

## Supplementary Material

### Subpopulation pattern of eel spermatozoa is affected by post-activation time, hormonal treatment and the thermal regimen

V. Gallego<sup>A</sup>, M. C. Vilchez<sup>A</sup>, D. S. Peñaranda<sup>A</sup>, L. Pérez<sup>A</sup>, M. P. Herráez<sup>B</sup>, J. F. Asturiano<sup>A</sup> and F. Martínez-Pastor<sup>B,C</sup>

<sup>A</sup>Grupo de Acuicultura y Biodiversidad, Instituto de Ciencia y Tecnología Animal, Universitat Politècnica de València, Camino de Vera s/n. 46022, Valencia, Spain.

<sup>B</sup>Institute for Cattle Development and Animal Health Institute for Cattle Development and Animal Health (INDEGSAL) and Molecular Biology, University of León, 24071 León, Spain.

<sup>C</sup>Corresponding author. Email: felipe.martinez@unileon.es

#### Materials and methods

##### *CASA analysis*

Sperm motility was analysed following the standardized protocol for European eel spermatozoa described by Gallego *et al.* (2013). The CASA system consisted of by a triocular optical phase contrast microscope (Eclipse E-400; Nikon, Tokyo, Japan), with a  $\times 10$  negative-contrast phase lens and an ISAS 782M camera at 60 fps, connected to a computer by an IEEE 1394 interface. To acquire motility, 2  $\mu\text{L}$  of undiluted sperm sample were diluted in 200  $\mu\text{L}$  of artificial sea water (Aqua Medic Meersalz, 37 g/L, 2% BSA (w/v), pH 8.2; Peñaranda *et al.* 2010). Four  $\mu\text{L}$  of this dilution were immediately charged in a SpermTrack-10® (Proiser R+D S.L., Paterna, Spain). At exactly 30 s post-activation (and at 60 and 90 s in Experiment 1), motility was acquired for 1 s at 60 frames/s using the ISAS ver. 1.2 software (Proiser, Paterna, Spain). The software was configured with 2 to 20  $\mu\text{m}^2$  for the head area and VCL  $> 10 \mu\text{m/s}$  to classify a spermatozoon as motile.

The software yielded the following parameters for each spermatozoa: three velocity parameters (VCL: velocity according to the actual path; VSL: velocity according to the straight path; VAP: velocity according to the smoothed path), three track linearity parameters (LIN: linearity; STR: straightness; WOB: wobble), the ALH (amplitude of the lateral displacement of the sperm head), the BCF (head beat-cross frequency), Dance and Dance Mean (measurements of the pattern of sperm motion). A brief description of each parameter is provided in Table S2.

##### *Subpopulation and statistical analyses*

Subpopulation and statistical analyses were carried out using the R statistical environment (R Core Team, 2013), a free software environment for statistical computing and graphics. Installed packages for carrying out the analyses are shown in Table S2.

First, motility data were processed to remove events appearing in less than 50 consecutive frames (broken or incomplete tracks, or tracks resulting from spermatozoa entering or leaving the field while acquiring). Samples with too few total spermatozoa or too few motile spermatozoa (less than 30 motile spermatozoa)

