1. Review by CAI Bibi:

- But no relevant paper is cited, discussed, or compared to the presented work.

→ Ans: Due to limited space in the poster, we did not include a discussion on relevant work. However, during the project, we discovered that previous work also found the better performance of random projection compared with other dimensionality reduction methods such as PCA (Fern & Brodley, 2003). This suggests that using the ARI score as a metric for evaluating the performance of dimensionality reduction methods is feasible and that random projection performs the best among our eight tested methods. Furthermore, there are several commercially available kits for ancestry prediction using SNPs, such as the 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population (Gettings, et al., 2014) and the 59-SNP assay in the Illumina ForenSeq kit for ancestry prediction in Singapore population (Ramani, et al., 2017). These applications demonstrate the feasibility of using SNPs for ancestry prediction and our results showed that clustering followed by dimensionality reduction of SNPs phenotype dataset is also capable of predicting population ancestry.

References

1. FERN, X. Z., & Brodley, C. E. (2003). Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach. In Proceedings of the Twentieth International Conference on International Conference on Machine Learning (ICML'03), 186–193.
2. Gettings, K. B., Lai, R., Johnson, J. L., Peck, M. A., Hart, J. A., Gordish-Dressman, H., Schanfield, M. S., & Podini, D. S. (2014). A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. Forensic Science International: Genetics, 8(1), 101–108. https://doi.org/10.1016/j.fsigen.2013.07.010
3. Ramani, A., Wong, Y., Tan, S. Z., Shue, B. H., & Syn, C. (2017). Ancestry prediction in Singapore population samples using the Illumina ForenSeq Kit. Forensic Science International: Genetics, 31, 171–179. https://doi.org/10.1016/j.fsigen.2017.08.013

2. Review by CHEN Zixin:

- As the authors utilized many alternative methods and in Part 2, RandProj performed best under two criteria, I am confused about the choice of t-SNE in Part 3. The potential reason might be t-SNE is better at visualization clusters, but some reasons are required to make the logic flow more coherent.
- The logic transition from part 2 to part 3 might be able to improve.

→ Ans: In our result, random projection performs the best, followed by t-SNE. However, in random projection, as shown in Figure 2B, the reduced dimension was set as 5957 for random projection considering the sample number and epsilon. The high number of dimensions required for prediction may not be met by our independent validation cohort. Therefore, we used the second best-performing method t-SNE.

3. Review by CUI Yiran

- Lack of theoretical/technical background of applied method so it is difficult for readers without much background.

→ Ans: Due to limited space in the poster, we did not include the theoretical background of each of the eight methods we utilized. Here, to briefly introduce the methods: PCA: linear

dimensionality reduction technique that identifies the most important features in a high-dimensional dataset and projects them onto a lower-dimensional space; MDS: technique for visualizing high-dimensional data by projecting it onto a lower-dimensional space while preserving the pairwise distances between the data points; t-SNE: non-linear dimensionality reduction technique that is particularly useful for visualizing high-dimensional data in two or three dimensions; ISOMAP: non-linear dimensionality reduction technique that uses a geodesic distance metric to preserve the intrinsic geometry of the data; LLE: non-linear dimensionality reduction technique that works by modeling the local linear relationships between the data points in the high-dimensional space and then projecting the data onto a lower-dimensional space while preserving these relationships; UMAP: non-linear dimensionality reduction technique that is particularly useful for preserving both the global and local structure of the data; Robust PCA: variant of PCA that is designed to handle datasets with outliers; Random Projection: simple and efficient linear dimensionality reduction technique that works by randomly projecting the high-dimensional data onto a lower-dimensional subspace.

4. Review by Gerry Windiarto Mohamad DUNDA

- Lack of explanation on the description of SNP dataset. Perhaps, explain the statistics of SNPs/person. Also, the task for the model is not completely described. How many labels and what is proportion of train-valid-test and so on. and preprocessing techniques. Lastly, the data format is not clear such as what is the size of the vector.
- Lack of discussion on potential limitations: The paper does not discuss potential limitations of the study, such as sample size, data quality, or statistical assumptions. This could limit the interpretability of the findings and the ability to generalize the results to other populations or datasets. A discussion of potential limitations could help to provide a more comprehensive understanding of the research.

