

PREDICCIÓN DE HAPLOGRUPOS MITOCONDRIALES DE *Bradybaena similaris* (GASTROPODA, CAMAENIDAE) EN ARGENTINA UTILIZANDO ENFOQUES DE MACHINE-LEARNING

E.N. Serniotti^{1,*}, A.I. Rau¹, A.M. Ferreyra², L. Iurinic¹, S.E. Scherf¹, L.B. Guzmán¹, R.E. Vogler¹ & A.A. Beltramino¹

¹ Laboratorio del Grupo de Investigación en Genética de Moluscos (GIGeMol), Instituto de Biología Subtropical (IBS), CONICET – Universidad Nacional de Misiones, Posadas, Misiones, Argentina. ² Laboratorio de Genética Molecular y Genómica, Instituto de Genética Humana, Hospital Escuela de Agudos Dr. R. Madariaga, Posadas, Misiones, Argentina.



Introducción

La introducción de especies invasoras en áreas no nativas se ha visto exacerbada en los últimos años debido al avance de la globalización, generando impactos a gran escala (Darrigran et al. 2020). Desde la pérdida de biodiversidad hasta el daño a cultivos, a la salud humana y animal, los moluscos constituyen excelentes ejemplos de invasiones biológicas de alto impacto que pocas veces son abordados desde un enfoque preventivo (Cowie & Robinson 2003). La caracterización genética de especies ya establecidas puede arrojar luces sobre las rutas de introducción existentes y colaborar en la generación de estrategias para evitar futuras introducciones (Crăstescu 2015). Sin embargo, los costos de secuenciación de ADN asociados con estudios poblacionales suele ser la mayor limitante en países de bajos recursos. Convenientemente, el desarrollo de algoritmos matemáticos para el aprendizaje automático ha hecho posible el aprovechamiento de caracteres morfológicos para la predicción de variabilidad genética (Barrow et al. 2021). En este trabajo aplicamos métodos de machine learning no supervisado para la predicción de haplogrupos mitocondriales de poblaciones argentinas de *Bradybaena similaris* (Férussac, 1822).

Discusión

Los análisis llevados a cabo utilizando medidas conquiliológicas no permitieron la discriminación de clusters, en tanto que los registros con datos genéticos no pudieron ser inequívocamente diferenciados del resto. Estos resultados podrían sugerir que: las medidas morfológicas utilizadas no tendrían relación con la divergencia genética observada; la muestra empleada no es representativa de la variabilidad conquiliológica de las poblaciones; o bien, la cantidad de registros analizados no fue suficiente para los modelos de machine learning.

Referencias

- Barrow LN, Masiero da Fonseca E, Thompson CEP, Carstens BC. 2021. Predicting amphibian intraspecific diversity with machine learning: Challenges and prospects for integrating traits, geography, and genetic data. *Molecular Ecology Resources* 21: 2818–2831.
Cowie RH, Robinson DG. 2003. Pathways of introduction of nonindigenous land and freshwater snails and slugs. En: Ruiz G, Carlton JT (eds). *Invasive species: vectors and management strategies*. Island Press, Washington DC, USA. pp. 93–122.
Crăstescu ME. 2015. Genetic reconstructions of invasion history. *Molecular Ecology* 24: 2212–2225.
Darrigran G, Aguado-Padrón I, Baez P et al. 2020. Non-native mollusks throughout South America: emergent patterns in an understudied continent. *Biological Invasions* 22: 853–871.