were removed from the subpopulation analysis to prevent the apparition of spurious clusters or biased parameters. Data were then processed to obtain the total motility and the median kinematic parameters of each sample. Spermatozoa were considered as motile if their VCL was higher than 10  $\mu\text{m/s}$ , and their number was divided by the total number of spermatozoa in each sample, presenting total motility as a percentage. The median (less biased than the mean by extreme values) of each of the motility variables was also calculated and used subsequently. These data were used for conventional motility analysis. A subpopulation analysis was carried out separately in each of the three datasets resulting from the three experiments, following statistical procedures for clustering CASA data (Martínez-Pastor *et al.* 2011). Prior to this, all variables were transformed in order to normalize variances. Arc sine transformation was used for percentages (LIN, STR and WOB), applying the formula  $\arcsin(\sqrt{\text{value}/100})$ . Natural logarithm transformation was used for other variables. The variables were also standardized. This process consists of centering (subtracting the mean) and scaling (dividing by the standard deviation), resulting in all variables having a mean of 0 and standard deviation of 1. Otherwise, variables with high values (e.g. VCL) would have a higher weight in the subsequent steps than variables with low values (e.g. ALH). Not all the kinematic variables were used in the clustering process. The variables were chosen by performing a hierarchical clustering (more similar variables, conveying similar information and thus being redundant, were clustered together), using the Hoeffding D statistic as a measure of independence (Fig. S1). A set of representative variables (minimising redundancy) was selected in each experiment. The clustering procedure was a modification of two-step methods proposed previously (Martínez-Pastor *et al.* 2005; Domínguez-Rebolledo *et al.* 2009). Two hierarchical clustering steps were used consecutively, and the reliability and stability of the solutions were checked as recommended (Martínez-Pastor *et al.* 2011). In the first step, the observations (individual spermatozoa) belonging to each sample were classified using an algorithm for agglomerative nesting processing (AGNES; Kaufman and Rousseeuw 1990). This kind of hierarchical clustering algorithm starts with individual observations and, in successive steps, it pairs them according to similarity, until all observations are grouped in a single cluster. The clustering algorithm requires a dissimilarity matrix that contains “distances” between observation pairs. This dissimilarity matrix was calculated from the raw data using euclidean metrics. Ward’s method was used for the clustering algorithm itself and performed satisfactorily with different CASA datasets.

The number of clusters ( $k$ ) was decided based on the *Silhouette information* for each  $k = [2, 8]$ , choosing a  $k$  such that the *Silhouette average width* was maximized. Once we clustered the data from each individual sample, we calculated the median values of the kinematic parameters for each of those clusters (from several hundred to more than 1000, depending on the experiment), and these median values were used in a second step using the same clustering method. The purpose of this second step was to join together similar clusters from different samples (in many cases, within the same sample), reaching a reduced number of clusters (Fig. S2). The cluster allocations obtained in the second step were lined up with the original clustering, ending up with the observations of each individual sample being clustered within a few subpopulations.

In order to analyse the subpopulation characteristics and their distribution throughout the spermiation period and across treatments, we obtained the median values for each subpopulation (using the whole dataset), and each sample was characterized according to the relative proportion of each subpopulation. Hypothesis testing on motility and subpopulation proportions were conducted as described in the main paper. Results were shown either as box-plots (Fig. S3) or as mean  $\pm$  s.e.m. plots.

## Results

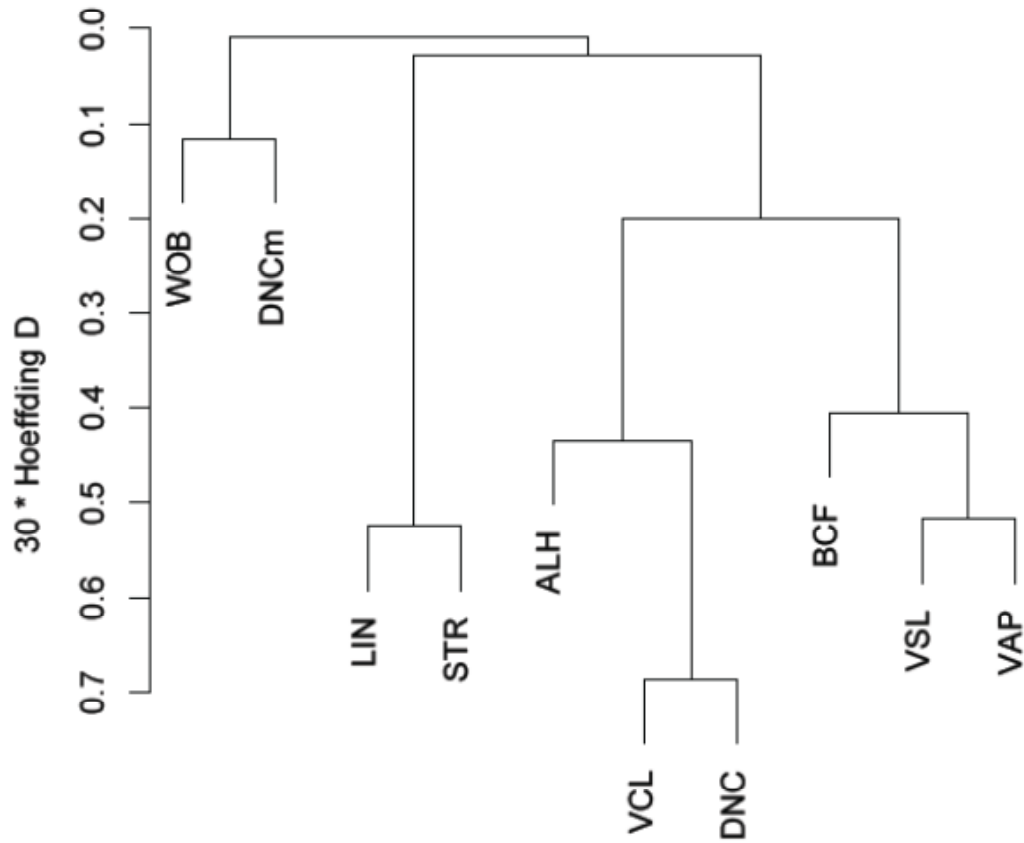
### *Eel body weight and sperm production in Experiment 3*