→ Ans: The dataset we used in this study is a cleaned dataset from Dr. Quanhua Mu and Dr.Yoonhee Nam, incorporating 488,890 SNPs from 1043 samples as well as the region information from patients. The SNPs data is in the format of a matrix with SNPs as row names and sample ID as column names. The genotype is denoted by 0(AA), 1(AC) and 2(CC). During data preprocessing, we transposed the matrix for convenience. And we used all SNPs for clustering after dimensionality reduction. We also extracted the region information of each sample as a label for ancestry prediction of each cluster. There are seven different regions in total, namely Africa, Europe, Oceania, America, East Asia, Middle East and Central/South Asia. In the case study, we used an independent cohort of 400 samples from the 1000 Genomes Project. We successfully predicted the ancestry with a high ARI of 0.709. Potential limitations: Firstly, we only used the Adjusted Rand Index (ARI) score to evaluate the performance of the different methods, which may not be sufficient to fully evaluate the performance of the methods. Future work could be done for comprehensive evaluation using other evaluation metrics, such as precision, recall, and F1 score. Secondly, in this report, we utilized a specific dataset for ancestry prediction, and the findings may not be generalizable to other datasets. It is important to consider the unique characteristics of the data being analyzed.

5. Review by HAO Yifan: NA

6. Review by Huang Zhi: NA

7. Review by HU Yueying:

- There are some wrong formats in the poster. Firstly, figures cannot be seen clearly. Secondly, the figures of case study dont have the chart names and number, such as shown as bottom fig. in the middle column.

→ Ans: Due to the limited space in the poster, the figures were made smaller. We will try to make them larger, especially the figures in part 1. In addition, we will give serial numbers and add titles for the figures in the case study to make them clearer. Figure 1: Applying eight dimensionality reduction methods on the SNPs dataset. Figure 2A: The adjusted rand indexes for separating the dataset into seven clusters after each of the eight dimensionality reduction methods. Figure 2B: The number of dimensions decreases with increasing distortion eps. Figure 3A: PCA shows no obvious batch effect between the original SNPs dataset and an independent validation cohort. Figure 3B: Region annotated PCA dimensionality reduction and k-means clustering result. Figure 3C: Region annotated t-SNE dimensionality reduction and k-means clustering result.

8. Review by HUANG Zhanmiao:

- Some methods and ideas are not explained so clearly. For instance, it says clustering using kNN in part 2, but the method in the contribution part is k-means. And in part 3, the idea that "gathered independent 400 samples" is not much understandable to readers, which may need a clearer and direct expression or explanation.

→ Ans: We used k-means for clustering after dimensionality reduction. We are sorry about the typo and we will modify our poster. In part 3, we used an independent cohort of 400 samples from the 1000 Genomes Project. The region information of these people was also given for us to check the prediction. We randomly selected 400 samples from the 1000 Genomes Project cohort and used them for validation.

9. Review by JIANG Tianshu: NA

10. Review by LAI Yanming:NA

11. Review by LIN Hangyu:NA

12. Review by LIU Chen:NA

13. Review by LI Yakun:NA

14. Review by LUO Yuanhui:

- How SNPs reveal ancestry information can be cleared since the two concepts are not exactly the same. Besides, prediction conducted on new testing data instead of training data could be better.

→ Ans: In this project, we intend to answer the questions whether SNPs can be useful for ancestry prediction, especially on its reduced features. Here, SNPs are different from direct ancestry information, we treat SNPs as dependent variables and ancestry information as

independent variable to construct the prediction method. After we selected the dimensionality reduction techniques, we indeed tried to predict ancestry information on one group of independent test data, not on the training data, as Part III showed.

15. Review by LI Haobo:

- The authors need to define their research problem more clearly; The legend of the figures is hard to read; The authors may need to introduce the methodology and analyze the result more.

→ Ans: In this project, we mainly focused on the research question that whether we can SNPs information to predict ancestry information, especially on the dimensional reduced SNPs data. For the figure-related concerns, we have answered the similar question under No.7. For the methodology-related and result comparison-related concerns, we have answered them under No.3 and No. 1 respectively.

16. Review by MA Ruochen:NA

17. Review by QIU Zhenyu: NA

18. Review by RUAN Yuyan: NA

19. Review by SHENG Rui:

- The author did not describe the reason why to validate the comparison results with a split of 7. Is the conclusion the same in other cases, such as 6, 8 or 9? Or is it most semantically effective to divide into 7 regions?

→ Ans: The reason is that our machine learning model was trained on a SNPs data which contains labels for 7 regions. Therefore, our model was designed to perform a 7-class classification.

20. Review by Shen Lue

- Lack model theory explanation.

