Original scientific paper

Principal component analysis and cluster analysis for fatty acid assessment of backfat in three pig breeds

Irina Chernukha[a], Marina Nikitina[a], Elena Kotenkova[a] and Liliya Fedulova[a]*

[a] *V. M. Gorbatov Federal Research Center for Food Systems of RAS, Talalikhina st., 26, Moscow 109316, Russia*

A B S T R A C T

Local pig breeds are bred by crossing pig breeds, and their carcass traits, technological, chemical and physical compositions differ. The fatty acid composition and content depend on both biological and handling specifity that can be assessed using statistical tools. Backfat composition of three pig breeds was studied: two were local (Livni and Altai), one was commercial (Duroc). The calculations were performed based on the results of fatty acids analysis using small datasets in the software R Studio 2022.07.2 Build 576. Principal component and cluster analysis revealed that the fatty acid composition of the two local pig breeds differed from each other and from the commercial breed, which resulted in clustering of the individual pigs studied into separate groups. Results of cluster analysis demonstrated similarity with PCA, but all breeds formed distinct groups; Altai pigs formed the most tight and distinct group, while Duroc and Livni pigs clustered in more expanded groups with overlapping regions. Non-parametric statistics (PCA and cluster analysis) were applicable to this small dataset.

1. Introduction

Biotechnological, biological and technological processes depend on numerous parameters that affect them the and cause difficulties associated with identifying the structure of the relationships of these parameters. Therefore, the researcher often must deal with stochastic, incomplete information. In this case, the use of multivariate statistical analysis is not only justified, but also essential (*Lin et al.*, 2009). Multivariate statistical methods, among a variety of possible probabilistic and statistical models, allow selection of the best model that corresponds to the initial statistical data, thereby characterizing the real behaviour of the studied set of objects. In turn, this allows us to assess the reliability and accuracy of conclusions made on the basis of limited statistical data. Methods of multivariate classification are designed to divide the considered sets of objects, subjects or phenomena into groups, in a certain sense homogeneous. It should be taken into account that each of the objects is characterized by a large number of different and stochastically related features. Problems of complex classification can be solved by cluster analysis (*Röttger*, 2016). The presence of many initial variables characterizing objects' functional processes makes it necessary to select just the most significant variables and study a smaller set of indicators. Most often, the initial variables undergo transformation, which ensures minimal loss of information. Such a solution can be provided by dimensionality reduction methods, which include factor and principal component analysis (PCA) (*Anowar*

et al., 2021; *Costello & Osborne*, 2019; *Everitt & Hothorn*, 2011; *Jolliffe & Cadima*, 2016).

Factor analysis is a set of methods that, based on objectively existing correlational relationships of features, make it possible to identify latent (hidden) generalizing characteristics of the structure of the studied objects and their properties (*Granato et al.*, 2018). The suitability assessment, i.e, how well the constructed model describes the structure of the initial variables, is carried out using the Kaiser–Meyer–Olkin criterion. Bartlett's sphericity criterion makes it possible to assess the correlation of the initial variables: with insufficient correlation, the use of factor analysis is impractical (*Abdi & Williams*, 2010). Usually, factor analysis PCA and/or cluster analysis are used in technological, biological, medical and biotechnological research (*Dong et al.*, 2014; *Hasan et al.*, 2021; *Kemsley et al.*, 2019; *Macchiato et al.*, 1992). In one study (*Ros-Freixedes et al.,* 2014), PCA was used to determine the relationship between meat quality traits, feeding patterns, scale activity, and number of conflict–avoidance interactions.

Pork is characterized by different quality parameters specific for different pig breeds. The pig herd in Russia has been steadily growing in recent years and reached the number of 28.3 million animals by March 2023. Russian consumers have a positive attitude towards pork, despite its high fat content (*Chernukha et al.,* 2023). Data about the negative impacts of cholesterol on the cardiovascular system and of red meat on the risk of morbid obesity and related diseases has led to a significant change in consumer requirements for food in general and pork in particular.

We have chosen and compared the backfat fatty acid content of three pig breeds raised in Russia nowadays: Livni breed (registered in 1949), Altai (2015) and Duroc — a globally known breed, that was brought to Russia and used to improve other local pig breeds. The empirical assessment of the fatty acid composition of the backfat of Livni, Altai and Duroc pigs was carried out on the basis of actual data obtained in laboratory studies using multidimensional statistical methods.

