Bioinformatics Summer School Report

Name: Omar Osama Rawas
Internship Organization: The Broad institute of MIT & Harvard
Supervisor's Name: Dr. Wail Balaawi
Project title: All of Us ancestry concordance across tools

1. Introduction

During my internship, I focused on the All of Us (AoU) project, specifically on evaluating the concordance between different tools used for ancestry prediction. The primary goal was to assess the reliability of categorical ancestry predictions compared to admixture estimates. This analysis is essential for accurately determining population-specific allele frequencies, which have significant implications for disease diagnosis and understanding genetic risk factors. By validating these methods, we aimed to improve the precision of genetic analyses in diverse populations.

2. Materials and Methodology

The project employed two main tools for ancestry inference, applied to a dataset of over 410,000 samples. This large-scale analysis was crucial for evaluating the robustness and reliability of the ancestry prediction methods:

- **Categorical Ancestry Prediction**: This was carried out using a Random Forest model trained on 1KG+HGDP data. The model classified samples into six ancestry groups (AFR, EUR, MID, EAS, SAS, AMR) with a minimum probability threshold of 75% for a valid label.

- **Admixture Estimates**: These were obtained using the Rye tool, which provided percentage-based ancestry proportions for each sample.

To evaluate the concordance between these methods, several statistical approaches were used:

- **Cosine Similarity**: Compared the vector of predicted ancestry probabilities to admixture estimates.

- **Cohen's Kappa**: Assessed agreement between discrete ancestry predictions from both methods.

- **Jaccard Index**: Measured similarity between predicted groups.

- **Confusion Matrix & F1 Score**: Evaluated prediction performance using both methods as potential ground truths.

- **Concordance Correlation Coefficient (CCC) and Pearson Correlation**: Measured correlation between the predictions for each ancestry group.

3. Results and Discussion

- **Cohen's Kappa**: Achieved a score of **0.84**, indicating strong concordance between the two tools.

- **F1 Score & Jaccard Index**: The F1 score was **0.91**, and the Jaccard index was **0.83**, both of which demonstrate strong alignment between the categorical ancestry predictions and admixture estimates. These metrics indicate robust performance in classifying the ancestry groups.

- **Cosine Similarity**: Averaged **0.94** across samples, reflecting high similarity.

- **Correlation**: Showed strong concordance across most ancestry groups, with the exception of the MID (Middle Eastern) group, which suffered from limited training data.

	AFR	AMR	EAS	EUR	MID	SAS
Pearson	0.99	0.91	0.98	0.94	0.56	0.92
CCC	0.97	0.89	0.98	0.92	0.32	0.91

These results suggest that both methods are reliable for ancestry inference in most populations, though the accuracy may be compromised for underrepresented groups like MID.

4. Skills and Knowledge Gained

Throughout this project, I worked extensively with the outputs of the Random Forest model and the Rye tool, which helped me strengthen my Python programming skills. I focused on refactoring code to handle larger datasets more efficiently and became familiar with Terra as a platform for managing and analyzing genomic data. Additionally, I worked on Jupyter notebooks and gained experience with Google Cloud Platform (GCP), which were essential for processing and visualizing the results of this project. I also enhanced my presentation skills by communicating the findings and implications of the analysis to both technical and non-technical audiences.

In addition to the technical skills developed, I gained valuable research experience by actively participating in group meetings at the Broad Institute. These engagements

deepened my understanding of the broader context of my work and allowed me to stay informed about the latest developments in the field, further enhancing my analytical and problem-solving skills.

5. Challenges and Solutions

One significant challenge was the inherent difference in the objectives of the two datasets being compared. The ancestry prediction model, based on a random forest approach, is designed for binary classification (assigning a 0 or 1 to each ancestry group), while the admixture estimates provide percentage-based proportions of ancestry. This discrepancy complicated the correlation analysis and made it challenging to interpret the p-values, as the differing data structures hindered the assessment of data normality and metric validity. To overcome this, I utilized multiple statistical methods to cross-validate the results, ensuring a robust evaluation of concordance.

6. Conclusion

This project successfully demonstrated the reliability of categorical ancestry predictions and admixture estimates for most population groups, with both methods showing strong concordance. However, the findings also underscore the importance of refining the random forest classification model to better capture genetic diversity within all populations.

Please note that due to the confidential nature of the data and code used in this project, I am unable to include the code with this submission.

7. Acknowledgments

I would like to express my gratitude to my mentors Dr. Wail Balaawi & Mr .Lee Lichtenstien at the Broad Institute for their guidance and support throughout this project. I am also thankful for the collaboration with the AoU project team.

Additionally, I extend my appreciation to the Data Science Platform (DSP Methods) team for providing a welcoming and supportive environment that greatly facilitated my learning and growth.

Finally, I would like to thank KAUST Academy for sponsoring and organizing this internship opportunity.

8. References

A. B. Conley et al., “Rye: genetic ancestry inference at biobank scale,” Nucleic Acids Research, vol. 51, no. 8, p. e44, Mar. 2023, doi: 10.1093/nar/gkad149.