Exploring Machine Learning for Species Delimitation in the *Trimerotropis pallidipennis* Complex (Othoptera: Acrididae)

Guzmán Noelia Verónica [1]

[1] Department of Ecology, Genetics and Evolution, Institute of Ecology, Genetics y Evolution Buenos Aires (IEGEBA), FCEyN, University of Buenos Aires, Argentine.

.UBA Universidad de Buenos Aires · **CONICET** · IEGEBA

INTRODUCTION & AIM

* The genus **Trimerotropis Stål (Orthoptera: Acrididae)** is distributed from Canada to Argentina, with 49 valid species mainly inhabiting dry and temperate regions (1), likely **originated in North America** and colonized South America during the Pleistocene, after the closure of the Isthmus of Panama (2-3).

* **North America has the largest number of species**, including *Trimerotropis pallidipennis* (Burmeister, 1838), which is the only one with an **amphitropical distribution** (3-4).

* In **South America,** the number of valid species was reduced from nine to four: ***T. pallidipennis, T. andeana, T. atacamensis,*** and ***T. ochraceipennis*** (with two subspecies) (1)

*Previuos molecular studies have revealed a lack of clear genetic boundaries among these taxa, with some groups comprising **multiple genetic lineages** (3,5). As a result, Husemann et al. (2013) proposed that the South American representatives of the genus be treated as a species complex, referred to as the *Trimerotropis pallidipennis* **species complex.** (Figure 1)

*One of these genetic lineages, hereafter referred to as ***Trimerotropis sp.*** (3), has a broad geographic and environmental distribution across Argentina. It is distinguished from the others by the presence of **polymorphic chromosomal inversions**, whose frequencies exhibit an altitudinal cline (2,4,6,7,8). Furthermore, Guzmán et al. (2020) identified two distinct genetic lineages within a single clinal transect and found evidence of **hybrid zones**.

The band-winged grasshopper *Trimerotropis pallidipennis* species complex (Orthoptera: Acrididae) has attracted attention for over a century due to its broad distribution, low morphological variation and complex evolutionary history.

AIM: Explore the evolutionary history of *T. pallidipennis* **through unsupervised machine learning (UML) and Convolutional Neural Network (CNN), and compared those with traditional model-based validation methods using previous genome-wide SNP datasets generated through ddRAD-seq**
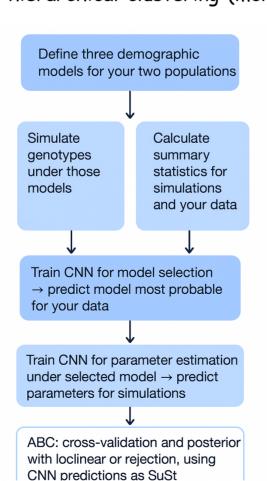
METHOD

New analyses were conducted on previously generated ddRAD-seq datasets from a total of 192 individuals from 24 localities. The samples correspond to individuals of *Trimerotropis ochraceipennis* from Chile, *T. andeana* from Bolivia and Peru, and *Trimerotropis* sp. from Argentina, collected across 34 localities (Fig. 1).
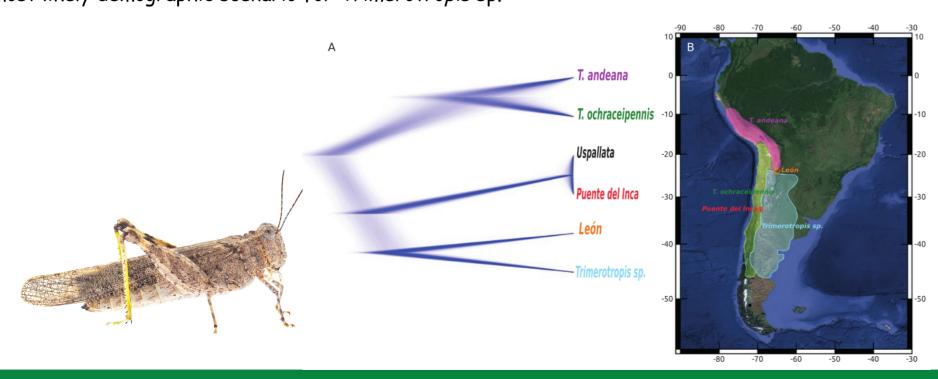
We applied two unsupervised machine learning (UML) approaches to the SNP datasets:

