Characterization of population structure from Genetic Variation in Lactose Intolerance Related Genes: An Exploratory Analysis.

Veedhi Solanki

2024-04-06

1. ABSTRACT

Researchers have observed that lactose intolerance varies across different populations of humans as certain populations evolved from pastoral groups who have survived by raising livestock and consuming their milk (Itan *et al*., 2010). In ancestrally pastoral groups, individuals possessing genes allowing for the consumption of dairy products were favoured by natural selection (Anguita-Ruiz *et al*., 2020). Various genes linked to lactose intolerance have been identified, such as *LCT, FUT2, FUT3* and *MCM6* (Anguita-Ruiz *et al*., 2020). Distinct populations within European (EAU) and African (AFR) super-populations in particular have been reported to differ in both phenotypic and genotypic lactose intolerance profiles (Tishkoff et al., 2007); however, the degree to which these populations might be differentiated based on variants within multiple associated lactose intolerance genes has been minimally investigated. Thus, this analysis aims to examine representative populations within EAU and AFR that are historically expected to exhibit differential lactose persistence. Namely, these are Great Britain (Silanikove *et al*., 2015) (GRB) and Finland (Vuorisalo *et al*., 2006) (FIN) within EAU, and Yoruba (Ransome-Kuti *et al*., 1971) (YRI) and Esan (Fan *et al*., 2019) (ESN) within Africa. Specifically, population structure will be explored with respect to the *LCT, FUT2, FUT3* and *MCM6* genes at the super-population level via clustering analysis, and classification models will be trained and tested to determine if super-population assignment can be accurately determined from variation within these genes

2. DATA DESCRIPTION

Variant Call Format (VCF) files were obtained from the 1000 Genomes Project, a publicly available resource including genotype information across 2,500 individuals from 26 global populations (Auton *et al*., 2015). Specifically, four VCF files were obtained and combined, corresponding to the *LCT*, *MCM6*, *FUT2* and *FUT3* genes, respectively. Population-level filtering was applied in the R programming language by cross-referencing sample IDs to corresponding population codes derived from a reference file available on the 1000 Genomes Project website. Representative populations within each super-population are shown in **Table 1**.

Population	**Super Population**
Finnish in Finland (FIN)	EAU
British in England and Scotland (GRB)	EAU
Yoruba in Ibadan, Nigeria (YRI)	AFR
Esan in Nigeria (ESN)	AFR

Table 1) *Populations and superpopulations from 10000 Genomes data included in analysis*

3. ANALYTICAL METHODS

3.1 Preprocessing and Filtering

VCF files were read into R and dataframe objects were extracted using the *vcfR* package. After combining VCF data for each gene, a final combined data frame was filtered to include only samples identified in each of the populations listed in Table 1. A series of quality control steps were then performed on the filtered VCF data. Initial filtering ensured that only variants labelled as single nucleotide polymorphisms (SNPs) and passing the quality threshold were included. As some variants labelled SNPs contained multiple alternative and reference alleles, additionally filtering was applied to retain only biallelic SNPs. A genotype matrix was constructed from the filtered VCF data frame, for which numeric encoding was applied to reflect the genotype for each variant as follows: homozygous recessive for the reference allele set to 0, heterozygous to 1, and homozygous dominant for the alternative allele to 2. Population parameters were then computed and used to assess assumptions of Hardy-Weinberg Equilibrium (HWE). For each variant, a Chi-squared test was conducted to detect significant deviations from HWE ($\alpha = 0.05$). Variants exhibiting significant deviations were excluded from further analysis.

3.2 Principal Component Analysis (PCA)

PCA was performed as the first clustering method to further explore the data and gain a better understanding of the structure of the dataset. PCA offers a powerful means to reduce the dimensionality of the complex dataset while preserving much of the original information. A scree plot was constructed to visualize the variations that each of the PCs explained. To enhance interpretability, data points were colour-coded to delineate distinct populations and super-populations within the dataset.

3.3 Hierarchical Clustering

Hierarchical clustering was applied to discern the structures within the dataset. Given the dataset's high-dimensional structure, an essential consideration in selecting an appropriate clustering method was the choice of distance metric. In this study, Manhattan distance was favoured over Euclidean distance due to its suitability for capturing dissimilarities in multi-dimensional data (López & Maldonado, 2018).

To further explore the dataset's clustering patterns, various hierarchical clustering methods were employed. These methods included both agglomerative and divisive approaches. Agglomerative methods, such as single linkage, complete linkage, and average linkage clustering, iteratively merge clusters based on proximity measures, effectively revealing hierarchical relationships among observations. In contrast, divisive clustering starts with a single cluster containing all data points and recursively divides it into smaller clusters. By employing both agglomerative and divisive strategies, this study aimed to identify the clustering method that would most effectively cluster the dataset.

