Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 10, Number 1, 2026
DOI: 10.30829/zero.v10i1.28505
Page: 157-167

Family Clustering Based On Y-Chromosome DNA Profile Using Unweighted Pair Group Method with Arithmetic Mean

[1]Meira Parma Dewi
Department of Mathematic, Universitas Negeri Padang, Padang, 25132, Indonesia

[2] Nurtami Soedarsono
Department of Oral Biology, University of Indonesia, Depok, 16424, Indonesia

Article Info

Article history:

Accepted 20 March 2026

Keywords:

STR DNA Profile;
Y-chromosome;
Unweighted Pair Group Method With Arithmetic Mean;

ABSTRACT

Indonesia is a diverse nation composed of numerous ethnic groups, each with distinct physical and genetic characteristics. Genetic similarities within ethnic populations can be examined through DNA profiling, particularly by analyzing Short Tandem Repeat loci. In Indonesia, DNA profiling has been widely applied in forensic identification and paternity testing. This study focuses on classifying the Javanese population into sub-tribes based on STR profile similarities using divisive hierarchical clustering. The optimal number of clusters was determined by the smallest Sum of Squared Errors (SSE), with the lowest value of 72583.12 and the highest Silhouatte coeffisien value is 0.78, yielding seven sub-tribe clusters. Subsequently, these sub-tribe clusters were further classified into family clusters using Y-chromosome STR (YSTR) data, which traces paternal lineage. The clustering process employed the Unweighted Pair Group Method with Arithmetic Mean, resulting in 21 family clusters. Compared to k-means clustering, divisive clustering produced sub-tribe clusters with more balanced population sizes. The establishment of sub-tribe and family clusters enhances the efficiency of individual identification, as DNA profile matching can be performed at the cluster level rather than across the entire population. This approach provides a more systematic framework for forensic applications and victim identification, particularly in cases involving male individuals where YSTR data is critical.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Meira Parma Dewi,
Department of Mathematic
Universitas Negeri Padang
Email: meiraparma@fmipa.unp.ac.id

1. INTRODUCTION

Indonesia is a diverse nation, each with distinct characteristics. Ethnic groups are groups of populations that share similar physical and genetic traits. The physical characteristics of each group are closely linked to genetic influences. Within the human DNA system, there are loci that map the shortest repeats of nucleotides, known as short tandem repeats (STRs). STRs are an excellent method for identifying DNA profiles because DNA STRs exhibit high levels of variation within a population [1], [2].

Structural methodology is a method used to identify relationships between one object and another within a population. The structural DNA method will identify the relationship between DNA profiles of a group of individuals within an ethnic group in Indonesia [3]. Within an ethnic group, genetic relationships can be identified through similarities in STR values.

In DNA profile identification, previously known as the Combined DNA Index System (CODIS), 16 STR loci were mapped [4], [5]. Loci are DNA markers that indicate the STR value of a DNA profile. However, there are times when 16 STR loci cannot provide biological information about a victim due to insufficient data or mixed interpretations during the identification process. This led to the development of the PowerPlex Fusion 6c System, which contains 27 STR loci which are CSF1PO, FGA, TH01, vWA, D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433 and D21S11 as known as Amelogenin and DYS391 as sex determinant, Penta D, Penta E, D22S1045, TPOX, SE33 Two Y-STRs repeat mutations are the DYS570 and DYS576 loci [5], [6].

The distribution of STR values in DNA profiles within a given population exhibits similarities that define the uniqueness of Indonesian ethnic groups. Several ethnic groups in Indonesia are genetically related [7]. This genetic relationship can arise through intermarriage between ethnic groups. Tracing an individual's origins based on tribal affiliation or closer family relationships is very important. In cases of disasters or crimes where unidentified victims are found, DNA data from the family or tribe is needed as a reference in the biological identification process.

The mathematical problem in this study can be formulated as follows: suppose there are n individuals from the Javanese population, each with a DNA profile represented as an STR vector $x_i=(x_{i1},x_{i2},...,x_{im})$ for $i=1,2,...,n$, where m is the number of observed STR loci. The first objective is to group these individuals into k sub-tribe clusters $(C1,C2,...,Ck)$ using divisive hierarchical clustering, with the optimal criterion being the minimization of the Sum of Squared Errors (SSE). Next, within each sub-tribe cluster, male individuals have YSTR vectors $y_i=(y_{i1},y_{i2},...,y_{ip})$, where p is the number of YSTR loci, and the second objective is to form family clusters using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method. Data will be grouped based on the distance from the individual STR value. Thus, the main mathematical problems of this study are (1) optimizing the number of sub-tribe clusters by minimizing SSE, and (2) forming family clusters based on YSTR data through hierarchical clustering using UPGMA.

