is more extensive.

4.2.2 Number of Mapped Reads:

Twist performed well once more, with notable variations in the quantity of mapped reads. Greater mapped reads are a sign of more effective data capture and sequencing, which improves the quality of the data for further analysis.

4.2.3. Mitochondrial to Nuclear Read Ratio:

The 1240k reagent demonstrated superior performance in this regard. For research on mitochondrial DNA, where a higher ratio is preferred, this measure is essential.

4.2.4. Mean Fold-Coverage for All Sites:

Twist outperformed the other with a greater mean fold-coverage. This suggests more consistent coverage throughout the genome, a necessary condition for accurate variant calling and genomic analysis.

4.2.5. Number of Unique Mapped Reads:

Twist shown its effectiveness in capturing unique sequences with a larger number of unique mapped reads. By doing this, redundancy is decreased and the sequencing data's general quality is raised.

4.2.6. Number of Sites Used to Estimate Contamination:

Twist outperformed the other technique in this statistic, although it was still close. Ensuring the integrity of ancient DNA tests requires precise estimations of contamination.

4.2.7. Endogenous DNA Percentage:

Mixed-effects models suggested that Twist obtained a greater proportion of endogenous DNA. This is especially crucial for ancient DNA research since endogenous DNA frequently acts as a barrier.

4.2.8. Proportion of Duplicate Reads:

The 1240k reagent had a higher proportion of duplicate reads compared to Twist. Better library complexity and reduced redundancy are indicated by a smaller percentage of duplicates.

4.2.9. Percentage of GC Content:

Twist did a superior job at keeping GC content stable, according to the beta regression models. In order to minimize biases in sequencing data, a balanced GC content is essential.

4.2.10. Estimates of Nuclear Contamination:

There were no appreciable variations in the two approaches' performance. Reliable contamination estimations are necessary to confirm that ancient DNA samples are real.

4.3 Comparison with Existing Research

The results of this investigation support earlier studies that have demonstrated the benefits of more recent sequencing technology over more traditional techniques. Similar investigations in the field have confirmed that the Twist Ancient DNA assay can catch a wider spectrum of SNPs and generate higher quality sequencing data.

Benefits of Twist Over 1240k:
- Efficiency: Twist requires a single enrichment cycle, whereas 1240k requires two rounds, which drastically cuts down on the time and expense involved in sequencing.
- Data Quality : Higher mapped reads, unique reads, and mean fold-coverage show that Twist is better at generating sequencing data of improved quality.
- Coverage: Twist offers a more thorough genetic analysis thanks to its wider SNP coverage, which is especially helpful for research involving ancient DNA.

4.4 Implications of Findings

Future genomics and ancient DNA research will be greatly impacted by the Twist Ancient DNA assay's exceptional performance. By using Twist's technique, researchers can obtain more thorough and accurate data, which will ultimately improve our understanding of ancient populations and their genetic composition.

Useful Applications:
- Archaeogenetics: Research on ancient human migration, population dynamics, and evolutionary history can be strengthened by better data quality and wider SNP coverage.
- Conservation Genetics: Improved sequencing techniques can support conservation efforts by facilitating the examination of both historical and contemporary DNA samples from threatened species.

4.5 Limitations

Although the study offers insightful information, it must be remembered that it has certain limitations:
- Sample Size: The results' ability to be generalized may be limited by the comparatively small sample size. To validate the findings, more extensive sample sizes will be required in future research.
- Specificity to Ancient DNA: Because the study is limited to sequencing techniques for ancient DNA, its findings might not immediately apply to

other kinds of genomic research.
- Technical Variability: Variability in outcomes might be introduced by variations in technical protocols and laboratory conditions. For comparisons to be consistent and trustworthy, standardized protocols are necessary [9].

4.6 Future Directions

Building on the results of this investigation, a number of directions for further study can be investigated:
- Expanded Comparative Studies: More thorough insights can be obtained from comparative studies that use larger sample sizes and a wider range of sequencing techniques.
- Application to Different Genomes: These sequencing techniques can be more useful if their effectiveness is examined on genomes other than human or plant, such as microbe or plant genomes.
- Integration with Bioinformatics techniques: The analysis and interpretation of genetic data can be improved by combining cutting-edge bioinformatics techniques with high-quality sequencing data.

