



Jan 27, 2026

Version 2

Systematic Method for Identifying Kinematic Sperm Subpopulations V.2

 [PLOS One](#)

DOI

<https://dx.doi.org/10.17504/protocols.io.6qpvr9e6zvmk/v2>

Cindy Rivas Arzaluz¹, María Elena Ayala Escobar², Andrés Aragón Martínez³

¹Instituto de Biotecnología, UNAM; ²FES Zaragoza, UNAM; ³FES Iztacala, UNAM

Andrés Aragón Martínez: Correspondence armandres@gmail.com;

Laboratory of Gametes ...



Cindy Rivas

Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.6qpvr9e6zvmk/v2>

External link: <https://doi.org/10.1371/journal.pone.0338399>

Protocol Citation: Cindy Rivas Arzaluz, María Elena Ayala Escobar, Andrés Aragón Martínez 2026. Systematic Method for Identifying Kinematic Sperm Subpopulations. [protocols.io https://dx.doi.org/10.17504/protocols.io.6qpvr9e6zvmk/v2](https://dx.doi.org/10.17504/protocols.io.6qpvr9e6zvmk/v2) Version created by [Cindy Rivas](#)

Manuscript citation:

Juárez-Contreras NA, Rivas-Arzaluz CU, Pérez OG, Ayala ME, Aragón-Martínez A (2025) Development of an open-hardware semen homogenizer and application to serotonin effects on sperm motility. PLOS One 20(12). doi: [10.1371/journal.pone.0338399](https://doi.org/10.1371/journal.pone.0338399)

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: January 27, 2026

Last Modified: January 27, 2026

Protocol Integer ID: 241663

Keywords: CASA system, kinematic sperm subpopulations, dataset, identifying kinematic sperm subpopulation, kinematic sperm subpopulations the identification, assisted sperm analysis, kinematic sperm subpopulation, kinematic sperm subpopulations in boar sperm data, processing of motility parameter dataset, boar sperm data, sperm within each subpopulation, motility parameter dataset, sperm, motility parameter, sperm cell, kinematic subpopulation, using principal component analysis, principal component analysis, number of motility parameter, motility parameter value, amplitude of lateral head displacement, principal component, pca, lateral head displacement, various statistical approach, identification process, curvilinear velocity

Funders Acknowledgements:

UNAM-PAPIIT

Grant ID: IN224925

Disclaimer

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to [protocols.io](#) is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with [protocols.io](#), can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

Abstract

The identification of kinematic sperm subpopulations typically relies on motility parameter values obtained using a Computer-Assisted Sperm Analysis (CASA) system. These parameters are calculated individually for each sperm cell. The number of motility parameters provided depends on the specific CASA system manufacturer. However, commonly evaluated parameters include curvilinear velocity (VCL), average path velocity (VAP), straight-line velocity (VSL), linearity (LIN), straightness (STR), beat cross-frequency (BCF), wobble (WOB), and amplitude of lateral head displacement (ALH).

Various statistical approaches utilize the motility parameter values generated by a CASA system to identify kinematic sperm subpopulations (Martínez-Pastor et al., 2011; Ramón and Martínez-Pastor, 2018). This identification process is typically carried out in two stages. In the first stage, the dataset's dimensionality is reduced using principal component analysis (PCA). In PCA, components with eigenvalues greater than 1 are retained. In the second stage, these principal components are used as input for clustering algorithms, such as K-means or hierarchical agglomerative clustering.

When employing hierarchical agglomerative clustering, the objective is to group data points based on their proximity to one another while maximizing the separation between groups. This is typically achieved using various distance metrics, with Euclidean distance being the most commonly applied. Selecting the appropriate number of clusters is critical and often involves the use of an objective criterion, such as inertia gain. Graphical analysis of inertia gain can be particularly helpful in this context.

