tapraid3/the-the/the-the/the99910/the0752d10z | elspr | S=1 | 12/30/11 | 8:19 | Art: 12005 | ce: 31

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Theriogenology

Theriogenology xx (2011) xxx

www.theriojournal.com

Taking advantage of the use of supervised learning methods for characterization of sperm population structure related with freezability in the Iberian red deer

AQ: au

M. Ramón[a,b], F. Martínez-Pastor[c,d], O. García-Álvarez[e], A. Maroto-Morales[a], A. Josefa Soler[a], P. Jiménez-Rabadán[e], M. Rocío Fernández-Santos[a], R. Bernabéu[f], J. Julián Garde[a,*]

[a] *Biology of Reproduction Group, National Wildlife Research Institute (IREC) UCLM-CSIC-JCCM, 02071, Albacete, Spain*
[b] *Department of Medicine, Animal Surgery, Faculty of Veterinary Medicine, University of Murcia, 30071, Murcia, Spain*
[c] *ITRA-ULE, INDEGSAL, University of León, 24071, León, Spain*
[d] *Molecular Biology (Cell Biology), University of León, 24071, León, Spain*
[e] *CERSYRA, ~~Castille and la Mancha~~, 13300, Valdepeñas, Spain*
[f] *Escuela Técnica Superior de Ingenieros Agrónomos, University of Castilla-La Mancha, 02071, Albacete, Spain*

Received 26 July 2011; received in revised form 8 November 2011; accepted 9 December 2011

Abstract

Using Iberian red deer as a model, this study presents a supervised learning method, the Support Vector Machines (SVM), to characterize sperm population structure related with freezability. Male freezability was assessed by evaluating motility, membrane status and mitochondrial membrane potential of sperm after a freezing-thawing procedure. The SVM model was generated using sperm motility information captured by computer-assisted sperm analysis (CASA) from thawed semen, belonging to six stags with marked differences on their freezability. A total of 1369 sperm tracks were recorded for seven kinematic parameters and assigned to four motility patterns based on them: weak motile, progressive, transitional and hyperactivated-like. Then, these data were split in two sets: the training set, used to train the SVM model, and the testing set, used to examine how the SVM method and three other unsupervised methods, a non-hierarchical, a hierarchical and a multistep clustering procedures, performed the sperm classification into subpopulations. The SVM was revealed as the most accurate method in the characterization of sperm subpopulations, showing all the sperm subpopulations obtained in this way high significant correlations with those sperm parameters used to characterize freezability of males. Given its superiority, the SVM method was used to characterize the sperm motile subpopulations in Iberian red deer. Sperm motile data from frozen–thawed semen belonging to 25 stags were recorded and loaded into the SVM model. The sperm population structure revealed that those males showing poor freezability were characterized by high percentages of sperm with a weak motility pattern. In opposite, males showing good freezability were characterized by higher percentages of sperm with a progressive and hyperactivated-like motility pattern and lower percentages of sperm with a weak motile pattern. We also identified a sperm subpopulation with a transitional motility pattern. This subpopulation increased as the freezability of males improved, and may be used as indicative of overall sperm motility.

Keywords: Sperm subpopulations; Support vector machines; Sperm freezability; Iberian red deer

1. Introduction

One of the most recognized characteristic of the mammalian ejaculates is its heterogeneity, reflected in the presence of different sperm subpopulations [1–9]. This heterogeneity has been related to different key

* Corresponding author. Tel: +34(8) 67 599200; fax: +34(8) 67 599238.
E-mail address: julian.garde@uclm.es (J. J. Garde).

0093-691X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.theriogenology.2011.12.011

issues of male reproductive performance. Thus, it has been found that ability to undergo capacitation and fertilize may vary depending on the subpopulation under consideration [10,11], and that freezability may vary significantly among sperm subpopulations [9,12]. Therefore, it is fundamental to obtain a prior deep knowledge of the population structure of semen, to study the relationships between the quality and the reproductive performance of sperm samples.

Cell cryopreservation has become an indispensable tool in biology. Biological materials can be safely kept and used after a very long period. In the case of spermatozoa, cryopreservation is used not only in research, but also in livestock management and in the conservation of wild and domestic species, as a complementary tool for managing live animals and preserving their genetic diversity. Sperm cryopreservation combined with artificial insemination (AI) is the assisted reproductive technology (ART) which possibly has been increasingly applied to deer species too [13].

One important problem for standardizing sperm cryopreservation protocols is that sperm from different individuals exhibit significant different responses to the same freezing treatment [14–16]. Thus, males may show different freezability depending on their sperm population structure. Therefore, it is of interest to identify those characteristics that favor the freezability of spermatozoa, and to characterize the distribution of sperm subpopulations of males as a way to predict their freezability.

Different statistical procedures have been used for the definition and identification of sperm subpopulations. Martinez-Pastor, et al. [17] provides references and a general discussion on the classification methods commonly used to identify sperm subpopulations. In the last years, other analytical and statistical methods for cluster analysis have been developed, which could improve the current analyses of sperm subpopulations. *Data Mining* and *Machine Learning* disciplines are becoming increasingly important tools that provide useful methods to reach those objectives. In a general way, we could say that the aim in *Data Mining* and *Machine Learning* is to design computer programs to solve a task not based on predefined rules provided by the user, but using relations that they "learned" from the information, data or feedback that they receive [18].

The learning processes can be roughly categorized as unsupervised or supervised.

In *Unsupervised Learning*, there is no outcome measure; we observe only the features and the goal is to describe the associations and patterns among a set of input measures. As examples of unsupervised learning methods, we have the hierarchical and non-hierarchical clustering methods, among others, which are the preferred methods currently used in sperm subpopulations analyses. In *Supervised Learning*, the goal is to predict the value of an outcome measure based on a number of input measures, so the presence of the outcome variable guide the learning process. Data are usually split in two sets: the *training set* and *test set*. Training set of data is used to observe the outcome and feature measurements for a set of objects. Using these data we build a prediction model, or learner, which will enable us to predict the outcome for new unseen objects, the test set. A good learner is one that accurately predicts such an outcome. As supervised methods we can found several references on literature: nearest neighbor methods [19], logistic regression [20], decision trees [21], support vector machines [22] or neural networks [23] among others. An extended explanation of these methods and other supervised methods is given by [18].