Figure S4a shows the variation of the body weight by week. The body weight influences the amount of hormone injected. It decreased from the beginning of spermiation, reflecting the absence of feeding of the adult eels, and it differed very little between hormonal treatments ( $P > 0.05$ ). Figure S4b shows the total recovery of spermatozoa (adding the amount obtained from each spermiating male) each week within each treatment, which overall was higher for hCG<sub>rec</sub>. Figure S4c shows the same visualization for SP3 spermatozoa, with a similar profile (even with a slightly higher advantage for hCG<sub>rec</sub>). It is clear from these profiles that hCG<sub>rec</sub> allowed the recovery of a higher quantity of spermatozoa and, more important, a higher quantity of SP3 (high-quality) spermatozoa. The amount of spermatozoa recovered by using hCG<sub>rec</sub> was larger in the first weeks of spermiation (8–13), but, in general, this treatment allowed us to obtain a

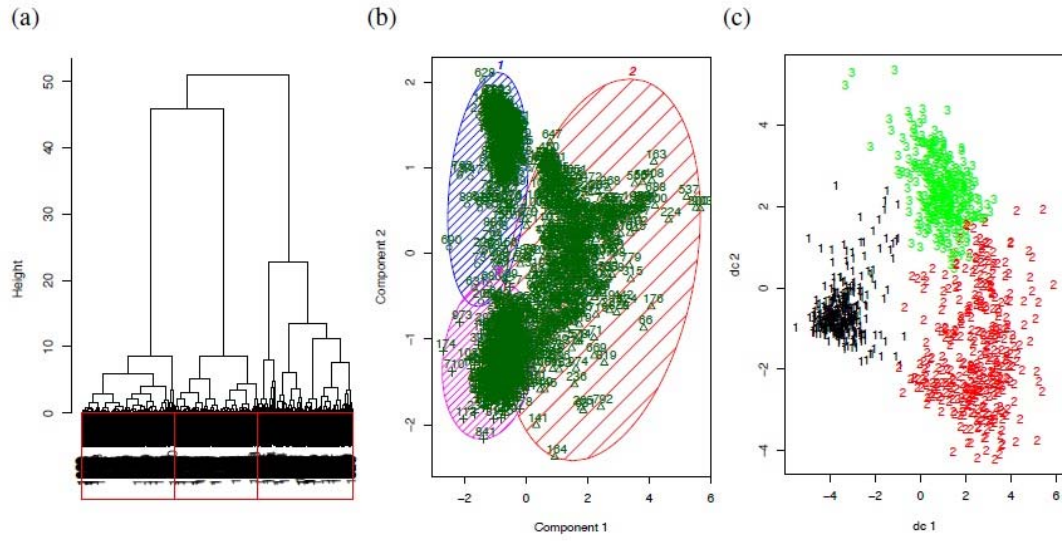
fairly high number of high-quality spermatozoa for most of the spermiation period. This resulted in the best yield/cost ratio for the three treatments, as shown in the main paper (Fig. 7).