2. Materials and methods

2.1. Sampling and fatty acids analysis

Three pig breeds were used: Livni (n=6); Duroc (n=7); Altai meat breed (n=5). Pigs were kept under the conditions of a commercial pig farm and consumed complete feed, were slaughtered at the live weight 110±10kg without electrical stimulation and the carcasses cooled at 0°C for 24 h in cold storage. Backfat (BF) in between 10th and 11th rib samples were removed and sent to the laboratory. Pieces of backfat 5×5 cm, with a depth from the subcutaneous fat surface to the muscle layer, and weighing approximately 70–150 g (±5 g), were cut from each carcass. The samples were obtained from at least three replicates from each carcass; average values of three replicates of fatty acid composition from each carcass were used for further data processing. Determination of the fatty acid composition was performed according to the reported method in the literature (*Ivankin et al.*, 2016), with the author's modification. Results are available at (*Chernukha et al.*, 2023).

2.2. Statistical analysis of fatty acids composition

The calculations were performed in the software R Studio 2022.07.2 Build 576. The R software is a freely distributed cross-platform software tool used for statistical calculations and data visualization. R distributions are available on the websites The Comprehensive R Archive Network (https://cran.r-project.org).

3. Results

We previously reported that backfat fatty acid composition was strongly differ between Livni, Altai and Duroc breeds, that affected on nutritional indices for assessing of fatty acids composition as ΣSFA, ΣUFA, ΣMUFA, ΣPUFA, ΣHUFA, ΣPUFA/ΣSFA, Σn–3 PUFA, Σn–6 PUFA, Σn–3 PUFA/Σn–6 PUFA, C 18:2/C 14, C 18:1/C 14, ΣFA short (from C4 to C 10), ΣFA medium (from C11 to C 16), ΣFA long (>C 17), and ΣC4-C16/ΣC17-C24, and the atherogenic index (IA) and thrombogenicity (IT). In brief, the highest UFA content, in particular omega 3 and omega 6 PUFA, was found in the backfat of Altai pigs with a minimum SFA content. The backfat of Livni pigs was characterized by the highest monounsaturated and medium chain fatty acid contents and the lowest short-chain fatty acid content; the atherogenicity index in was close to this of Duroc. On contrary, thrombogenicity index in Livni backfat was even lower than the Duroc one (*Chernukha et al.*, 2023).

PC1 described 39.34% of the variance of the original dataset, while PC1 and PC2 together explained approximately 72.1% of the variance of the original dataset. The eight main components of PC1-PC6 described about 93.79% of the initial dataset. The weighting coefficients by which the original variables were included in the PCA were calculated by d_pca$rotation and are presented in Table 1.

The resulting matrix shows how the principal component axes were shifted relative to the original ones. Each component was calculated as the sum of the multiplication of the weighting coefficients with the corresponding fatty acid value. The most significant variables (fatty acids) in PC1 were C20:2w6, C14:1, C16:1 and C18:1 with weighing coefficients of −0.321, 0.315, 0.300 and 0.292, respectively, while C22:1w9 was included in PC1 with a very small weighing coefficient of −0.017. The most significant variables in PC2 were C18:0, C20:0, C15:0, C20:3w6, C20:3w3 and C20:4w6 with weighing coefficients of −0.351, 0.309, −0.306, −0.285, −0.281 and −0.275, respectively, while C18:1 was included in PC2 with a weighing coefficient only of −0.068. The most significant variables in PC3 were C20:1w9, C22:1w9, C20:3w3, C12:0 and C18:1 with weighing coefficients of −0.506, −0.418, −0.313, 0.305 and −0.300, respectively, while C17:0 was included in PC3 with a very small weighing coefficient of −0.003. Interestingly, the weighing coefficients of C21:0 and C22:1w9 in PC4 were very high and averaged −0.715 and −0.459, respectively. The highest weighing coefficients for the major fatty acids, C18:0 and C18:1, were included in PC2 and PC1, respectively, and averaged 0.351 and 0.292, respectively, while for C16:0 and C18:2w6, weighing coefficients were approximately equal in PC1 and PC2 and