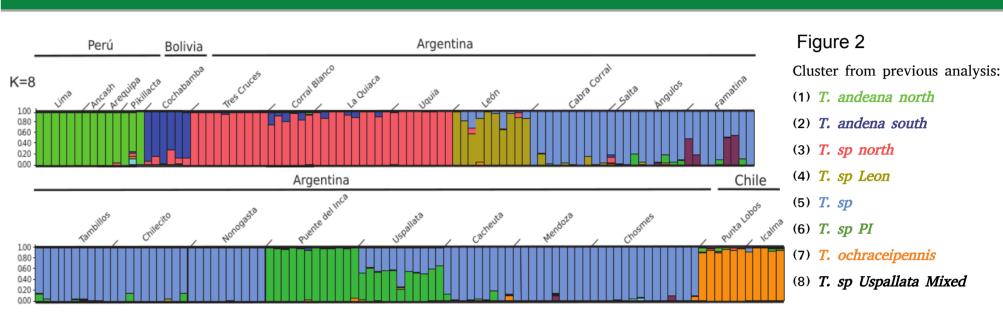(1) **Random Forest (RF)** on scaled DAPC data → proximity matrix for **cMDS** and **isoMDS** (10,11).

(2) **t-SNE** on PCA-reduced SNPs (12). Three clustering approaches were applied to UML outputs to verify consistency with DAPC and STRUCTURE results: (1) PAM clustering using cluster (K from DAPC) (13); (2) PAM with K determined by gap statistic(14); and (3) hierarchical clustering (mclust, (15)) using PCDimension (16).
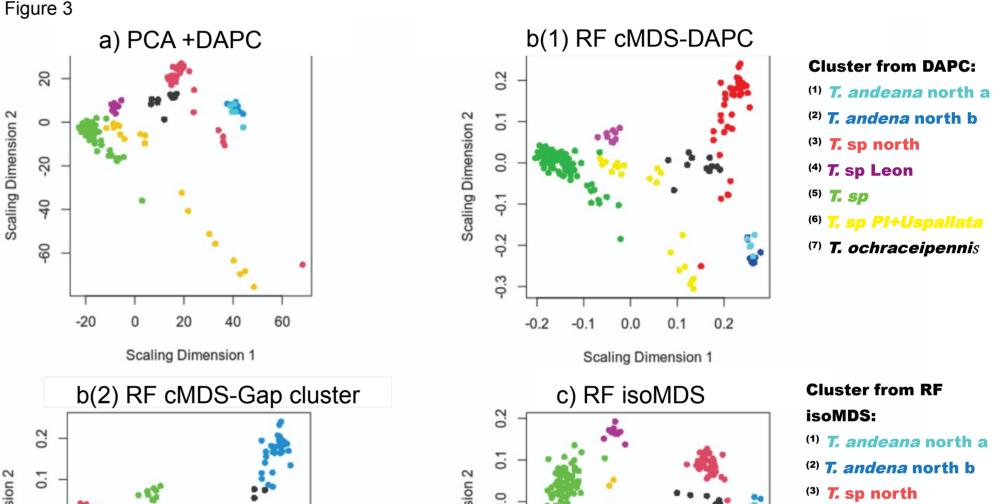
To explore the demographic history of *Trimerotropis* sp. lineages, we tested three evolutionary scenarios (Parallel expansion, one lineage expansion only, and No expansion) using **convolutional neural networks (CNNs)**. We performed 10,000 coalescent simulations per model with species-specific parameters and incorporated the same proportion of missing data observed in empirical SNP datasets. The CNN architecture was adapted from Kirschner et al. (2022) and calibrated using temperature scaling to improve model prediction. The trained network was then applied to empirical SNP data to identify the most likely demographic scenario for *Trimerotropis* sp.
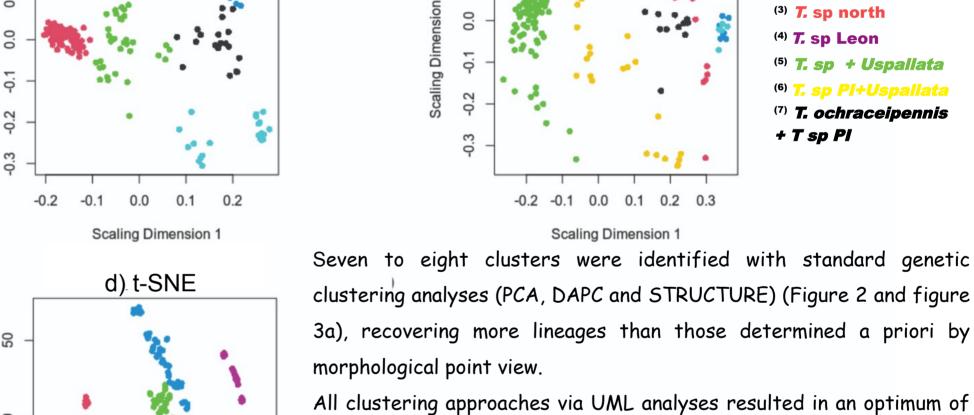
Define three demographic models for your two populations
↓
Simulate genotypes under those models → Calculate summary statistics for simulations and your data
↓
Train CNN for model selection → predict model most probable for your data
↓
Train CNN for parameter estimation under selected model → predict parameters for simulations
↓
ABC: cross-validation and posterior with loclinear or rejection, using CNN predictions as SuSt

