ACTA MEDICA LITUANICA. 2019. Vol. 26. No. 4. P. 211–216


# A comparative analysis of mathematical methods for homogeneity estimation of the Lithuanian population

**Alma Molytė**[1, 2],

**Alina Urnikytė**[1],

**Vaidutis Kučinskas**[1]

[1] *Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, Vilnius, Lithuania*

[2] *Department of Information Systems, Faculty of Fundamentals Sciences, Vilnius Gediminas Technical University, Vilnius, Lithuania*

**Background.** Population genetic structure is one of the most important population genetic parameters revealing its demographic features. The aim of this study was to evaluate the homogeneity of the Lithuanian population on the basis of the genome-wide genotyping data. The comparative analysis of three methods – multidimensional scaling, principal components, and principal coordinates analysis – to visualize multidimensional genetics data was performed. The results of visualization (mapping images) are also presented.

**Materials and methods.** The data set consisted of 425 samples from six ethnolinguistic groups of the Lithuanian population. Genomic DNA was extracted from whole venous blood using either the phenol-chloroform extraction method or the automated DNA extraction platform TECAN Freedom EVO. Genotyping was performed at the Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, Lithuania, with the Illumina HumanOmniExpress-12 v1.1 and the Infinium OmniExpress-24. For the estimation of homogeneity of the Lithuanian population, PLINK data file was obtained using PLINK v1.07 program. The Past3 software was used to visualize the genotype data with multidimensional scaling and principal coordinates methods. The SmartPCA from EIGENSOFT 7.2.1 program was used in the principal component analysis to determine the population structure.

**Conclusions.** Methods of multidimensional scaling, principal coordinate, and principal component for the genetic structure of the Lithuanian population were investigated and compared. The principal coordinate and principal component methods can be used for genotyping data visualization, since any essential differences in the results obtained were not observed and compared to multidimensional scaling. The Lithuanian population is homogenous whereas the points are strongly close when we use the principal coordinates or principal component methods.

**Keywords:** genotyping, multidimensional scaling, principal components, principal coordinate analysis, genotypes data visualization

* Correspondence to: Alma Molytė, Faculty of Medicine, Vilnius University, M. K. Čiurlionio 21, Vilnius 03101, Lithuania.
Email: alma.molyte@mf.vu.lt

## INTRODUCTION

Nowadays population genetic structure is one of the most important parameters in analysing population research. Different genetic models based on genetic markers are used to evaluate the population structure. Appropriate mathematical methods have been developed for different genetic models to obtain information from genetic marker data to explore the population structure.

There is a large class of methods that have been developed for multidimensional data visualization (1, 2). The visual presentation of the data enables seeing the data structure, clusters, outliers, and other properties of multidimensional data. Direct data visualization is a graphical presentation of a data set providing a quality understanding of the information contents in a natural and direct way.

There exist numerous methods that can be used for reducing the dimensionality and particularly for visualizing the *n*-dimensional data: principal component analysis (PCA) (3), multidimensional scaling (MDS) (4), locally linear embedding (LLE) (5), etc. These methods can be used to visualize the data set provided that a sufficiently small output dimensionality ($d = 2$, $d = 3$) is chosen.

In this study, methods of multidimensional scaling, principal coordinates, and principal components were used for detecting population genetic structure and potential outliers of the Lithuanian population. Our core task was to determine the accuracy of each method for genome-wide genotyping data visualization.

## MATERIALS AND METHODS

### Samples and genotyping

The data set consisted of 425 samples from unrelated Lithuanian individuals. The samples were collected randomly from six ethnolinguistic groups of Lithuania: three groups of Aukštaičiai (from the region in the north-eastern part of the country) and three groups of Žemaičiai (from the ethnic region in the north-western part of Lithuania) (Table 1).

**Table 1.** Sample size from six ethnolinguistic groups

| Ethnolinguistic group | Sample size |
|---|---|
| East Aukštaitija (EA) | 79 |
| South Aukštaitija (SA) | 67 |
| West Aukštaitija (WA) | 79 |
| North Žemaitija (NZ) | 79 |
| South Žemaitija (SZ) | 78 |
| West Žemaitija (WZ) | 43 |
| Total | 425 |

Genomic DNA was extracted from whole venous blood using either the phenol-chloroform extraction method or the automated DNA (extraction platform TECAN Freedom EVO (TECAN Group Ltd., Männedorf, Switzerland), based on paramagnetic particle method. DNA concentration and quality were measured by NanoDropR ND-1000 spectrophotometer (NanoDrop Technologies Inc., US).

