BINF 6970 – Assignment 3

Name: Abelhard Jauwena

Student ID: 1040413

Background

Cancer is a disease caused by both genetic and environmental factors. The environmental factors that contribute to cancer are well understood and include cigarette smoking, alcohol consumption, stress, and a sedentary lifestyle, among others (Anand et al., 2008). Similarly, research in population genetics has also helped elucidate the genetic risk factors for cancer, most of which involve single nucleotide polymorphisms (SNPs) (Huang et al., 2015). Three of the most prominent SNPs associated with the onset of cancer occur in the zinc finger protein 23 (ZNF23), family with sequence similarity 135 member B (FAM135B), and RNA binding motif protein 38 (Rbm38) genes (Huang et al., 2015). ZNF23 has been shown to induce apoptosis in ovarian cancer cells, and so mutations that downregulate ZNF23 activity may promote ovarian cancer cell survival (Huang et al., 2008). Additionally, ZNF23 may also play a role in the development of hepatocellular carcinoma (Shi et al., 2011). On the other hand, mutations in the FAM135B gene may contribute to the onset of esophageal squamous cell carcinoma (ESCC) (Song et al., 2014). Finally, Rbm38 modulates the expression of p53, a nuclear transcription factor that induces apoptosis in cells with DNA damage (Huang et al., 2015; Ozaki & Nakagawara, 2011). As such, mutations in the Rbm38 gene may inhibit tumor suppression (Xue et al., 2014).

Data from ethnically diverse countries – such as the United States (US), Canada, and England – show that various cancers seem to affect some superpopulations more than others (*Cancer Stat*, n.d.; Delon et al., 2022; Hwee & Bougie, 2021). In the US, for instance, non-Hispanic Black males have the highest rates of new cancer diagnoses and deaths, while non-Hispanic Asian and Pacific Islander males have the lowest (*Cancer Stat*, n.d.). Conversely, in Canada and England, both White males and females have the highest cancer incidence rates (Delon et al., 2022; Hwee & Bougie, 2021). Moreover, the lowest cancer incidence rates in Canada and England occur in the South Asian and Mixed/Multiple ethnic groups, respectively (Delon et al., 2022; Hwee & Bougie, 2021). Granted, most of these discrepancies may be

attributed to lifestyle differences, as only 5-10% of cancer cases can be ascribed to genetic mutations (Anand et al., 2008). Nevertheless, findings from genome-wide association studies (GWAS) have identified more than 700 risk loci for cancer, which are found to be unevenly distributed across superpopulations (Park et al., 2018). Specifically, around 80% of these loci were discovered in European descendants, 15% in East Asian descendants, 3% in multiethnic descendants, and less than 1% in African and Latin descendants (Park et al., 2018). It might therefore be useful to investigate whether SNPs in these loci can act as markers for differentiating individuals belonging to different superpopulations.

One of the ways we can distinguish individuals in different superpopulations based on SNPs is clustering – a technique used in machine learning for grouping points in a data set into clusters using some similarity measure (Omran et al., 2007). Broadly speaking, clustering techniques can be categorized as either hierarchical or partitional (Omran et al., 2007). Hierarchical clustering techniques focus on generating a "clustering tree" or "dendogram," which is a visual representation showing how related the clusters are to each other (Omran et al., 2007; Zappia & Oshlack, 2018). Hierarchical clustering techniques can also be further divided into two categories based on which methods they use to generate dendograms; divisive clustering algorithms use splitting methods, while agglomerative clustering algorithms use merging methods (Omran et al., 2007). On the other hand, partitional clustering techniques focus on iteratively assigning a set of data points into a specified number of clusters (Kutbay, 2018). Some examples of partitional clustering techniques include the K-means, fuzzy C-means, Gaussian Expectation-Maximization (EM), and Partitioning Around Medioids (PAM) algorithms (Domino, 2019; Omran et al., 2007). Nonetheless, despite their different approaches, all clustering techniques aim to assign points in a data set into distinct groups.

Objectives

The objectives of this report are twofold. First, this report aims to investigate whether SNPs in the ZNF23, FAM135B, and Rbm38 genes can be used as markers to differentiate between individuals belonging to the African and East Asian superpopulations. Second, this report aims to compare the efficacies of four clustering algorithms in differentiating between individuals belonging to the African and East Asian superpopulations using SNPs in these genes.

Specifically, I will compare the efficacies of agglomerative hierarchical, divisive hierarchical, K-means, and PAM clustering algorithms.

I hypothesize that the SNPs in these genes can indeed be used to distinguish individuals belonging to the African and East Asian superpopulations, and that all clustering algorithms represent viable options for carrying out this analysis.