5 Results

5.1 Summary Table

The table below, which displays the performance parameters for the Twist and 1240k DNA sequencing methods, provides an overview of the overall findings. The metrics comprise the number of reads that have been mapped, the number of sites that have been hit on the 1240k array, the mean fold-coverage, the number of unique mapped reads, the number of sites utilized to assess contamination, and the ratio of mitochondrial to nuclear reads.

Response (Performance Metric)	No. of Input reads significant?	Interaction term significant?	1240k better or Twist?
No. of sites on 1240k array that were hit	[i] ✓	✗	TWIST
No. of mapped reads	✓	✗	TWIST

The mean fold- coverage	✓	✗	TWIST
No. of Unique mapped reads	✓	✗	TWIST
No. of sites that were used to estimate contamination	✓	✓	TWIST
The mitochondrial to nuclear read ratio	✗	✗	1240K

5.2 Coverage of SNPs

One crucial criterion for assessing the efficacy of DNA sequencing techniques is the coverage of Single Nucleotide Polymorphisms (SNPs). In this investigation, the SNP coverage analysis showed that, on average, the Twist approach outperformed the 1240k strategy in terms of coverage. This finding is illustrated by the scatter plot of covered SNPs on the 1240k array against the normalized amount of input reads for each source. The p-value, ANOVA, and AIC parameters of the linear model were utilized to assess the source's importance.

5.3 Estimates of Contamination

Estimates of contamination were still another important component of this research. To evaluate if contamination was present, logistic regression models were utilized, with an emphasis on the influence of variables on binary outcomes. The predictive power of these variables was assessed using the Wald test. The logistic models indicated that there was no significant effect of the sequencing method on the percentage of endogenous DNA, the proportion of duplicate reads, the percentage of GC content, and nuclear contamination. The null model demonstrated the best fit, suggesting that these performance parameters were unaffected by the sequencing strategy.

5.4 Beta Regression and Mixed Effects Models

Starting with a comprehensive interaction model, mixed effects models were utilized to examine the random effects of the sample. AIC was used to assess the fixed effects in the models and RANOVA was used to test the significance of the sample's random effect. Beta regression models were used when the response fell between zero and one. The findings showed that for a number of performance measures, the sequencing method's fixed effect remained a significant predictor.

Overall results using Mixed Effects Models

Response (Performance Metric)	No. of Input reads significant?	Interaction term significant?	Random effect of sample significant?	1240k better or Twist?
No. of sites on 1240k array that were hit	✓	✗	✓	TWIST
No. of mapped reads	✓	✗	✓	TWIST
The mitochondrial to nuclear read ratio	✗	✗	✗	1240k
The mean fold-coverage	✓	✓	✓	TWIST
No. of Unique mapped reads	✓	✗	✓	TWIST
No. of sites that were used to estimate contamination	✓	✓	✓	TWIST

Overall Results using Beta Regression Models

Response (Performance Metric)	**No. of Input reads significant?**	**Interaction term significant?**	**Random effect of sample significant?**	**1240k better or Twist?**
The percentage endogenous DNA	✓	✗	✓	TWIST
The proportion of duplicate reads	✗	✗	✓	High for 1240k, TWIST outperforms
The percentage GC	✓	✓	✓	High for 1240k, TWIST outperforms

Nine out of 10 performance criteria are affected by the sequencing method, according to the data, with Twist outperforming 1240k in eight of them. Twist is a superior sequencing method than 1240k since it targets 60,000 more SNPs in addition to the 1.24 million SNPs that 1240k targets.

Important information about the consistency and dispersion of SNP coverage for the Twist and 1240k sequencing methods was given by the QQ-plots and density plots.