## References

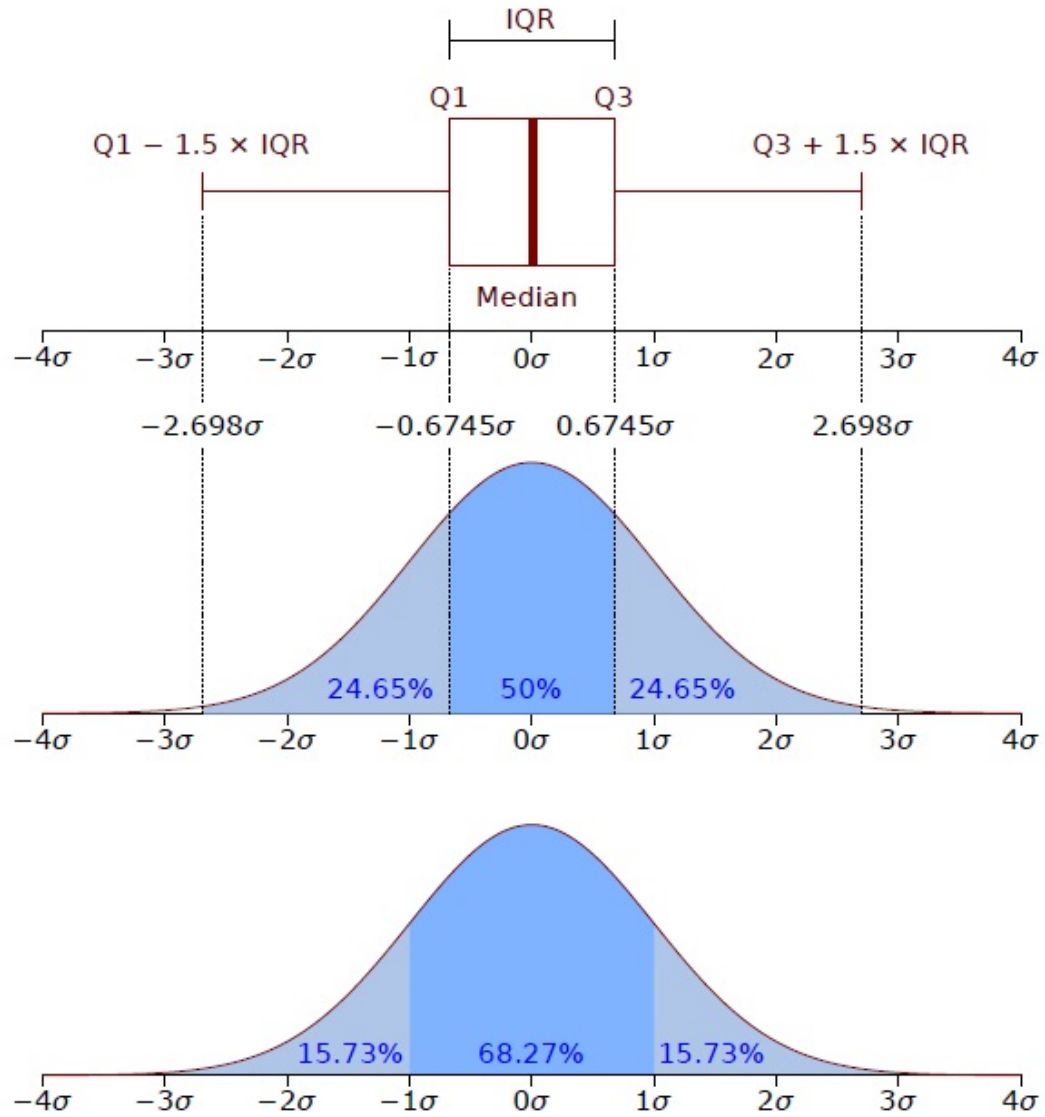
- Domínguez-Rebolledo, A. E., Fernández-Santos, M. R., García-Alvarez, O., Maroto-Morales, A., Garde, J. J., and Martínez-Pastor, F. (2009). Washing increases the susceptibility to exogenous oxidative stress in red deer spermatozoa. *Theriogenology* **72**, 1073–1084.
- Gallego, V., Carneiro, P. C. F., Mazzeo, I., Vílchez, M. C., Peñaranda, D. S., Soler, C., Pérez, L., and Asturiano, J. F. (2013). Standardization of European eel (*Anguilla anguilla*) sperm motility evaluation by CASA software. *Theriogenology* **79**, 1034–1040.
- Hennig, C. (2013). ‘fpc: Flexible procedures for clustering.’ R package version 2.1-5.
- Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biom. J.* **50**, 346–363.
- Kaufman, L., and Rousseeuw, P. (1990). ‘Finding Groups in Data: An Introduction to Cluster Analysis.’ (Wiley: New York).
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2013). ‘cluster: Cluster Analysis Basics and Extensions.’ R package version 1.14.4.
- Martínez-Pastor, F., García-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.
- 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.
- Morales, M., with code developed by the R Development Core Team, with general advice from the R-help listserv community, and especially Duncan Murdoch (2012). ‘sciplot: Scientific Graphing Functions for Factorial Designs.’ R package version 1.1-0.
- Peñaranda, D. S., Marco-Jiménez, F., Pérez, L., Gallego, V., Mazzeo, I., Vicente, J. S., Jover, M., and Asturiano, J. F. (2010). Evaluation of different diluents for short-term storage of European eel sperm under air-limited conditions. *J Appl. Ichthyol.* **26**, 659–664.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013). ‘nlme: Linear and Nonlinear Mixed Effects Models.’ R package version 3.1-109.
- R Core Team (2013). ‘R: A Language and Environment for Statistical Computing.’ R Foundation for Statistical Computing, Vienna, Austria.
- Venables, W. N., and Ripley, B. D. (2002). ‘Modern Applied Statistics with S.’ (Springer: New York), fourth edition.
- Warnes, G. R., includes R source code and/or documentation contributed by Bolker, B., Lumley, T., and Johnson, R. C. (2012). ‘gmodels: Various R programming tools for model fitting.’ R package version 2.15.3.
- Warnes, G. R., with contributions from Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., Rogers, J., and others (2013). ‘gdata: Various R programming tools for data manipulation.’ R package version 2.12.0.2.
- Wickham, H. (2009). ‘ggplot2: elegant graphics for data analysis.’ (Springer: New York).
- Wickham, H. (2011). The split-apply-combine strategy for data analysis. *J. Stat. Softw.* **40**, 1–29.