Details and Description of Data

I obtained the data in two steps. Firstly, I obtained the positions of the ZNF23, FAM135B, and Rbm38 genes in the human genome from the National Center for Biotechnology Information (NCBI). Secondly, I inputted these positions into the "Data Slicer" tool in the "GRCh37" human genome assembly to obtain information for all three genes for individuals belonging to both the African and East Asian superpopulations (*Get to*, 2018). The information for all these genes were formatted as gzipped VCF files. As such, I downloaded a total of six gzipped VCF files, where each file contains information for one gene in one of the superpopulations. After all filtering steps, the data set contains a total of 1,052 individuals from both the African and East Asian superpopulations.

Data Analysis Methods

The R script I used to conduct my analyses is submitted along with this report in the CourseLink dropbox for this assignment.

I split my analysis into three parts, all of which were done in R. In the first part, I obtained the Hardy-Weinberg Equilibrium (HWE) values for the ZNF23, FAM135B, and Rbm38 genes. In the second part, I processed the data so that it is ready for use in clustering analyses. In the third part, I performed clustering on all individuals based on SNP data for the three genes. Again, I used the agglomerative hierarchical, divisive hierarchical, K-means, and PAM clustering algorithms to group these individuals into either the African or East Asian superpopulations, and I evaluated the performances of these algorithms by looking at their silhouette coefficients.

Part 1: Obtaining HWE Values

I first used the "readVcf" function to read in all gzipped VCF files into R as "CollapsedVCF" objects. Afterward, I used the "snpSummary" function to count the SNP distribution statistics for the three genes in each superpopulation, and then output the results as data frames. I then combined the data frames for each superpopulation by genes, obtaining a total of three data frames. Finally, I used these data frames to create a table displaying the HWE values for each gene (see *Tab. 1* below).

Part 2: Processing the Data

To begin preparing the data for clustering analyses, I reread the gzipped VCF files into R as "vcfR" objects using the "read.vcfR" function. Then, I used the "extract.gt" function to extract the numerically encoded genotypes of the three genes and output the results as matrices. I also added appropriate suffixes to the columns of these matrices to distinguish between individuals belonging to the two superpopulations, where "_AFR" denotes "African" and "_EAS" denotes "East Asian." After adding these suffixes, I combined the matrices for each superpopulation by genes and converted the resulting outputs into data frames. After going through all these steps, I obtained a total of three data frames (i.e., one for each gene).

I proceeded by calculating the allele frequencies in all three data frames and filtering out sites that have allele frequencies of less than 0.001 as well as any NAs. Afterward, I combined all three data frames back into a single matrix, then transposed the matrix such that the chromosome sites make up the column names and the individual IDs make up the row names. I also omitted any columns in the matrix that have a variance of zero to allow for clustering analyses. Lastly, I calculated the means and standard deviations of the data contained in the matrix, then scaled the matrix using those means and standard deviations. The data is now ready for use in clustering analyses.

Part 3: Performing Clustering and Evaluating the Algorithms

I first used the average silhouette method to find out the optimal number of clusters that each algorithm should group the data points by. I chose the average silhouette method over the elbow method because it is more reliable for finding the optimal number of clusters (the elbow method relies on empirical observations) (*Hierarchical Cluster*, n.d.; Kumar, 2020). According

to this method, the optimal number of clusters is two (see *Fig. 1* below). Moreover, I also generated a scaled dissimilarity matrix, which acts as an input for all the four clustering algorithms used in my analysis.

Before running the agglomerative hierarchical clustering algorithm, I computed the coefficients for various agglomerative methods. The generalized average method ("gaverage") showed the highest agglomerative coefficient; unfortunately, it is not available as input to the "eclust" function (i.e., a hierarchical clustering function). As such, I chose Ward's method instead, which has the second highest agglomerative coefficient and is a viable input. I then performed agglomerative hierarchical clustering using the "eclust" function with Ward's method, specifying that I need the function to group the data points into two clusters. Finally, I visualized the clustering results using a scatter plot and validated the clusters using a silhouette plot.

The divisive hierarchical, K-means, and PAM clustering algorithms do not require a prespecified method. Therefore, I only provided the desired number of clusters into the "eclust" function to run these algorithms. Similarly, I visualized the clustering results and validated the clusters using scatter plots and silhouette plots, respectively.

Summary of Results

HWE Values

	Gene	HWE_Count
1	Zinc Finger Protein 23	316
2	Family With Sequence Similarity 135 Member B	8666
3	RNA Binding Motif Protein 38	343

Tab. 1. The HWE Values for the ZNF23, FAM135B, and Rbm38 genes in individuals belonging to the African and East Asian Superpopulations

Tab. 1 shows that the FAM135B gene exhibits the highest variation among the three genes (Abramovs et al., 2020). The ZNF23 and Rbm38 also exhibit similar levels of variation.