In this document, we present the strategy and procedure for identifying kinematic sperm subpopulations in boar sperm data subjected to different pH conditions (Rivas et al., 2022). The proposed workflow enables the processing of motility parameter datasets obtained via a CASA system to identify kinematic subpopulations in two systematic steps. The final output is a dataset containing the original input data alongside an additional column indicating the cluster assignment for each data point. This resulting dataset facilitates the determination of the proportion of sperm within each subpopulation and allows for assessing the effects of different treatments on these subpopulations.



Guidelines

The dataset used in this workflow is a dataframe consisting of 11,620 rows of data. The file is called motility_results_ph.csv and will be stored in an object named mr. However, for the purposes of this document, a sample of 500 rows will be used.

Object mr:

- Object Type: Data frame
- Number of Rows: 11,620
- Number of Variables: 11
- Description of Variables:

1. ID1 (Factor with 75 levels): A unique identifier that represents date and number of video. Examples of values include "220219-1", "220219-10", etc.

2. ID2 (Factor with 5 levels): Each level is a value of pH ("pH.70", "pH.74", etc.)

3. ID3 (Factor with 6 levels): Each level is the name of a semental ("balder", "berthrand", etc.)

4. VCL (num): Curvilinear Velocity, a measure of velocity along a curved path.

5. VAP (num): Average Path Velocity, the velocity calculated along a smoothed path.

6. VSL (num): Straight Line Velocity, velocity between two points in a straight line.

7. LIN (num): Linearity of movement, calculated as VSL/VCL . It is a value between 0 and 1.

8. STR (num): Straightness of movement, calculated as VSL/VAP . Also a value between 0 and 1.

9. WOB (num): Wobble, a measure of the oscillation of the movement, calculated as VAP/VCL .

10. BeatCross (num): Beat Cross Frequency, the frequency with which the trajectory crosses the midpoint reference.

11. ALH (num): Amplitude of Lateral Head displacement, a measure of lateral movement of head.



Before start

Libraries:

To replicate the procedure presented in this document, ensure that the FactoMineR and dplyr libraries are installed. These can be downloaded and installed directly in R if not already available: <https://cran.r-project.org/>

Data:

Download the dataset 230819_pluginnuevo3.tab and save it to your working directory. The dataset is available at <https://dataverse.harvard.edu> (<https://doi.org/10.7910/DVN/CRVVS1>) (Rivas et al., 2022).

Next steps:

Download the following file directly from R:

Command

new command name

```
download.file(url =  
"https://dataverse.harvard.edu/api/access/datafile/4549661",  
destfile="motility_results_ph.csv")
```

Create an object named mr and store the contents of the downloaded dataset in it

Command

new command name

```
mr<-read.csv("motility_results_ph.csv", header=TRUE, sep=",",  
stringsAsFactors=TRUE)
```

The file *motility_results_ph.csv* contain 11,620 lines of data, but for the purpose of this document, a sample of 500 lines of data will be used.



Command

new command name

```
mr_sample <- mr[sample(nrow(mr), 500, replace = FALSE), ]
```

Stage 1. Acquisition and creation of objects

- 1 The dataset used in this workflow is a dataframe that contains 11,620 lines of data, for the purpose of this document, a sample of 500 lines of data will be used.
We can access it and create the working object in two ways:

1.1 Option 1

The dataset used in this workflow can be downloaded from our account on the Harvard Dataverse site at <https://doi.org/10.7910/DVN/CRVVS1>
Save the dataset in your working directory and create an object to work with a sample of the dataset (500 rows). Review the 'Before Start' section.

1.2 Option 2

In the R software, download the dataset using the following code

Command

new command name

```
download.file(url =  
"https://dataverse.harvard.edu/api/access/datafile/4549661",  
destfile="230819_pluginuevo.tab")
```

- 1.3 Create an object named mr and use the read.csv function to read the previously downloaded file

Command

new command name

```
mr<- read.csv("230819_pluginuevo3.tab", header=TRUE, sep="\t",  
stringsAsFactors=TRUE)
```