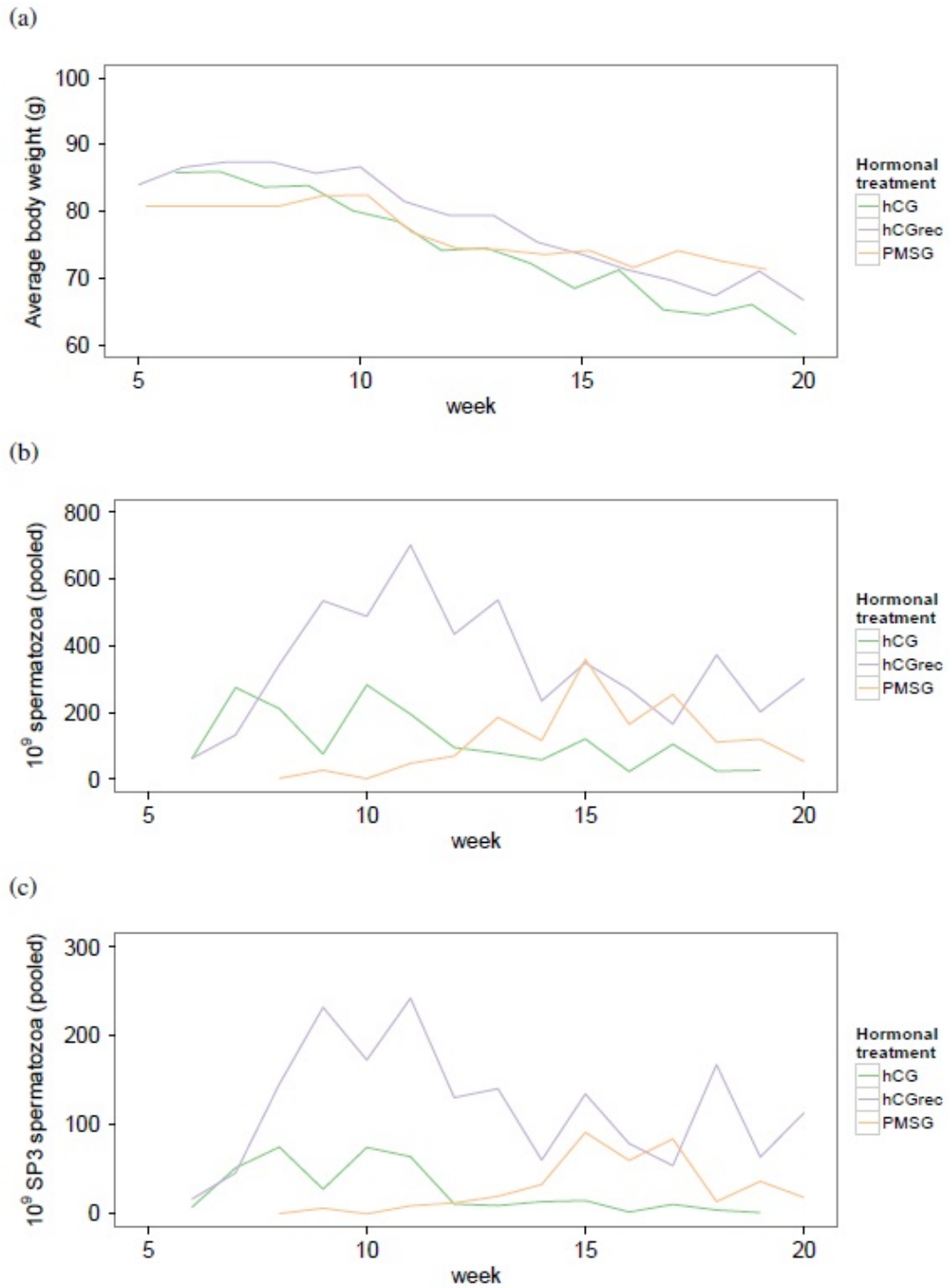
**Fig. S1.** Different visualisations of the solution for the second clustering step for Experiment 3. Panel (a) shows the tree resulting from the hierarchical clustering carried out by the AGNES algorithm using Ward's method. The algorithm works "bottom-up", joining observations in successive steps. The lower part of the plot is blurred because of the high number of starting observations, which are the clusters obtained in the first step (1045 in this experiment). In the upper part of the top, the three "branches" representing the three subpopulations obtained are evident. Panel (b) plots the clusters from the first step in a bidimensional space defined by the two first principal components obtained from the variables used in the clustering (98% of the variance explained). The ellipses enclose the observations belonging to each final subpopulation. Panel (c) displays a similar visualization, but using discriminant coordinates ("canonical variables") obtained from the variables used in the clustering. These plots were obtained using the R functions `plot.agnes()`, `clusplot()` and `plotcluster()`, respectively.