In this study we used the Support vector machines (SVM) procedure. SVM's are a set of related supervised learning methods used for classification and regression [18,22,24]. SVM has been successfully used for several purposes [25,26]. In the field of biological sciences, we can found references about the use of this procedure for protein sequence comparisons [27,28], classification of genes and proteins [29,30], microarray gene expression analysis [31] or cancer classification [32,33]. As other supervised learning techniques, the SVM procedure involves separating data into *training* and *testing sets*. Each instance in the training set contains one "target value" (i.e., the class labels) and several "attributes" (i.e., the features of the observed variables). The goal of the SVM is to produce a model, based on the *training data*, which predicts the target values of the test data only given the test data attributes. Thus, having learned the features of one class, the SVM could recognize new objects as members or non-members of that class based on their attributes.

The purpose of this study has been to characterize those sperm subpopulations based on motile characteristics that could be related with the freezability of males in the Iberian reed deer, using the SVM methodology. For this study, semen samples from Iberian red deer were used as model. The advantage of using wild animals is that males are not artificially selected for fertility. Thus, they are expected to exhibit considerable diversity in sperm characteristics and fertility, as well as being an excellent model to study the eventual as-

sociations between sperm characteristics and reproduction performance.

2. Materials and methods

2.1. Animals and sperm collection

Animal manipulations were performed in accordance with the Spanish Animal Protection Regulation, RD1201/2005, which conforms to European Union Regulation 2003/65. This study included a total of 25 Iberian red deer (*Cervus elaphus hispanicus*) stags hunted during the mating season in the south of Spain, coinciding with their reproductive season (end of September to December) [34]. Both testes and epididymes were removed (in the scrotum) and transported at 20 to 21 °C to the laboratory. Elapsed time between animal death and sperm analyses ranged from 3 to 6 h [35]. At the laboratory, testes and epididymides were removed from the scrotum. Spermatozoa were collected by cutting the distal proportion of the epididymides with a surgical blade, and diluted in PBS (pH 7.5; 320 mOsm/kg). After dilution, sperm motility was assessed subjectively and only those epididymal semen samples with a minimum quality were cryopreserved, as a way to assure that all sperm samples showed good quality before freezing. Thus, only semen samples with a sperm motility subjectively assessed over 80% (SM; 0–100%) and a quality movement over 4 (quality of movement (QM); on a scale of 0–5, where 0 is no motility and 5 is vigorous progressive movement) were freeze. Cryopreservation was performed as described by Soler, et al. [16], and frozen semen was stored in liquid nitrogen (−196 °C) for a minimum period of 6 mo before thawing. Thawing was performing by dropping the straws in a water bath with saline serum at 37 °C for 20 s.

2.2. Assessment of thawed sperm quality

After thawing, semen samples were incubated in a water bath at 37 °C during 2 h previously to semen quality assessment. Percentage of motile spermatozoa (SM) and the QM were subjectively assessed and a resume measure, the *Sperm Motility Index* (SMI) was calculated as described by Comizzoli, et al. [36]:

$$SMI = \frac{SM + (20 \times QM)}{2}$$

Membrane stability with YO–PRO-1, the viability with propidium Iodide (PI) and the mitochondrial membrane potential with Mitotracker Deep Red were assessed by flow cytometry [37]. Thus, the YO–PRO-1-/PI-ratio is the proportion of viable spermatozoa with a stable membrane, the YO–PRO-1+/PI+ ratio is the proportion of death spermatozoa and the Mitotraker+/YO-PRO-1-ratio is the proportion of spermatozoa with high mitochondrial membrane potential.

We also objectively assessed the motility characteristics of sperm by Computer Assisted Semen Analysis (CASA) as described in Martínez-Pastor, et al. [38]. Analyses were carried out using the Sperm Class Analyzer software (SCA 2002, Microptic, Barcelona, Spain) and the following motility descriptors were recorded: curvilinear velocity (VCL, μm/sec), average path velocity (VAP, μm/sec), straight line velocity (VSL, μm/sec), linearity (LIN, %), straightness (STR, %), amplitude of lateral head displacement (ALH, μm) and beat cross frequency (BCF, Hz). A total of 6542 spermatozoa belonging to the 25 stags were recorded, with a minimum of 200 spermatozoa per male being assessed.

2.3. Characterization of sperm motile subpopulations

2.3.1. Classification methods

This study made use of four different classification procedures: the non-hierarchical (*k-means*) clustering, the hierarchical clustering, and a multistep procedure that used both clustering methods jointly [12] as unsupervised methods, and the support vector machines procedure as supervised learning method.

2.3.1.1. Non-hierarchical (k-means) clustering. A non-hierarchical clustering method performing a dis-

Table 1
Seminal parameters after freezing and thawing of spermatozoa from Iberian red deer stags.[ac]

	n	SMI (%)	YO–PRO-1-/PI- (%)	YO–PRO-1+/PI + (%)	Mt+/YO–PRO-1- (%)
All males	25	39.30 ± 2.93	39.67 ± 2.63	41.22 ± 2.98	38.98 ± 3.03
Bad freezers	3	15.00 ± 2.50	20.65 ± 1.37	61.61 ± 2.29	20.17 ± 1.27
Good freezers	3	52.50 ± 1.44	55.76 ± 0.25	25.13 ± 1.71	55.76 ± 4.04

[a] Data are mean ± SEM.
[c] All parameters differed significantly among different males' groups ($P < 0.05$).

joint cluster analysis on the basis of Euclidean distances has been used. The optimal numbers of clusters to keep, four in this study, was set by using the Silhouette Average Width (SAW) criterion [39]. The kmeans function from the STATS R package [40] was used.

2.3.1.2. Hierarchical clustering. In this study, distance matrix was computed by using the *Euclidean* distance measure and the *Ward*'s minimum variance method was set to classify the data. To determine the final number of clusters, the *Hubert* Γ coefficient [41] criterion was considered. To perform the analysis, we used the *hclust* function in the STATS package [40].

2.3.1.3. Multistep procedure. Non-hierarchical and hierarchical clustering methods were used jointly in a multistep procedure. Multistep procedures have been used successfully to classify sperm subpopulations [7,12]. Generally, non-hierarchical methods are employed as the first step. The clusters produced by the non-hierarchical method are then merged in the second step by an agglomerative hierarchical method. The first step may also be used to identify outliers or special clusters, allowing continuation to the second step with an optimized set of clusters. In this study, a total of 10 clusters were obtained from the non-hierarchical step and the merged in a final number of four clusters in the hierarchical clustering step. The criteria considered to define the optimal number of clusters in each step were then same describe above for the non-hierarchical and hierarchical clustering methods.

2.3.1.4. Support vector machines. The SVM procedure involved the split of data into a training set used to train the SVM and a testing data set to evaluate the accuracy of these SVM. This testing set was used to evaluate the accuracy of the other non-supervised methods as well. The SVM equation obtained from the training step were used to characterize the different sperm subpopulations in the Iberian red deer. A further explanation of these steps is given below.

