CSIRO PUBLISHING

Reproduction, Fertility and Development
https://doi.org/10.1071/RD17479

Review

Implementation of novel statistical procedures and other advanced approaches to improve analysis of CASA data

M. Ramón[A] *and F. Martínez-Pastor*[B,C]

[A]CERSYRA-IRIAF, Junta de Comunidades de Castilla-La Mancha, Valdepeñas, Spain.
[B]INDEGSAL and Department of Molecular Biology (Cell Biology), Universidad de León, 24071 León, Spain.
[C]Corresponding author. Email: felipe.martinez@unileon.es

Abstract. Computer-aided sperm analysis (CASA) produces a wealth of data that is frequently ignored. The use of multiparametric statistical methods can help explore these datasets, unveiling the subpopulation structure of sperm samples. In this review we analyse the significance of the internal heterogeneity of sperm samples and its relevance. We also provide a brief description of the statistical tools used for extracting sperm subpopulations from the datasets, namely unsupervised clustering (with non-hierarchical, hierarchical and two-step methods) and the most advanced supervised methods, based on machine learning. The former method has allowed exploration of subpopulation patterns in many species, whereas the latter offering further possibilities, especially considering functional studies and the practical use of subpopulation analysis. We also consider novel approaches, such as the use of geometric morphometrics or imaging flow cytometry. Finally, although the data provided by CASA systems provides valuable information on sperm samples by applying clustering analyses, there are several caveats. Protocols for capturing and analysing motility or morphometry should be standardised and adapted to each experiment, and the algorithms should be open in order to allow comparison of results between laboratories. Moreover, we must be aware of new technology that could change the paradigm for studying sperm motility and morphology.

Additional keywords: clustering, computer-aided sperm analyses, spermatozoon, subpopulations, support vector machines (SVM).

Received 9 November 2017, accepted 14 March 2018, published online xx xxxxx xxxx

Introduction

Computer-aided sperm analysis (CASA-Mot for motility and CASA-Morph for morphometry) systems are able to produce a huge amount of data (Amann and Waberski 2014). However, this wealth of information remains underutilised because of computer limitations and disregard of the possibilities hidden in those data.

Automated sperm analysis yields the two-dimensional coordinates of tracks (for motility) or head boundaries (for morphometry) of several hundred spermatozoa per sample (usually summarised by eight to 12 parameters per cell; Verstegen *et al*. 2002). Other approaches use the contour coordinates of the head (Varea Sánchez *et al*. 2013), whereas some studies have focused on the dimensions of the midpiece and principal piece (Malo *et al*. 2006). For a long time, these analyses were limited to producing a few average parameters per sample. Although an efficient approach (CASA systems directly provide the results), it misses the natural variability of samples, which potentially conceals valuable information and the presence of special or valuable spermatozoa in the sample. Currently, the features offered by standard microcomputers allows the average researcher to perform multiparametric analyses in large databases, taking advantage of the amount of data provided by image analysis of sperm samples. However, the challenge here is to choose the right tools to analyse these data.

The aim of this review is to present an overview of the possibilities of CASA data analysis to the spermatologist, especially regarding the study of sperm subpopulations (data clustering). We have omitted a myriad of important details on both automated sperm motility or morphometry analysis, and on the statistics of clustering datasets. Readers seeking further information should use this review as a starting point to a more specialised bibliography on either the settings, software and interpretation of CASA-Mot and CASA-Morph (Verstegen *et al*. 2002; Castellini *et al*. 2011; Amann and Waberski 2014) or statistical algorithms and data manipulation (linkage and clustering methods, data before and after processing, cluster description etc.; Xu and Wunsch 2005; Leonard and Droege 2008; Martínez-Pastor *et al*. 2011; Yániz *et al*. 2015*b*, 2016; Maroto-Morales *et al*. 2016).

Sperm heterogeneity: when differences make the difference

If anything has been confirmed by the use of CASA systems, it is the existence of clear sperm heterogeneity. The existence of morphological diversity among species is widely assumed, although this diversity seems less clear as we go deeper at the individual level (Birkhead *et al*. 2008). Sperm heterogeneity has been related to different key issues of male reproductive performance (Martinez-Pastor *et al*. 2005; Petrunkina *et al*. 2007; Ramón *et al*. 2013; Maroto-Morales *et al*. 2015). It is therefore necessary to characterise this heterogeneity in a detailed and precise way to increase our chances of finding associations between sperm features and outcomes of the fertilisation process. Nevertheless, for a long time, characterisation of sperm features was limited to producing a few average parameters per sample, with the consequent loss of valuable information about the natural variability of the samples. Ramón *et al*. (2014) highlighted the disadvantages of characterising an ejaculate using only average values. As an example, in that paper Ramón *et al*. (2014) showed six ejaculates exhibiting similar mean values for two sperm head shape parameters (head length and the perimeter to area factor, p2a) but with clear differences in subpopulation structure. Considering only mean values did not lead to any association with the fertility of the males. However, when the subpopulation structure (i.e. sperm heterogeneity) was considered, strong associations with fertility were observed (Ramón *et al*. 2014).