The Optimal Number of Clusters

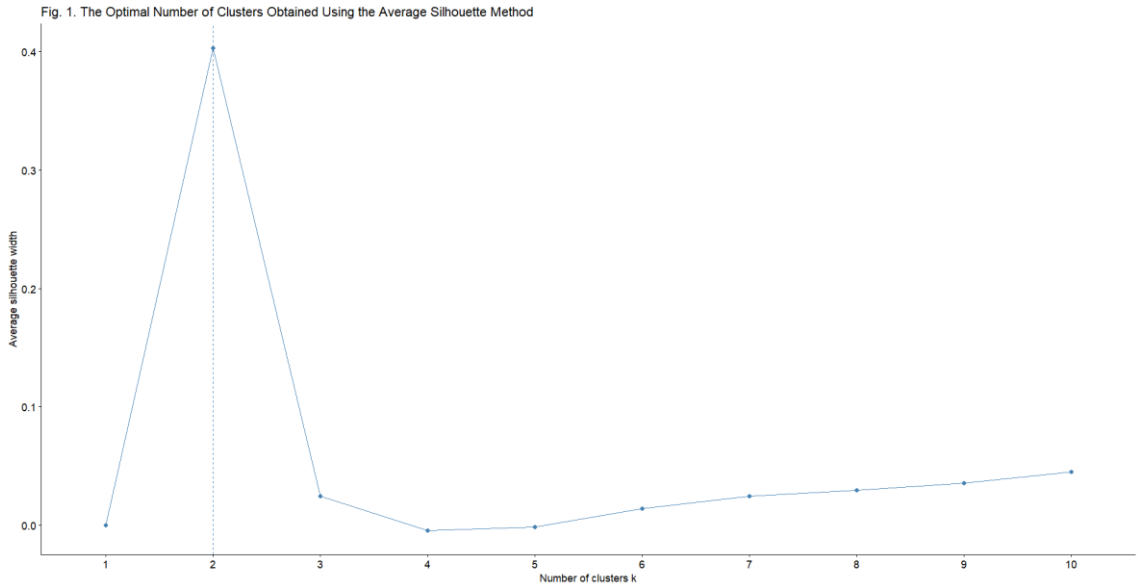

Fig. 1. The Optimal Number of Clusters Obtained Using the Average Silhouette Method

Fig. 1 shows that the optimal number of clusters is two according to the average silhouette method.

Clusters Obtained Using Each Algorithm

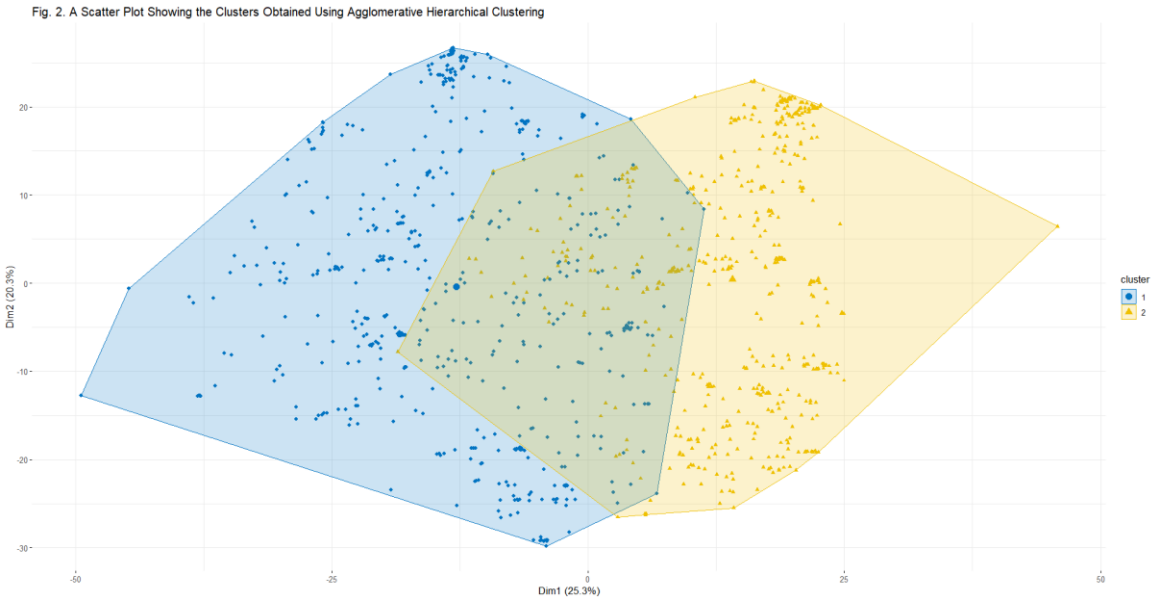

Fig. 2. A Scatter Plot Showing the Clusters Obtained Using Agglomerative Hierarchical Clustering

