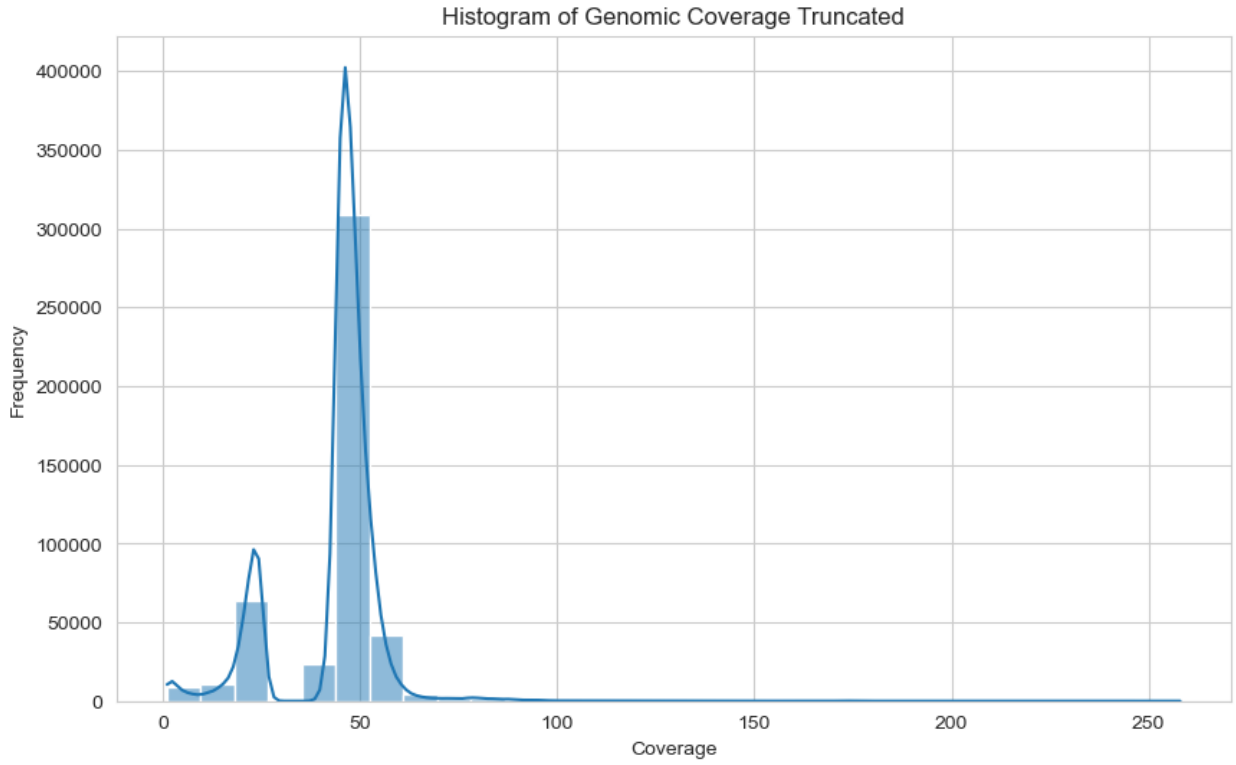
Genomic Coverage Analysis Using Machine Learning and Statistical Modeling