3.4 Classification Models

This study used k-nearest neighbours (KNN) and Random Forest (RF) classification methods to evaluate the accuracy of predicting super-populations from genetic variant data.

RF classification began with a default model of 500 trees to classify genetic samples into super-populations based on SNP data. The model's accuracy was assessed through predictions on a test dataset, providing an initial effectiveness measure. To optimize the model, we examined the error rate against the number of trees and conducted a grid search over the `mtry` parameter, which dictates the number of variables considered at each split, to enhance model accuracy. This process involved 10-fold cross-validation to ensure robustness, utilizing the `caret` package for execution. The optimal `mtry` value was then applied to train a final RF model, which was again tested for accuracy and evaluated via a confusion matrix to understand classification performance comprehensively. The optimization highlighted the SNPs most indicative of population structure, offering insights into genetic diversity. The final model underscored the RF algorithm's capability to manage complex genetic data, balancing accuracy with computational efficiency in uncovering population differences.

KNN algorithm for the classification of genetic samples into specified super-populations, based on data extracted from VCF files targeting genes such as LCT, FUT3, MCM6, and FUT2. We then normalized the feature contribution by scaling the data, partitioning it into a 75% training set and a 25% testing set to assess the KNN model's predictive accuracy. We varied the number of neighbours (k) ranging from 1 to 20, to determine the optimal configuration for classification accuracy. The model's performance was evaluated based on its ability to accurately assign samples to their respective super-populations, using a confusion matrix for comparison against true metadata-derived assignments. This approach highlighted the model's effectiveness in leveraging SNP data for population structure analysis.

4. RESULTS

4.1 Population Clustering

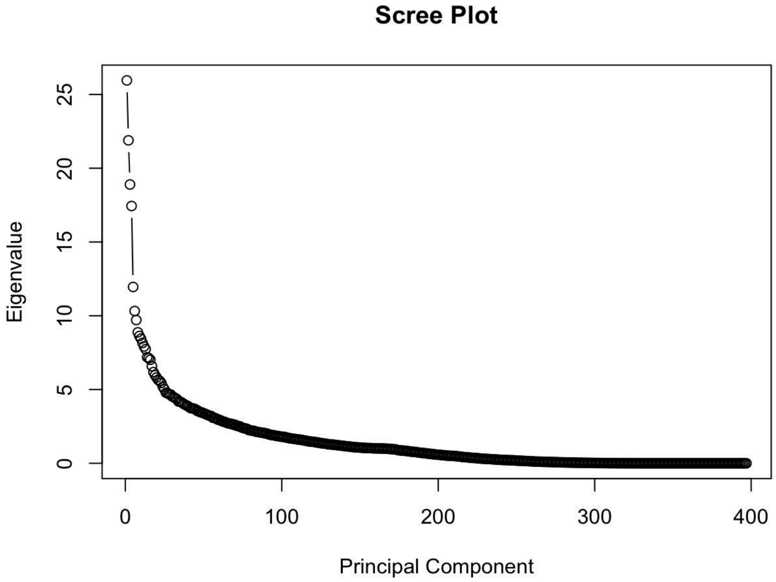

Figure 1: *Scree plot for the PCA showing the eigenvalues of the PCs. This represents how much variation in the data each of the PCs accounts for;*