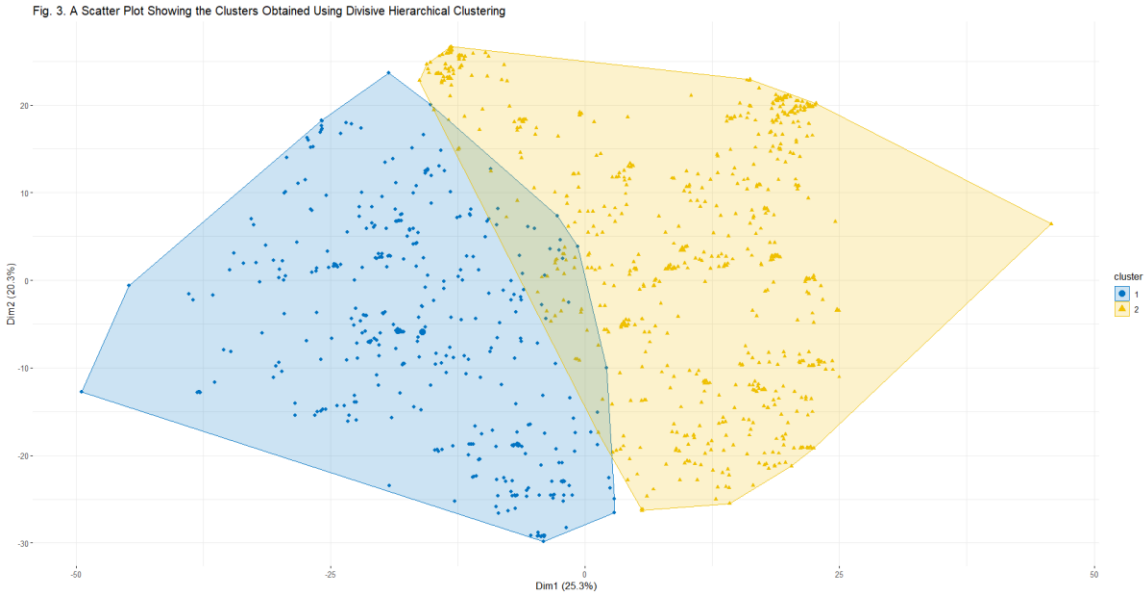

Fig. 3. A Scatter Plot Showing the Clusters Obtained Using Divisive Hierarchical Clustering