Genotyping was performed at the Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, Lithuania, with Illumina HumanOmniExpress-12v1.1 (296 samples) and the Infinium OmniExpress-24 (129 samples) arrays (Illumina, San Diego, CA, USA), with overlap of 707,138 SNPs genome-wide distributed. Quality control of the genotyping data was performed according to the manufacturer's standard recommendations. Individuals with call rate <98% and standard deviation (SD) of Log R ratio >0.3 were excluded from further analysis. GenomeStudio v2011.1 program (Illumina, USA) was used to distinguish the genotypes from the sample and to export the data in PED/MAP format.

For the estimation of homogeneity of the Lithuanian population, PLINK data file (binary format) was obtained using PLINK v1.07 program (3). Individuals or SNPs with >10% missing data, minor allele frequency (MAF) <0.01, and Hardy-Weinberg equilibrium (HWE) test *P*-value of less than $10^{-4}$ were excluded. SNPs in linkage disequilibrium were removed with the indep-pairwise option of PLINK v1.07 using a window size of 50 SNPs, a step size of 5, and an $r^2$ threshold of 0.5.

## VISUALIZATION METHODS

In this paper, we performed an analytic investigation of multidimensional scaling, principal components, and principal coordinates methods, which are used for multidimensional data visualization.

If we have the dataset $X = \{X_1, X_2, \ldots, X_m\}$ in the n-dimensional space, where $X_i = (X_{i1}, X_{i2}, \ldots, X_{in})$, $i = 1, \ldots, m$, we desire to get the dataset $Y = \{Y_1, Y_2, \ldots, Y_m\}$ in $d$-dimensional space, where $Y_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $i = 1, \ldots, m$ and $d < n$. If a sufficiently small output dimensionality $d = 2$ or $d = 3$ is chosen, two or three dimensional vectors obtained may be presented in a scatter plot.

The Past3 program was used to visualize the genotype data with multidimensional scaling and principal coordinates methods. The SmartPCA from EIGENSOFT 7.2.1 is one of the basic programs used in the principal component analysis to determine homogeneous or heterogeneous population structure. The method was developed for the samples not related to the population structure.

**Multidimensional scaling**

Multidimensional scaling (MDS) refers to a group of methods that are widely used for dimensionality reduction and visualization of multidimensional data (4). The starting point of MDS is a matrix consisting of pairwise proximities of the data. The proximities are similarity or dissimilarity. The main goal of multidimensional scaling is to find lower-dimensional data $Y_i$, $i = 1, \ldots, m$, such that the distances between the data in the lower-dimensional space were be as close to the original distances (or other proximities) as possible (4). The stress function $E_{MDS}$ must be minimized.

$E_{MDS} = \Sigma_{i<j} w_{ij}(\delta(Y_i, Y_j) - d(Y_i, Y_j))^2$, where $w_{ij}$ is a weight; $\delta(Y_i, Y_j)$ is the value of proximity between the $n$-dimensional data $X_i$ and $X_j$, $d(Y_i, Y_j)$ is the distance (usually, Euclidean) between two-dimensional data $Y_i$ and $Y_j$. If the proximity is the Euclidean distance, then $\delta(Y_i, Y_j) = d(Y_i, Y_j)$. The multidimensional scaling method is based on a distance matrix computed with distance measures. The results of MDS depend on the initial values of two-dimensional vectors if the MDS stress is minimized in an iterative way.

**Principal coordinates analysis**

Principal coordinates analysis (PCO) is another method also known as metric multidimensional scaling. The algorithm is taken from Davis (1986). The main idea of this method is finding the eigenvalues and eigenvectors of a matrix containing the distances or similarities between all data points. Giving a measure of the variance accounted for by the corresponding eigenvectors (coordinates), the eigenvalues are given for the first four most important coordinates (or fewer if there are fewer than four data points). The percentages of variance accounted for by these components are also given (6).

Before eigenanalysis, the values of similarity and distance index values can be raised to the power c. C is the "transformation exponent", which can be 1, 2, 4 and 6 (7). We needed principal coordinates analysis with the standard value $c = 2$.

**Principal components**

Principal components analysis (PCA) is one of the powerful and popular statistical linear projection methods. Linear transformation is widely used for a dimensionality reduction, feature extraction, and visualization of multidimensional data. The main goal of this method is finding the trend with the largest variance. The input data is a matrix of multivariate data, with items in rows and variates in columns. The eigenvectors and their eigenvalues were calculated by the singular value decomposition algorithm (SVD).