2.3.2. Training and testing data sets

To identify the sperm subpopulation structure that best correlates with the freezability of males, individual sperm tracks from six stags with marked differences on their freezability were assessed (Table 1, Fig. 1, closed circles). A total of 1369 sperm tracks were recorded for the seven kinematic parameters defined above (Section 2.2) and assigned to four different motility patterns, as described in Goodson, et al. [42] (Table 2, Fig. 2). Then, this database was split and used to generate the SVM equations and to test how the SVM method and other clustering procedures perform the spermatozoa classification into subpopulations. Thus, the *training data set* consisted of 720 sperm tracks (120 per male) randomly chosen and was used to generate the SVM equations, while the *testing data set* consisted of the other 649 sperm tracks and was used to test the performance of the clustering procedures. The kinematic parameters VCL, VSL,

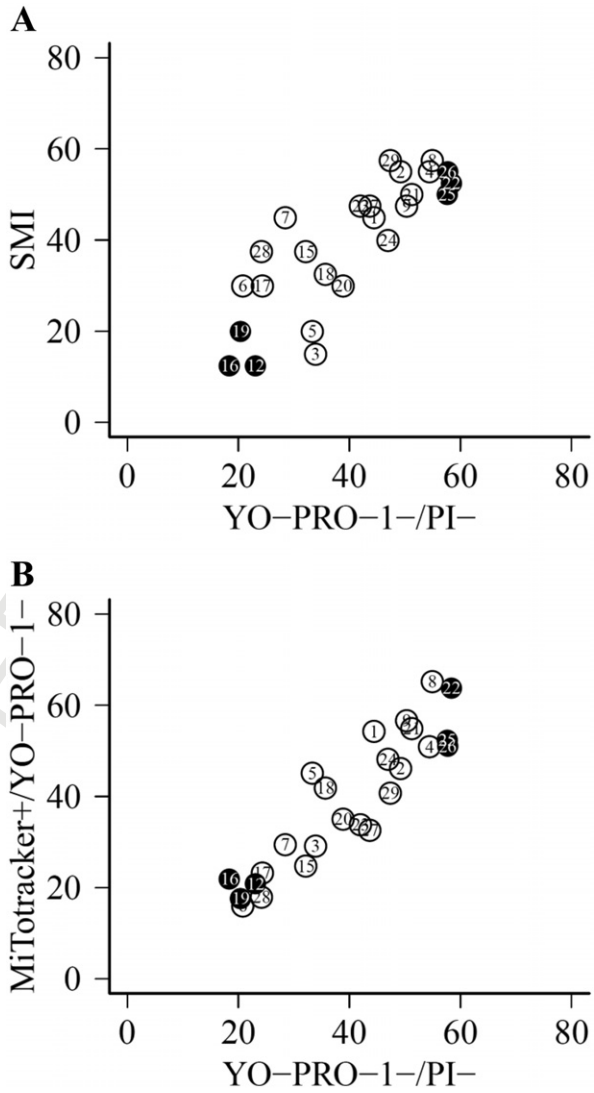

Fig. 1. Characterization of semen quality of males after the freezing-thawing procedure in the Iberian red deer. Numbers indicate males used in this study. Six males with marked differences on their freezability (closed circles) were used to train de SVM model and to test the accuracy of different statistical methods. A, Relation between the proportion of viable spermatozoa with a stable membrane (YO–PRO-1-/PI–) and the SMI. B, Relation between the proportion of viable spermatozoa with a stable membrane and proportion of spermatozoa with high mitochondrial membrane potential (Mitotraker+/YO-PRO-1–).

Table 2
Kinematics parameters for the four sperm subpopulations defined in the SVM training step.[ac]

	VCL (μm/sec)	VSL (μm/sec)	VAP (μm/sec)	Lin (%)	STR (%)	ALH (μm)	BCF (Hz)
SP1	31.44 ± 0.80	7.87 ± 0.27	14.13 ± 0.36	24.29 ± 0.51	53.25 ± 0.75	1.72 ± 0.03	3.79 ± 0.13
SP2	110.39 ± 2.85	60.93 ± 1.52	78.41 ± 1.88	56.12 ± 0.79	78.07 ± 0.84	3.97 ± 0.11	9.10 ± 0.27
SP3	114.63 ± 1.85	35.44 ± 0.79	61.02 ± 0.95	31.17 ± 0.50	58.41 ± 0.97	4.69 ± 0.09	9.52 ± 0.27
SP4	168.77 ± 1.96	34.59 ± 0.93	102.01 ± 0.98	20.48 ± 0.49	33.80 ± 0.81	6.42 ± 0.10	9.24 ± 0.22

[a] Data are mean ± SEM.
[c] SP1, weakly motile; SP2, progressive; SP3, transitional; SP4, hyperactivated-like.

VAP, ALH and BCF were loaded into the SVM procedure to generate SVM equations that were able to distinguish among sperm belonging to different subpopulations. We discarded the use of LIN and STR parameters because of they are linear combination of the other motility parameters. Once the SVM equations were constructed, the accuracy of the classification of spermatozoa into different subpopulations by the SVM procedure, as well as, by the k-means, hierarchical and multistep clustering procedures were tested using data in the *testing set*. For the latter clustering methods, the same five kinematic traits, VCL, VSL, VAP, ALH and BCF were used.

2.3.3. Characterization of sperm motile subpopulations in the Iberian red deer

Finally, we examined the sperm population structure of the Iberian red deer using the SVM equations obtaining from the *training set*, with the aim to find a subpopulations distribution being related with freezability of males. Thus, 25 Iberian red deer males showing different freezability were used, and a total of 6542 sperm tracks were recorded and used to characterize the sperm motile subpopulations.

2.4. Statistical analysis

All statistical analyses in this study were conducted with the R statistical software [41]. To implement the SVM methodology, we used the svm function from the e1071 R package [43]. Package e1071 provides an interface to libsvm [44], a robust and fast implementation of the most popular SVM formulations (C and ν classification, ε and ν regression, and novelty detection).

Results obtained from the use of different classification methods to characterize the subpopulation in the test set were presented graphically (Fig. 3) and as a confusion matrix (Table 3). In this matrix, each row represents the instances in a predicted class, while each column represents the instances in an actual class. Values on diagonal (in bold) represents events that have been well-classified.