This example highlights the importance of examining sperm heterogeneity when conducting a study; otherwise, we may fail in our attempt to find functional associations. The statistical procedures for the assessment of sperm heterogeneity have been reviewed previously (Martínez-Pastor *et al*. 2008, 2011; Ramón *et al*. 2014; Yániz *et al*. 2016) and will be discussed succinctly in the following two sections, but some general recommendations are presented here. First, it is important to consider at which level sperm heterogeneity is going to be assessed; that is, whether an intraspecific or an intraindividual (from the same population) comparison is going to be investigated. For the most general case, namely the interspecific comparison, an approach characterising sperm samples with mean values and a relatively small sample size may be enough to identify existing differences. However, as we go deeper and look for differences within the same species, or even within the same individual, this characterisation must be more detailed and a larger sample will be required to ensure that we have a representative sample of the variability of the population under investigation (unsupervised clustering methods might be the choice for initial studies; see below). Second, in most cases the graphical representation of the data will be useful to determine the degree of heterogeneity within the samples and to decide which statistical procedure will be adequate to analyse the data. Third, when conducting a clustering procedure, the choice of the variables to be used (and the weight that each will have in the analysis) is as important as the clustering method. For the selection of the variables, the graphical exploration recommended before may be useful, but variables should be also selected according to the objectives of the study. Variable selection leads to our last recommendation: whenever possible, we should take advantage of previous results about the processes under investigation in order to maximise our chances of finding relevant functional relationships. Thus, implementation of supervised clustering methods (see below) is presented as a good option for the assessment of sperm heterogeneity considering other sources of prior information.

Statistical analysis of CASA data: unsupervised clustering

Unsupervised clustering of data refers to the lack of *a priori* criteria for grouping observations (Everitt *et al*. 2011). That is, the results of the clustering will depend on the characteristics of the dataset alone. Thus, although this approach is useful for learning about sperm subpopulations and defining the clustering structure of datasets obtained from different species and treatments, these approaches should be considered as a first step. The use of supervised methods (with criteria established from prior experiences) is more computationally efficient and more adequate for practical deployment of this kind of analysis (e.g. embedded into CASA software).

Unsupervised clustering has been used in most studies on sperm motility and morphometry subpopulations. Two main clustering strategies are available in most studies: hierarchical and non-hierarchical (partitional) methods (Xu and Wunsch 2005). Non-hierarchical methods (the *k*-means method being the most well known) are based on the initial partitioning of the data in a predefined number of clusters, followed by iterations in order to reassign observations to the 'correct' cluster. The initial number of clusters (k) must be specified, either by the researcher (based on a sensible guess) or by the algorithm (optimisation). Some algorithms are relatively fast (even with large datasets) and simple to use, but the main problem is deciding on the number of clusters before the partition. Because the number of sperm populations is generally reported to be between three and five, it is feasible to explore the partitioning results in this narrow margin. Indeed, the *k*-means algorithm, or versions of it, have been highly popular, especially for classifying motility data (Davis *et al*. 1995; Rivera *et al*. 2005; Quintero-Moreno *et al*. 2007; Martínez-Pastor *et al*. 2008).

Hierarchical methods work by successively organising the data into a hierarchical structure. The resulting tree-like structure (plotted as a dendrogram) allows the immediate investigation of different clustering results, depending on the level the dendrogram is cut. Moreover, this kind of representation clearly shows the clustering structure and the relationship among different observations. The main drawback of this method is that hierarchical algorithms work, at the very least, in quadratic time, making direct analysis of large datasets (e.g. CASA-Mot or CASA-Moprh data) prohibitive. Nonetheless, algorithm refinement (e.g. parallelisation) and the increasing power of modern desktop computers (fast processors, large memory, 64-bit architecture) allow for the use of hierarchical methods with these data. Indeed, some studies have already used hierarchical algorithms in a single step to cluster moderately large CASA datasets (Henning *et al*. 2014).

Hierarchical algorithms are also used for variable clustering, helping identify relationships between variables. This information

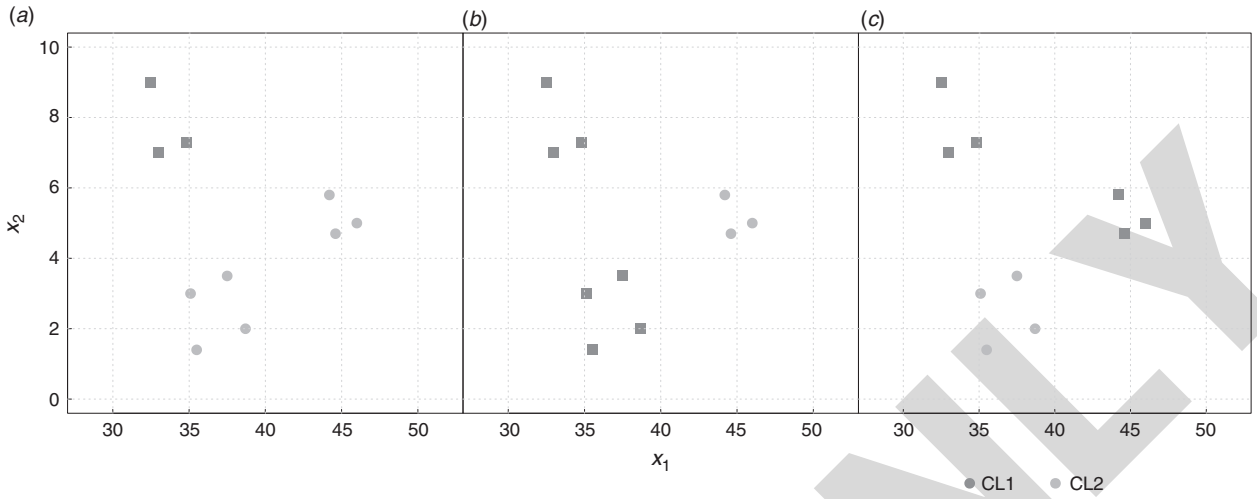

Fig. 1. Three possible ways of classifying 10 points (e.g. males) into two groups (CL1 and CL2) based on values of two variables (x_1 and x_2). (*a*) Variable x_2 drives the classification, with points with higher values classified as a unique group. (*b*) Variable x_1 drives the classification. (*c*) Variable x_2 drives the classification, but in this case points with lower values were classified as a unique group.

can be used to select a variable set with minimal redundancy, for data clustering (Flores *et al*. 2009; Gallego *et al*. 2015).