Principal component analysis is a tool widely used in genomics and statistical genetics, employed to infer cryptic population structure from genome-wide data such as single nucleotide polymorphisms (SNPs) (8), and/or to identify outlier individuals which may need to be removed prior to further analyses, such as genome-wide association studies (GWAS) (9).

**The similarity and distance indices**

We used 20 similarity and distance indices (Euclidean, Gower, Chord, Manhattan, Bray-Curtis, Cosine, Morisita, Horn, Correlation, Rho, Dice, Jaccard, Ochiai, Kukczynski, Simpson, Hamming/p-distance, Jukes-Cantor, Kimura, User-supplied similarity, User-supplied distance) for the multidimentional scaling method and principal coordinates analysis. These similarities and distance measures are described in the publication PAST PAleontological STatistics Version 3.18 by Øyvind Hammer.

In this paper, we analysed only Euclidean and Gower similarity and distance indices in greater detail since with these similarities we achieved the best possible results. The basic Euclidean distance means the distance between the two points in a plane. Gower is a distance measure that

averages the difference over all variables, each term normalized for the range of $X_k = (x_{k1}, x_{k2}, \ldots, x_{kn})$ and $X_l = (x_{l1}, x_{l2}, \ldots, x_{ln})$ variables calculated as follows:

$$d_{kl} = \frac{1}{2}\sum_i \frac{|x_{ki} - x_{li}|}{\max\limits_s x_{si} - \min\limits_s x_{si}}.$$

The Gower measure is similar to the Manhattan distance but with the range normalization (6).

## RESULTS

The aim of this study was to explore the most suitable method for inferring the Lithuanian population structure using genotype data: multidimensional scaling, principal coordinate, and principal component.

The visualization results of the genotype data of six ethnolinguistic groups of Lithuania visualized by multidimensional scaling methods are presented in Fig. 1, by principal coordinates analysis in Fig. 2, and visualized by principal components in Fig. 3.

In order to estimate the quality of mapping, the stress function was calculated.

The research results show that the stress function values are smaller and approximately equal when we used the similarity of the Euclidean distance (Stress = 1.371) and Gower (Stress = 1.372) than other similarities of multidimensional scaling method (Fig. 1). The stress function values are larger when we used other similarities (except Euclidean and Gower).

The results of visualization obtained by the principal coordinates are presented in Fig. 2. The Coordinate 1 explained 0.57% and Coordinate 2 0.55% of the genetic variation among the studied samples (424) of the Lithuanian population when the similarity index was the Euclidean distance and Coordinate 1 explained 0.77%, Coordinate 2 0.76% of genetic variation, provided that the similarity index was Gower.

The results of the data analysis show that Coordinate 1 explained 11.25% and Coordinate 2 11.11%, when the principal components method was used with the Euclidean distance similarity index (Fig. 3).

The investigation results show that the principal components method is more suitable to analyze the population genetic structure than the methods of multidimensional scaling or principal coordinates. On the other hand, the principal coordinates method is more suitable as compared to the multidimensional scaling (Fig. 1 and Fig. 2).

In Figs.1-3, the West Žemaičiai (WZ) denote a green circle, South Aukštaičiai (SA) – a grey triangle, West Aukštaičiai (WA) – a red star, North Žemaičiai (NZ) – a light brown square, South Žemaičiai (SZ) – a yellow dagger, East Aukštaičiai (EA) – a blue diamond.

It can be seen that since the similarity index is the Euclidean distance, the points obtained by principal coordinate and principal component, are clustered very strongly, but the points obtained by the multidimensional scaling method are dispersed, and it is difficult to evaluate the population structure. It is evident that the outliers are more visible when we use the principal coordinates and principal components methods provided that the similarity index remains the same, i.e., the Euclidean distance.