Table 1. The weighing coefficients of variables for principal component analysis of fatty acids in pig backfat

| Fatty acid	PC1	PC2	PC3	PC4	PC5	PC6
C12:0	0.25254	−0.15083	0.30451	−0.15960	−0.02860	0.05781
C14:0	0.28244	−0.13437	0.22284	−0.09715	−0.07601	−0.07513
C14:1	0.31499	−0.11740	−0.01659	0.06955	−0.09928	0.14156
C15:0	0.14525	−0.30626	0.08470	0.13551	−0.23892	0.27058
C16:0	0.21224	0.22819	0.26237	−0.00315	0.01579	−0.17230
C16:1	0.30046	−0.13484	0.08210	−0.11432	0.23817	−0.10021
C17:0	0.20505	−0.26119	−0.00272	0.08141	−0.03581	0.38339
C17:1	0.25919	−0.15177	−0.17790	0.06388	0.12658	0.38071
C18:0	−0.07992	0.35111	0.01583	0.15142	−0.16620	0.13819
C18:1	0.29187	0.06843	−0.30043	0.07666	0.11940	−0.05156
C18:2w6	−0.27045	−0.20982	0.16991	−0.11273	−0.02205	0.00950
C18:3w3	−0.19321	−0.26028	0.01163	0.11557	−0.34489	0.17560
C20:0	−0.11828	0.30860	−0.08270	−0.03140	0.00115	0.47631
C20:4w6	−0.21027	−0.27504	−0.00589	−0.04711	0.26284	−0.16111
C20:3w6	−0.21367	−0.28545	−0.01876	−0.11578	0.07526	0.00053
C20:2w6	−0.32092	−0.11952	−0.07649	−0.00566	−0.10877	−0.00028
C20:3w3	−0.10897	−0.28113	−0.31268	0.16291	−0.26812	−0.10250
C20:1w9	0.11012	0.17256	−0.50606	0.22402	−0.06591	−0.04578
C21:0	0.12679	0.10449	−0.18884	−0.71457	−0.57225	−0.06937
C22:0	0.20575	−0.20635	−0.21556	0.22062	−0.15703	−0.47643
C22:1w9	−0.01658	−0.14635	−0.41755	−0.45889	0.41945	0.12727
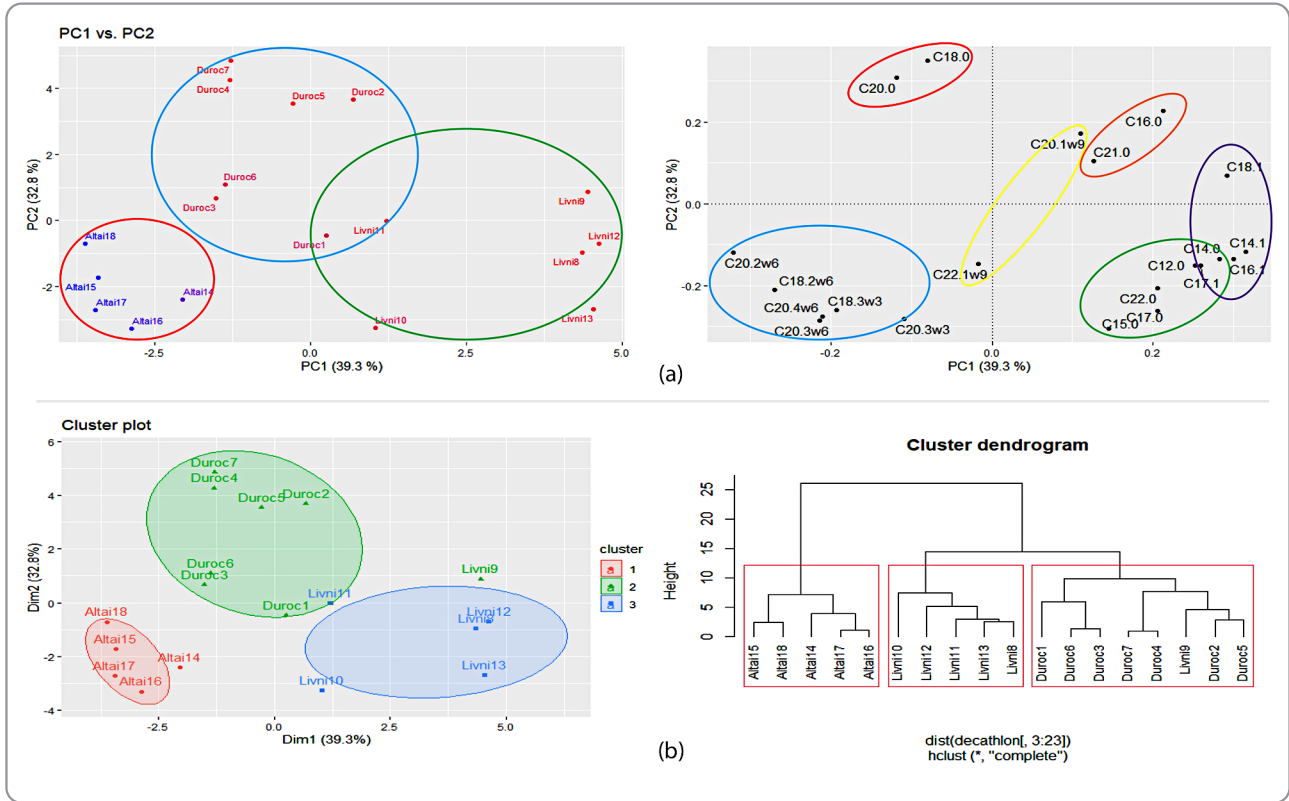
Figure 1. Results of PCA (a) and cluster (b) analysis.