Project Overview

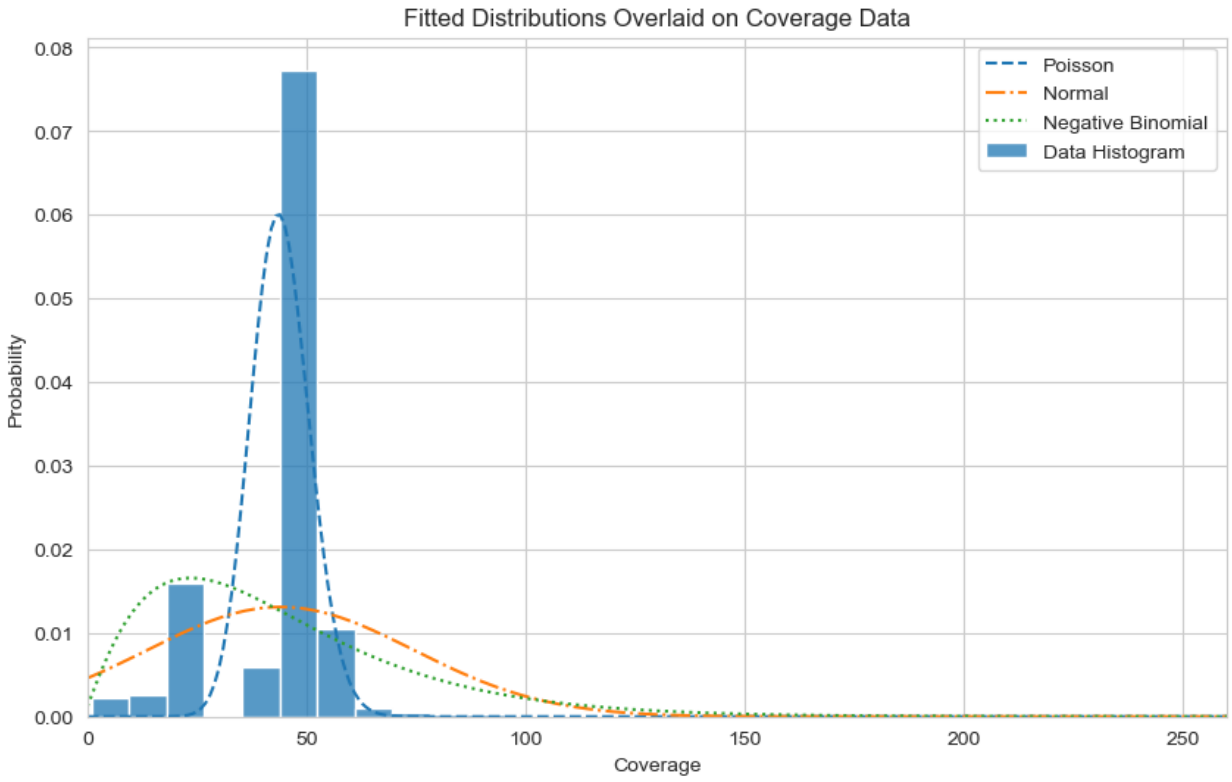
This project focused on applying machine learning and statistical modeling techniques to analyze regions of copy number variability in the human genome. Utilizing data from the HG002 human public genome, the study involved preprocessing coverage values from BAM files and implementing various statistical models to comprehend the variability in genomic coverage.

Methodology

- Data Preprocessing: The initial step involved preprocessing the genomic coverage data, including the exclusion of zero coverage values.
- Descriptive Statistics: Key metrics such as mean coverage (44.10), standard deviation (30.44), and maximum observed coverage (1731) were computed, indicating substantial variability.



- Distribution Fitting and Analysis: The Poisson model was initially applied, with mean coverage as the lambda parameter. Additionally, Kullback-Leibler (KL) divergence analysis was conducted to assess the fit of empirical data to the Poisson model (KL divergence: 0.0396) and the negative binomial model (KL divergence: 2.49).



- Regression Analysis: A generalized linear model (GLM) using negative binomial distribution was employed. Key results include a pseudo R-squared (CS) of 0.05042 and a significant coefficient for scaled coverage (0.5285). The Mean Squared Error (MSE) for the model stood at 0.5772.

Results and Findings

- Data Variability: The data demonstrated significant variability and overdispersion, challenging the modeling process.
- Model Fit and Challenges: While the Poisson model showed a reasonable fit, it failed to fully capture the data distribution, particularly in the tails. The negative binomial model, more suited for overdispersed data, nonetheless showed limitations in explaining variability in copy numbers.
- Predictive Accuracy: Despite establishing significant relationships, the models exhibited challenges in predictive accuracy, as indicated by the high MSE and low pseudo R-squared values.

Challenges and Limitations

- Modeling Limitations: Both models, while statistically significant, were inadequate in fully capturing the distribution of the data, especially in the tails.
- Predictive Power: The regression models, although significant, demonstrated limitations in their predictive power for copy number variability.

Conclusions

The analysis successfully applied machine learning and statistical modeling techniques to genomic coverage data, offering valuable insights into the distribution and variability of coverage in the human genome. However, it also highlighted the complexity of such data and the necessity for robust modeling approaches to enhance predictive accuracy.