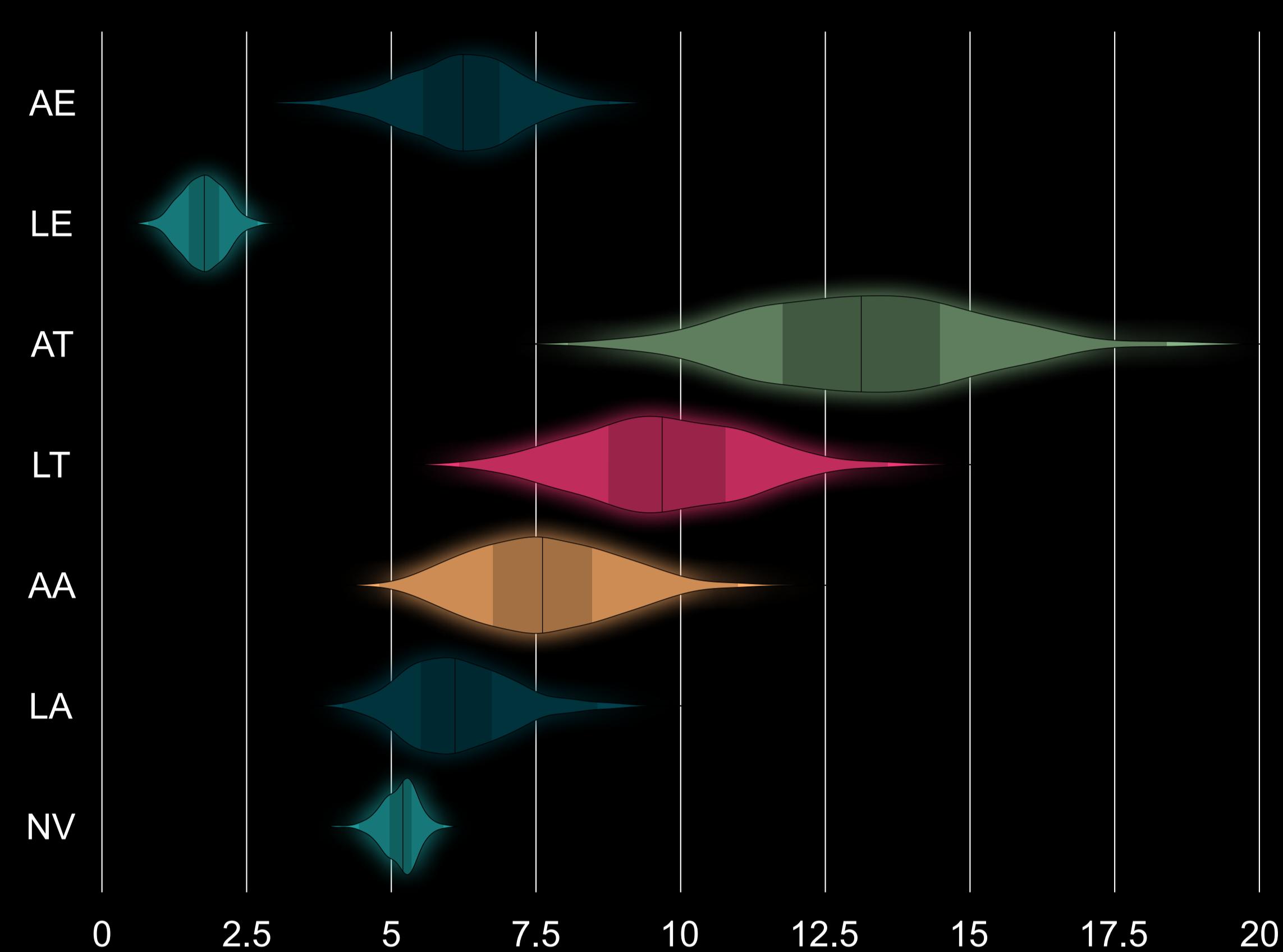


Figura 1. Violin plot de las medidas relevadas para los 503 individuos empleados en este estudio. Eje Y: abreviaturas referidas en la Tabla I. Eje X: medidas en mm.