**Fig. S2.** Proportion of each sperm subpopulation (Table 1) in each acquisition time ( $x$ -axes) (Experiment 1). Different letters indicate that groups (acquisition times) differ  $P < 0.05$ . The proportions of the three subpopulations differed significantly ( $P < 0.05$ ). See Fig. S3 in the supplementary material for a description of box-plots.



**Fig. S3.** Box-plots allow the visualization of the distribution of data. The top schematics show a box-plot in a horizontal position, built from a normal distribution (graphs below), in order to compare their elements with standard deviation ( $\sigma$ ) values. The box is drawn by extracting the first and third quartiles ( $Q1$  and  $Q3$ , also called Tukey's lower-hinge and higher-hinge of the data), and therefore it encloses 50% of the data. The length of the box is the interquartile range ( $IQR$ ). The  $IQR$  times 1.5 is the length of the protruding segments, called "whiskers". If the whisker limits are farther than the highest or lowest observation, the whiskers are drawn only up to that observation (e.g., Fig. 1a in the main manuscript). Usually, observations beyond the whisker limits are plotted as dots (not drawn here, but shown in Fig. 1b of the main manuscript, for instance). The line inside the box shows the localization of the median. The distributions below the box-plot help to understand how the data is distributed within each element of the box-plot, compared to the familiar  $\pm\sigma$  of the normal distribution. Since this box-plot was built from a normal distribution, it is symmetrical and the median is located just in the middle of the box-plot, but this is generally not the case with experimental data. Indeed, one of the strengths of box-plots is that it allows us to easily visualize the distribution of the data in a simple way, while obtaining many useful statistics (such as Tukey's five numer summary: minimum,  $Q1$ , median,  $Q3$  and maximum).