- 1.4 Load the dplyr library in the R software using the following code

Command

new command name

```
library(dplyr)
```

Expected result

```
##  
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':  
##  
##   filter, lag
```

```
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

- 1.5 Take a sample of 500 rows from the dataset and store them in a new object named `mr_sample`

**Command****new command name**

```
mr_sample<- mr %>% group_by(ID2) %>% slice_sample(n=100,  
replace=FALSE)
```

- 1.6 Convert the mr_sample object into a dataframe and verify its structure

Command**new command name**

```
mr_sample<-as.data.frame(mr_sample)  
str(mr_sample)
```



Expected result

```
## 'data.frame': 500 obs. of 11 variables:
## $ ID1 : Factor w/ 75 levels "220219-1","220219-10",...:
69 28 58 46 70 6 7 28 69 58 ...
## $ ID2 : Factor w/ 5 levels "pH.70","pH.74",...: 1 1 1 1
1 1 1 1 1 1 ...
## $ ID3 : Factor w/ 6 levels "", "balder", "berthrand",...:
2 3 2 4 2 6 6 3 2 2 ...
## $ VCL : num 265.6 47.2 356.8 239.3 94.3 ...
## $ VAP : num 108.6 36.6 143.3 86 65.6 ...
## $ VSL : num 68.9 28 57 55.4 41.3 ...
## $ LIN : num 0.259 0.592 0.16 0.232 0.438 ...
## $ STR : num 0.635 0.764 0.398 0.644 0.631 ...
## $ WOB : num 0.409 0.775 0.402 0.359 0.695 ...
## $ BeatCross: num 30.8 38.7 28.2 38.1 29 ...
## $ ALH : num 8.62 1.96 12.42 6.7 4.69 ...
```

The `mr_sample` object has 500 rows and 11 columns. Column 1 (ID1) represents the identifier of the capture routine video. Column 2 (ID2) corresponds to the treatment, indicating different pH values. Column 3 (ID3) represents the identifier of the donor male.

Stage 2: Principal Component Analysis (PCA)

- 2 Load the required libraries

Command

new command name