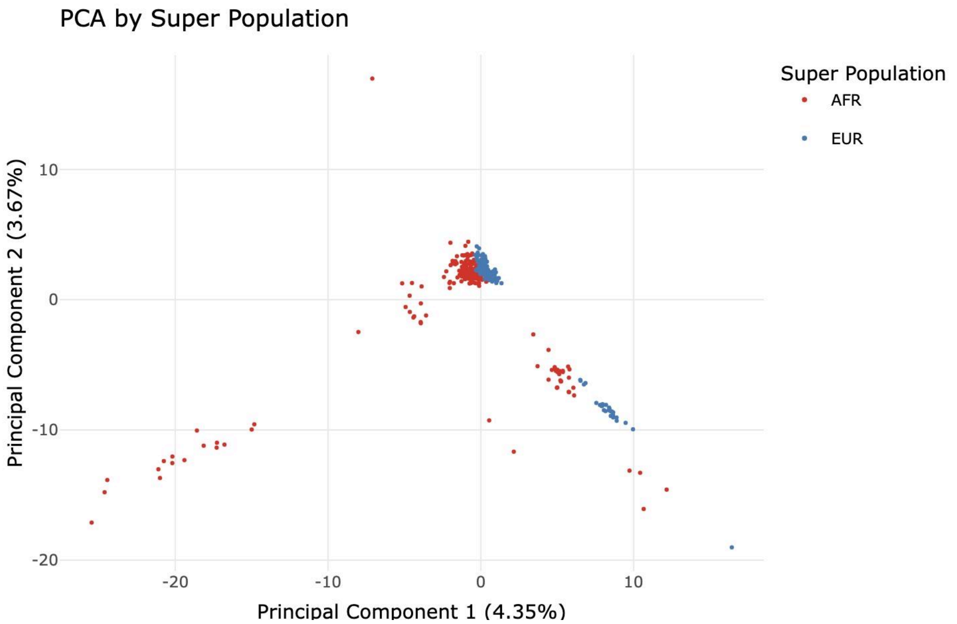

Figure 2: *PCA plot showing the basic structure of the data. Visible yet weak clustering of points by superpopulation.*

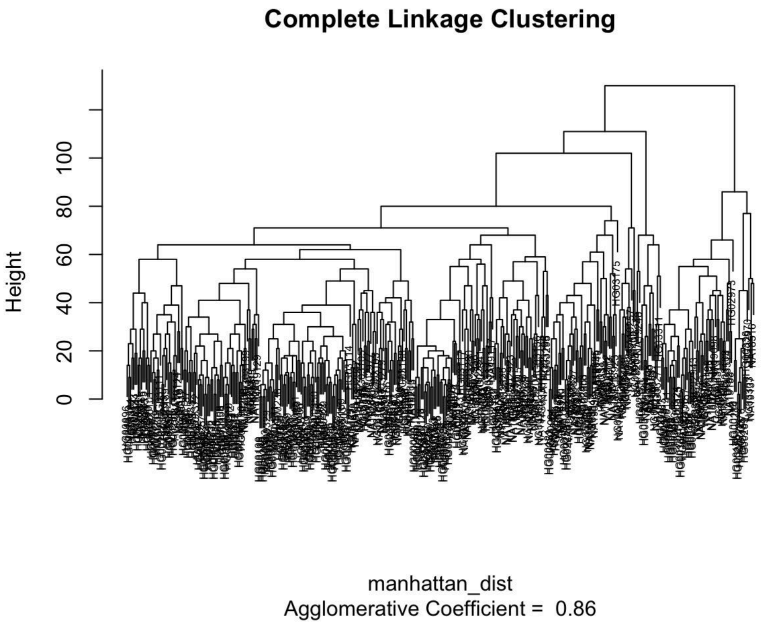

Figure 3: *Complete linkage clustering dendrogram. Shows the hierarchical splitting and clustering of observation in the dataset. There are distinct clusters of observations.*

cutcomp	ESN	FIN	GBR	YRI	AFR	EAU
1	85	77	83	93	178	160
2	14	22	8	15	29	30

Table 2: *Complete linkage clustering output for dendrogram cut to two clusters. The data were clustered into two groups, however, the two super-populations (as well as the subpopulations) did not cluster separately into the two groups.*

4.2 Population Classification

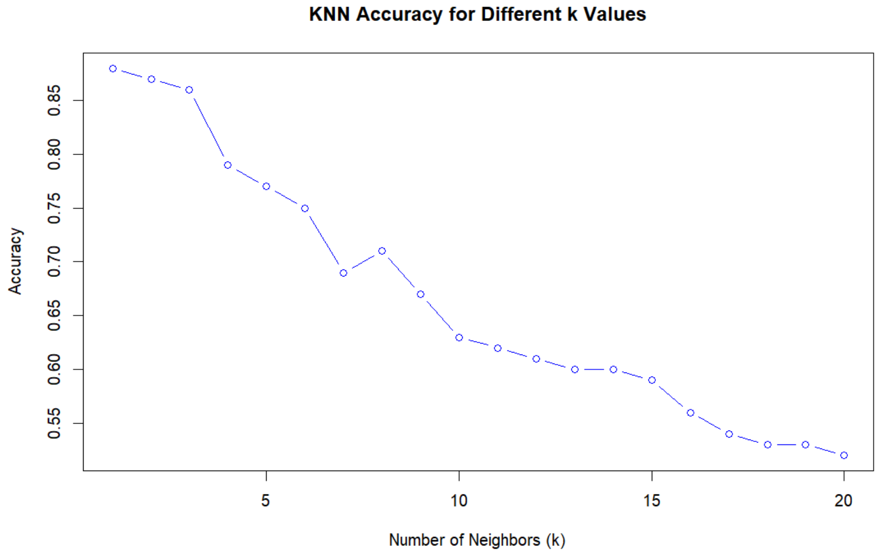

Figure 4: *KNN accuracy plot showing the relationship between the number of k (1-20) and the prediction accuracy for the super-population classification. The plot illustrates that the optimal k value for the highest accuracy is at k - 1. The KNN classification accuracy decreases as the value of k increases indicating the possibility of overfitting with fewer ka and underfitting with more k.*