Most clustering attempts, with both CASA-Mot and CASA-Morph data, have relied on a compromise between the celerity of non-hierarchical methods and the flexibility and information provided by hierarchical methods. In two-step methods, the dataset is first partitioned into a predefined number of clusters. This first step is performed by a fast, non-hierarchical method and the resulting cluster centroids are fed into a hierarchical algorithm. This second step performs the final classification in a reasonable time, and is also used to determine the final number of clusters and to perform exploratory analyses on the classification. Two-step methods have been very popular for classifying CASA-Mot data (Abaigar *et al*. 1999; Martinez-Pastor *et al*. 2005; Martínez *et al*. 2006; Yániz *et al*. 2015*a*) and particularly CASA-Morph data (Peña *et al*. 2005; Esteso *et al*. 2009; Maroto-Morales *et al*. 2012, 2015). Recently, we proposed a variation of this methodology, in which a first hierarchical step was performed in individual samples, resulting in three to eight clusters per sample, and then a second hierarchical step reclassified the resulting centroids, reassigning the initial clusters to three to four subpopulations (Gallego *et al*. 2015; Fernández-Gago *et al*. 2017; Ledesma *et al*. 2017). This method is fast and allows exploration of the individual hierarchical classification within samples, but it requires a fairly high number of observations in each sample.

The use of unsupervised methods has yielded promising results on the detection and characterisation of sperm subpopulations depending on motility or morphometry parameters. However, whereas the studies shed light on the effects of capacitation, cryopreservation, individual variability etc. on sperm motility and morphology, there was a lack of association between the cluster structure and sperm fertility. Recently, some efforts to relate sperm subpopulations with field fertility have yielded fruitful results (Santolaria *et al*. 2015; Yániz *et al*. 2015*a*), adding practical meaning to this research area.

Nevertheless, the researcher must always keep in mind this advice, when using unsupervised methods: 'Clustering finds patterns in data – whether they are there or not' (Altman and Krzywinski 2017).

Statistical analysis of CASA data: statistical learning (supervised methods)

The advantage of unsupervised methods is that they allow for the categorisation of sperm heterogeneity in an efficient manner and without the need for any other prior information. Nevertheless, this advantage limits their applicability, especially when looking for functional associations of sperm heterogeneity with fertility or sperm cryoability (the resilience to withstand cryopreservation, also called freezability), among others. This limitation is illustrated in Fig. 1: this figure shows the three possible ways of classifying 10 points (e.g. males) into two groups (e.g. high and low fertility) depending on the value of two parameters (any morphometric or motility parameter; x_1 and x_2 in this example). Obtaining one classification or another will depend on the weight of the variables used for classification, on the clustering methods or the points chosen as starting values etc.; that is, on methodological aspects more than on physiological and/or functional aspects. Indeed, although articles reporting unsupervised methods reach a similar number of subpopulations, the characteristics of these subpopulations vary more or less widely among studies. To overcome this limitation (and also with the aim of implementing efficient and repeatable sperm classification protocols), prior information from other studies could be used. Following the example above, suppose that results from previous studies have shown that values of x_2 below 4 units are related with low fertility. Considering this information as a prior would lead us to the classification shown in Fig. 1*c*, and this would result in a signification association with the feature of interest (in this example, fertility).

The use of prior information guiding the clustering process is what characterises supervised methods. Supervised methods represent a step forward to unsupervised methods, in which prior information guides the prediction processes, and where outcomes from previous analysis can be used to update the inferred function and to predict new events. As Hastie *et al*. (2017) state, supervised learning makes use of inputs (a set of variables that are measured or preset and have some effect on one or more outputs) to predict the value of the outputs. In their book, Hastie *et al*. (2017) provide an in-depth review of machine learning methods and their applications in several research fields that may be of interest to those readers who want to start in this type of analysis.

The use of this type of analysis in reproductive biology, and specifically in CASA, is still scarce. One of the first studies implementing these methods was conducted by Goodson *et al*. (2011), who used support vector machines (SVM) to classify spermatozoa based on motility features throughout the transition from a progressive to hyperactive pattern in mice, and developed a software that uses SVM equations to classify individual sperm motility patterns automatically (this software can be requested from these authors). The same methodology was used by Ramón *et al*. (2012) to classify spermatozoa based on motility features in relation to sperm cryopreservation. Ramón *et al*. (2012) compared SVM with the unsupervised methods commonly used to assess sperm subpopulations (hierarchical, non-hierarchical and the multistep method proposed by Martinez-Pastor *et al*. 2005), and showed how SVMs were superior to classical methods. The use of supervised learning methods allowed associations to be found between the structure of subpopulations obtained from that analysis and male cryoability. In another study, Sahoo and Kumar (2014) compared five data-mining techniques on a fertility database to evaluate seminal quality and to predict whether the patient was either normal or had altered fertility based on environmental and lifestyle parameters or features. Focusing on morphometric features, Mirsky *et al*. (2017) used an SVM classifier to automatically classify spermatozoa as having good or bad morphology based on three-dimensional (3D) morphology information obtained by interferometric phase microscopy, as a prior step to the selection of sperm cells to be used for IVF. In another study, Chang *et al*. (2017) compared four supervised learning methods to characterise spermatozoa based on morphometric measures of sperm heads. Chang *et al*. (2017) emphasised the need to use automated methods given the high degree of inter-expert variability in the assessment of morphological sperm characteristics.