The Javanese population is grouped based on the STR values of DNA profiles to obtain clusters grouped based on genetic proximity. Cluster formation can be done using several clustering methods, including k-means and divisive clustering [8] , [9], [10]. These two clustering methods differ in the number of clusters formed and the population size of each cluster. From the formed sub-tribe clusters, family clusters will then be formed based on YSTR using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [11]. UPGMA will measure the similarity of STR values from DNA profiles between individuals in each sub-tribe. This cluster will indicate genetic proximity based on paternal lineage [2], [12], [13].

Extensive research has been conducted on DNA profile similarity, where measuring the similarity between two DNA profiles serves as a method for biological identification of individuals by employing fuzzy similarity measures (FSM) [14], [15]. Previous studies have developed fuzzy models capable of accommodating shifts in STR values within DNA profiles, allowing two STR values—one of which has undergone a shift—to still be categorized as similar if the shift remains within a defined range. Other studies have utilized YSTR markers to assess the similarity of DNA profiles between two male individuals [6]. In this context, FSM has been applied to measure the similarity between the DNA profile of a son and those of his paternal lineage, such as his father, grandfather, and paternal uncles [16]. Additionally, UPGMA has been widely adopted as a method for clustering genetic data [17], [18], [19], [20].

The results of this study are expected to make a significant contribution to the identification of victims of disasters or criminal acts. Building upon these foundations, the present study emphasizes the clustering of DNA profiles based on STR values to identify sub-ethnic groups within the Javanese population that share higher genetic similarity. The absence of a national DNA database in Indonesia poses significant challenges in determining the biological identity of disaster or crime victims when family information is unavailable. By measuring the similarity between the STR values of victims and those of sub-ethnic DNA profiles, this research aims to facilitate biological identification. A strong match with a particular sub-ethnic group can significantly narrow the search process, which can then be refined further using STR DNA profile information from family clusters.

2. RESEARCH METHOD

Clustering DNA profile data presents several computational challenges, particularly when dealing with large-scale forensic or population datasets. The high dimensionality of genetic data, where each individual profile may contain dozens of loci and allelic pairs, requires algorithms capable of handling complex multivariate structures. Moreover, the inherent heterogeneity of STR (Short Tandem Repeat) variation across individuals complicates the definition of consistent genetic distance measures, which are critical for accurate clustering. As datasets

expand to include thousands of individuals, issues of scalability emerge, demanding efficient algorithms that balance computational speed with analytical precision.

A further complication arises from noise and incomplete data. Missing alleles or incomplete loci are common in forensic practice due to technical limitations in DNA extraction and amplification. Such missing values can distort clustering outcomes, leading to biased or unstable groupings if not properly addressed. Consequently, robust computational strategies are essential to ensure that clustering results remain reliable despite data imperfections.

To mitigate these challenges, several approaches can be employed. Statistical imputations, such as Expectation-Maximization or multiple imputation methods, can be used to estimate missing alleles with reasonable accuracy.

2.1 DNA Profile

The DNA profile is an individual DNA structure that describes their biological identity. DNA profiles of somebody consist of 27 loci, each of which maps STR from the specifications of each locus. A person's DNA profile is identified by examining their biological evidence, also known as DNA evidence, which can be obtained from several parts of the body, such as blood, saliva, bones, muscles, sperm, teeth, hair, or body fluids such as urine and sweat [2], [5].

Human DNA profiles can be identified through the examination of STR markers. STR is a repeating pattern of nucleotides that usually consists of 2–6 bases with the same pattern without any other sequence or intervention of a different sequence. Currently, the PowerPlex Fusion 6C System has been developed which contains 27 STR loci, namely CSF1PO, FGA, TH01, vWA, D1S1656, D2S1338, D2S441, D3S1358, D5S818, D7S820, D8S1179, D10S1248, D12S391, D13S317, D1816S5392119, D13S5392119 and known as amelogenin and DYS391 as sex determinants, penta D, penta E, D22S1045, TPOX, SE33 two repeating mutations of Y-STRs, i.e., locus DYS570 and DYS576. Each locus consists of a pair of alleles, each of which is inherited from the biological father and mother.