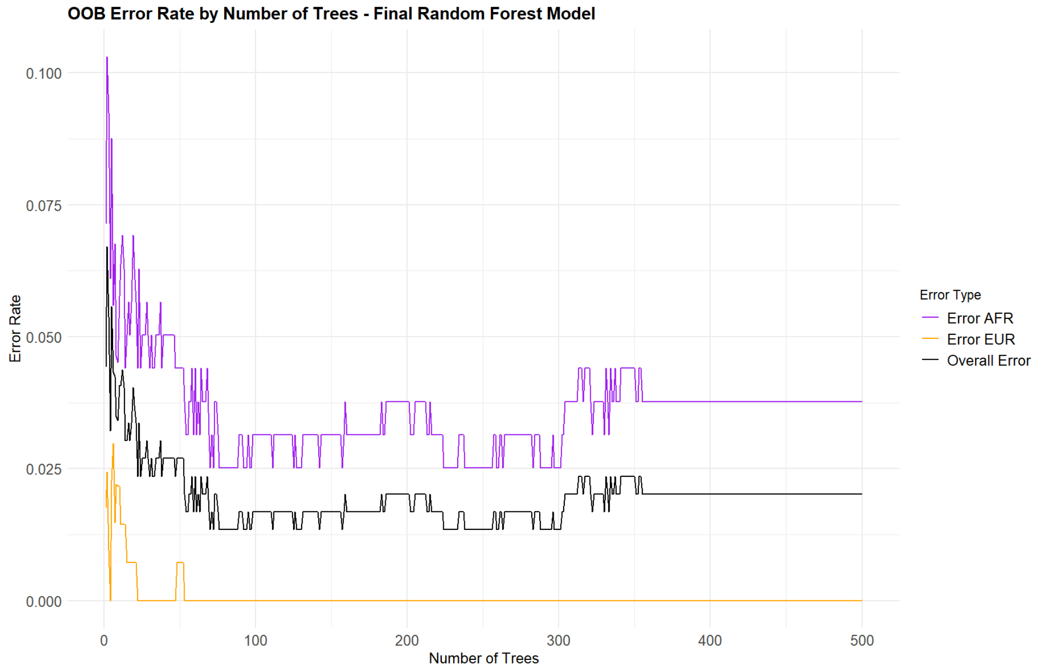

Figure 5: *Out-of-bag (OOB) error rate versus number of decision trees for final optimized randomForest classification model.*

5. DISCUSSION

5.1 Population Clustering

The PCA allowed us to visualize the basic structure of the data as the plot shows some clear clustering of the data based on super-populations. However, it should be noted that PC1 and PC2 account for only 4.35% and 3.67% of the variability in the data. This is underscored by the scree plot (refer to Figure 1), which illustrates that the majority of the dataset's variance remains unaccounted for by PC1 and PC2. This is likely due to the constraints of the dataset's ordinal nature, where values are confined to a narrow range of 0, 1, or 2, therefore PCA was unable to effectively reduce the dimensions of the highly dimensional dataset used in this analysis.

The hierarchical clustering methods, despite employing various linkage strategies, faced challenges in effectively partitioning the dataset into cohesive clusters. Notably, complete linkage emerged as the most promising method, yielding distinguishable clusters. However, upon examination of the output (refer to Table 2), it becomes apparent that these clusters failed to align with either the overarching super-populations or their respective subpopulations. The inability of the clustering algorithms to discern meaningful patterns within the data points to the complexity inherent in the dataset's structure, but could also suggest an incompatibility between the data format and the hierarchical clustering algorithms.

5.2 Population Classification

KNN and RF provided insight into the accuracy of predicting super-populations from the variants. The accuracy of KNN was heavily dependent on the number of k as shown in **Figure #,** there is a clear inverse relationship between the value of k and the accuracy of the classifier. The best result was at the point k = 1 to the potential of nearest-neighbour influence in genetic classification but the quality of our dataset can expose the KNN model to overfitting risk due to a drop in accuracy at larger k values. The accuracy of the KNN model is 77%. The confusion matrix shows that while the model perfectly identified all 52 EUR samples, it struggled with AFR super-population, misclassifying 23 out of 48 AFR samples as EUR. This can indicate that EUR samples have a similarity with AFR samples.