**Fig. S4.** Summary of the clustering analysis of Experiment 2 (water temperature), showing the proportions of subpopulation 1 (“slow swimmers”), 2 (“circular swimmers”) and 3 (“fast swimmers”) (Table 2), along time (weeks,  $x$ -axis) and within each treatment (see Fig. 3 for the description of treatments and plot elements). For T15, the proportion of subpopulation 1 fitted a negative quadratic model, whereas subpopulation 3 data fitted a positive linear model. For T20, subpopulation 1 data fitted a negative quadratic model, and subpopulations 2 and 3 fitted positive quadratic models.

**Table S8. Definitions of standardised abbreviations of the kinematic parameters provided by CASA (Boyers *et al.* 1989)**

Only some of them were used in our study

Name	Acronym	Formula (if derived)	Units	Interpretation
Curvilinear velocity	VCL		$\mu\text{m/s}$	The time-average velocity of the sperm head along its actual trajectory (as rendered by the software by joining the position of the sperm head in successive images).
Straight-line velocity	VSL		$\mu\text{m/s}$	The time-average velocity of the sperm head along a straight line from its first detected position to its last detected position.
Average-path velocity	VAP		$\mu\text{m/s}$	The time-average velocity of the sperm head along its average trajectory. The average trajectory is computed by smoothing the actual trajectory.
Linearity	LIN	$\text{VSL/VCL}$	%	Similarity of the rendered trajectory to a straight line.
Straightness	STR	$\text{VSL/VAP}$	%	Similarity of the averaged trajectory to a straight line.
Wobble	WOB	$\text{VAP/VCL}$	%	Similarity of the averaged trajectory to the rendered trajectory, a measure of the oscillation of the sperm head across its main direction of movement.
Amplitude of lateral head displacement	ALH		$\mu\text{m}$	The amplitude of the lateral movement trajectory, actually, the distance of the sperm head to the averaged trajectory in each image. Usually the mean ALH is reported.
Beat-cross frequency	BCF		Hz	The time-average rate at which the actual sperm trajectory crosses the average path trajectory, and an indirect measurement of the flagellar beat frequency.
Dance	DNC	$\text{VCL} \times \text{ALH}$	$\mu\text{m}^2/\text{s}$	A measure of the pattern of sperm motion, ideally the area swept by the head along its main direction of movement (a combination of the lateral and forward movement of the head).
Dance Mean	DNCm	$\text{VCL} \times \text{ALH/VSL}$ or $\text{ALH/LIN}$	$\mu\text{m}$	A measure of the pattern of sperm motion, with a higher value indicating a more zigzagging path.



**Table S2. R packages (libraries) used in this study (packages loaded as “dependencies” are not included)**

Package name	Title	Reference
cluster	Cluster analysis (extended from Rousseeuw <i>et al.</i> )	Maechler <i>et al.</i> (2013)
fpc	Flexible procedures for clustering	Hennig (2013)
gdata	Various R programming tools for data manipulation	Warnes <i>et al.</i> (2013)
gmodels	Various R programming tools for model fitting	Warnes <i>et al.</i> (2012)
ggplot2	An implementation of the Grammar of Graphics	Wickham (2009)
MASS	Support Functions and Datasets for Venables and Ripley’s MASS	Venables and Ripley (2002)
multcomp	Simultaneous Inference in General Parametric Models	Hothorn <i>et al.</i> (2008)
nlme	Linear and Nonlinear Mixed Effects Models	Pinheiro <i>et al.</i> (2013)
plyr	Tools for splitting, applying and combining data	Wickham (2011)
sciplot	Scientific Graphing Functions for Factorial Designs	Morales <i>et al.</i> (2012)