2.2 Y-Chromosome (YSTR)

YSTR is a Y chromosome found in the human DNA profile in male lineages. YSTR is the STR locus that is only owned by men. Thus, the measurement of DNA profile similarity using YSTR should be conducted on male queries and male references. YSTR analysis is commonly used for forensic purposes, paternity and genealogical DNA testing. YSTR is acquired from the paternal lineage. Therefore, a boy will have the same YSTR value as his biological father and his male sibling [6]. YSTR is the male STR locus. Thus, the measurement of DNA profile similarity using YSTR should be conducted on male queries and male references. The measurement of DNA profile similarity using YSTR was conducted on male queries with reference to the DNA profile of the male victim's biological family. In this case, the reference used was the victim's grandfather's STR DNA profile from the father's side.

The references selected to compare DNA profile data if the similarity is measured based on YSTR are all males who have a biological relationship with the victim from the father's lineage Candidates for reference include the father, brother, father's brother, grandfather from the father's side, male cousin, and nephew. The selection of Y-STR as a variable in family cluster grouping is based on its unique inheritance pattern. Unlike autosomal STRs, which undergo recombination and are inherited from both parents, Y-STRs are transmitted exclusively through the paternal lineage and remain largely unchanged across generations. This stability makes Y-STRs particularly valuable in forensic genetics, as they enable the tracing of male-specific lineages with high reliability. Consequently, the use of Y-STRs facilitates the identification of male individuals in cases where DNA information from immediate family members is unavailable. By complementing autosomal STR analysis, Y-STR clustering enhances the resolution of population substructure and strengthens applications in both criminal investigations and disaster victim identification.

2.3 Unweighted Pair Group Method With Aritmetic Mean

Unweighted pair group method with aritmetic mean (UPGMA) is a method that can be used to cluster data [21], [22], [23]. The UPGMA technique can be classified as a hierarchical clustering technique because it allows us to see the hierarchy of a cluster formed from smaller clusters. To find relationships within data clusters, UPGMA first constructs a dendrogram. A dendrogram is a tree-shaped diagram that depicts the arrangement of clusters.

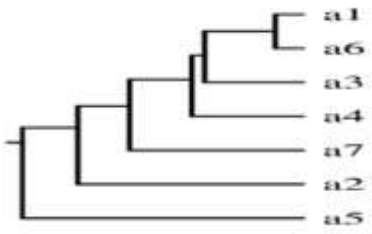

Figure 1. Dendogram

To find relationships within data clusters, UPGMA first constructs a dendrogram. A dendrogram is a tree-shaped diagram that depicts the arrangement of clusters.

$$\frac{1}{|A| .|B|}\sum_{x\in A}\sum_{y\in B} d(x,y) \quad (1)$$

In forming the similarity matrix, the Euclidean distance method will be used

$$d = \sqrt{\sum_{i=1}^{n}(X_{ji-}X_{ki})^2} \quad (2)$$

2.4 Data

This study utilizes DNA profile data obtained from the forensic laboratory of POLRI, comprising 192 individual records. Each profile contains 27 loci, with each locus represented by two alleles. The dataset reflects individuals of Javanese ethnicity, randomly sampled from Central and East Java Provinces. To ensure demographic diversity, samples were collected across mountainous, lowland, and coastal regions. Such geographic variation was intended to capture differences in STR values, thereby improving the clustering quality of the dataset. While the present study focuses on the Javanese population, the approach is adaptable to other ethnic groups in Indonesia and to global populations, offering potential insights into biological relationships among diverse communities. An example of an individual DNA profile is provided in the following table.