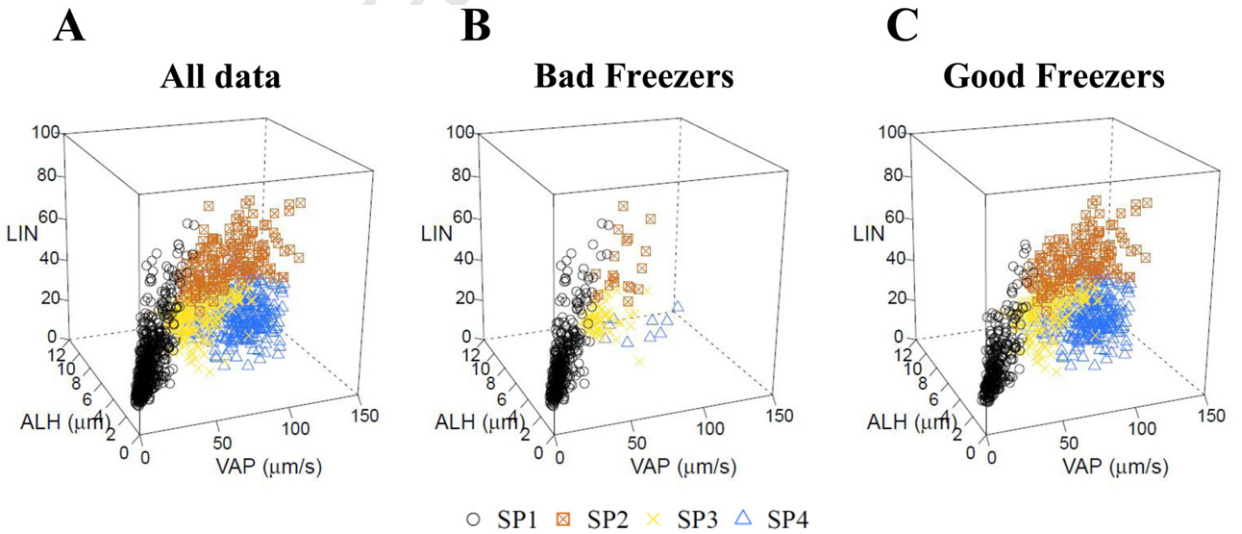

Fig. 2. Sperm motility characteristics for the six Iberian red deer stags used to train the SVM and to test the accuracy of different statistical methods (A). Four different motility patterns are identified: weak motile (SP1), progressive (SP2), transitional (SP3) and hyperactivated-like (SP4). Data are also represented for the group of males with poor (B) and good (C) freezability, separately.

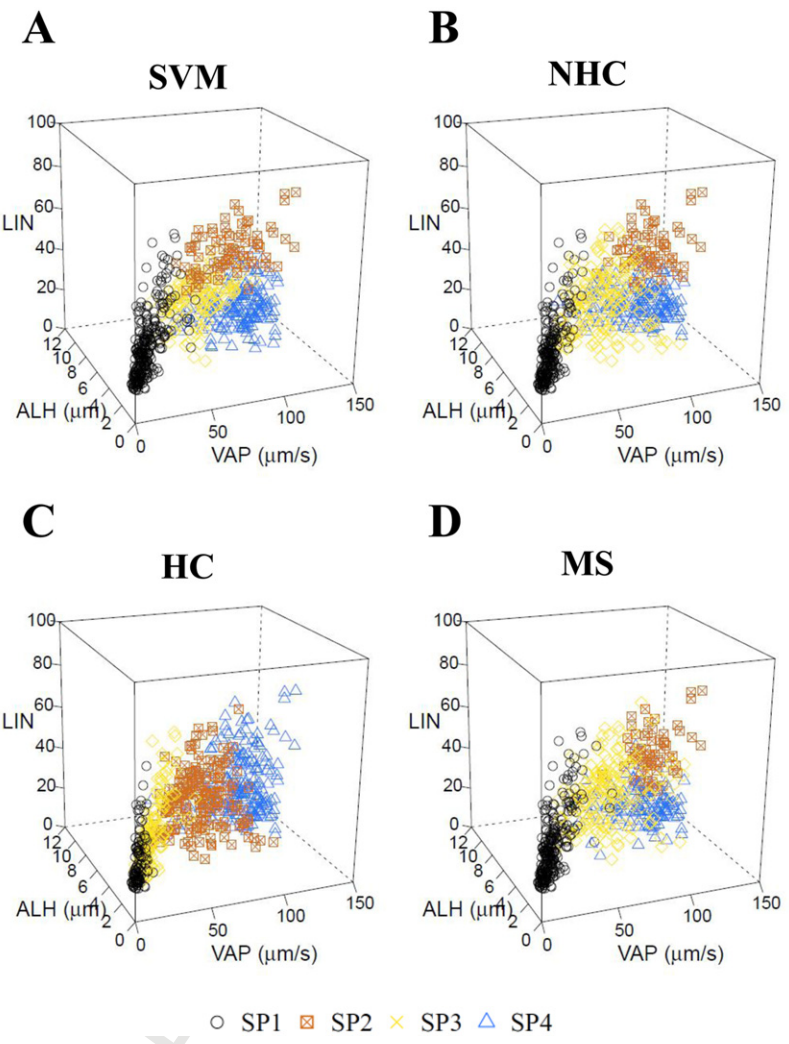

Fig. 3. Sperm subpopulation characterization of the data in the *testing set* using the SVM method (A) and the non-hierarchical (B), hierarchical (C) and multistep clustering procedures (D). Four different motility patterns are identified: weak motile (SP1), progressive (SP2), transitional (SP3) and hyperactivated-like (SP4).

To evaluate how different methods performed the classification, the overall accuracy rate has been calculated. The overall accuracy is defined as the sum of the diagonal of the confusion matrix divided by the total number of events. For this study, accuracy has been presented as an error rate, that is, one minus the overall accuracy. In addition, correlations between the sperm parameters used to determine the freezability of a male (that is, the SMI and the cytometry parameters) and the percentages of subpopulations of males were calculated for each classification method. This allowed us to examine which of these methods perform a sperm subpopulation characterization that best correlates with freezability of males.

The method that performed the most accuracy classification, in this study the SVM method, was then used to characterize the sperm subpopulations distribution in the Iberian red deer. Once sperm population structure was characterized, we performed a principal component analysis (PCA) to examine the relations between the different sperm subpopulations defined for the Iberian red deer and to explore how an overall measure of the sperm population distribution could be useful to characterize freezability of males.