All the studies mentioned above performed classifications based on several sperm features but, more importantly, guided this classification according to functional aspects that helped find associations. It is expected that the use of these methodologies will increase in the future. The development of dedicated software for the classification process would contribute to the widespread use of these analyses while allowing automatisation of the procedure.

Going beyond the CASA systems

As pointed out in the Introduction, CASA-Mot and CASA-Morph systems have caused a revolution in the field of spermatology by allowing, in an objective way, the collection of a large amount of information about the morphological and motility characteristics of spermatozoa. The use of this information has revealed new associations between sperm characteristics and their functionality, which has ultimately allowed us to better understand the complex mechanism of the fertilisation process. Conversely, the implementation of these systems in the daily routine of assisted reproduction centres has allowed a better characterisation of sperm quality and an increase in fertility and prolificacy (Holt *et al*. 1997; Broekhuijse *et al*. 2015). Beyond these advantages, new technologies and the large amount of data they generate have led to new challenges, such as how to manage and interpret these data. Moreover, in order to manage and interpret these data, we need to deepen our understanding of the mechanisms that condition the fertilisation process.

CASA systems yield two-dimensional coordinates of several motility and morphometric features, and the use of mathematical formulas allows calculation of derived parameters for a better characterisation of the motility track or morphological dimensions. This procedure works well if the shape of the object we want to capture is simple, but fails if there are complexities in the shape, such as the sperm head apical hook in rodents. Furthermore, the measures provided by CASA systems do not allow consideration of the fact that spermatozoa swim in a 3D space or the fact that size and shape are not always equivalent. The implementation of geometric morphometrics (GM) analysis has been proposed to deal with some of these limitations. The core of these methods lies in the landmark-based approach in which the exact spatial position of a given anatomical structure is specified. Thus, GM methods allow the morphometry of an object to be assessed in a more precise way, considering all the particular characteristics that define that object in a way that is not affected by subjective aspects like scaling, rotation or translation (i.e. in a more generalisable way; Rohlf and Slice 1990; Bookstein 1997). Within the field of biological sciences, studies using GS methods have increased in the past decade, usually aimed at addressing questions in evolutionary morphology (Zelditch *et al*. 2012; Mcnulty and Vinyard 2015). More specifically, GM has been used to characterise the sperm head apical hook in mice and the role of sperm competition in modulating its shape (Firman and Simmons 2009; Firman *et al*. 2011). In a more recent study, Varea Sánchez *et al*. (2013) applied the principles of morphometrics to analyse rodent sperm head morphometry and compared this method with two traditional morphometric methods. All these studies highlight the potential of GM analysis, as well as the difficulties in interpreting GM results and the need for the integration of this analysis with other functional analyses. A technological innovation that tries to fill this functional gap is imaging flow cytometry (Basiji *et al*. 2007). This type of analysis couples the collection of high-throughput data with streamlined image analysis. Information on sperm features such as size and shape, granularity, intensity, radial distribution and texture can be obtained (Blasi *et al*. 2016) in a large sperm population. The main advantage of this technique, making it unique, is the ability to simultaneously evaluate morphometric and physiological parameters in the same cell. As for GM analysis, the main

challenge in imaging flow cytometry is the management and analysis of the data gathered. The use of machine learning methods discussed in this section may provide a useful framework for this propose, as already reported (Blasi *et al*. 2016).

Role of sperm morphometry and motility: how to reveal functional associations between sperm design and sperm function

The information obtained from CASA systems has proved useful in identifying relationships between sperm characteristics and functional aspects. Thus, different studies have reported relationships between sperm morphometry and motility and their role in fertility or survival following cryopreservation (Garde *et al*. 2006; Fitzpatrick *et al*. 2010; Ramón *et al*. 2013; Simpson *et al*. 2014). Although these studies had the same objective, the methodological approaches differed. In their study of red deer (*Cervus elaphus hispanicus*), Malo *et al*. (2006) based their findings on the small within-male and considerable between-male variation observed in sperm dimensions, which allowed the correct characterisation of individual sperm samples using mean values and their correspondence with differences in fertility. However, when low within-male variability and high between-male variability are not present, the use of average values is not suitable and characterisation of sample heterogeneity is required. This was the case in the study of Ramón *et al*. (2013), who characterised the subpopulation structure of sperm samples based on morphometric and motility parameters and made use of supervised learning methods to determine relationships between these two features and cryoability. Fitzpatrick *et al*. (2010), in fish, and Simpson *et al*. (2014) went further in the search for relationships between sperm morphometry and sperm motility, dealing with the intramale variation in a more efficient way, studying three internally and three externally fertilising species. These authors measured multiple morphological and motility traits from the same cell in order to look for correlations between sperm size and velocity, making use of high-definition video and image processing systems that allowed them to capture the shape and trajectories of each sperm cell in a detailed way. This approach represents a valuable improvement in the assessment of the relationships between sperm morphometry and motility, allowing the simultaneous evaluation of sperm heterogeneity and maximising our chances of finding functional relationships between these two features. The generalisation of this type of analysis may contribute to a better understanding of the mechanisms determining the fertilisation process and the role of different sperm traits in it. In this vein, the development of new analytical tools, such as imaging flow cytometry, will contribute to the expansion of these analyses.