Table 1. DNA Profile Dataset

Loci	Alelle 1	Alelle 2
AMELOGENIN	X	Y
D3S1358	15	16
D1S1656	16	17
D2S441	10	11
D10S1248	13	15
D13S317	10	10
Penta E	12	12
D16S539	9	13
D18S51	13	15
D2S1338	17	26
CSF1PO	11	11
Penta D	7	9
TH01	7	10
vWA	17	17
D21S11	30	31.2
D7S820	8	8
D5S818	12	12
TPOX	8	8
D8S1179	11	13
D12S391	19	19
D19S433	13	14
SE33	28.2	28.2
D22S1045	15	16
DYS391	9	9
FGA	22	24.2
DYS576	19	19
DYS570	22	22

2.5 Data Analysis Technique

In this study, the DNA profile under investigation comprises a total of 27 loci. To construct sub-ethnic clusters, 23 loci were selected as the primary variables, resulting in 46 derived features. This selection was intended to maximize discriminatory power while maintaining computational efficiency in cluster formation. In contrast, the construction of family-level clusters utilized Y-chromosomal short tandem repeat (Y-STR) markers, which consisted of 3 loci and generated 6 features. The use of Y-STRs is particularly relevant for tracing paternal lineages, as these markers exhibit male-specific inheritance patterns and provide finer resolution in distinguishing familial relationships. Together, these methodological choices allow for a hierarchical clustering approach, where broader sub-ethnic structures are first identified, followed by more granular family-level groupings.

The proposed method is divided into two stages. The first is to group the Javanese population into several sub-tribe clusters formed using the Divisive clustering method. At this stage, sub-tribe clusters of the Javanese tribe will be generated, where each sub-tribe is a cluster that has a close kinship relationship based on the similarity or proximity of the STR values of 23 DNA profile loci. In the second stage, from each sub-tribe that has been formed, the sub-tribe population will be clustered based on the Y-STR value of the DNA profile of male individuals using the UPGMA clustering method. This clustering aims to build a smallest cluster of the Javanese population that can be called a genetic genealogy (family) of male descendants.

To form a family genetic genealogy DNA cluster, STR values from a set of Y-STR DNA data (haplogroups) are used as the basis for data grouping. The Y-STR loci from the DNA profile consist of loci DYS391, DYS570, and DYS576.

Table 2. YSTR Locus

LOCUS	Alelle	Sampel 1	Sample 2	Sampel 3
DYS391	Alel 1	11	10	10
	Alel 2	11	10	10
DYS570	Alel 1	17	19	19
	Alel 2	17	19	19
DYS576	Alel 1	20	17	15
	Alel 2	20	17	15

2.6 Sum Square Error

The Sum of Squared Errors (SSE) is a mathematical measure used to evaluate the quality of clustering. It calculates how far each data point is from the centroid of its assigned cluster, then sums up all those squared distances. Formally, if we have k clusters (C1,C2,...,Ck) and each cluster has a centroid μj, the SSE is defined as:

$$SSE = \sum_{k=0}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2 \quad (3)$$

where xi represents the data point (in this case, the STR DNA profile of an individual) and μj is the centroid of cluster Cj. A smaller SSE indicates that the data points are closer to their cluster centroids, meaning the clustering is tighter and more representative. Conversely, a larger SSE suggests that the clusters are more dispersed and less effective. In the context of this study, different values of k were tested, and the optimal number of clusters was chosen based on the smallest SSE, it is making this the most efficient grouping of STR DNA profiles for the Javanese population.

3. RESULT AND ANALYSIS

The divisive method is a hierarchical method that aims to add one piece of data to a single data group until all the data is combined into clusters formed from similar characteristics [23], [24]. Two clustering experiments using the divisive method with 192 data sets yielded different numbers of clusters. The first experiment produced 7 subtribe clusters, while the second experiment produced 21 family clusters. To select the best number of clusters, a data accuracy test using the sum of square errors (SSE) will be performed. SSE was chosen as the data accuracy testing technique because the test is conducted through an internal evaluation, meaning the test is conducted only with data within the cluster without any additional external information [25].

In the data validation test with SSE, the best cluster will be selected from the smallest SSE value. From the SSE measurements of few DNA data clustering experiments, the following SSE values were obtained.

Table 3. SSE Score

Experiment Number	Number of Cluster	SSE Score
1	4	89348,09
2	5	73412,32
3	7	72583,12
4	10	97731,25

Table 3 above shows that four experiments yielded SSE values for different numbers of clusters. The smallest SSE value was 72583.12 for a cluster size of 7. Beyond calculating the SSE value, the selection of the optimal number of clusters can also be determined using the silhouette coefficient, which serves to evaluate the quality of the clusters formed. Based on four experiments conducted with varying numbers of clusters, the following SC values were obtained.