3. Results

3.1. Assessment of thawed sperm quality

After the freezing-thawing procedure, semen samples of Iberian red deer showed, on average (mean ±

Table 3
Confusion matrix for predictions in the testing set. For different classification methods, data on diagonal (in bold) represents events (no. of spermatozoa) that were correctly labeled. The error rate is also presented.[ac]

	Non-hierarchical				Hierarchical				Multistep				SVM			
	Actual				Actual				Actual				Actual			
Predicted	SP1	SP2	SP3	SP4	SP1	SP2	SP3	SP4	SP1	SP2	SP3	SP4	SP1	SP2	SP3	SP4
SP1	239	0	2	0	118	0	0	0	256	1	9	0	266	2	8	0
SP2	0	54	6	7	19	42	98	48	0	43	5	17	0	81	1	1
SP3	28	37	83	20	130	1	3	18	11	46	90	35	1	6	106	7
SP4	0	0	28	145	0	48	18	137	0	1	15	120	0	2	4	164
Error[c]	0.197				0.537				0.216				0,049			

[a] Error = 1 – *(sum of confusion matrix diagonal/Number of observations.*
[c] SP1, weakly motile; SP2, progressive; SP3, transitional; SP4, hyperactivated-like.

SEM), an SMI of 39.3 ± 2.9%, a percentage of YO–PRO-1-/PI- and YO–PRO-1+/PI+ sperm of 41.2 ± 2.9% and 39.7 ± 2.6%, respectively, and Mitotracker+/ YO–PRO-1-of 39.0 ± 3.0% (Table 1). Highly significant differences were observed between the two groups of males with different freezability. Thus, males with poor freezability showed an SMI of 15.0 ± 2.5%, a YO–PRO-1-/PI-of 20.7 ± 1.4%, a YO–PRO-1+/PI+ of 61.6 ± 2.3% and a Mitotracker+/YO-PRO-1-of 20.2 ± 1.3%, while for the group of three males with good freezability, a percentage of 52.5 ± 1.4% for the SMI, a 55.8 ± 0.3% and a 25.1 ± 1.7% for the YO–PRO-1-/PI- and the YO–PRO-1+/PI+, respectively, and a 55.8 ± 4.0% for the Mitotracker+/YO-PRO-1-, were observed.

3.2. SVM model training

The 1369 sperm tracks from six stags with large differences on their freezability were manually classified into subpopulations based on their motility characteristics observed at 2 h of incubation after thawing (Table 2, Fig. 2). Four motility patterns were clearly defined: a subpopulation of *weak motile* spermatozoa (SP1) characterized by its low velocity (VAP = 14.13 ± 0.36 μm/sec) and very low linearity (LIN = 24.29 ± 0.51%) with low lateral head displacement from the path of movement (ALH = 1.72 ± 0.03 μm); a subpopulation of spermatozoa with a vigorous *progressive* movement (SP2: VAP = 78.41 ± 1.88 μm/sec; LIN = 56.12 ± 0.79%; ALH = 3.97 ± 0.11 μm); a *transitional* subpopulation (SP3) that showed decreasing speed (VAP = 61.02 ± 0.95 μm/sec) and linearity (LIN = 31.17 ± 0.50%) comparing with SP2, but with an increasing lateral head movement (ALH = 4.69 ± 0.09 μm); and a *hyperactivated-like* subpopulation (SP4) characterized by fast spermatozoa (VAP = 102.01 ± 0.98 μm/sec) with low linearity (LIN = 20.48 ± 0.49%) and a considerable lateral head movement (ALH = 6.42 ± 0.10 μm). Total sperm tracks were characterized as follows (mean ± SD): 55.0 ± 27.2% as weak motile, 10.2 ± 7.7% as progressive, 18.8 ± 11.3% as transitional and 16.0 ± 16.4% as hyperactivated-like. When we compared between males with poor and good freezability, the sperm distribution was (mean ± SD): for the poor freezers, 77.7 ± 17.2% as SP1, 3.8 ± 1.8% as SP2, 16.6 ± 17.2% as SP3 and 2.0 ± 2.1% as SP4; for the good freezers, 32.4 ± 5.4% as SP1, 16.6 ± 4.8% as SP2, 21.0 ± 2.7% as SP3 and 30.0 ± 9.0% as SP4. Differences between both poor and good freezers in the distribution of all the subpopulations were significant ($P < 0.05$). Characterization of sperm into these four subpopulations can be graphically observed on Fig. 2A, as well as for the group of males with poor (Fig. 2B) and good (Fig. 2C) freezability, separately. Once sperm tracks were visually assigned to the different subpopulations based on their motility patterns, the database was split into the *training set* and the *testing set*. A total of 720 sperm tracks (120 per male) were randomly chosen to be the *training set* and then used to generate the SVM equations.

3.3. Accuracy of sperm classification by different mathematical approaches

A *testing set*, including 649 sperm tracks belonging to the six stags with different freezability were used to test the accuracy of the SVM method and the k-means, hierarchical and multistep clustering procedures. This *testing set* included 267 weakly motile, 91 progressive, 119 transitional and 172 hyperactivated-like sperm tracks. Table 3 presents the how different statistical procedures performed the sperm classification. The SVM method performed the most accurate classification showing an error rate of only 0.049 (~ 5%). Thus,

almost all the sperm were well-classified into their subpopulations using the SVM model generated from data in the *training set* (Fig. 3A). The sperm characterization obtained using the non-hierarchical clustering and the multistep methods were quite similar, with error rates of 0.197 and 0.216, respectively. These two methods well-identified the sperm with a weakly movement, and the transitional and hyperactivated-like subpopulations, but showed problems when identifying the sperm characterized by progressive movement (Fig. 3B, 3D). The hierarchical clustering was the method that performed the less accurate classification, with an error rate of 0.537. This method showed an erratic classification of sperm, with more than half of sperm tracks being miss-classified (Fig. 3C). Correlations between sperm parameters used to evaluate the freezability of a male and percentages of different subpopulations of males were calculated for each classification method (Table 4). The method showing the highest correlations was the SVM method, being the four subpopulations significantly correlated ($P < 0.05$) with the SMI and the sperm parameters evaluated by flow cytometry. Meanwhile, non-hierarchical clustering methods showed lees strong correlations and only the SP4 showed significant correlations with the sperm parameters used to determine the freezability of males. Therefore, the SVM method resulted to be clearly superior to the unsupervised clustering methods when looking for a subpopulations distribution that best correlates with freezability of males.

3.4. Characterization of the sperm motile subpopulations in the Iberian red deer

Finally, we made used of the SVM model previously obtained to characterize the motile subpopulation structure in the Iberian red deer and related with its freezability. The 6542 sperm tracks recorded from the 25 Iberian red deer stags were loaded into the SVM model and then classified into four subpopulations, obtaining the following average distribution within a male (mean ± SD and range): 56.0 ± 19. 7% [27–92%] of sperm tracks were classified as weak motile; 10.8 ± 6.8% [1–24%] were classified as progressive; 20.8 ± 9.5% [5–41%] were classified as transitional; and 11.4 ± 11.6% [0 – 40%] were classified as hyperactivated-like. The characteristics of the four motility patterns identified in the Iberian red deer (Tabla 5) were similar to those described in the *training set* (Table 2).