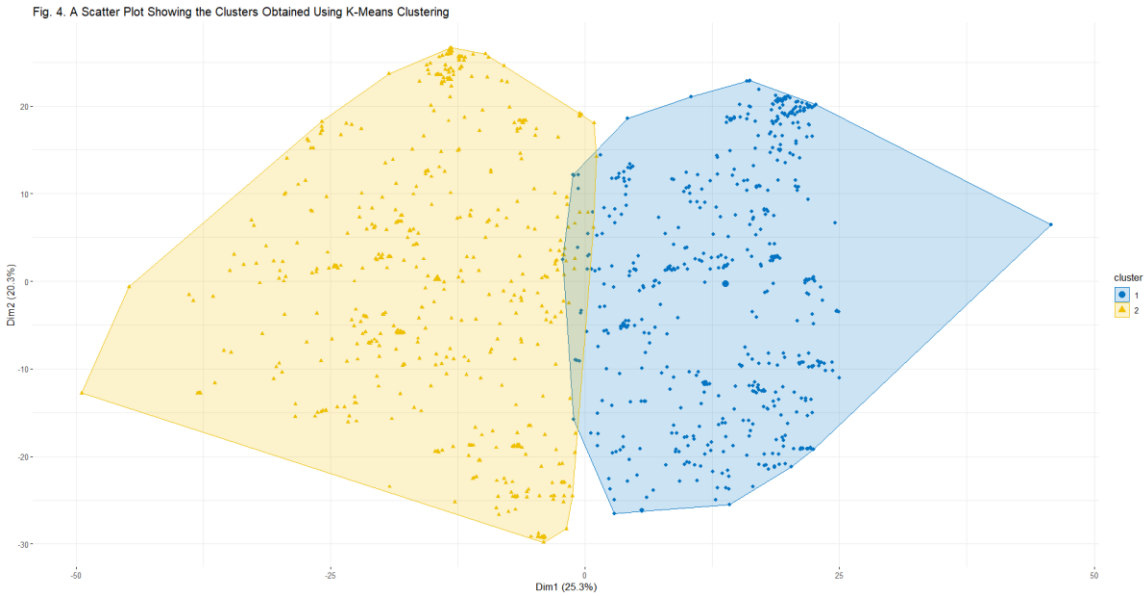

Fig. 4. A Scatter Plot Showing the Clusters Obtained Using K-Means Clustering

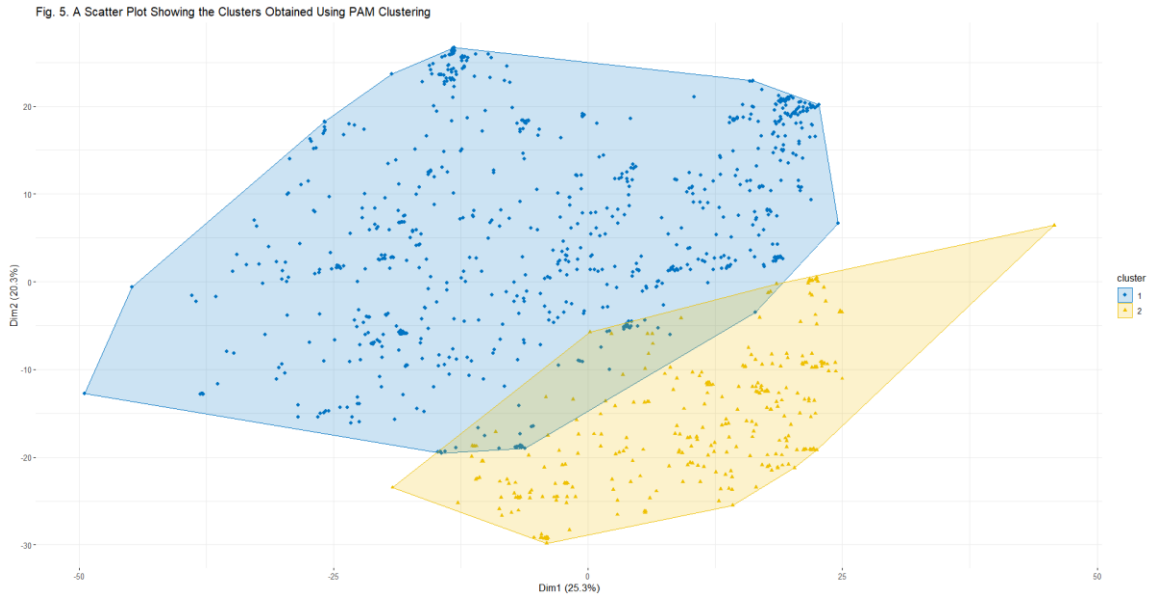

Fig. 5. A Scatter Plot Showing the Clusters Obtained Using PAM Clustering

Fig. 2, *Fig. 3*, *Fig. 4*, and *Fig. 5* display the clusters obtained using the agglomerative hierarchical, divisive hierarchical, K-means, and PAM clustering algorithms. The clusters produced by all four algorithms show clear separations. Nonetheless, the clusters produced using agglomerative hierarchical clustering overlap quite significantly.

Evaluations of Each Algorithm

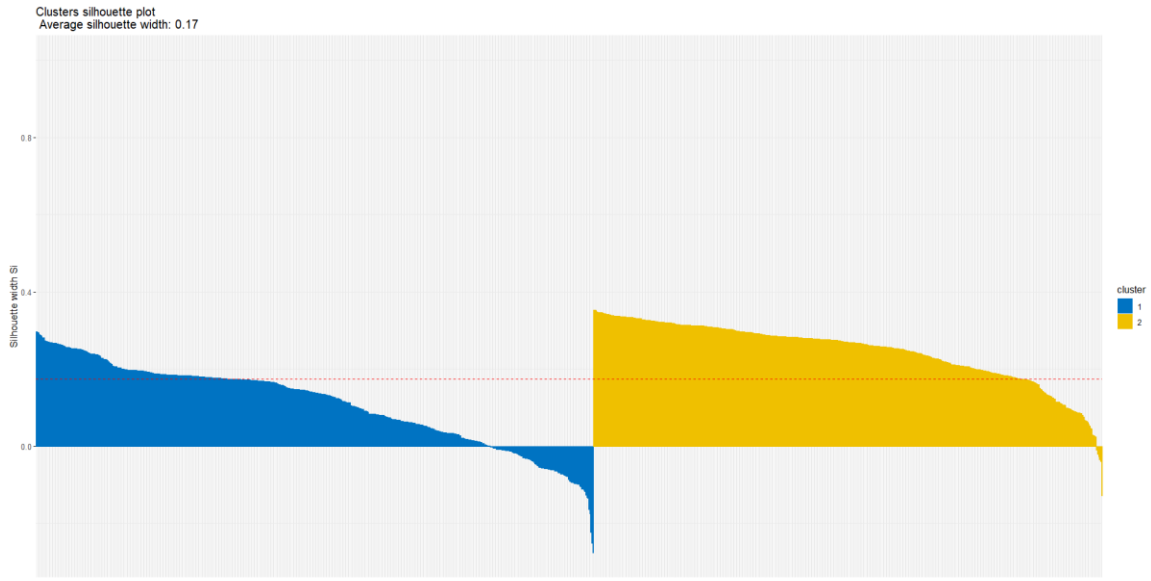

Fig. 6. A Silhouette Plot Showing the Average Silhouette Coefficient of the Clusters Obtained Using Agglomerative Hierarchical Clustering