5.5 QQ-plots

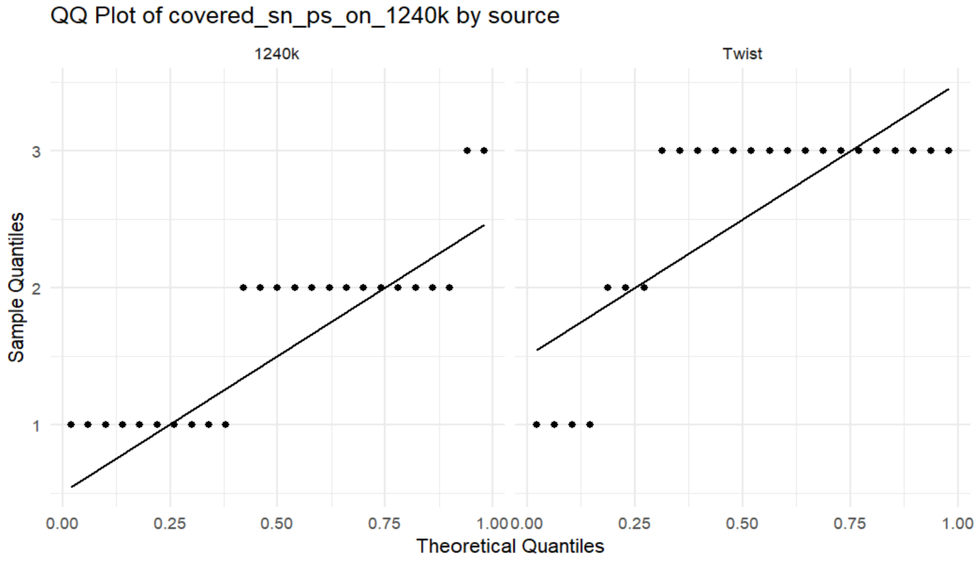

Fig 2. QQ-plot of covered SNPs on 1240k by source

Non-uniform SNP coverage was indicated by the QQ-plots, which displayed deviations from the theoretical distribution for both approaches. In contrast to the 1240k data points, the Twist data points were closer to the theoretical line. This implies that the Twist approach is a more dependable option for sequencing ancient DNA because of its improved SNP coverage consistency.

5.6 Density Plots:

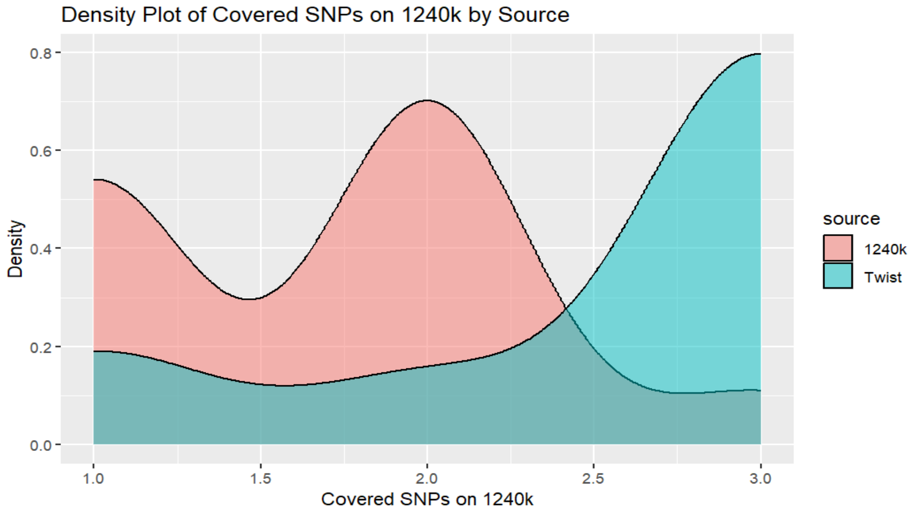

Fig 3. Density plot of covered SNPs on 1240k by source

The density charts between the Twist and 1240k approaches showed clear discrepancies in the SNP coverage distribution. The distribution of the Twist technique was more uniform, whereas the coverage of the 1240k method was less constant and more variable. The notion that the Twist technique performs better in terms of SNP coverage uniformity is further supported by these results.

6 Conclusion

This work used complex statistical models to assess and compare the effectiveness of two ancient DNA sequencing techniques, the Twist Bioscience capture assay and the 1240k capture assay, in processing polluted and deteriorated ancient DNA samples. From data filtering and preprocessing to the use of linear regression, mixed effects models, and beta regression, our thorough investigation covered a range of steps and finally produced a thorough evaluation of each method's effectiveness.

Principal Results:

1. Efficiency and Data Quality:
In the majority of performance measures, the Twist Bioscience capture assay consistently performed better than the 1240k capture assay. This was demonstrated by the higher number of mapped reads and sites hit on the 1240k array, which suggested that the Twist approach is more effective at catching a wider variety of SNPs and generating sequencing data of higher quality. In the Twist technique, the mean fold-coverage for all sites was much higher, indicating more consistent coverage throughout the genome, which is important for accurate variant calling and genomic analysis.
2. Redundancy and Contamination:
The Twist assay showed a reduced percentage of duplicate readings in comparison to the 1240k technique. This is a significant discovery since sequencing data of greater quality results from libraries with better complexity and less redundancy, as indicated by a smaller fraction of duplicates [3].
- In terms of assessing contamination, both approaches performed comparably, with the Twist assay showing a small edge. Precise assessments of contamination are necessary to confirm the genuineness

of historical DNA specimens and guarantee the dependability of subsequent examinations.