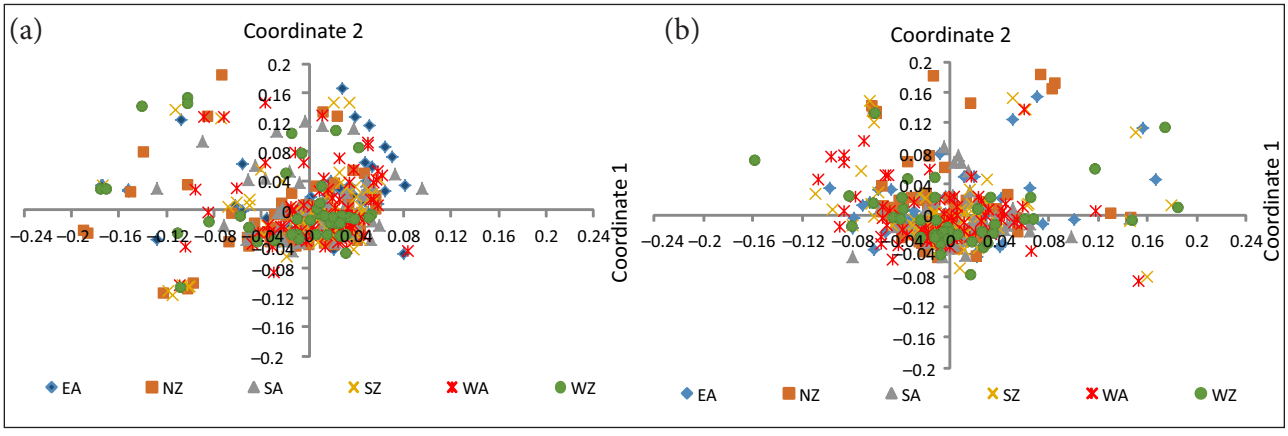


**Fig. 1.** Population structure by the multidimensional scaling: a) Euclidean distance b) Gower

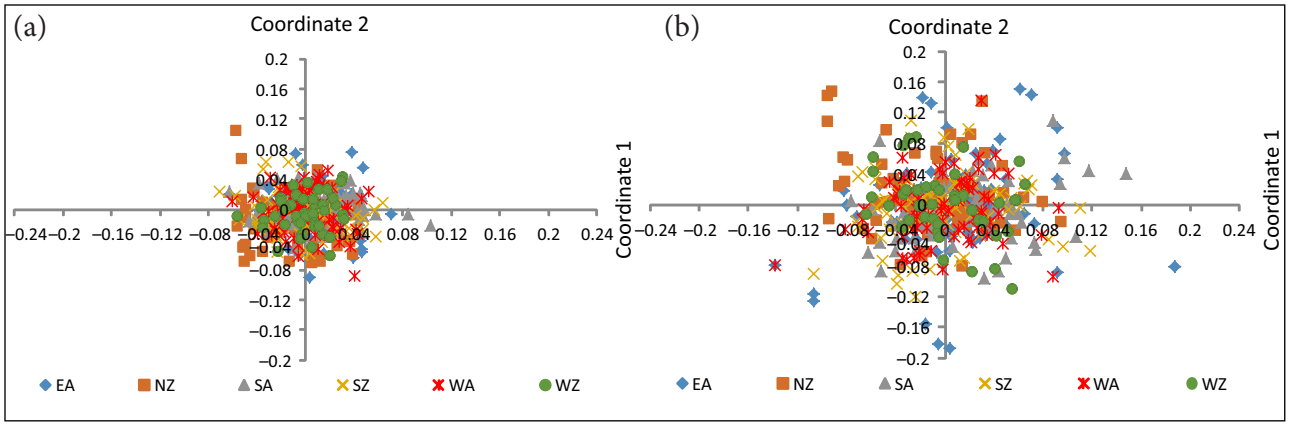


**Fig. 2.** Population structure by the principal coordinate analysis: a) Euclidean distance b) Gower

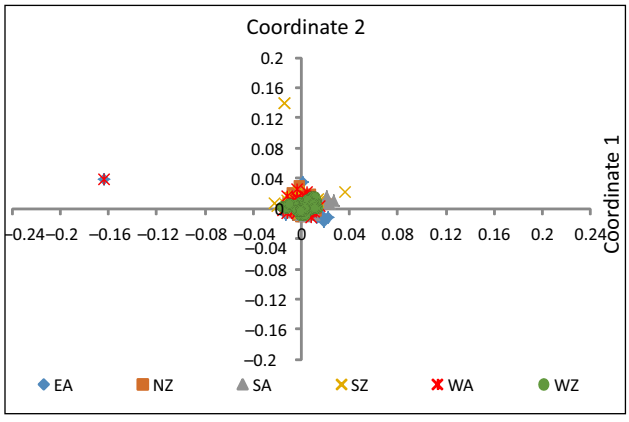


**Fig. 3.** Population structure by the principal components analysis (Euclidean distance)

Figure 2a and Figure 3 show that all six ethnolinguistic groups form one general cluster, and therefore we can conclude that the Lithuanian population is homogeneous.