Conclusions and practical recommendations on the statistical assessment of sperm motility and morphometry

Throughout this review we have tried to highlight the advantages of using advanced statistical tools to find patterns in databases obtained from sperm image analysis. The possibilities are enormous, and with improvements in microscopes, cameras and computers, richer data and more informative algorithms may be used.

Researchers must be aware of some caveats, some of which have been explained in more depth in other articles in this special volume (Bompart *et al*. 2018; Yániz *et al*. 2018*a*, 2018*b*; Yeste *et al*. 2018). Adequate equipment and standardised protocols for sample preparation and image acquisition are compulsory, but details are frequently overlooked (e.g. adequate quality controls), which may lead to within- and between-laboratory variability (Owen and Katz 1993). A typical example is the need of high camera frame rates when capturing motile spermatozoa (Castellini *et al*. 2011), well above those reported in most studies. Another warning deals with the variability of algorithms, for both the acquisition of CASA-Mot and CASMA-Morph data and the clustering of data and subsequent analysis (mostly proprietary software, with algorithms unknown to researchers). It is desirable to join other fields of biology in the adoption of open software (Swedlow and Eliceiri 2009), which can be examined and developed by any other researcher. Some authors have already contributed with open software for CASA-Mot (Wilson-Leedy and Ingermann 2007; Purchase and Earle 2012; Elsayed *et al*. 2015; Giaretta *et al*. 2017) and CASA-Morph (Butts *et al*. 2011). Moreover, the use of open platforms for performing statistical analyses, such as R (https://www.r-project.org/, accessed 1 April 2018) or Python (https://www.python.org/, accessed 1 April 2018) would allow for direct comparison of results between laboratories.

We must be also aware of new technological advances, or new uses for old ones, that may result in paradigm shifts, such as the use of fluorescence for the morphological study of the sperm nucleus (Vicente-Fiel *et al*. 2013), the aforementioned imaging flow cytometry or the use of 3D analysis. Thus, Mirsky *et al*. (2017) used an SVM classifier to automatically classify spermatozoa as having good or bad morphology based on 3D morphology information obtained using interferometric phase microscopy. Similarly, sperm analysis could be considerably enhanced by studying the motility of cells allowed to swim in any direction, as demonstrated recently (Su *et al*. 2013).

We have also highlighted the needed to integrate these systems with other tests and to take advantage of new statistical approaches to reveal functional associations. Therefore, in parallel with the technological developments described above, it is essential that statistical methodology and software be developed that allow the management and analysis of all these data, through the generalisation of its use.

Finally, we apologise for not citing all the relevant studies on this topic. The reference list provided is ample, and we invite researchers willing to implement and develop these methods to explore not only spermatology-related articles, but also general books on data clustering and machine learning.

Conflicts of interest

The authors declare no conflicts of interest.

References

Abaigar, T., Holt, W. V., Harrison, R. A., and del Barrio, G. (1999). Sperm subpopulations in boar (*Sus scrofa*) and gazelle (*Gazella dama mhorr*) semen as revealed by pattern analysis of computer-assisted motility assessments. *Biol. Reprod.* **60**, 32–41. doi:10.1095/BIOLREPROD60.1.32

Altman, N., and Krzywinski, M. (2017). Points of significance: clustering. *Nat. Methods* **14**, 545–546. doi:10.1038/NMETH.4299

Amann, R. P., and Waberski, D. (2014). Computer-assisted sperm analysis (CASA): capabilities and potential developments. *Theriogenology* **81**, 5–17.e3. doi:10.1016/J.THERIOGENOLOGY.2013.09.004

Basiji, D. A., Ortyn, W. E., Liang, L., Venkatachalam, V., and Morrissey, P. (2007). Cellular image analysis and imaging by flow cytometry. *Clin. Lab. Med.* **27**, 653–670. doi:10.1016/J.CLL.2007.05.008

Birkhead, T. R., Hosken, D. J., and Pitnick, S. S. (2008). 'Sperm Biology: An Evolutionary Perspective.' (Academic Press: Burlington, MA.)

Blasi, T., Hennig, H., Summers, H. D., Theis, F. J., Cerveira, J., Patterson, J. O., Davies, D., Filby, A., Carpenter, A. E., and Rees, P. (2016). Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nat. Commun.* **7**, 10256. doi:10.1038/NCOMMS10256

Bompart, D., García-Molina, A., Valverde, A., Caldeira, C., Yániz, J., Núñez de Murga, M., and Soler, C. (2018). CASA-Mot technology: how results are affected by the frame rate and counting chamber. *Reprod. Fertil. Dev.* . doi:10.1071/RD17551

Bookstein, F. L. (1997). 'Morphometric Tools for Landmark Data: Geometry and Biology.' (Cambridge University Press: Cambridge.)

Broekhuijse, M. L. W. J., Gaustad, A. H., Bolarin Guillén, A., and Knol, E. F. (2015). Efficient boar semen production and genetic contribution: the impact of low-dose artificial insemination on fertility. *Reprod. Domest. Anim.* **50**(Suppl. 2), 103–109. doi:10.1111/RDA.12558

Butts, I. A. E., Ward, M. A. R., Litvak, M. K., Pitcher, T. E., Alavi, S. M. H., Trippel, E. A., and Rideout, R. M. (2011). Automated sperm head morphology analyzer for open-source software. *Theriogenology* **76**, 1756–1761.e1–3. doi:10.1016/J.THERIOGENOLOGY.2011.06.019

Castellini, C., Dal Bosco, A., Ruggeri, S., and Collodel, G. (2011). What is the best frame rate for evaluation of sperm motility in different species by computer-assisted sperm analysis? *Fertil. Steril.* **96**, 24–27. doi:10.1016/J.FERTNSTERT.2011.04.096