In contrast, RF classification demonstrates more robustness in classifying super-population. The initial model fit with default parameters exhibited an accuracy of 96% in classifying super-population on the testing data, which was improved to 98% with ‘mtry’ parameter tuning, correctly predicting 52 out of 54 EUR samples and 44 out of 46 AFR samples. **Figure 5** demonstrates a plateau in OOB error highlighting an optimal number of trees at around 400, beyond which improvement in the model is minimal. Of note are differential OOB error rates between EAU and AFR super populations, with higher error occurring in AFR. This may reflect different levels of model accuracy in classifying these populations owing to several factors, such as a lower sample size or greater genetic diversity within the AFR population. In general, however, the RF classifier performed well and demonstrated a greater ability to determine super-population class from genetic variation in lactose intolerance genes as compared to KNN.

5.3 Limitations and Future Considerations

Given the focus on numerous variants within a few genes linked to lactose intolerance, we chose not to include linkage disequilibrium (LD) analysis in our filtering approach. Due to the close proximity of these variants within each gene, a high likelihood of LD among them is anticipated. Although our analysis does not aim to differentiate the impacts of specific SNPs, define haplotype structures, or identify recombination events, employing LD-based filtering could potentially have enriched our understanding of genetic population structure related to the genes under study. Another limitation is a limited investigation of model overfitting, ,specifically with the application of cross-validation and training versus testing performance metrics.

Future studies might benefit from broadening the scope to include more genes and variants impacting lactose metabolism, alongside conducting detailed linkage disequilibrium analyses as discussed above. In addition, the sample size was relatively low and future studies should expand beyond 1000 Genomes data as well as leverage further genotyping efforts in understudied populations for which substantial genetic diversity is observed. Finally, this analysis inherently fails to account for the interactions between genetic predispositions and environmental factors as they pertain to lactose persistence phenotypes; accordingly, genotypic and phenotypic assumptions are greatly simplified and the findings should be interpreted accordingly.

6. REFERENCES

1. Itan, Y., Jones, B. L., Ingram, C. J., Swallow, D. M., & Thomas, M. G. (2010). A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evolutionary Biology*, *10*(1), 36. https://doi.org/10.1186/1471-2148-10-36
2. Anguita-Ruiz, A., Aguilera, C. M., & Gil, Á. (2020). Genetics of lactose intolerance: An updated review and Online Interactive World Maps of phenotype and genotype frequencies. *Nutrients*, *12*(9), 2689. https://doi.org/10.3390/nu12092689
3. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007 Jan;39(1):31-40. doi: 10.1038/ng1946. Epub 2006 Dec 10. PMID: 17159977; PMCID: PMC2672153.
4. Vuorisalo T, Arjamaa O, Vasemägi A, Taavitsainen JP, Tourunen A, Saloniemi I. High lactose tolerance in North Europeans: a result of migration, not in situ milk consumption. Perspect Biol Med. 2012;55(2):163-74. doi: 10.1353/pbm.2012.0016. PMID: 22643754.
5. Silanikove N, Leitner G, Merin U. The Interrelationships between Lactose Intolerance and the Modern Dairy Industry: Global Perspectives in Evolutional and Historical Backgrounds. Nutrients. 2015 Aug 31;7(9):7312-31. doi: 10.3390/nu7095340. PMID: 26404364; PMCID: PMC4586535.
6. Ransome-Kuti, O., Kretchmer, N., Hurwitz, R. *et al.* Absorption of lactose by various Nigerian ethnic groups. *Pediatr Res* 5, 388–389 (1971). https://doi.org/10.1203/00006450-197108000-00074
7. Fan S, Kelly DE, Beltrame MH, Hansen MEB, Mallick S, Ranciaro A, Hirbo J, Thompson S, Beggs W, Nyambo T, Omar SA, Meskel DW, Belay G, Froment A, Patterson N, Reich D, Tishkoff SA. African evolutionary history inferred from whole

genome sequence data of 44 indigenous African populations. Genome Biol. 2019 Apr 26;20(1):82. doi: 10.1186/s13059-019-1679-2. Erratum in: Genome Biol. 2019 Oct 9;20(1):204. PMID: 31023338; PMCID: PMC6485071.

8. 1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature*. 2015 Oct 1;526(7571):68-74. doi: 10.1038/nature15393. PMID: 26432245; PMCID: PMC4750478.
9. López, J., & Maldonado, S. (2018). Redefining nearest neighbour classification in high-dimensional settings. *Pattern Recognition Letters*, *110*, 36–43. https://doi.org/10.1016/j.patrec.2018.03.023