To explore the relations between these four sperm subpopulations, we performed a principal-component analysis. We retained the first two principal compo-

Table 4
Correlations among those sperm characteristics used to classify the six males in the training set as "*poor*" or "*good*" freezers and the percentages of the sperm motile subpopulations of males.[a,*]

	SVM				NHC			
	SP1 (%)	SP2 (%)	SP3 (%)	SP4 (%)	SP1 (%)	SP2 (%)	SP3 (%)	SP4 (%)
SMI (%)	−0.75	0.69	0.81	0.81	−0.58	0.81	0.26	0.81
YO–PRO-1-/PI- (%)	−0.83	0.58	0.89	0.81	−0.31	0.64	−0.20	0.81
Mitotracker+/YO–PRO-1- (%)	−0.94	0.81	0.77	0.93	−0.09	0.75	−0.20	0.93
	HC				MS			
	SP1 (%)	SP2 (%)	SP3 (%)	SP4 (%)	SP1 (%)	SP2 (%)	SP3 (%)	SP4 (%)
SMI (%)	−0.58	0.53	−0.58	0.81	−0.58	0.89	0.46	0.81
YO–PRO-1-/PI- (%)	−0.31	0.12	−0.43	0.81	−0.31	0.76	0.09	0.81
Mitotracker+/YO–PRO-1- (%)	−0.09	−0.06	−0.03	0.93	−0.09	0.70	−0.03	0.93

* Significant correlations ($P < 0.05$) are represented in bold.
[a] Spearman correlation coefficients are presented.

Table 5
Sperm characteristics of the four motile subpopulations identified in the Iberian red deer.[ac]

	N (%)[d]	VCL (μm/sec)	VSL (μm/sec)	VAP (μm/sec)	Lin (%)	STR (%)	ALH (μm)	BCF (Hz)
SP1	56.99 ± 19.67	34.73 ± 0.36	9.13 ± 0.13	16.41 ± 0.18	25.68 ± 0.23	53.42 ± 0.31	1.83 ± 0.02	4.42 ± 0.06
SP2	10.84 ± 6.75	105.80 ± 1.19	59.23 ± 0.64	75.27 ± 0.77	56.96 ± 0.38	79.08 ± 0.39	3.67 ± 0.04	9.82 ± 0.13
SP3	20.77 ± 9.45	114.57 ± 0.74	35.02 ± 0.33	60.40 ± 0.39	30.87 ± 0.23	58.35 ± 0.43	4.75 ± 0.04	9.74 ± 0.10
SP4	11.40 ± 11.56	159.15 ± 1.06	31.75 ± 0.52	100.53 ± 0.53	10.79 ± 0.28	31.44 ± 0.47	5.92 ± 0.05	9.72 ± 0.13

[a] Data are mean ± SEM.
[c] SP1, weakly motile; SP2, progressive; SP3, transitional; SP4, hyperactivated-like.
[d] Summary of the percentages of subpopulations found on each male (mean ± SD).

nents based on Kaiser criterion (Fig. 4). The first principal component accounted for 64.1% of the total variance, and which could be interpreted as an indicator of non-vigorous movement, so the greater this value is, the less vigorous the movement is. This principal component allowed differentiating between males with higher percentages of SP1, and males with higher percentages of SP2 and SP4, the latter closely related (Fig. 4). The subpopulation with a transitional motility pattern (SP3) was mainly reflected on the second principal component which accounted for 25.7% of the total variance. This second component allows differentiating among males with an average freezability, and could be interpreted as indicator of overall motility. Thus, males with higher percentages of motile sperm showed greater values for this second component.

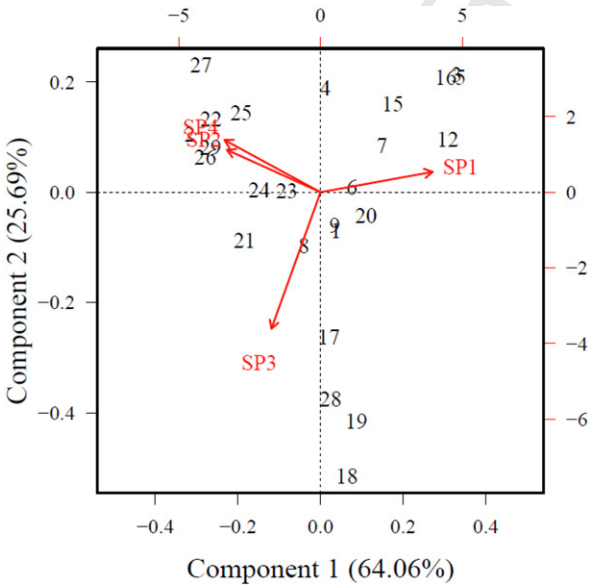

Fig. 4. Males distribution in the multidimensional ordination space defined by the first two principal components from the PCA analysis. Numbers indicate males used in this study. Arrows indicate vectors representing the four motility patterns/subpopulations.

4. Discussion

In the present study, we characterize the population structure of motile epididymal spermatozoa in the Iberian red deer, and we suggest that this distribution could help to explain the sperm freezability of different males. Contrasting with the statistical methods commonly used for the characterization of sperm subpopulations, we propose a supervised learning method, the SVM procedure, and we show the superiority of this method over traditional ones.

This study has been aimed to characterize the sperm population structure in the Iberian stags by finding some subpopulations based on motile characteristics that maximize the correlation with the freezability of these males. A number of studies have addressed the characterization of sperm motile subpopulations in thawed samples in several species [6,9,12,44–48]. Most of them have used unsupervised statistical methods [17], so the characterization of sperm subpopulation has been conducted without considering any prior information on freezability of males. This could lead to the sperm population structure defined by using those clustering methods was no optimal. To our knowledge, few references are found in the literature on the use of supervised learning methods for sperm analyses. For instance, Holt [50] used discriminant analysis (a supervised classification system) to assign cluster memberships to unclustered datasets, using an initial dataset that had been classified using cluster analysis. In other two studies, Vulcano, et al. [51] and, more recently, Goodson, et al. [42] examined sperm motility patterns under capacitating conditions for ram and for mouse sperm, respectively, using the SVM methodology.