Table 4. Silhouatte Coeffisien Score

Number of Clusters	Nilai SC
4	0,51
5	0,39
7	0,78
10	0,47

From the SC values presented in the table above, it can be observed that the cluster with k=7 achieves an SC score of 0.78. This relatively high value indicates that the data points are well grouped within their respective clusters and that the separation between clusters is clear. In other words, the clustering structure formed at k=7 is both distinct and stable, suggesting that this configuration provides the most meaningful representation of the dataset. Compared to lower SC values obtained in other trials, the score of 0.78 reinforces the conclusion that seven clusters yield the optimal balance between cohesion within clusters and separation between them.

The clustering process using the divisive hierarchical clustering method will group DNA STR value data that have similar STR values, with very small distances obtained between one datum and another. The results of measuring the closeness of STR values form a dendrogram as shown in the following figure 2, consisting of 7 clusters.

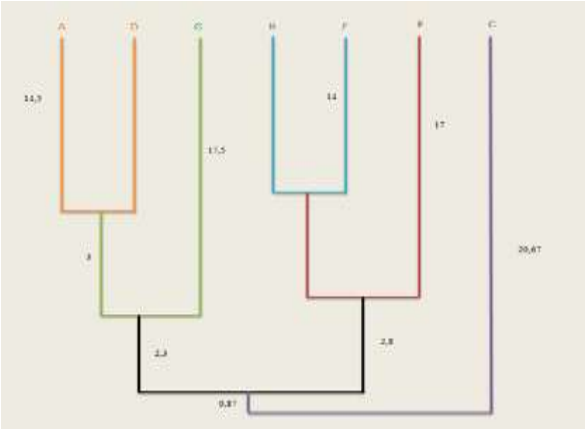

Figure 2. Dendogram for Javanese DNA Profile

Among the seven sub-tribe clusters identified, population sizes exhibited variation as a function of the genetic proximity reflected in the DNA profile short tandem repeat (STR) values. The distribution of individuals within each cluster was determined through the application of k-means and divisive hierarchical clustering algorithms [26], [27], [28]. The resulting population sizes for each sub-tribe cluster are presented below.

Table 5.a. Divisive Clustering results

Cluster	Population Size
A	27
B	31
C	26
D	21
E	23
F	29
G	35
Amount	192

Table 5.b. K-means Clustering Result

Cluster	Population Size
A	20
B	40
C	51
D	11
E	50
F	15
G	5
Amount	192

Based on the distribution of populations within each cluster obtained through the two clustering approaches, notable differences in population balance were observed. The clusters generated using the k-means algorithm exhibited uneven population sizes, indicating that the method tends to partition individuals into groups of varying densities when applied to DNA profile STR values. In contrast, the divisive hierarchical clustering method produced clusters with more uniform population sizes. This outcome suggests that divisive clustering is better suited for capturing the hierarchical genetic relationships among individuals, as it accounts for lineage-based proximity rather than relying solely on centroid-based partitioning. The observed uniformity highlights the influence of hierarchical structures in genetic data, where STR profiles reflect inherited relationships that naturally align with divisive clustering. These findings underscore the importance of selecting clustering algorithms that are consistent with the biological characteristics of the dataset, particularly when analyzing population substructures and familial groupings.

The population in the subtribe clusters formed in the first stage using divisive clustering will then be re-clustered to form genetic genealogical (family) clusters based on the STR values on the Y chromosome (Y-STR). The purpose of forming these family clusters is to group males with similar Y-STR values into a specific group. This is very useful for tracing the family relationships of one male to another.

UPGMA is used as a clustering method to construct a cluster within a sub-tribe of the Javanese tribe based on the STR value on the Y chromosome (Y-STR) in the male population. The formation of new clusters with UPGMA is based on the shortest distance from the distance matrix [23]. The result of combining data based on Y-STR is the formation of clusters that map the family's genetic genealogy where one cluster represents one descendant from the male line.