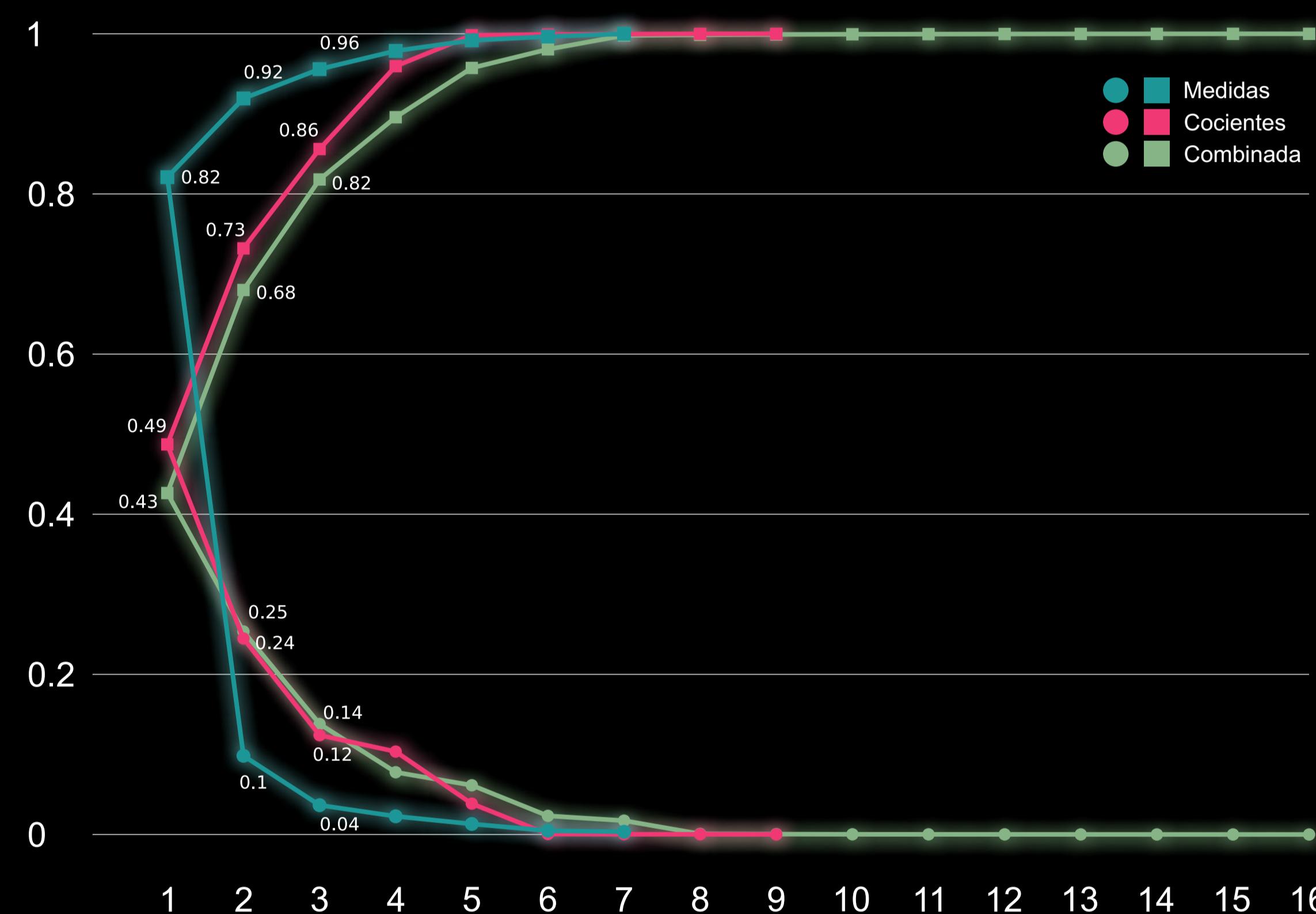


Figura 2. Varianza total explicada por cada componente de cada PCA realizado. Eje Y: varianza total explicada. Eje X: componentes principales. (●) varianza individual, (■) varianza acumulada.