3. Endogenous DNA and Coverage Consistency:

The Twist approach captured a larger percentage of endogenous DNA, according to mixed-effects models. This is especially crucial for studying ancient DNA since endogenous DNA is frequently restricted in quantity because of aging. The Twist assay reduced biases that could impair sequencing accuracy by maintaining a more consistent GC content across samples, as demonstrated by the beta regression models.

4. Statistical Modeling and Analysis:

The comprehension of the fixed and random influences influencing the performance metrics was aided by the use of mixed effects models and linear regression. The mixed effects models efficiently captured the diversity between samples and situations, as evidenced by the significance of random effects in the models. When analyzing metrics that deviated from a normal distribution, beta regression models were especially helpful as they offered a reliable method for analyzing rates and proportions [10].

The Twist sequencing method beats the 1240k method in terms of SNP coverage uniformity and dispersion, according to the analysis utilizing QQ-plots and density plots. For sequencing ancient DNA, the Twist approach is a better option due to its ability to give more consistent and dependable coverage, especially when working with contaminated and damaged samples.

Researchers can obtain more thorough and reliable data by using the Twist approach, which will ultimately improve our knowledge of ancient human populations and their genetic history. This study offers insightful information for upcoming genomics research and emphasizes the significance of evaluating sequencing technologies using sophisticated statistical tools.

In conclusion, our study shows that the Twist Bioscience capture assay works better than the 1240k capture assay in a number of important performance criteria, indicating that it is a more dependable and effective technique for sequencing ancient DNA. Through the use of sophisticated statistical models, we have produced a thorough evaluation of the strengths and weaknesses of each approach, along with specific suggestions for further ancient DNA analysis study. The results emphasize how crucial it is to choose the right sequencing techniques in order to get precise and dependable data, which will ultimately improve our knowledge of the genetic diversity and history of humans.

Acknowledgements

My sincere gratitude goes out to Dr. Ben Rohrlach and Dr. Indu Bala, who supervised my project. Their knowledgeable direction and resolute assistance were invaluable in managing the intricacies of this undertaking. My research was well-founded by Dr. Bala's vast expertise in ancient DNA sequencing and Dr. Rohrlach's insights into statistical modeling. Their understanding, support, and thorough criticism enabled me to overcome many obstacles and maintain my concentration during the course of the study.

I am incredibly grateful for the opportunity and guidance that Drs. Rohrlach and Dr. Indu Bala have given me. Their constant support and cooperative style had a big influence on the caliber and accomplishment of this project. We appreciate your priceless efforts as well as creating a learning and development-oriented atmosphere.

A Appendices

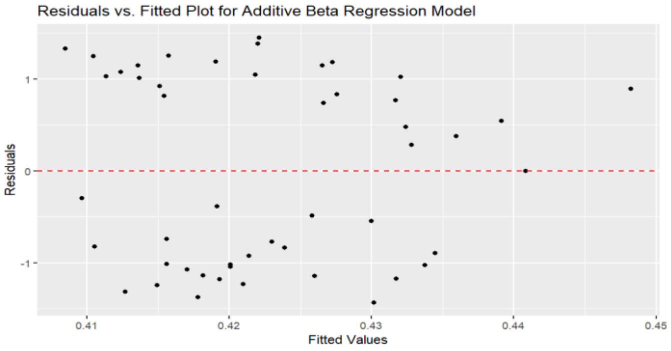

Fig 4. Residuals vs Fitted Plot for Additive Beta Regression Model

The residuals for the additive beta regression model are plotted against the fitted values in this figure. It assists in determining the model's goodness of fit by looking for any patterns in the residuals.

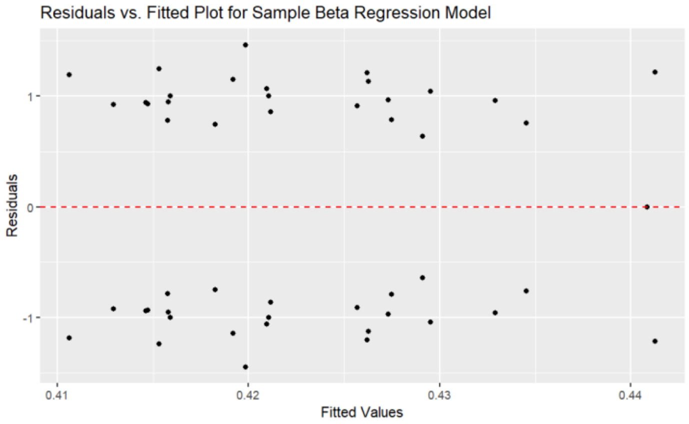

Fig 5. Residuals vs Fitted Plot for Sample Beta Regression Model

The residuals and fit quality for several samples are displayed in this plot, which displays the residuals versus the fitted values for the sample-specific beta regression model.