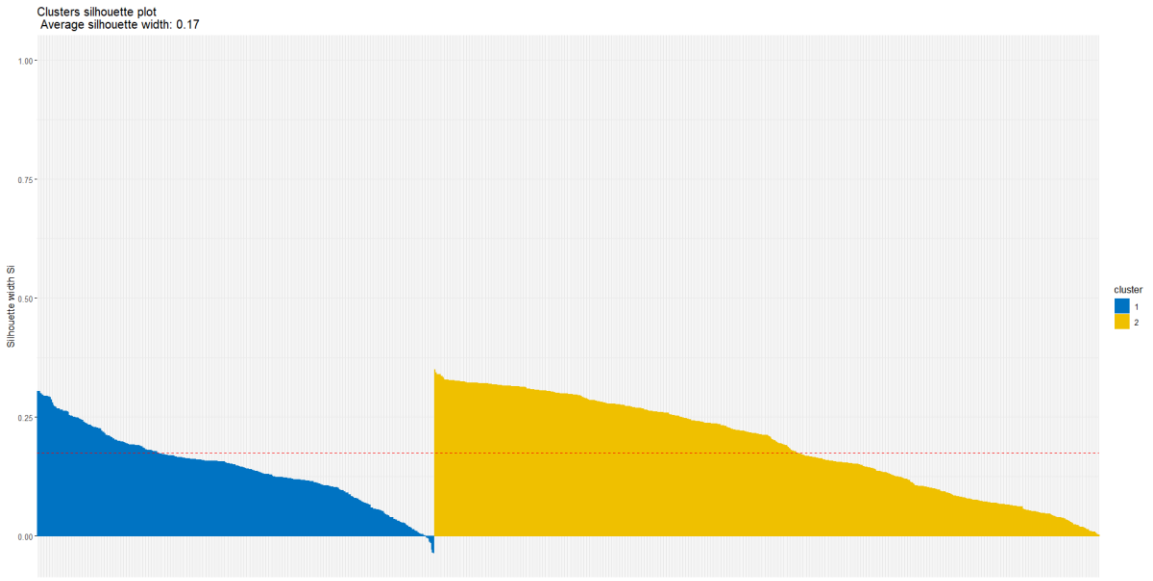

Fig. 7. A Silhouette Plot Showing the Average Silhouette Coefficient of the Clusters Obtained Using Divisive Hierarchical Clustering

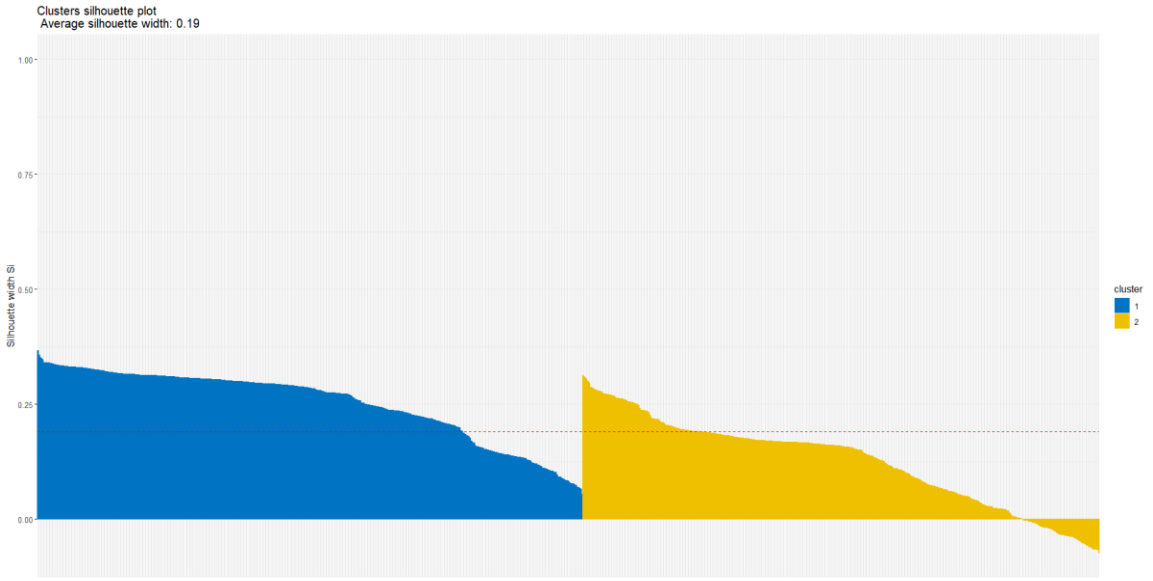

Fig. 8. A Silhouette Plot Showing the Average Silhouette Coefficient of the Clusters Obtained Using K-Means Clustering