To form a distance matrix using UPGMA, the mean square deviation method is used. The MSD is used to measure the distance between two research data sets by considering the differences in values from the population or sample data. The MSD method is calculated using the following equation

$$MSD_{jk} = \frac{\sum_{i=1}^{n}(X_{ji}-X_{ki})^2}{n} \quad (4)$$

he sub-tribe clusters that were formed have varying members consisting of male and female individuals. The number of males in each sub-tribe cluster is shown in the following table

Table 6. The number of males in each sub-tribe cluster

Sub-tribe Cluster	Number of Male
A	11
B	13
C	17
D	10
E	12
F	16
G	23

The male population in each sub-tribe cluster will be regrouped to obtain family groups. Each sub-tribe produces a different number of family clusters depending on the similarity of STR values on the male population's Y-STR. In grouping the population based on Y-STR, the dendrogram is limited to level 2. This is aimed at obtaining tighter and more accurate cluster results. From the clustering of population data for each sub-tribe cluster based on Y-STR values, the family clusters formed in each sub-tribe are as follows

Table 7. The number of Family Cluster in each sub-tribe cluster

Sub-tribe Cluster	Number of Family Clusters
A	2
B	3
C	4
D	2
E	3
F	3
G	4

The clustering of populations based on DNA profiles, both at the sub-ethnic and family levels, carries significant implications for the process of individual identification, whether in disaster victim recovery or in cases of violent crime. In this approach, the DNA profile of a victim can be compared against the established sub-ethnic and family clusters to assess genetic similarity. By assigning a similarity score to the victim's DNA profile, the analysis reveals the degree of closeness between the victim and the reference population clusters. This framework not only enhances the accuracy of identification but also provides a systematic method for tracing genetic relationships when immediate family DNA information is unavailable.

The dataset analyzed in this study is composed solely of samples from the pure Javanese population, without inter-ethnic admixture, which constitutes a notable limitation. Intermarriage across ethnic groups can increase variability in STR values and elevate the likelihood of outliers. Outlier dapat disebabkan oleh nilai STR dari data dihasilkan melalui perkawinan antar suku. Outliers can greatly affect the results of hierarchical clustering using the UPGMA method, as they may drastically shift the distances between clusters and make the dendrogram less stable. The presence of one or two extreme data points can alter the tree structure, leading to biased interpretations of relationships among groups. If not addressed, a UPGMA dendrogram may display misleading relationships between populations. Such outliers, when deviating from the main cluster, often form separate clusters, thereby influencing the total number of clusters produced and impacting evaluation metrics such as the Silhouette Coefficient (SC) and Sum of Squared Errors (SSE).

This phenomenon also introduces computational challenges, as the expansion in the number of loci increases the dimensionality of the feature space. To mitigate these issues, dimensionality reduction techniques are expected to enhance clustering performance by reducing redundancy and improving the interpretability of results. Cluster robustness and sensitivity can be evaluated through multiple approaches. Bootstrap resampling, which involves repeatedly reconstructing dendrograms, provides insight into the consistency of clustering outcomes. In addition, the Dunn index serves as a quantitative measure of cluster stability, enabling systematic evaluation of the reliability of the clustering structure. These methods have not been applied in this research, but can be implemented in further research.

This study introduces an innovative approach by clustering DNA profiles based on STR values to identify sub-ethnic groups within the Javanese population. However, several limitations should be noted. The dataset is restricted to samples from Central and East Java, which may not fully capture the genetic diversity of the broader Javanese population or other ethnic groups in Indonesia. The absence of a national DNA database also constrains validation against larger and more heterogeneous datasets. Furthermore, the analysis relies primarily on STR markers, whereas incorporating additional genetic markers such as SNPs or mtDNA could enhance comprehensiveness. In particular, mtDNA, which is maternally inherited, offers distinct advantages for tracing biological identity in female individuals and may serve as a complementary basis for family-level clustering.

Despite these limitations, the novelty of this research lies in its application of fuzzy similarity measures and clustering methods (UPGMA) to accommodate STR value shifts and to identify sub-ethnic genetic similarities. In addition to the current findings, this study can be further developed as a foundation for measuring the similarity between an individual's DNA profile and the DNA profiles of ethnic or sub-ethnic groups. Such an extension would involve the application of probabilistic models derived from population-level DNA data, enabling more rigorous comparisons between individual and group profiles. By incorporating probability-based approaches, the analysis could provide a quantitative measure of genetic similarity, thereby improving both the accuracy and reliability of population-based identification. This methodological framework not only contributes to victim identification in disaster or crime contexts but also opens pathways for exploring biological relationships among ethnic groups in Indonesia and potentially worldwide.