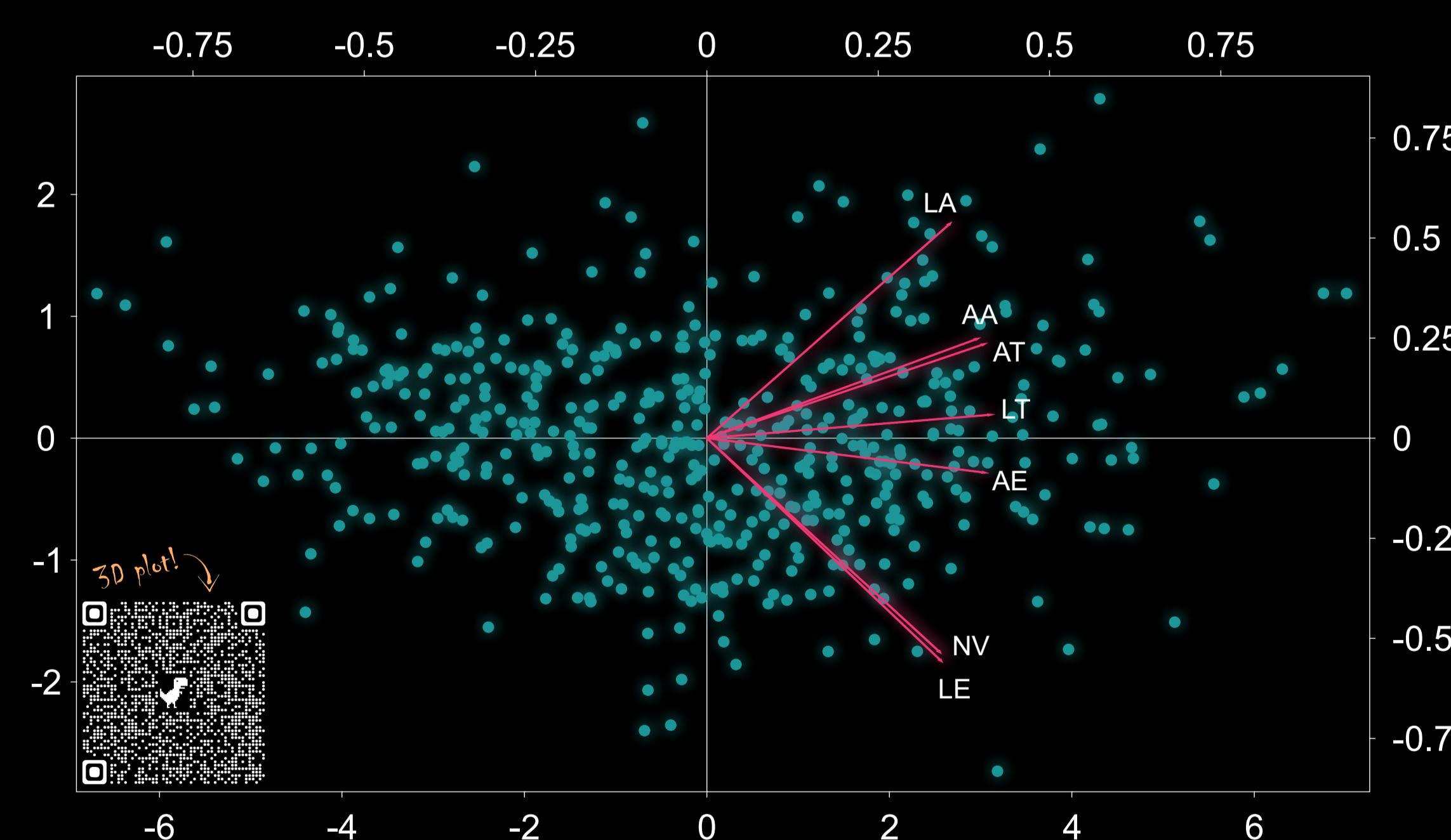


Figura 3. PCA de las medidas relevadas. Ejes inferior e izquierdo: desvíos estándar. Ejes superior y derecho: pesos de las variables (vectores en fucsia).