## CONCLUSIONS

In this paper, the multidimensional scaling, principal coordinate, and principal component methods for the Lithuanian population genetic structure have been investigated and compared. We conclude that the principal coordinate and principal component methods can be used for genotyping data visualization, since any essential differences in the results obtained have not been observed and compared to multidimensional scaling. The results show that the Lithuanian population is homogeneous as the points are clustered very strongly when we use the principal coordinates or principal component methods.

## ACKNOWLEDGEMENTS

This study is a part of the ANELGEMIA project. This work was supported by the Research Council of Lithuania, grant no. S-MIP-20-34.

Received 12 June 2019
Accepted 3 July 2019

## References

1. Chen C, Hardle W, Unwin A. Handbook of data visualization. Berlin: Springer; 2008.
2. Dzemyda G, Kurasova O, Medvedev V. Dimension reduction and data visualization using neural networks. In: Maglogiannis I, Karpouzis K, Wallace M, Soldatos J., editors. Emerging artificial intelligence applications in computer engineering. Real world AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies. Frontiers in artificial intelligence and applications. Amsterdam: IOS Press; 2007. p. 160, 25–49.
3. Jolliffe IT. Principal component analysis. 2nd edition. Springer Series in Statistics. New York: Springer; 2002.
4. Borg I, Groenen P. Modern multidimensional scaling. New York: Springer; 2005.
5. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000; 290(5500): 2323–6.
6. PAST Paleontological Statistics Version 3.18. Reference manual Øyvind Hammer. Oslo: Natural History Museum, University of Oslo; 1999–2017.

7. Podani J, Miklos I. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. Ecology. 2002; 83: 3331–43.
8. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38: 904–9.
9. Abraham G, Inouye M. Fast principal component analysis of large-scale genome-wide data. Editor *PLoS One*. 2014; 9(4): e93766.

**Alma Molytė, Alina Urnikytė, Vaidutis Kučinskas**

## MATEMATINIŲ METODŲ, TAIKOMŲ LIETUVOS POPULIACIJOS HOMOGENIŠKUMUI NUSTATYTI, LYGINAMOJI ANALIZĖ

*Santrauka*

**Įvadas.** Vienas svarbiausių populiacijos genetikos parametrų yra populiacijos genetinė struktūra, atskleidžianti demografinius populiacijos ypatumus. Tyrimo tikslas – nustatyti Lietuvos populiacijos homogeniškumą remiantis viso genomo skenavimo duomenimis. Daugiamačiams genetiniams duomenims vizualizuoti buvo atlikta lyginamoji trijų metodų analizė: daugiamačių skalių, pagrindinių komponenčių ir pagrindinių koordinačių. Taip pat pateikti vaizdai, gauti vizualizavimo metu.

**Medžiaga ir metodai.** Duomenų imtį sudarė 425 asmenys iš šešių Lietuvos populiacijos etnolingvistinių grupių. Tiriamųjų asmenų DNR buvo išskirta iš kraujo leukocitų fenolio–chloroformo ekstrakcijos metodu bei automatizuota DNR išskyrimo sistema Tecan Freedom EVO. DNR genotipavimas atliktas naudojant VNP Illumina HumanOmniExpress-12 *v1.1* ir Infinium OmniExpress-24 lustus Vilniaus universiteto Biomedicinos instituto Žmogus ir medicininės genetikos katedroje. Lietuvos populiacijos homogeniškumui įvertinti buvo naudojamas PLINK duomenų failas pasitelkus PLINK v1.07 programą. Genotipo duomenys buvo vizualizuoti daugiamačių skalių ir pagrindinių komponenčių metodu PAST3 programa. Populiacijos genetinei struktūrai nustatyti pagrindinių komponenčių metodu buvo naudojama The SmartPCA from EIGENSOFT 7.2.1 programa.

**Išvados.** Vertinant Lietuvos populiacijos genetinę struktūrą buvo ištirti ir palyginti daugiamačių skalių, pagrindinių koordinačių ir pagrindinių komponenčių metodai. Gauti rezultatai parodė, kad genotipo duomenų vizualizavimui geriau naudoti pagrindinių koordinačių ir pagrindinių komponenčių metodus, nes gauti rezultatai yra panašūs, palyginti su daugiamačių skalių metodu. Lietuvos populiacija yra homogeniška, o vizualizuoti duomenys yra glaudžiai susiję, kai naudojami pagrindinių koordinačių arba pagrindinių komponenčių metodai.

**Raktažodžiai:** genotipavimas, daugiamatės skalės, pagrindinės komponentės, pagrindinės koordinatės, genotipinių duomenų vizualizavimas