in sum averaged 0.440 and −0.480, respectively. For most part, the highest weighing coefficients for the minor fatty acids were located in PC1 and PC2, while C12:0 and C20:3w3 were in PC3, but the weighting coefficients were approximately equal to those in PC1 and PC2, respectively. The highest weighting coefficients of C21:0 and C22:1w9 were in PC4, C20:0 was in PC6, C18:3w3 was in PC5, but the sum of C18:3w3 coefficients in PC1 and PC2 exceeded the weighting coefficient for this fatty acid in PC5. Figure 1 shows the results of PCA (a) and cluster (b) analysis. PCA visualization of PC1 and PC2 demonstrated that according to fatty acid composition, the pig breeds divided in three groups. Altai pigs gathered in a more tight and distinct group, while Duroc and Livni pigs clustered in more expanded groups which have an overlapping regions. Interestingly, the ω3 and ω6 fatty acids formed a tight and distinct group, while ω9 fatty acids were located in two quadrants. Saturated fatty acids were divided in three groups and located in different quadrants: C18:0 and C20:0; C16:0 and C21:0 pairs were each groups of one major and one minor fatty acid. The certain saturated fatty acids (C12:0, C14:0, C15:0, C17:0 and C22:0) formed a tight and distinct group, which was overlapped with the group of monounsaturated fatty acids and located in two quadrants. Results of cluster analysis demonstrated similarity with PCA, but all pig breeds formed more tight and distinct groups. With cluster analysis, Altai pigs gathered in one tight and distinct group, while Duroc and Livni pigs were in separate, more expanded groups, although one Livni pig was joined to the Duroc group.

4. Discussion

Pig breeds have different fat deposition, fat-specific metabolic characteristics, fatty acid compositions and various other different properties (*Poklukar et al.*, 2020; *Popova & Nakev*, 2019). *Zappaterra et al.* used PCA to study fatty acids composition in BF of 798 individuals Italian Large White heavy pig in order to assess environmental factors and carcass features associated with changes in fatty acids composition (*Zappaterra et al.*, 2022). *Piasentier et al.* used smaller dataset sizes that averaged 24–50 individuals in each group, where diet effect on the lard composition was weighted using the PCs scores as covariates in a tri-factorial (genotype, carcass leanness, sex) covariance design (*Piasentier et al.*, 2009). *Rocchetti et al.* investigated the impact of different diets on the lipidomic profile of pork using 36 animals and applied both PCA and cluster analysis (*Rocchetti et al.*, 2022). *Petroman et al.* applied

gas chromatography-mass spectrometry and PCA to evaluate the dissimilarity of Mangalitza lipid fractions in different layers of backfat with or without heat treatment (*Petroman et al.*, 2021). In that study, a small sample size was used, but clear discrimination was obtained between the raw and thermally processed fat as well as for fat layers (*Petroman et al.*, 2021). *Nizar et al.* revealed the clustering of lard, chicken fat, beef fat into four subclasses using PCA and only three replicates (*Nizar et al.,* 2013). We used Livni (n=6), Duroc (n=7) and Altai (n=5) meat breed pigs, and, despite the small sample, also obtained a clear, breed-specific distribution based on the results of backfat fatty acid composition. The application of both cluster analysis (*Ahn et al.*, 2012; *Dalmaijer et al.*, 2022; *McNeish & Harring*, 2017) and PCA (*Björklund,* 2019; *Shaukat et al.*, 2016; *Yata & Aoshima*, 2010) for small datasets revealed both advantages and disadvantages. Cluster analysis assumes, instead of grouping the initial variables by their correlation into factors (PCA), clustering of animal breeds, in accordance with their fatty acid characteristics, thereby contributing to the overall assessment of the backfat. Thus, the purpose of this method is to distribute pig breeds into groups (clusters) in which fatty acids are relatively homogeneous. Results of cluster analysis demonstrated similarity with PCA, but all breeds formed relatively tight and distinct groups. However, Altai pigs gathered in a very tight and distinct group, while Duroc and Livni pigs grouped in more expanded groups, and one Livni individual pig was joined to the Duroc group.