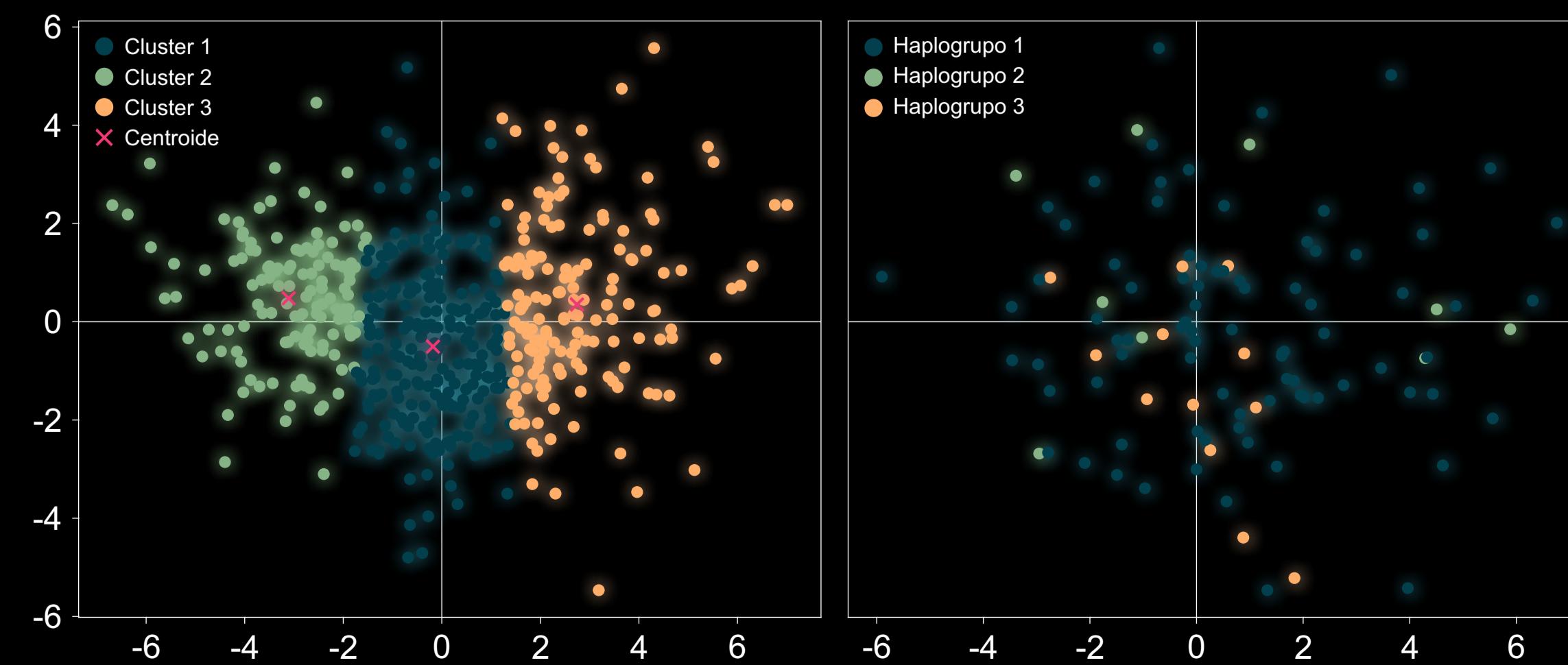


Figura 4. Clusterización de k-means sobre el PCA de las medidas relevadas. Izquierda: coloración de k-means; derecha: coloración de haplogrupos

Materiales y métodos

Se tomaron ocho medidas conquiliológicas (alto y ancho de la concha, la espira y la abertura, número de vueltas y patrón de coloración) a partir de 503 individuos de *B. similaris* presentes en la Colección Malacológica del Instituto de Biología Subtropical (IBS-Ma), de los cuales 106 presentaban información genética mitocondrial (*cox1* y *rrnL*). Las mediciones hechas sobre fotografías de los individuos mediante el programa ImageJ. Las longitudes medidas fueron relacionadas mediante cocientes creando nueve variables adicionales y los haplotipos fueron divididos en tres haplogrupos para facilitar la visualización de los análisis. Los datos fueron analizados utilizando diversas librerías disponibles en Python 3.8 por medio del entorno JupyterLab aplicado en Visual Studio Code. Las dimensiones fueron reducidas empleando PCA y t-SNE, mientras que los clusters fueron aproximados mediante k-means.

Resultados

Las medidas relevadas para los 503 individuos presentaron una distribución aproximadamente normal (Fig. 1, Tabla I). El conjunto de variables que mayor cantidad de varianza acumuló en los componentes principales (PCs) 1 y 2 fue el de las medidas originales excluyendo los morfotípos, con 82 y 10% respectivamente (Fig. 2). En este PCA, el largo total (LT) fue la variable con mayor peso para el PC1, mientras que el largo de la abertura (LA) lo fue para el PC2 (Fig. 3). Por último, no fue posible diferenciar los distintos haplogrupos utilizando el método de clusterización de k-means sobre el dataset con dimensiones reducidas con PCA o t-SNE (Fig. 4).

Tabla I. Estadísticas de resumen de las medidas relevadas para los 503 individuos empleados en este estudio.

	AE	LE	AT	LT	AA	LA	NV
Media	6.21	1.76	13.17	9.73	7.65	6.19	5.16
Desvío	0.97	0.37	1.93	1.44	1.22	0.91	0.30
Mínimo	3.20	0.71	8.05	6.17	4.79	4.15	4.02
Q1	5.55	1.50	11.77	8.75	6.76	5.51	4.97
Mediana	6.24	1.77	13.12	9.68	7.60	6.10	5.20
Q2	6.87	2.03	14.48	10.77	8.47	6.74	5.35
Máximo	9.04	3.13	19.09	14.45	11.95	9.72	6.00

AE, ancho de la espira; LE, largo de la espira; AT, ancho total; LT, largo total; AA, ancho de la abertura; LA, largo de la abertura; NV, número de vueltas

Agradecemos a la Facultad de Ciencias Exactas, Químicas y Naturales; Universidad Nacional de Misiones (Proyectos 16/Q1227-PI y 16/Q1559-PI) y a la Agencia I+D+i (PICT-2019-03294) por financiar este estudio.