```
library(FactoMineR)
```

- 2.1 Create a new object named `mr_sample2` that contains only the selected variables (columns)

Command

new command name

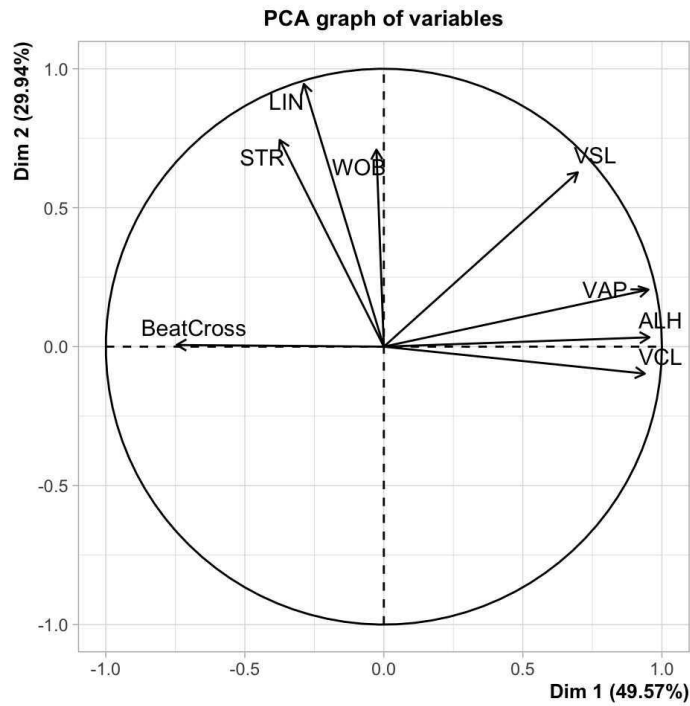
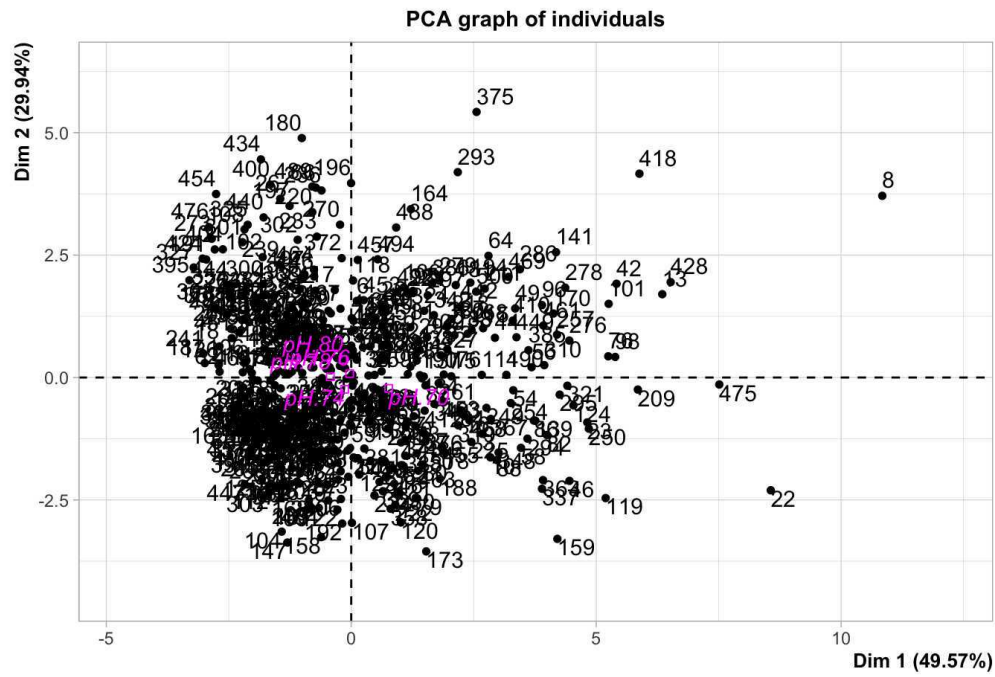
```
mr_sample_2<-mr_sample[,c(2,4:11)]
```

- 2.2 Create a new object named `res.pca` to store the results of the Principal Component Analysis (PCA) function. Specify the following arguments in the PCA function: scale the data, calculate three principal components, include a qualitative variable with no weight in the results, and generate a plot of the results.

Command

new command name

```
res.pca = PCA(mr_sample_2, scale.unit=TRUE, ncp=3, quali.sup=1,  
graph=TRUE)
```



The PCA function generates two plots:

1. PCA Graph of Individuals: This plot displays the values of the first and second principal components for each individual, with the corresponding labels observed next to each point.
2. PCA Graph of Variables: This plot presents a correlation circle, where each arrow represents the magnitude and direction (positive or negative) of the correlation of a variable with the first and second principal components. For instance, the variable VCL shows a strong positive correlation with the first principal component and a weak correlation with the second principal component

- 2.3 Type `res.pca` in the console to displays the list of contained objects, followed by their description.

Note

The `res.pca` object is of the list type, and its contents can be accessed using the `$` operator.

Command

new command name

`res.pca`

Expected result

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 500 individuals, described by
## 9 variables
## *The results are available in the following objects:
##
##      name                description
## 1  "$eig"                "eigenvalues"
## 2  "$var"                "results for the variables"
## 3  "$var$coord"          "coord. for the variables"
## 4  "$var$cor"            "correlations variables - dimensions"
## 5  "$var$cos2"           "cos2 for the variables"
## 6  "$var$contrib"        "contributions of the variables"
## 7  "$ind"                "results for the individuals"
## 8  "$ind$coord"          "coord. for the individuals"
## 9  "$ind$cos2"           "cos2 for the individuals"
## 10 "$ind$contrib"        "contributions of the individuals"
## 11 "$quali.sup"          "results for the supplementary
categorical variables"
## 12 "$quali.sup$coord"    "coord. for the supplementary
categories"
## 13 "$quali.sup$v.test"   "v-test of the supplementary
categories"
## 14 "$call"               "summary statistics"
## 15 "$call$centre"        "mean of the variables"
## 16 "$call$ecart.type"    "standard error of the variables"
## 17 "$call$row.w"         "weights for the individuals"
## 18 "$call$col.w"         "weights for the variables"
```