→ Ans: Due to limited space in the poster, we did not include the theoretical background of each of the eight methods we utilized. Here, to briefly introduce the methods: PCA: linear dimensionality reduction technique that identifies the most important features in a high-dimensional dataset and projects them onto a lower-dimensional space; MDS: technique for visualizing high-dimensional data by projecting it onto a lower-dimensional space while preserving the pairwise distances between the data points; t-SNE: non-linear dimensionality reduction technique that is particularly useful for visualizing high-dimensional data in two or three dimensions; ISOMAP: non-linear dimensionality reduction technique that uses a geodesic distance metric to preserve the intrinsic geometry of the data; LLE: non-linear dimensionality reduction technique that works by modeling the local linear relationships between the data points in the high-dimensional space and then projecting the data onto a lower-dimensional space while preserving these relationships; UMAP: non-linear dimensionality reduction technique that is particularly useful for preserving both the global and local structure of the data; Robust PCA: variant of PCA that is designed to handle datasets with outliers; Random Projection: simple and efficient linear dimensionality

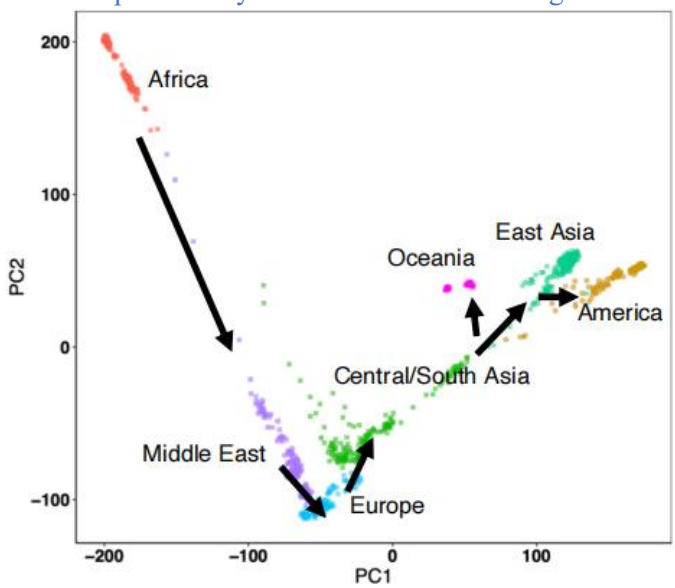
reduction technique that works by randomly projecting the high-dimensional data onto a lower-dimensional subspace.

- Only uses region-level labels, while the dataset contains 4 granualarity levels.

→ Ans: Since region information has shown to be related to genomic information, we select region as our main subject. In addition, since it is a mini-project, we prefer to focus on one major factor.

- The indication on the ancestry prediction from the results are not explained.

→ Ans: Thanks for the comment. Yes, it is interesting to analyze the ancestry information from the dimension reduction graphs. Here is some analysis. We observed that geographic information is well reflected in PCA and MDS, for example, European and Asian populations are located closely. For another thing, such geographic information is related to human ancestry. If we assume that humans originated in Africa, we can clearly trace their migration from Africa to the Middle East, Europe, Central/South Asia, and then diverged into Oceania and East Asia, and America. We think these results can serve as genomic proof for current literature and even as some preliminary results for further investigation into new findings.



- Lack explaination for the visualizations of the results.

→ Ans: Due to the limited space in the poster, the figures were made smaller. We will try to make them larger, especially the figures in part 1. In addition, we will give serial numbers and add titles for the figures in the case study to make them clearer. Figure 1: Applying eight dimensionality reduction methods on the SNPs dataset. Figure 2A: The adjusted rand indexes for separating the dataset into seven clusters after each of the eight dimensionality reduction methods. Figure 2B: The number of dimensions decreases with increasing distortion eps. Figure 3A: PCA shows no obvious batch effect between the original SNPs dataset and an independent validation cohort. Figure 3B: Region annotated PCA dimensionality reduction and k-means clustering result. Figure 3C: Region annotated t-SNE dimensionality reduction and k-means clustering result.

21. Review by SHIHUA Mingzhi

- The reason why some methods perform better than others are not well explained.

→ Ans: Thanks for the suggestion. It is interesting to analyze why the performance varies. However, due to the limited space, we omitted that part. Here are some analysis for your reference.