Future research should address these limitations by incorporating more diverse datasets that include individuals with inter-ethnic admixture, thereby capturing the broader genetic variability of the Javanese and other Indonesian populations. The presence of outliers resulting from high STR variation highlights the need for advanced computational strategies. As the number of loci increases, the dimensionality of the feature space expands, posing challenges for clustering accuracy and interpretability. To overcome this, dimensionality reduction methods such as Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), or Uniform Manifold Approximation and Projection (UMAP) could be employed. These techniques

not only reduce redundancy but also enhance the visualization of genetic relationships, thereby improving the robustness of clustering outcomes. Integrating such approaches with STR, SNP, and mtDNA markers would provide a more comprehensive framework for population clustering and forensic identification, particularly in contexts where genetic diversity and admixture are significant.

4. CONCLUSION

The formation of Javanese sub-ethnic groups was carried out based on the similarity of STR DNA profiles among individuals within the population. The grouping process employed a divisive clustering method (top-down approach) to obtain smaller sub-ethnic clusters. Subsequently, the resulting sub-ethnic clusters were regrouped using Y-STR markers to identify family clusters along the paternal lineage, applying the UPGMA clustering method. By forming family clusters based on male lineage, it is hoped that it will facilitate the identification of an individual based on the STR value of the DNA profile, especially using the Y-chromosome (YSTR). YSTR is a chromosome that is passed down to male offspring without experiencing changes, so it is very effective for finding the biological relationship of a male individual with his biological male family. From 192 Javanese DNA samples, seven sub-ethnic clusters and twenty-one family clusters were generated. The determination of the optimal number of sub-ethnic clusters was based on the SSE value of 72583.12 and a Silhouette Coefficient of 0.78. The subdivision of the population into sub-ethnic and family clusters is expected to facilitate the identification of disaster or crime victims in cases where DNA profile information from the victim's immediate family is unavailable. Future studies should expand DNA clustering beyond Y-chromosome markers by including mtDNA and considering inter-ethnic admixture. This broader approach will improve clustering accuracy and provide deeper insights into the genetic diversity and interconnectedness of Indonesia's populations.

ACKNOWLEDGEMENT

The author wishes to convey appreciation to the forensic laboratory of the Indonesian National Police (POLRI) for supplying the DNA profile dataset. In addition, the author gratefully acknowledges all individuals and institutions whose support facilitated the successful completion of this research and the preparation of this article

5. REFERENCES

[1] W. Goodwin Editor, "Forensic DNA Typing Protocols Second Edition Methods in Molecular Biology 1420." [Online]. Available: http://www.springer.com/series/7651

[2] M. P. Dewi, "Pengukuran Kemiripan Profil DNA Manusia Menggunakan Algoritma Apriori Dan Fuzzy Inference System," *Disertation*, University of Indonesia. 2025.

[3] M. S. Anggreainy, M. R. Widyanto, B. H. Widjaja, and N. Soedarsono, "Gaussian Fuzzy Number for STR-DNA Similarity Calculation Involving Familial and Tribal Relationships," *Adv. Bioinformatics*, vol. 2018, 2018, doi: 10.1155/2018/8602513.

[4] M. P. Dewi, A. M. Arymurthy, S. Setiawan, and N. Soedarsono, "HUMAN DNA PROFILE IDENTIFICATION USING DNA DATABASE SYSTEM," *8th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia, 2022, pp. 1-4, doi: 10.1109/ICST56971.2022.10136305.

[5] J. M. Butler, "FORENSIC DNA TYPING," 2005.

[6] "Forensic Science Regulator Guidance Y-STR Profiling," 2021.

[7] M. P. Dewi, A. M. Arymurthy, S. Setiawan, and N. Soedarsono, "Triangular fuzzy number for similarity measurement of Y chromosome DNA profile," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 519–528, Feb. 2024, doi: 10.11591/eei.v13i1.5304.

[8] V. Ramazanova, M. Sambetbayeva, A. Tokhmetov, Zh. Lamasheva, and S. Serikbayeva, "APPLICATION OF AGGLOMERATIVE CLUSTERING FOR FORMING SKILL COMMUNITIES OF JOB VACANCIES," *Вестник КазУТБ*, vol. 4, no. 25, Dec. 2024, doi: 10.58805/kazutb.v.4.25-646.