- 2.4 The eigenvalues and cumulative variance values reported in Rivas et al. (2022) can be found in the `res.pca$eig` object

Command**new command name**

res.pca\$eig

Expected result

```
##          eigenvalue percentage of variance cumulative
percentage of variance
## comp 1 3.96582357          49.5727946
49.57279
## comp 2 2.39521837          29.9402297
79.51302
## comp 3 0.99221034          12.4026292
91.91565
## comp 4 0.42550528           5.3188160
97.23447
## comp 5 0.12786974           1.5983717
98.83284
## comp 6 0.04754866           0.5943582
99.42720
## comp 7 0.03306028           0.4132535
99.84045
## comp 8 0.01276377           0.1595471
100.00000
```

According to the Kaiser criterion, select the principal components with eigenvalues greater than 1.

2.5

Save the coordinate values of each principal component to generate clean and customized graphics

**Command****new command name**

```
dind<-as.data.frame(res.pca$ind$coord)
names(dind)
```

Expected result

```
## [1] "Dim.1" "Dim.2" "Dim.3"
```

Append these columns to the mr_sample_2 object. For example:

Command**new command name**

```
tab2<- cbind(mr_sample_2$ID2,dind)
head(tab2)
```



Expected result

```
##      mr_sample_2$ID2      Dim.1      Dim.2      Dim.3
## 1      pH.70  1.155339564  0.06880405  0.8731306
## 2      pH.70 -2.785419405  2.61534418 -1.2313661
## 3      pH.70  2.671369221 -1.14843568  0.2633520
## 4      pH.70 -0.006226783 -0.46870882  1.4537604
## 5      pH.70 -0.884580976  1.51072808 -1.5240606
## 6      pH.70  0.112930932  1.56553832 -0.4460328
```

- 2.6 Convert the `tab2` object into a dataframe using `as.data.frame()`. After conversion, verify the column names, and preview the first few rows of the dataset. Rename the first column to `pH` and confirm the updated column names again.

Command

new command name

```
tab2<-as.data.frame(tab2)
names(tab2)
```

Expected result

```
## [1] "mr_sample_2$ID2" "Dim.1"          "Dim.2"
      "Dim.3"
```

**Command**

new command name

```
head(tab2)
```

Expected result

```
## mr_sample_2$ID2      Dim.1      Dim.2      Dim.3
## 1      pH.70  1.155339564  0.06880405  0.8731306
## 2      pH.70 -2.785419405  2.61534418 -1.2313661
## 3      pH.70  2.671369221 -1.14843568  0.2633520
## 4      pH.70 -0.006226783 -0.46870882  1.4537604
## 5      pH.70 -0.884580976  1.51072808 -1.5240606
## 6      pH.70  0.112930932  1.56553832 -0.4460328
```