It was found that linear methods like PCA and MDS (using Manhattan distance) and non-linear methods that preserve local and global similarity and dissimilarity like t-SNE and Sammon projection show better performance than methods based on nearest neighborhood like ISOMAP and LLE. Meanwhile, linear methods reveal more geographic information, for example, European and Asian population are located closely, which cannot be observed in t-SNE and Umap. On the other hand, clear boundaries among the 7 fine-grained populations can be observed in t-SNE, which indicates that its result may be more suitable as inputs of a classifier.

22. Review by TENG Fei:

- It is quite confusing that why divide people by regions. Will the experiment results vary if selecting other dividing criteria?

→ Ans: Because recent studies have suggested that people from different regions may have different ancestors, which can be reflected in genomic information. Therefore, we chose to analyze the data in terms of this factor.
The experiment results may vary if the target factor is not related to genomics. For example, genomic data may not be able to reflect sex informaiton. In this case, our results may not serve as good input for the classifier. To solve this kind of problem, we may need to collect other data instead of SNPs.

23. Review by WANG Zhiwei: NA

24. Review by Xia_Ruizhe: NA

25. Review by XIA_Wencan:

- lack of mathematical statement of algorithms.

→ Ans: Due to limited space in the poster, we did not include the theoretical background of each of the eight methods we utilized. Here, to briefly introduce the methods: PCA: linear dimensionality reduction technique that identifies the most important features in a high-dimensional dataset and projects them onto a lower-dimensional space; MDS: technique for visualizing high-dimensional data by projecting it onto a lower-dimensional space while preserving the pairwise distances between the data points; t-SNE: non-linear dimensionality reduction technique that is particularly useful for visualizing high-dimensional data in two or three dimensions; ISOMAP: non-linear dimensionality reduction technique that uses a geodesic distance metric to preserve the intrinsic geometry of the data; LLE: non-linear dimensionality reduction technique that works by modeling the local linear relationships between the data points in the high-dimensional space and then projecting the data onto a lower-dimensional space while preserving these relationships; UMAP: non-linear dimensionality reduction technique that is particularly useful for preserving both the global and local structure of the data; Robust PCA: variant of PCA that is designed to handle datasets with outliers; Random Projection: simple and efficient linear dimensionality reduction technique that works by randomly projecting the high-dimensional data onto a lower-dimensional subspace.

26. Review by Yan Ningyu:

- It would be better if they add more discussion about the weakness of methods and give some reasons.

- The whole contents are well organized and logical. It could be better if the figures have their identifiers.

→ Ans: In this project, we indeed don't analyze much about the shortage of each method. But we do compare the result of each method and select the best 2 results. Here is the weakness of the methods:
MDS (Multidimensional Scaling): MDS is sensitive to the choice of the dissimilarity measure, and it may produce poor results if the measure doesn't accurately reflect the structure of the data. It's also computationally expensive for large datasets. **PCA (Principal Component Analysis)**: PCA assumes that the data lies in a linear subspace, which may not always be true. It's also sensitive to the scaling of the variables, and it may not preserve local structures in the data. **UMAP (Uniform Manifold Approximation and Projection)**: UMAP is sensitive to the choice of hyperparameters (e.g., number of neighbors, min_dist), and it may produce different results depending on the parameter choices. It may also struggle to preserve global structures in some cases. **LLE (Locally Linear Embedding)**: LLE assumes that the data lies on a manifold and that local linear approximations can be used to describe it. However, it may not work well for datasets with noise or where the manifold assumption doesn't hold. LLE can also suffer from local minima during the optimization process. **ISOMAP**: ISOMAP is sensitive to the choice of neighborhood size and is computationally expensive for large datasets. It may also struggle with noisy data and is not robust to missing data. Sammon Mapping: Sammon Mapping can be sensitive to the choice of the initial embedding, leading to different results. It's also computationally expensive due to the iterative nature of the algorithm.
In contrast, RandProj, ROSPCA, and t-SNE have certain advantages:

1. **RandProj (Random Projection)**: RandProj is computationally efficient and can handle large datasets. It also doesn't require any tuning of hyperparameters.
2. **ROSPCA (Robust PCA)**: ROSPCA is more robust to outliers and noise compared to PCA. It can handle non-linear structures in the data by using a kernel function.
3. **t-SNE (t-Distributed Stochastic Neighbor Embedding)**: t-SNE is particularly good at preserving local structures and clustering in the data. It's also relatively robust to the choice of hyperparameters.

27. Review by Yan Bokai: NA

28. Review by ZHANG Fa: NA

29. Review by ZHONG GuangZheng: NA

30. Review by ZHOU Qiqi: NA