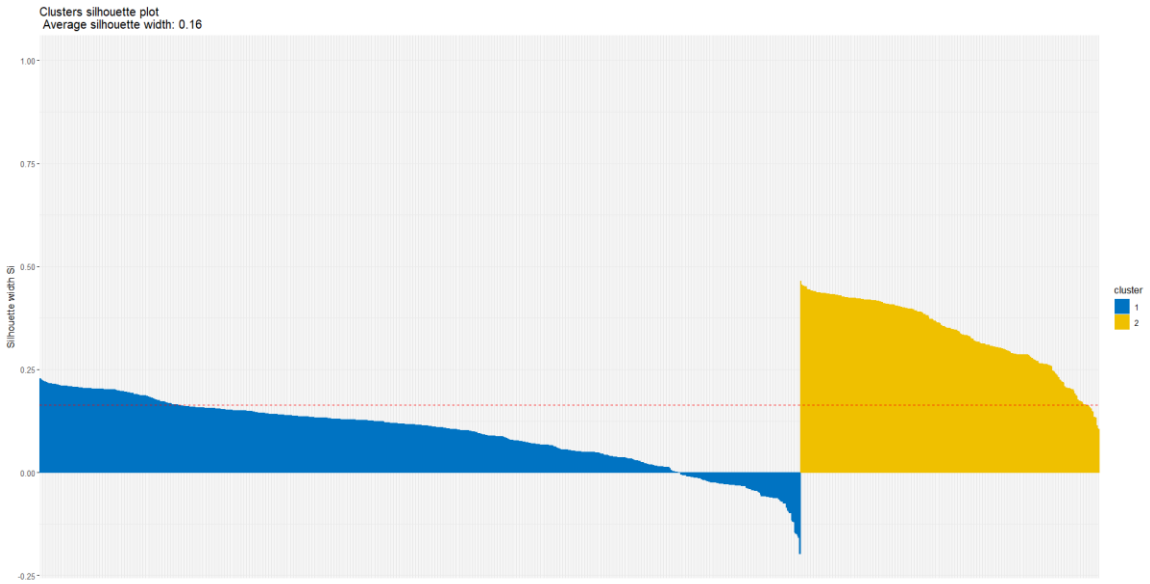

Fig. 9. A Silhouette Plot Showing the Average Silhouette Coefficient of the Clusters Obtained Using PAM Clustering

Fig. 6, *Fig. 7*, *Fig. 8*, and *Fig. 9* show that all four clustering algorithms had low silhouette coefficients. Specifically, both the agglomerative and divisive hierarchical clustering algorithms had a coefficient of 0.17, the K-means clustering algorithm had a coefficient of 0.19, and the PAM clustering algorithm had a coefficient of 0.16.

Discussions and Conclusions

The SNPs in the ZNF23, FAM135B, and Rbm38 genes prove to be useful markers for distinguishing between individuals belonging to the African and East Asian superpopulations. The clusters produced by all the four clustering algorithms show a clear separation between them (see *Fig. 2*, *Fig. 3*, *Fig. 4*, and *Fig. 5*). These three genes are linked to the development of various kinds of cancers; as such, the findings from this report corroborate the results from various surveys, which show that cancer affects certain superpopulations more than others (*Cancer Stat*, n.d.; Delon et al., 2022; Hwee & Bougie, 2021). However, there is currently no consensus as to whether the incidence of cancer is higher in the African or East Asian superpopulation globally. Future studies can therefore investigate whether there are discrepancies in cancer incidences and mortality rates between these two superpopulations, seeing that there seems to be quite significant variations in the ZNF23, FAM135B, and Rbm38 genes between them.

Surprisingly, none of the clustering algorithms tested in this report showed satisfactory performance when clustering individuals into either the African or East Asian superpopulations based on the SNPs in the ZNF23, FAM135B, and Rbm38 genes. The silhouette coefficients for all four algorithms are close to zero (see *Fig. 6*, *Fig. 7*, *Fig. 8*, and *Fig. 9*), which indicates that the data points are not well clustered and that the two clusters produced by each algorithm are not significantly far apart (Bhardwaj, 2020; Quinn, n.d.). These low silhouette coefficients do not at all reflect the disparate clusters produced by all four algorithms; therefore, it may be that there are data processing- or code-related mistakes that significantly hamper the performances of these algorithms. Future studies should reinvestigate the efficacies of these four algorithms in clustering data points while making sure to rigorously process the data and use the appropriate R functions for the correct tasks.