Here, we examine the sperm distribution of males showing different freezability as a first step in the identification of sperm subpopulations, and then use this information to generate an SVM model for the characterization of other semen samples, different from those used to generate the model. Thus, this study

proposes the use of different sources of information, in this case the SMI, membrane integrity and mitochondrial activity, as a prior knowledge for the characterization of sperm subpopulations. Individual sperm tracks from six stags with marked differences on their freezability have been assessed and assigned to different motility patterns, as described in Goodson, et al. [42]. This information has been then used to generate an SVM model that clearly identifies and quantifies four distinct patterns of sperm movement in populations of Iberian red deer sperm: weak motile, progressive, transitional and hyperactivated-like.

We have evaluated how the SVM and the non-hierarchical clustering methods performed the characterization of sperm subpopulations. The SVM has been the most accurate method, with $< 5\%$ of sperm being miss-classified, and being the four subpopulations obtained from this method high significant correlated with the SMI, membrane integrity and mitochondrial membrane potential used to characterize the freezability of males. By contrast, non-hierarchical methods have showed errors above 20%, and only one of the subpopulations (SP4) have showed significant correlations with those sperm parameters used to characterize the freezability of males. The characterization provided by these methods has not taken into account information on the differences in the distribution of sperm between poor and good freezers and, for that reason, has resulted to be a little different from the characterization performed by the SVM method. Thus, although we could find differences in the sperm population structure defined in this way between males showing poor and good freezability, this association would be less strong than that expected from the use of the supervised learning method in which information on sperm motility distribution of males have been used to develop the SVM model.

Thus, the SVM method has been used to characterize the sperm motile population structure of frozen–thawed semen in the Iberian red deer, to find a subpopulations distribution that best correlates with freezability of males. Motility data from 25 stags showing differences on freezability were recorded and loaded into the SVM model previously generated in the training step. The same four subpopulations with a weak motile, progressive, transitional and hyperactivated-like motility patterns have been clearly represented in the sperm population. Significant differences on the distribution of sperm among these four subpopulations have been observed between males with different freezability. Thus, sperm characterized by a weak motility pattern (SP1) were predominant in those males with poor freezability, while for those males showing better freezability, higher percentages of progressive and hyperactivated-like sperm were observed. Concerning to the transitional subpopulation, the percentages increased as the sperm quality at thawing increase.

To further explore the relations between the four sperm subpopulations, and to explore how an overall measure of the sperm population distribution could be useful to characterize the freezability in the Iberian red deer, we have performed a PCA. The PCA rendered a first principal component accounting for 64.06% of the total variance, which could be interpreted as an indicator of non-vigorous movement. This factor would be very useful to discriminate between males showing great differences on their freezability, but could be less efficient in differentiating between males showing an average freezability. To the latter, consider the percentage of sperm belonging to the transitional subpopulation (SP3) would be of interest. This percentage has been mainly reflected on the second principal component, which has accounted for 25.69% of the total variance and could be interpreted as indicator of the overall sperm motility. Thus, higher values of this second component have been found in those males showing better freezability.

5. Conclusions

In conclusion, SVM has demonstrated to be very useful tools when we look for functional correlations between spermatozoa characteristics and freezability of males. In this study, the SVM method has performed the most accurate classification, being the subpopulations distribution obtained high significant correlated with those sperm characteristics used to characterize the sperm freezability of males. The characterization of the population structure of motile spermatozoa in the Iberian red deer using an SVM method has resulted in the identification of four subpopulations characterized by different motility patterns. These subpopulations have showed different distribution among males showing differences on their quality on thawed semen and, therefore, could be useful to characterize the freezability of males.

Here, we have used the SVM method to characterize the sperm motile population structure related with freezability. However, this method can also be useful for other purposes, among which highlight the study of fertility potential of males. Sperm fertility studies are now focused on analyses that incorporate multiple vari-

ables to examine how different sperm parameters interact to determine fertility. Semen samples must be subjected in parallel to several different tests and their outcome should be subjected to multiparametric analyses to provide the highest level of fertility prediction. And it will be within this field where the SMV method proposed in this study will provide and important support.

Acknowledgments

This study was supported in part by the Spanish Ministry of Science and Innovation (Project AGL2007-60 271/GAN), and by the Education and Science Council of the Junta de Castilla y León, Spain (Project LE019A10-2). Felipe Martínez-Pastor was supported by the Ramón y Cajal program (Ministry of Science and Innovation, Spain).

Uncited references

This section consists of references that are included in the reference list but are not cited in the article text. Please either cite each of these references in the text or, alternatively, delete it from the reference list. If you do not provide further instruction for this reference, we will retain it in its current form and publish it as an "un-cited reference" with your article [49].

References

[1] Petrunkina AM, Töpfer-Petersen E. Heterogeneous osmotic behaviour in boar sperm populations and its relevance for detection of changes in plasma membrane. Reprod Fertil Dev 2000; 12:297–305.

[2] Paasch U, Grunewald S, Fitzl G, Glander HJ. Deterioration of plasma membrane is associated with activated caspases in human spermatozoa. J Androl 2003;24:246–52.

[3] Pérez-Llano B, Yenes-García P, García-Casado P. Four subpopulations of boar spermatozoa defined according to their response to the short hypoosmotic swelling test and acrosome status during incubation at 37 degrees C. Theriogenology 2003; 60:1401–7.

[4] Buffone MG, Doncel GF, Marín Briggiler CI, Vazquez-Levin MH, Calamera JC. Human sperm subpopulations: relationship between functional quality and protein tyrosine phosphorylation. Hum Reprod 2004;19:139–46.

[5] Esteso MC, Soler AJ, Fernández-Santos MR, Quintero-Moreno AA, Garde JJ. Functional significance of the sperm head morphometric size and shape for determining freezability in iberian red deer (Cervus elaphus hispanicus) epididymal sperm samples. J Androl 2006;27:662–70.

[6] Núñez-Martínez I, Moran JM, Peña FJ. A three-step statistical procedure to identify sperm kinematic subpopulations in canine ejaculates: changes after cryopreservation. Reprod Domest Anim 2006;41:408–15.

[7] Martínez-Pastor F, Cabrita E, Soares F, Anel L, Dinis MT. Multivariate cluster analysis to study motility activation of Solea senegalensis spermatozoa: a model for marine teleosts. Reproduction 2008;135:449–59.

[8] Druart X, Gatti JL, Huet S, Dacheux JL, Humblot P. Hypotonic resistance of boar spermatozoa: sperm subpopulations and relationship with epididymal maturation and fertility. Reproduction 2009;137:205–13.

[9] Muiño R, Peña AI, Rodríguez A, Tamargo C, Hidalgo CO. Effects of cryopreservation on the motile sperm subpopulations in semen from asturiana de los Valles bulls. Theriogenology 2009;72:860–8.

[10] Harrison RA. Capacitation mechanisms, and the role of capacitation as seen in eutherian mammals. Reprod Fertil Dev 1996; 8:581–94.