5. Conclusion

Backfat fatty acid contents of three pig breeds raised in Russia nowadays (Livni, Altai and Duroc) were compared. According to PCA, PC1 and PC2 explain approximately 72.1% of the original data. The highest weighted coefficients for the major and most of the minor fatty acids were presented in PC2 and PC1 separately or in sum, but some minor fatty acids were described by PC4 and PC6, i.e., C21:0, C22:1w9 and C20:0. Fatty acids ω3 and ω6 gathered in a common group, while ω9 were located in two quadrants. Saturated fatty acids were divided in three groups and located in different quadrants: two groups were small and wide, and were overlapped by the monounsaturated fatty acid group. Cluster analysis and PCA demonstrated similarities: all breeds formed groups, although the Altai pigs formed the most tight and distinct group, while Duroc and Livni pigs were in more expanded groups with overlapping regions. Non-parametric statistics (PCA and cluster analysis) were shown to be applicable to this small dataset.

References

Abdi, H. & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459, https://doi.org/10.1002/wics.101

Ahn, J., Lee, M. H. & Yoon, Y. J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica*, 22, 443–464, http://dx.doi.org/10.5705/ss.2010.148

Anowar, F., Sadaoui, S. & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review*, 40, Article 100378 https://doi.org/10.1016/j.cosrev.2021.100378

Björklund, M. (2019). Be careful with your principal components. *Evolution*, 73(10), 2151–2158, https://doi.org/10.1111/evo.13835

Chernukha, I., Kotenkova, E., Pchelkina, V., Ilyin, N., Utyanov, D., Kasimova, T., Surzhik, A. & Fedulova, L. (2023). Pork Fat and Meat: A Balance between Consumer Expectations and Nutrient Composition of Four Pig Breeds. *Foods,* 12, Article 690, https://doi.org/10.3390/foods12040690

Costello, A. B. & Osborne, J. W. (2019). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation,* 10, Article 7, https://doi.org/10.7275/jyj1-4868)

Dalmaijer, E. S., Nord, C. L. & Astle, D. E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics*, 23, Article 205, https://doi.org/10.1186/s12859-022-04675-1

Dong, J.-J., Li, Q.-L., Yin , H., Zhong, C., Hao, J.-G., Yang, P.-F., Tian, Y.-H. & Jia, S.-R. (2014). Predictive analysis of beer quality by correlating sensory evaluation with higher alcohol and ester production using multivariate statistics methods. *Food Chemistry*, 161, 376–382, https://doi.org/10.1016/j.foodchem.2014.04.006

Everitt, B. & Hothorn, T. (2011). An Introduction to Applied Multivariate Analysis with R. New York: Springer, https://doi.org/10.1007/978-1-4419-9650-3

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L. & Maggio R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science and Technology*, 72, 83–90, https://doi.org/10.1016/j.tifs.2017.12.006

Hasan, M. M., Sood, V., Erkinbaev, C., Paliwal, J., Suman, S. & Rodas-Gonzalez, A. (2021). Principal component analysis of lipid and protein oxidation products and their impact on color stability in bison *longissimus lumborum* and *psoas major* muscles. *Meat Science*, 178, Article 108523, https://doi.org/10.1016/j.meatsci.2021.108523

Ivankin, A. N., Oliferenko, G. L., Kulikovskii, A. V., Chernukha, I. M., Semenova, A., Spiridonov, K. I. & Nasonova, V. (2016). Determination of unsaturated fatty acids with a migrating double bond in complex biological matrices by gas chromatography with flame ionization and mass spectrometry detection. *Journal of Analytical Chemistry*, 71, 1131–1137, https://doi.org/10.1134/s1061934816110046

Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374, 20150202, https://doi.org/10.1098/rsta.2015.0202

Kemsley, E. K., Defernez, M. & Marini, F. (2019). Multivariate statistics: considerations and confidences in food authenticity problems. *Food Control*, 105, 102–112, https://doi.org/10.1016/j.foodcont.2019.05.021