References

Abramovs, N., Brass, A., & Tassabehji, M. (2020). Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Frontiers in Genetics, 11*. **https://doi.org/10.3389/fgene.2020.00210**

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B., & Aggarwal, B. B. (2008). Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharmaceutical Research, 25*(9), 2097–2116. **https://doi.org/10.1007/s11095-008-9661-9**

Bhardwaj, A. (2020, May 26). *Silhouette Coefficient*. Towards Data Science. **https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c**

Cancer Stat Facts: Cancer Disparities. (n.d.). SEER. Retrieved April 4, 2023, from **https://seer.cancer.gov/statfacts/html/disparities.html**

Delon, C., Brown, K., Payne, N., Kotrotsios, Y., Vernon, S., & Shelton, J. (2022). Differences in cancer incidence by broad ethnic group in England, 2013–2017. *British Journal of Cancer, 126*(12), 1765–1773. **https://doi.org/10.1038/s41416-022-01718-5**

Domino. (2019, October 10). *Clustering in R*. Domino Data Lab. **https://www.dominodatalab.com/blog/clustering-in-r**

Get to Know Your Reference Genome (GRCh37 vs GRCh38). (2018, April 4). BiteSizeBio. Retrieved April 5, 2023, from **https://bitesizebio.com/38335/get-to-know-your-reference-genome-grch37-vs-grch38/**

Hierarchical Cluster Analysis. (n.d.). UC Business Analytics R Programming Guide. Retrieved April 5, 2023, from **https://uc-r.github.io/hc_clustering**

Huang, C., Yang, S., Ge, R., Sun, H., Shen, F., & Wang, Y. (2008). *ZNF23 induces apoptosis in human ovarian cancer cells. Cancer Letters, 266*(2), 135–143. **https://doi.org/10.1016/j.canlet.2008.02.059**

Huang, T., Shu, Y., & Cai, Y. (2015). Genetic differences among ethnic groups. *BMC Genomics, 16*(1). **https://doi.org/10.1186/s12864-015-2328-0**

Hwee, J., & Bougie, E. (2021). Do cancer incidence and mortality rates differ among ethnicities in Canada? *Health Reports*. **https://doi.org/10.25318/82-003-x202100800001-eng**

Kumar, S. (2020, October 18). *Silhouette Method – Better than Elbow Method to find Optimal Clusters*. Towards Data Science. **https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891**

Kutbay, U. (2018). Partitional Clustering. *InTech EBooks*. **https://doi.org/10.5772/intechopen.75836**

Omran, M. G. H., Engelbrecht, A. P., & Salman, A. A. (2007). An overview of clustering methods. *Intelligent Data Analysis, 11*(6), 583–605. **https://doi.org/10.3233/ida-2007-11602**

Ozaki, T., & Nakagawara, A. (2011). Role of p53 in Cell Death and Human Cancers. *Cancers, 3*(1), 994–1013. **https://doi.org/10.3390/cancers3010994**

Park, S. L., Cheng, I., & Haiman, C. A. (2018). Genome-Wide Association Studies of Cancer in Diverse Populations. *Cancer Epidemiology, Biomarkers & Prevention, 27*(4), 405–417. **https://doi.org/10.1158/1055-9965.epi-17-0169**

Quinn, R. (n.d.). *A Quick Introduction to Clustering in R*. A Quick Introduction to Clustering in R. **https://rstudio-pubs-static.s3.amazonaws.com/375287_5021917f670c435bb0458af333716136.html**

Shi, Y., Zheng, L., Luo, G., Wei, J., Zhang, J., Yu, Y., Feng, Y., Li, M., & Xu, N. (2011). Expression of Zinc Finger 23 Gene in Human Hepatocellular Carcinoma. *Anticancer Research, 31*(10), 3595-3599. **https://ar.iiarjournals.org/content/31/10/3595.long**

Song, Y., Li, L., Ou, Y., Gao, Z., Li, E., Li, X., Zhang, W., Wang, J., Xu, L., Zhou, Y., Ma, X., Liu, L., Zhao, Z., Huang, X., Fan, J., Dong, L., Chen, G., Ma, L., Yang, J., Chen, L., He, M., Li, M., Zhuang, X., Huang K., Qiu, K., Yin, G., Guo, G., Feng, Q., Chen, P., Wu, Z., Wu, J., Ma, L., Zhao, J., Luo, L., Fu, M., Xu, B., Chen, B., Li, Y., Tong, T., Wang, M., Liu, Z., Lin, D., Zhang, X., Yang, H., Wang, J., & Zhan, Q. (2014). Identification of genomic alterations in oesophageal squamous cell cancer. *Nature, 509*(7498), 91–95. **https://doi.org/10.1038/nature13176**

Xue, J., Xia, T., Liang, X., Zhou, W., Cheng, L., Shi, L., Wang, Y., & Ding, Q. (2014). RNA-binding protein RNPC1: acting as a tumor suppressor in breast cancer. *BMC Cancer, 14*(1). **https://doi.org/10.1186/1471-2407-14-322**

Zappia, L., & Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience, 7*(7). **https://doi.org/10.1093/gigascience/giy083**