Command

new command name

```
colnames(tab2)[1]<-"pH"
names(tab2)
```

Expected result

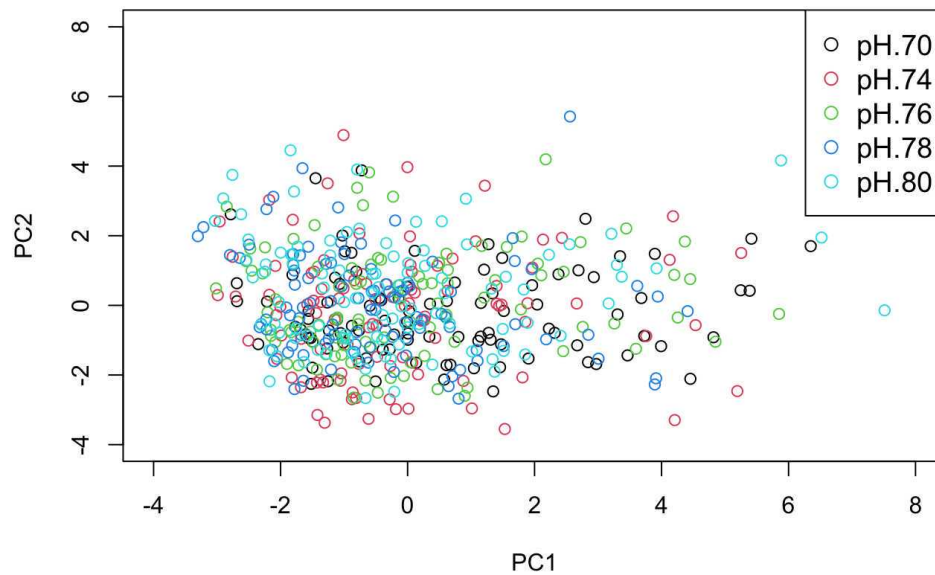
```
## [1] "pH"      "Dim.1" "Dim.2" "Dim.3"
```

- 2.7 Construct the graph using the prepared `tab2` object, ensuring all relevant variables are correctly represented for visualization

Command

new command name

```
plot(tab2$Dim.1, tab2$Dim.2, col=tab2$pH, xlab = "PC1", ylab = "PC2",  
xlim = c(-4,8), ylim = c(-4,8))  
legend("topright", legend = levels(tab2$pH), col =  
1:length(levels(tab2$pH)), pch = 1, cex = 1.2)
```



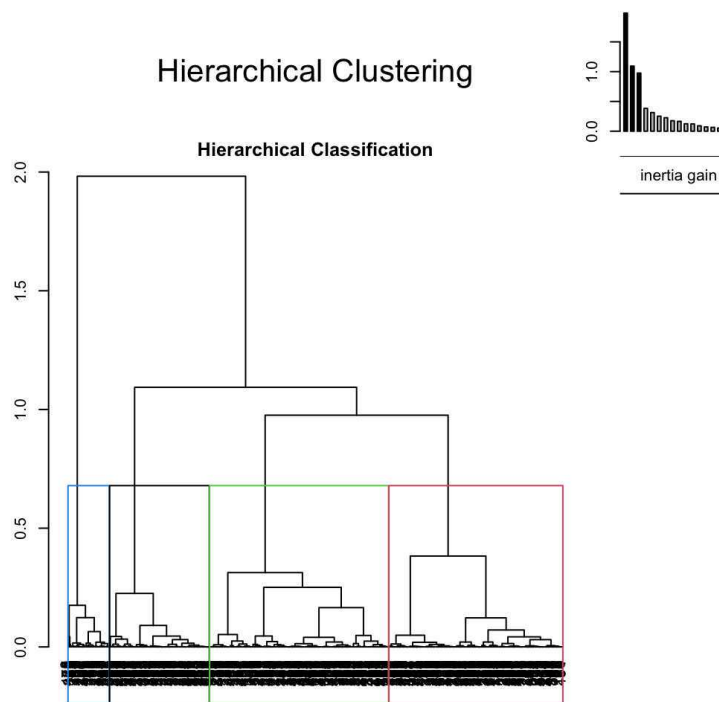
Stage 3: Hierarchical clustering on principal components

- 3 After performing the principal component analysis, use the `res.pca` object as input for the HCPC function.
- 3.1 Create a new object named `res.hcpc` to store the results of the HCPC function