Chang, V., Garcia, A., Hitschfeld, N., and Härtel, S. (2017). Gold-standard for computer-assisted morphological sperm analysis. *Comput. Biol. Med.* **83**, 143–150. doi:10.1016/J.COMPBIOMED.2017.03.004

Davis, R. O., Drobnis, E. Z., and Overstreet, J. W. (1995). Application of multivariate cluster, discriminate function, and stepwise regression analyses to variable selection and predictive modeling of sperm cryosurvival. *Fertil. Steril.* **63**, 1051–1057. doi:10.1016/S0015-0282(16)57547-5

Elsayed, M., El-Sherry, T. M., and Abdelgawad, M. (2015). Development of computer-assisted sperm analysis plugin for analyzing sperm motion in microfluidic environments using Image-J. *Theriogenology* **84**, 1367–1377. doi:10.1016/J.THERIOGENOLOGY.2015.07.021

Esteso, M. C., Fernández-Santos, M. R., Soler, A. J., Montoro, V., Martínez-Pastor, F., and Garde, J. J. (2009). Identification of sperm-head morphometric subpopulations in Iberian red deer epididymal sperm samples. *Reprod. Domest. Anim.* **44**, 206–211. doi:10.1111/J.1439-0531.2007.01029.X

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). 'Cluster Analysis.' (Wiley: Chichester.)

Fernández-Gago, R., Álvarez-Rodríguez, M., Alonso, M. E., González, J. R., Alegre, B., Domínguez, J. C., and Martínez-Pastor, F. (2017). Thawing boar semen in the presence of seminal plasma improves motility, modifies subpopulation patterns and reduces chromatin alterations. *Reprod. Fertil. Dev.* **29**, 1576–1584. doi:10.1071/RD15530

Firman, R. C., and Simmons, L. W. (2009). Sperm competition and the evolution of the sperm hook in house mice. *J. Evol. Biol.* **22**, 2505–2511. doi:10.1111/J.1420-9101.2009.01867.X

Firman, R. C., Cheam, L. Y., and Simmons, L. W. (2011). Sperm competition does not influence sperm hook morphology in selection lines of house mice. *J. Evol. Biol.* **24**, 856–862. doi:10.1111/J.1420-9101.2010.02219.X

Fitzpatrick, J. L., Garcia-Gonzalez, F., and Evans, J. P. (2010). Linking sperm length and velocity: the importance of intramale variation. *Biol. Lett.* **6**, 797–799. doi:10.1098/RSBL.2010.0231

Flores, E., Fernández-Novell, J. M., Peña, A., and Rodríguez-Gil, J. E. (2009). The degree of resistance to freezing–thawing is related to specific changes in the structures of motile sperm subpopulations and mitochondrial activity in boar spermatozoa. *Theriogenology* **72**, 784–797. doi:10.1016/J.THERIOGENOLOGY.2009.05.013

Gallego, V., Vílchez, M. C., Peñaranda, D. S., Pérez, L., Herráez, M. P., Asturiano, J. F., and Martínez-Pastor, F. (2015). Subpopulation pattern of eel spermatozoa is affected by post-activation time, hormonal treatment and the thermal regimen. *Reprod. Fertil. Dev.* **27**, 529–543. doi:10.1071/RD13198

Garde, J. J., Martinez-Pastor, F., Gomendio, M., Malo, A. F., Soler, A. J., Fernandez-Santos, M., Esteso, M. C., Garcia, A. J., Anel, L., and Roldan, E. R. S. (2006). The application of reproductive technologies to natural populations of red deer. *Reprod. Domest. Anim.* **41**(Suppl. 2), 93–102. doi:10.1111/J.1439-0531.2006.00773.X

Giaretta, E., Munerato, M., Yeste, M., Galeati, G., Spinaci, M., Tamanini, C., Mari, G., and Bucci, D. (2017). Implementing an open-access CASA software for the assessment of stallion sperm motility: relationship with other sperm quality parameters. *Anim. Reprod. Sci.* **176**, 11–19. doi:10.1016/J.ANIREPROSCI.2016.11.003

Goodson, S. G., Zhang, Z., Tsuruta, J. K., Wang, W., and O'Brien, D. A. (2011). Classification of mouse sperm motility patterns using an automated multiclass support vector machines model. *Biol. Reprod.* **84**, 1207–1215. doi:10.1095/BIOLREPROD.110.088989

Hastie, T., Tibshirani, R., and Friedman, J. (2017). 'The Elements of Statistical Learning: Data Mining, Inference, and Prediction.' 2nd edn. (Springer: New York, NY.)

Henning, H., Petrunkina, A. M., Harrison, R. A. P., and Waberski, D. (2014). Cluster analysis reveals a binary effect of storage on boar sperm motility function. *Reprod. Fertil. Dev.* **26**, 623–632. doi:10.1071/RD13113

Holt, C., Holt, W. V., Moore, H. D., Reed, H. C., and Curnock, R. M. (1997). Objectively measured boar sperm motility parameters correlate with the outcomes of on-farm inseminations: results of two fertility trials. *J. Androl.* **18**, 312–323.