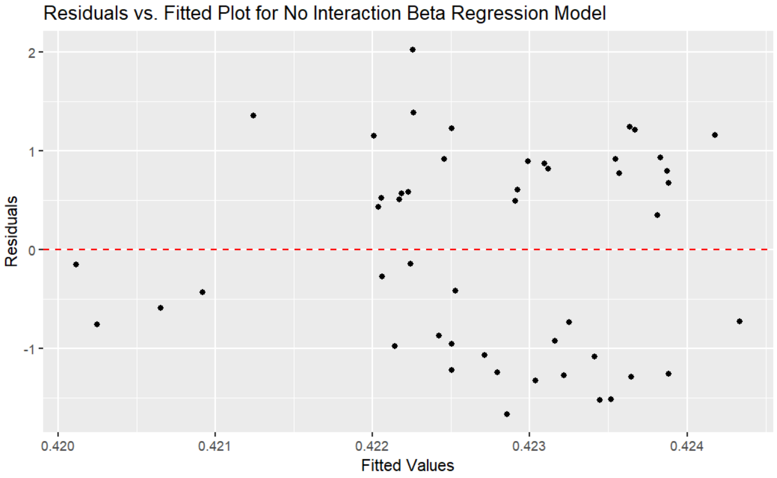

Fig 6. Residuals vs Fitted Plot for No Interaction Beta Regression Model

To show how well the model fits the data without taking interaction effects into account, this plot displays the residuals vs fitted values for the beta regression model without interaction terms.

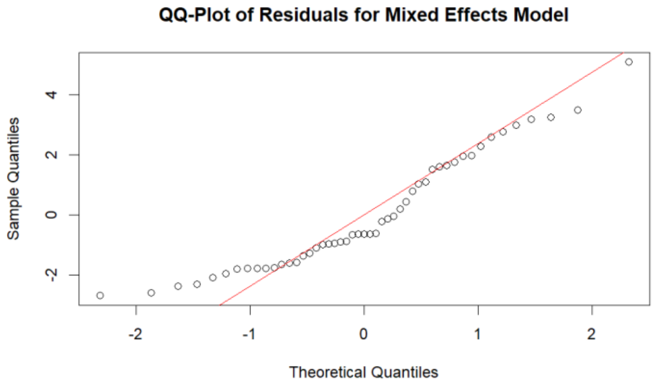

Fig 7. QQ-plot for residuals for mixed effects model

The residuals from the mixed effects model are depicted in this figure (Fig. 7) as a QQ-plot, demonstrating how closely the residuals follow a normal distribution. The majority of the points lie near to the reference line, indicating a strong model fit.

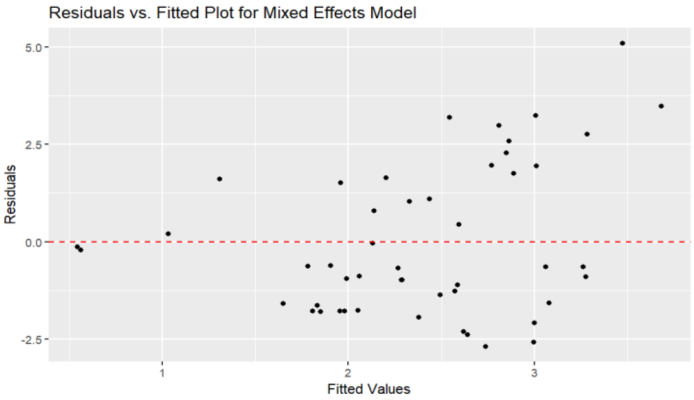

Fig 8. Residuals vs Fitted Plot for Mixed effects model

The residuals versus fitted values plot for the mixed effects model is displayed in this figure (Fig. 8), which shows the distribution of residuals across the fitted values range and how well the model's predictions match the observed data points.