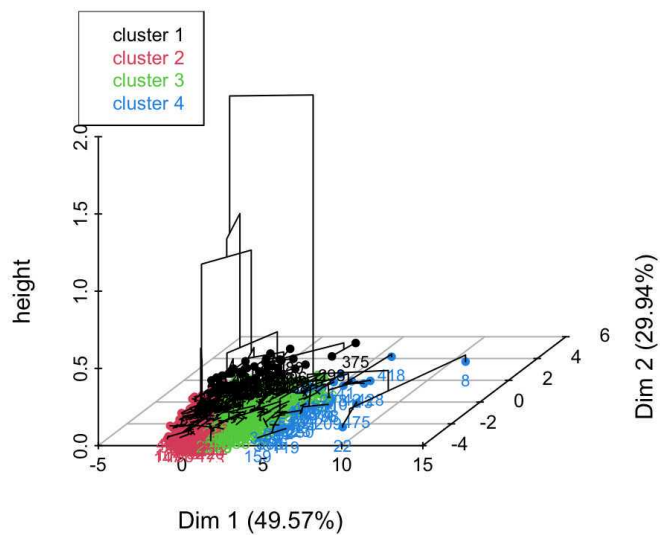
Command

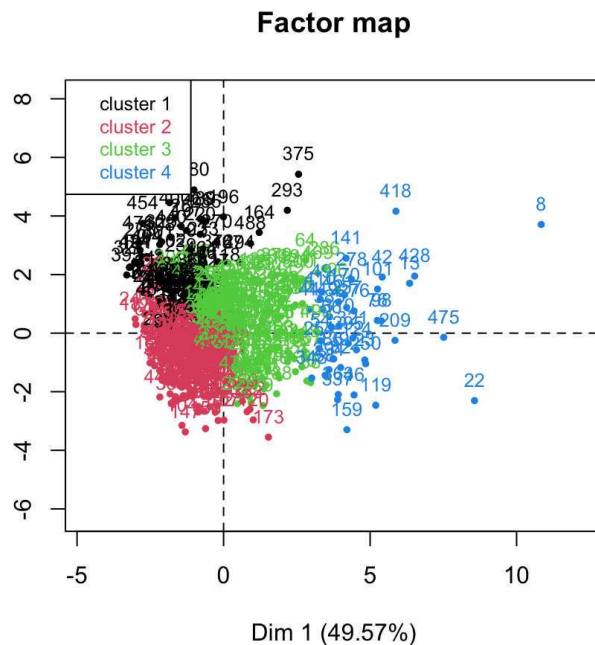
new command name

```
res.hcpc <- HCPC(res.pca, nb.clust=-1, conso=0, min=3, max=10,  
metric="euclidean", method="ward")
```



Hierarchical clustering on the factor map





The HCPC function generates three plots:

1. Hierarchical Clustering Plot:

This plot features a dendrogram with a horizontal line suggesting the cutoff point for group formation. The plot also includes an inset with a bar chart indicating the inertia gain. By clicking on the bar or another level of the graph, two additional plots are generated. For this document, the `nb.clust` argument was set to -1 for automatic cutoff determination, but setting it to 0 allows the user to select the cutoff height manually.

2. Hierarchical Clustering on the Factor Map:

This plot provides a three-dimensional representation, with the first and second principal components on the x and y axes, respectively, and the hierarchical dendrogram projected in two dimensions along the z axis.

3. Factor Map Plot:

This graph displays the values of the first and second principal components, with points colored according to the formed groups or clusters. These clusters represent the kinematic subpopulations.

- 3.2 Type `res.hcpc` in the console to display the list of contained objects, followed by their descriptions. Access the contents of the `res.hcpc` object using the `$` operator.

Note

The `res.hcpc` object is of the list type, and its contents can be accessed using the `$` operator.



Command

new command name

res.hcpc

Expected result

```
## **Results for the Hierarchical Clustering on Principal
Components**
##      name
## 1  "$data.clust"
## 2  "$desc.var"
## 3  "$desc.var$quanti.var"
## 4  "$desc.var$quanti"
## 5  "$desc.var$test.chi2"
## 6  "$desc.axes$category"
## 7  "$desc.axes"
## 8  "$desc.axes$quanti.var"
## 9  "$desc.axes$quanti"
## 10 "$desc.ind"
## 11 "$desc.ind$para"
## 12 "$desc.ind$dist"
## 13 "$call"
## 14 "$call$t"
##      description
## 1  "dataset with the cluster of the individuals"
## 2  "description of the clusters by the variables"
## 3  "description of the cluster var. by the continuous var."
## 4  "description of the clusters by the continuous var."
## 5  "description of the cluster var. by the categorical var."
## 6  "description of the clusters by the categories."
## 7  "description of the clusters by the dimensions"
## 8  "description of the cluster var. by the axes"
## 9  "description of the clusters by the axes"
## 10 "description of the clusters by the individuals"
## 11 "parangons of each clusters"
## 12 "specific individuals"
## 13 "summary statistics"
## 14 "description of the tree"
```