Ledesma, A., Zalazar, L., Fernández-Alegre, E., Hozbor, F., Cesari, A., and Martínez-Pastor, F. (2017). Seminal plasma proteins modify the distribution of sperm subpopulations in cryopreserved semen of rams with lesser fertility. *Anim. Reprod. Sci.* **184**, 44–50. doi:10.1016/J.ANIREPROSCI.2017.06.015

Leonard, S. T., and Droege, M. (2008). The uses and benefits of cluster analysis in pharmacy research. *Res. Social Adm. Pharm.* **4**, 1–11. doi:10.1016/J.SAPHARM.2007.02.001

Malo, A. F., Gomendio, M., Garde, J., Lang-Lenton, B., Soler, A. J., and Roldan, E. R. S. (2006). Sperm design and sperm function. *Biol. Lett.* **2**, 246–249. doi:10.1098/RSBL.2006.0449

Maroto-Morales, A., Ramón, M., García-Álvarez, O., Soler, A. J., Fernández-Santos, M. R., Roldan, E. R. S., Gomendio, M., Pérez-Guzmán, M. D., and Garde, J. J. (2012). Morphometrically-distinct sperm subpopulations defined by a multistep statistical procedure in ram ejaculates: intra- and interindividual variation. *Theriogenology* **77**, 1529–1539. doi:10.1016/J.THERIOGENOLOGY.2011.11.020

Maroto-Morales, A., Ramón, M., García-Álvarez, O., Montoro, V., Soler, A. J., Fernández-Santos, M. R., Roldan, E. R. S., Pérez-Guzmán, M. D., and Garde, J. J. (2015). Sperm head phenotype and male fertility in ram semen. *Theriogenology* **84**, 1536–1541. doi:10.1016/J.THERIOGENOLOGY.2015.07.038

Maroto-Morales, A., García-Álvarez, O., Ramón, M., Martínez-Pastor, F., Fernández-Santos, M. R., Soler, A. J., and Garde, J. J. (2016). Current status and potential of morphometric sperm analysis. *Asian J. Androl.* **18**, 863–870. doi:10.4103/1008-682X.187581

Martínez, I. N., Moran, J. M., and Pena, F. J. (2006). Two-step cluster procedure after principal component analysis identifies sperm subpopulations in canine ejaculates and its relation to cryoresistance. *J. Androl.* **27**, 596–603. doi:10.2164/JANDROL.05153

Martinez-Pastor, F., Garcia-Macias, V., Alvarez, M., Herraez, P., Anel, L., and de Paz, P. (2005). Sperm subpopulations in Iberian red deer epididymal sperm and their changes through the cryopreservation process. *Biol. Reprod.* **72**, 316–327. doi:10.1095/BIOLREPROD.104.032730

Martínez-Pastor, F., Cabrita, E., Soares, F., Anel, L., and Dinis, M. T. (2008). Multivariate cluster analysis to study motility activation of *Solea senegalensis* spermatozoa: a model for marine teleosts. *Reproduction* **135**, 449–459. doi:10.1530/REP-07-0376

Martínez-Pastor, F., Tizado, E. J., Garde, J. J., Anel, L., and de Paz, P. (2011). Statistical series: opportunities and challenges of sperm motility subpopulation analysis. *Theriogenology* **75**, 783–795. doi:10.1016/J.THERIOGENOLOGY.2010.11.034

Mcnulty, K. P., and Vinyard, C. J. (2015). Morphometry, geometry, function, and the future. *Anat. Rec. (Hoboken)* **298**, 328–333. doi:10.1002/AR.23064

Mirsky, S. K., Barnea, I., Levi, M., Greenspan, H., and Shaked, N. T. (2017). Automated analysis of individual sperm cells using stain-free interferometric phase microscopy and machine learning. *Cytometry A.* **91**, 893–900. doi:10.1002/CYTO.A.23189

Owen, D. H., and Katz, D. F. (1993). Sampling factors influencing accuracy of sperm kinematic analysis. *J. Androl.* **14**, 210–221.

Peña, F. J., Saravia, F., García-Herreros, M., Núñez-martínez, I., Tapia, J. A., Johannisson, A., Wallgren, M., and Rodríguez-Martínez, H. (2005). Identification of sperm morphometric subpopulations in two different portions of the boar ejaculate and its relation to postthaw quality. *J. Androl.* **26**, 716–723. doi:10.2164/JANDROL.05030

Petrunkina, A. M., Waberski, D., Günzel-Apel, A. R., and Töpfer-Petersen, E. (2007). Determinants of sperm quality and fertility in domestic species. *Reproduction* **134**, 3–17. doi:10.1530/REP-07-0046

Purchase, C. F., and Earle, P. T. (2012). Modifications to the IMAGEJ computer assisted sperm analysis plugin greatly improve efficiency and fundamentally alter the scope of attainable data. *J. Appl. Ichthyol.* **28**, 1013–1016. doi:10.1111/JAI.12070

Quintero-Moreno, A., Rigau, T., and Rodriguez-Gil, J. E. (2007). Multivariate cluster analysis regression procedures as tools to identify motile sperm subpopulations in rabbit semen and to predict semen fertility and litter size. *Reprod. Domest. Anim.* **42**, 312–319. doi:10.1111/J.1439-0531.2006.00785.X

Ramón, M., Martínez-Pastor, F., García-Álvarez, O., Maroto-Morales, A., Soler, A. J., Jiménez-Rabadán, P., Fernández-Santos, M. R., Bernabéu, R., and Garde, J. J. (2012). Taking advantage of the use of supervised learning methods for characterization of sperm population structure related with freezability in the Iberian red deer. *Theriogenology* **77**, 1661–1672. doi:10.1016/J.THERIOGENOLOGY.2011.12.011

Ramón, M., Soler, A. J., Ortiz, J. A., García-Alvarez, O., Maroto-Morales, A., Roldan, E. R. S., and Garde, J. J. (2013). Sperm population structure and male fertility: an intraspecific study of sperm design and velocity in red deer. *Biol. Reprod.* **89**, 110. doi:10.1095/BIOLREPROD.113.112110