[9] F. Haque, "Emission Trajectories in the Trade Context: A Comprehensive Machine Learning Approach Using K-Means and ARIMA," 2025. [Online]. Available: https://journals.e-palli.com/home/index.php/jsere

[10] Y. A. Badr, K. T. Wassif, and M. Othman, "Automatic Clustering of DNA Sequences with Intelligent Techniques," *IEEE Access*, vol. 9, pp. 140686–140699, 2021, doi: 10.1109/ACCESS.2021.3119560.

[11] Y. Li and L. Xu, "Unweighted multiple group method with arithmetic mean," in *Proceedings 2010 IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications, BIC-TA 2010*, 2010, pp. 830–834. doi: 10.1109/BICTA.2010.5645232.

[12] S. Dogan, A. Asic, I. Muhovic, L. Besic, and D. Marjanovic, "Overview of the genetic STR clustering among worldwide human populations," *Int. J. Hum. Genet.*, vol. 14, no. 3–4, pp. 131–142, Sep. 2014, doi: 10.1080/09723757.2014.11886236.

[13] G. Qu, Z. Yan, and H. Wu, "Clover: Tree structure-based efficient DNA clustering for DNA-based data storage," *Brief. Bioinform.*, vol. 23, no. 5, Sep. 2022, doi: 10.1093/bib/bbac336.

[14] C. M. Ruitberg, D. J. Reeder, and J. M. Butler, "STRBase: a short tandem repeat DNA database for the human identity testing community," 2001. [Online]. Available: http://www.cstl.nist.gov/biotech/strbase/.

[15] M. R. Karim *et al.*, "Deep learning-based clustering approaches for bioinformatics," *Brief. Bioinform.*, vol. 22, no. 1, pp. 393–415, Jan. 2021, doi: 10.1093/bib/bbz170.

[16] V. K. Malhotra, H. Kaur, and M. A. Alam, "An Analysis of Fuzzy Clustering Methods," 2014.

[17] O. Eric U. and O. Michael O., "Overview of Agglomerative Hierarchical Clustering Methods," *British Journal of Computer, Networking and Information Technology*, vol. 7, no. 2, pp. 14–23, Jun. 2024, doi: 10.52589/bjcnit-cv9poogw.

[18] S. Sivasankari and D. S. Sukumaran, "Fuzzy Membership Partition Based Effective Hierarchical Agglomerative Flat Clustering Method for High Dimensional Data," 2024.

[19] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (N. Y).*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[20] G. J. Hua, C. L. Hung, C. Y. Lin, F. C. Wu, Y. W. Chan, and C. Y. Tang, "MGUPGMA: A Fast UPGMA Algorithm With Multiple Graphics Processing Units Using NCCL," Oct. 03, 2017, *SAGE Publications Ltd.* doi: 10.1177/1176934317734220.

[21] R. Tibshirani, T. Hastie, M. Eisen, D. Ross, D. Botstein, and P. Brown, "Clustering methods for the analysis of DNA microarray data," 1999.

[22] L. Fu and E. Medico, "FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data," *BMC Bioinformatics*, vol. 8, Jan. 2007, doi: 10.1186/1471-2105-8-3.

[23] K. A. Jain and R. C. Dubes, "Algorithm for Clustering Data," 1988.

[24] N. I. Boyko and O. A. Tkachyk, "Hierarchical clustering algorithm for dendrogram construction and cluster counting," *Informatics and mathematical methods in simulation*, vol. 13, no. 1–2, pp. 5–15, Apr. 2023, doi: 10.15276/imms.v13.no1-2.5.

[25] D. Wei, Q. Jiang, Y. Wei, and S. Wang, "A novel hierarchical clustering algorithm for gene sequences," *BMC Bioinformatics*, vol. 13, no. 1, Jul. 2012, doi: 10.1186/1471-2105-13-174.

[26] X. Xiong, K. L. Chan, and K. L. Tan, "Similarity-Driven Cluster Merging Method for Unsupervised Fuzzy Clustering."

[27] T. Ardyanti and Mhd. Furqan, "Implementation of the Agglomerative Hierarchical Clustering Method in Ordering Hijab Products," *sinkron*, vol. 8, no. 4, pp. 2479–2489, Oct. 2024, doi: 10.33395/sinkron.v8i4.14156.

[28] Etidal Altom Kunna, M. N. Omar, and M. F. Zolkipli, "Recent Advances in DDoS Attacks for SDN Controller: A Review," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 65, no. 1, pp. 1–13, Mar. 2025, doi: 10.37934/araset.65.1.113.