Lin, T. I., Ho, H. J. & Chen, C. L. (2009). Analysis of multivariate skew normal models with incomplete data. *Journal of Multivariate Analysis*, 100 (10), 2337–2351, https://doi.org/10.1016/j.jmva.2009.07.005

Macchiato, M. F., Ragosta, M., Cosmi, C. & Lo Porto, A. (1992). A method in multivariate statistics to analyze ecosystems starting from their species composition. *Ecological Modelling*, 62(4), 295–310, https://doi.org/10.1016/0304-3800(92)90004-x

McNeish, D. M. & Harring, J. R. (2017). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics — Simulation and Computation*, 46(2), 855–869, https://doi.org/10.1080/03610918.2014.983648

Microsoft R Application Network (2023). Retrieved from https://mran.microsoft.com/download. Accessed May 10, 2023.

Nizar, N. N. A., Marikkar, J. M. N. & Hashim, D. M. (2013). Differentiation of Lard, Chicken Fat, Beef Fat and Mutton Fat by GCMS and EA-IRMS Techniques. *Journal of Oleo Science*, 62(7), 459–464, https://doi.org/10.5650/jos.62.459

Petroman, C., Popescu, G., Szakal, R.-N., Păunescu, V., Drăghia, L. P., Bujancă, G. S., Chirilă, C. A., Hădărugă, D. I., Văduva, L., Hădărugă, N. G. & Petroman, I. (2021). Fatty Acid Profile of Lipid Fractions of Mangalitza (*Sus scrofa domesticus*) from Northern Romania: A GC-MS-PCA Approach. *Foods*, 10, 242, https://doi.org/10.3390/foods10020242

Piasentier, E., Di Bernardo, N., Morgante, M., Sepulcri, A. & Vitale, M. (2009). Fatty acid composition of heavy pig back fat in relationship to some animal factors. *Italian Journal of Animal Science*, 8(sup2), 531–533, https://doi.org/10.4081/ijas.2009.s2.531

Poklukar, K., Čandek-Potokar, M., Batorek Lukač, N., Tomažin, U. & Škrlep, M. (2020). Lipid Deposition and Metabolism in Local and Modern Pig Breeds: A Review. *Animals*, 10, 424, https://doi.org/10.3390/ani10030424

Popova, T. & Nakev, J. (2019). Fatty acid profile of the backfat layers in four pig breeds. *Food Science and Applied Biotechnology*, 2(1), 24–29, https://doi.org/10.30721/fsab2019.v2.i1.66

Rocchetti, G., Vitali, M., Zappaterra, M., Righetti, L., Sirri, R., Lucini, L., Dall'Asta, C., Davoli, R. & Galaverna, G. (2022). A molecular insight into the lipid changes of pig Longissimus thoracis muscle following dietary supplementation with functional ingredients. *PLoS One*, 17(3), e0264953, https://doi.org/10.1371/journal.pone.0264953

Ros-Freixedes, R., Sadler, L. J., Onteru, S. K., Smith, R. M., Young, J. M., Johnson, A. K., Lonergan, S. M., Huff-Lonergan, E., Dekkers, J. C. M. & Rothschild, M. F. (2014). Relationship between gilt behavior and meat quality using principal component analysis. *Meat Science*, 96(1), 264–269, https://doi.org/10.1016/j.meatsci.2013.07.004

Röttger, R. (2016). Clustering of Biological Datasets in the Era of Big Data. *Journal of Integrative Bioinformatics*, 13(1), 300, https://doi.org/10.1515/jib-2016-300

Shaukat, S. S., Rao, T. A. & Khan, M. A. (2016). Impact of sample size on principal component analysis ordination of an environmental data set: effects on eigenstructure. *Ekológia* (Bratislava), 35(2), 173–190, https://doi.org/10.1515/eko-2016-0014

The Comprehensive R Archive Network. Retrieved from https://cran.r-project.org. Accessed May 10, 2023.

Yata, K. & Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis*, 101(9), 2060–2077, https://doi.org/10.1016/j.jmva.2010.04.006

Zappaterra, M., Catillo, G., Lo Fiego, D. P., Belmonte, A. M., Padalino, B. & Davoli, R. (2022). Describing backfat and *Semimembranosus* muscle fatty acid variability in heavy pigs: Analysis of non–genetic factors. *Meat Science*, 183, 108645, https://doi.org/10.1016/j.meatsci.2021.108645