Ramón, M., Jimenez-Rabadan, P., Garcia-Alvarez, O., Maroto-Morales, A., Soler, A. J., Fernandez-Santos, M. R., Perez-Guzman, M. D., and Garde, J. J. (2014). Understanding sperm heterogeneity: biological and practical implications. *Reprod. Domest. Anim.* **49**, 30–36. doi:10.1111/RDA.12404

Rivera, M. M., Quintero-Moreno, A., Barrera, X., Palomo, M. J., Rigau, T., and Rodríguez-Gil, J. E. (2005). Natural Mediterranean photoperiod does not affect the main parameters of boar-semen quality analysis. *Theriogenology* **64**, 934–946. doi:10.1016/J.THERIOGENOLOGY.2005.01.001

Rohlf, F. J., and Slice, D. (1990). Extensions of the Procrustes method for the optimal superimposition of landmarks. *Syst. Zool.* **39**, 40–59. doi:10.2307/2992207

Sahoo, A. J., and Kumar, Y. (2014). Seminal quality prediction using data mining methods. *Technol. Health Care* **22**, 531–545. doi:10.3233/THC-140816

Santolaria, P., Vicente-Fiel, S., Palacín, I., Fantova, E., Blasco, M. E., Silvestre, M. A., and Yániz, J. L. (2015). Predictive capacity of sperm quality parameters and sperm subpopulations on field fertility after artificial insemination in sheep. *Anim. Reprod. Sci.* **163**, 82–88. doi:10.1016/J.ANIREPROSCI.2015.10.001

Simpson, J. L., Humphries, S., Evans, J. P., Simmons, L. W., and Fitzpatrick, J. L. (2014). Relationships between sperm length and speed differ among three internally and three externally fertilizing species. *Evolution* **68**, 92–104. doi:10.1111/EVO.12199

Su, T.-W., Choi, I., Feng, J., Huang, K., McLeod, E., and Ozcan, A. (2013). Sperm trajectories form chiral ribbons. *Sci. Rep.* **3**, 1664. doi:10.1038/SREP01664

Swedlow, J. R., and Eliceiri, K. W. (2009). Open source bioimage informatics for cell biology. *Trends Cell Biol.* **19**, 656–660. doi:10.1016/J.TCB.2009.08.007

Varea Sánchez, M., Bastir, M., and Roldan, E. R. S. (2013). Geometric morphometrics of rodent sperm head shape. *PLoS One* **8**, e80607. doi:10.1371/JOURNAL.PONE.0080607

Verstegen, J., Iguer-Ouada, M., and Onclin, K. (2002). Computer assisted semen analyzers in andrology research and veterinary practice. *Theriogenology* **57**, 149–179. doi:10.1016/S0093-691X(01)00664-1

Vicente-Fiel, S., Palacín, I., Santolaria, P., Hidalgo, C. O., Silvestre, M. A., Arrebola, F., and Yániz, J. L. (2013). A comparative study of the sperm nuclear morphometry in cattle, goat, sheep, and pigs using a new computer-assisted method (CASMA-F). *Theriogenology* **79**, 436–442. doi:10.1016/J.THERIOGENOLOGY.2012.10.015

Wilson-Leedy, J. G., and Ingermann, R. L. (2007). Development of a novel CASA system based on open source software for characterization of zebrafish sperm motility parameters. *Theriogenology* **67**, 661–672. doi:10.1016/J.THERIOGENOLOGY.2006.10.003

Xu, R., and Wunsch, D. C., II. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**, 645–678. doi:10.1109/TNN.2005.845141

Yániz, J. L., Palacín, I., Vicente-Fiel, S., Sánchez-Nadal, J. A., and Santolaria, P. (2015*a*). Sperm population structure in high and low field fertility rams. *Anim. Reprod. Sci.* **156**, 128–134. doi:10.1016/J.ANIREPROSCI.2015.03.012

Yániz, J. L., Soler, C., and Santolaria, P. (2015*b*). Computer assisted sperm morphometry in mammals: a review. *Anim. Reprod. Sci.* **156**, 1–12. doi:10.1016/J.ANIREPROSCI.2015.03.002

Yániz, J. L., Vicente-Fiel, S., Soler, C., Recreo, P., Carretero, T., Bono, A., Berné, J. M., and Santolaria, P. (2016). Comparison of different statistical approaches to evaluate morphometric sperm subpopulations in men. *Asian J. Androl.* **18**, 819–823. doi:10.4103/1008-682X.186872

Yániz, J. L., Silvestre, M. A., Santolaria, P., and Soler, C. (2018*a*). CASA-Mot in mammals: an update. *Reprod. Fertil. Dev* . doi:10.1071/RD17432

Yániz, J. L., Palacín, I., Caycho, K. S., Soler, C., Silvestre, M. A., and Santolaria, P. (2018*b*). Determining the relationship between bull sperm kinematic subpopulations and fluorescence groups using an integrated sperm quality analysis technique. *Reprod. Fertil. Dev.* . doi:10.1071/RD17441

Yeste, M., Bonet, S., Rodríguez-Gil, J. E., and Rivera Del Álamo, M. M. (2018). Evaluation of sperm motility with CASA-Mot: which factors may influence our measurements? *Reprod. Fertil. Dev.* . doi:10.1071/RD17479

Zelditch, M. L., Swiderski, D. L., and Sheets, H. D. (2012). 'Geometric Morphometrics for Biologists: A Primer.' (Academic Press: New York.)

AUTHOR QUERIES

AQ1: There are two Yániz papers so I have inserted both – is this correct?