The `res.hcpc$data.clust` object contains the data from `mr_sample_2`, and includes a column named `clust` at the end, which indicates the cluster to which each data line belongs.

**Command****new command name**`head(res.hcpc$data.clust)`**Expected result**

```
##      ID2      VCL      VAP      VSL      LIN      STR
WOB BeatCross
## 1 pH.70 265.64667 108.57502 68.93514 0.259499 0.634908
0.408720 30.79646
## 2 pH.70 47.19655 36.58363 27.96065 0.592430 0.764294
0.775134 38.72727
## 3 pH.70 356.77786 143.29716 56.96533 0.159666 0.397533
0.401643 28.17391
## 4 pH.70 239.27805 86.00893 55.40223 0.231539 0.644145
0.359452 38.08696
## 5 pH.70 94.31265 65.55947 41.34975 0.438433 0.630721
0.695129 28.96552
## 6 pH.70 165.31839 97.35522 67.48124 0.408190 0.693145
0.588895 28.44828
##      ALH clust
## 1 8.619123      3
## 2 1.959340      1
## 3 12.417836      3
## 4 6.697638      3
## 5 4.690355      3
## 6 5.007826      3
```

This object is crucial as it allows for the analysis of the dataset based on treatments and clusters.

Store the contents of `res.hcpc$data.clust` in a new object for subsequent analysis.



Command

new command name

```
mr_sub<-res.hcpc$data.clust
```

The `res.hcpc$desc.axes` object contains the values for each cluster based on quantitative variables (motility parameters).

Stage 4: Subpopulations in the dataset

- 4 One of the objectives of identifying kinematic sperm subpopulations is to describe the effect of experimental treatments on the structure of these subpopulations, starting by analyzing the proportion of sperm from each treatment within each subpopulation. Similarly, it is important to describe the effect of each treatment on each cluster. This aspect is addressed in another protocol (<https://dx.doi.org/10.17504/protocols.io.14egn6p4ql5d/v1>).

Protocol



NAME

Optimizing CASA Data: Transforming Sperm Coordinates into Long Format for Enhanced Machine Learning Analysis

CREATED BY

Cindy Rivas

Preview

Conclusions

- 5 With the proposed workflow in this document, you can process a file of motility parameters obtained from a CASA system to identify kinematic subpopulations in two steps. The outcome is an object that contains the original input data along with an additional column indicating the cluster to which each data line belongs. Analyzing



the resulting object allows for identifying the proportion of sperm in each subpopulation and assessing the effect of each treatment on each subpopulation.

Protocol references

Martínez-Pastor F, Tizado EJ, Garde JJ, Anel L and Paz P de (2011) Statistical Series: Opportunities and challenges of sperm motility subpopulation analysis. *Theriogenology* **75** 783–795.

Ramón M and Martínez-Pastor F (2018) Implementation of novel statistical procedures and other advanced approaches to improve analysis of CASA data. *Reproduction, Fertility, and Development* **30** 860–866.

Rivas AC, Ayala EME and Aragon MA (2022) Effect of various pH levels on the sperm kinematic parameters of boars. *South African Journal of Animal Science* **52** 693–704.

Acknowledgements

This work was supported by UNAM-PAPIIT IN224925.

This work was supported by the Programa de Becas Posdoctorales de la Dirección General de Asuntos del Personal Académico (DGAPA), UNAM to the postdoctoral fellow Cindy Ursula Rivas Arzaluz.