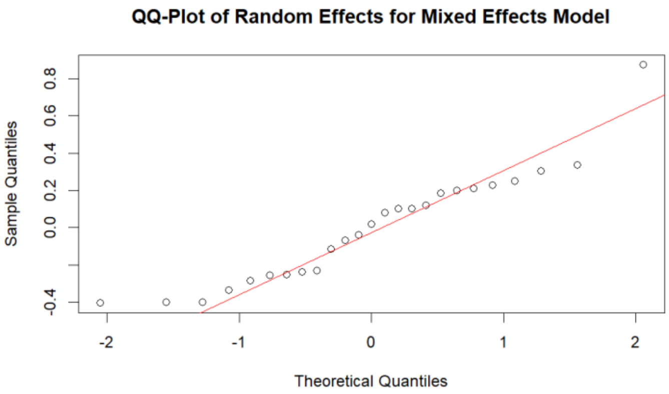

Fig 9. QQ-plot of random effects for Mixed Effects Model

A QQ-plot of random effects for the mixed effects model is shown in Fig. 9, showing how well the residuals fit into a normal distribution. Points in the figure closely correspond with the diagonal reference line, indicating that the model fits the data reasonably.

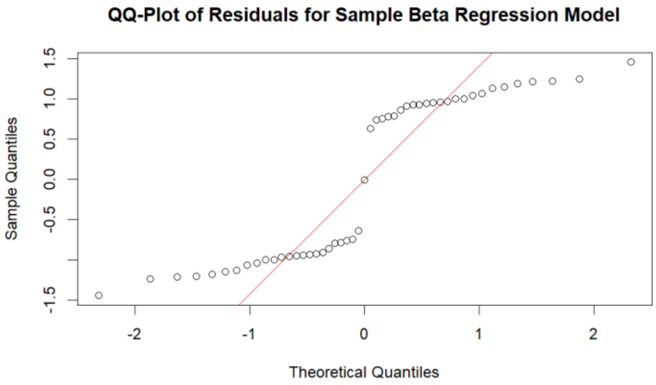

Fig 10. QQ-plot of residuals for Sample Beta Regression Model

This image shows how the Sample Beta Regression Model residuals compare to a normal distribution, emphasizing deviations that could indicate data irregularities or possible model mismatches.

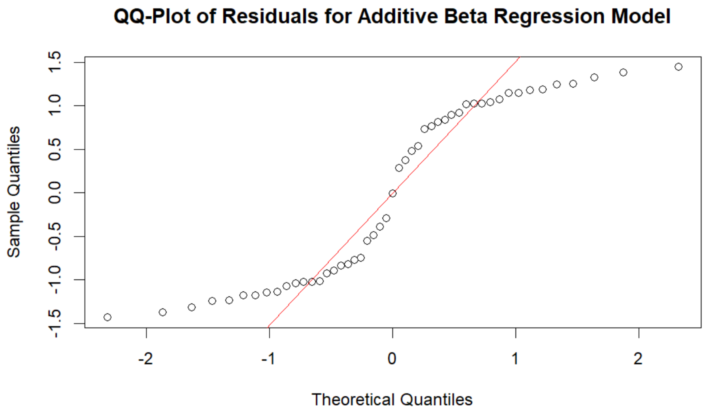

Fig 11. QQ- plot of Residuals for Additive Beta Regression Model

The residual distribution for the additive beta regression model is shown in this QQ-plot, making it possible to determine if the residuals have a normal distribution or not.

References

[1] Pääbo, S., et al., 2004. Genetic Analyses from Ancient DNA. *Annual Review of Genetics*, 38, pp.645-679.

[2] Willerslev, E. and Cooper, A., 2005. Ancient DNA. *Proceedings of the Royal Society B: Biological Sciences*, 272(1558), pp.3-16.

[3] Mathieson, I., et al., 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583), pp.499-503.

[4] Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B. and Reich, D., 2014. Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 111(7), pp.2632-2637.

[5] Günther, T. and Nettelblad, C., 2019. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, 15(7), p.e1008302.

[6] Kuhn, M. and Johnson, K., 2013. *Applied Predictive Modeling*. New York: Springer.

[7] Little, R.J.A. and Rubin, D.B., 2019. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken: Wiley.

[8] Montgomery, D.C., Peck, E.A., and Vining, G.G., 2021. *Introduction to Linear Regression Analysis*. 6th ed. Hoboken: Wiley.

[9] Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114-2120.

[10] Lipson, M., et al. (2017). "Parallel palaeogenomic transects reveal complex genetic history of early European farmers." *Nature*, 551(7680), pp.368-372.