[11] Petrunkina AM, Waberski D, Günzel-Apel AR, Töpfer-Petersen E. Determinants of sperm quality and fertility in domestic species. Reproduction 2007;134:3–17.

[12] Martinez-Pastor F, Garcia-Macias V, Alvarez M, Herraez P, Anel L, de Paz P. Sperm subpopulations in Iberian Red deer epididymal sperm and their changes through the cryopreservation process. Biol Reprod 2005;72:316–27.

[13] Asher GW, Berg DK, Evans G. Storage of semen and artificial insemination in deer. Anim Reprod Sci 2000;62:195–211.

[14] Holt WV. Fundamental aspects of sperm cryobiology: the importance of species and individual differences. Theriogenology 2000;53:47–58.

[15] Thurston LM, Watson PF, Holt WV. Sperm cryopreservation: a genetic explanation for species and individual variation. Cryo Lett 2002;23:255–62.

[16] Soler AJ, García AJ, Fernández-Santos MR, Esteso MC, Garde JJ. Effects of thawing procedure on postthawed in vitro viability and in vivo fertility of red deer epididymal spermatozoa cryopreserved at −196 degrees C. J Androl 2003;24:746–56.

[17] Martínez-Pastor F, Tizado EJ, Garde JJ, Anel L, de Paz P. Statistical Series: Opportunities and challenges of sperm motility subpopulation analysis. Theriogenology 2011;75:783–95.

[18] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction, second edition. Springer Series in Statistics, 2009.

[19] Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. J Machine Learn Res 2009;10:207–44.

[20] Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: a learning curve analysis. The J Machine Learn Reseach 2003;4:211–55.

[21] Esmeir S, Markovitch S. Anytime learning of decision trees. J Machine Learn Res 2007;8:981–33.

[22] Scheinberg K. An efficient implementation of an active set method for SVMs. J Machine Learn Res 2006;7:2237–57.

[23] Ripley BD. Pattern recognition and neural networks. Cambridge University Press; 2008.

[24] Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565–7.

[25] Cristianini N, Shawe-Taylor J. An introduction to support Vector machines and other kernel-based learning methods. Cambridge, UK: Cambridge University Press; 2000.

[26] Noble WS. Support vector machine applications in computational biology. In: Schoelkopf B, Tsuda K, Vert JP. Kernel methods in computational biology. Cambridge2004 p. 71 92 MIT Press; MA

[27] Liao L, Noble WS. Combining pairwise sequence similarity and support vector machines for remote protein homology detection. In: Proceedings of the Sixth Annual International Conference on Computational Molecular Biology; 2002, p. 225–232.
[28] Leslie C, Kuang R. Fast string kernels using inexact matching for protein sequences. J Machine Learn Res 2004;5:1435–55.
[29] Vert JP. A tree kernel to analyze phylogenetic profiles. Bioinformatics 2002;18:276–84.
[30] Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics 2004;20:467–76.
[31] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 2000;97:262–7.
[32] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531–7.
[33] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46:389–422.
[34] García AJ, Landete-Castillejos T, Garde JJ, Gallego L. Reproductive seasonality in female Iberian red deer (*Cervus elaphus hispanicus*). Theriogenology 2002;58:1553–62.
[35] Garde JJ, Ortiz N, García AJ, Gallego L, Landete-Castillejos T, López A. Postmortem assessment of sperm characteristics of the red deer during the breading season. Arch Andol 1998;41:195–202.
[36] Comizzoli P, Mauget R, Mermillod P. Assessment of in vitro fertility of deer spermatozoa by heterologous IVF with zona-free bovine oocytes. Theriogenology 2001;56:261–74.
[37] Martínez-Pastor F, Fernández-Santos MR, Del Olmo E, Domínguez-Rebolledo AE, Esteso MC, Montoro V, et al. Mitochondrial activity and forward scatter vary in necrotic, apoptotic and membrane-intact spermatozoan subpopulations. Reprod Fertil Dev 2008;20:547–56.
[38] Martínez-Pastor F, Martínez F, Alvarez M, Maroto-Morales A, García-Alvarez O, Soler AJ, et al. Cryopreservation of Iberian red deer (*Cervus elaphus hispanicus*) spermatozoa obtained by electroejaculation. Theriogenology 2009;71:628–38.
[39] Silhouettes RPJ. A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[40] R Development Core Team. R: A language and environment for statistical computing, R.Vienna, Austria: Foundation for Statistical Computing. 2010.
[41] Haldiki M, Batistakis Y, Vazirgiannis M. Cluster validity methods, SIGMOD. Record;2002;31:40–5.
[42] Goodson SG, Zhang Z, Tsuruta JK, Wang W, O'Brien DA. Classification of mouse sperm motility patterns using an automated multiclass support vector machines model. Biol Reprod 2011;84:1207–15.
[43] Dimitriadou E, Hornik K, Leisch F, Meyer D, Weingessel A. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.5-22. 2009. Available at: http://CRAN.R-project.org/package=e1071; accessed:. AQ: 1
[44] LIBSVM. A library for support vector machines; 2001. Available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm; accessed:. AQ: 2
[45] Thurston LM, Watson PF, Mileham AJ, Holt WV. Morphologically distinct sperm subpopulations defined by Fourier shape descriptors in fresh ejaculates correlate with variation in boar semen quality following cryopreservation. J Androl 2001;22:382–94.
[46] Cremades T, Roca J, Rodriguez-Martinez H, Abaigar T, Vazquez JM, Martinez EA. Kinematic changes during the cryopreservation of boar spermatozoa. J Androl 2005;26:610–8.
[47] Martinez-Pastor F, Diaz-Corujo AR, Anel E, Herraez P, Anel L, de Paz P. Post mortem time and season alter subpopulation characteristics of Iberian red deer epididymal sperm. Theriogenology 2005;64:958–74.
[48] Beirão J, Cabrita E, Pérez-Cerezales S, Martínez-Páramo S, Herráez MP. Effect of cryopreservation on fish sperm subpopulations. Cryobiology 2011;62:22–31.
[49] Dorado J, Alcaráz L, Duarte N, Portero JM, Acha D, Hidalgo M. Changes in the structures of motile sperm subpopulations in dog spermatozoa after both cryopreservation and centrifugation on PureSperm(®) gradient. Anim Reprod Sci 2011;125: 211–8.
[50] Holt WV. Can we predict fertility rates? Making sense of sperm motility. Reprod Domest Anim 1996;31:17.
[51] Vulcano GJ, Moses DF, Valcárcel A, de las Heras MA. A lineal equation for the classification of progressive and hyperactive spermatozoa. Math Biosci 1998;149:77–93.