- T. andeana
- T. ochraceipennis
- Uspallata
- Puente del Inca
- León
- Trimerotropis sp.

RESULTS & DISCUSSION

Figure 2

Cluster from previous analysis:
(1) *T. andeana north*
(2) *T. andena south*
(3) *T. sp north*
(4) *T. sp Leon*
(5) *T. sp*
(6) *T. sp PI*
(7) *T. ochraceipennis*
(8) *T. sp Uspallata Mixed*

Figure 3

a) PCA +DAPC

b(1) RF cMDS-DAPC

Cluster from DAPC:
(1) *T. andeana north a*
(2) *T. andena north b*
(3) *T. sp north*
(4) *T. sp Leon*
(5) *T. sp*
(6) *T. sp PI+Uspallata*
(7) *T. ochraceipennis*

b(2) RF cMDS-Gap cluster

c) RF isoMDS

Cluster from RF isoMDS:
(1) *T. andeana north a*
(2) *T. andena north b*
(3) *T. sp north*
(4) *T. sp Leon*
(5) *T. sp + Uspallata*
(6) *T. sp PI+Uspallata*
(7) *T. ochraceipennis + T sp PI*

d) t-SNE

Seven to eight clusters were identified with standard genetic clustering analyses (PCA, DAPC and STRUCTURE) (Figure 2 and figure 3a), recovering more lineages than those determined a priori by morphological point view.

All clustering approaches via UML analyses resulted in an optimum of K = 7, except for gap statics clustering in Random forest of cMDS (K = 5, where T. sp from Uspallata join with T. sp and Leon, and T. sp from PI con *T. ochraceipennis*).

All K = 7 UML clustering assignments were similar, differing only in whether the hybrid population was grouped with T. sp PI or T. sp, and in distinguishing (or not) the *T. andeana* populations.

The scenario depicting expansion in both lineage was selected as the most explanatory demographic model with a posterior probability (PP) higher than 0.99.

Figure 2: Clustering results for the *Trimerotropis pallidipennis* complex based on SNP dataset from STRUCTURE plot previously obtained (Guzman et al. 2024). **Figure 3**: Clustering results for the *Trimerotropis pallidipennis* complex based on SNP dataset (a) PCA plot with DAPC clusters (K = 7) (b1) Random forest cMDS plot, with clusters identified via PAM and mclust (K= 7), and (b2) gap statistic (K=5). (c) Random forest isoMDS plot, all clustering algorithms favored K= 7 (d) t- SNE plot, all clustering algorithms favored K = 7 except for gap algorithms that favoured K 1 and a suboptimal in K 7.

CONCLUSION

Unsupervised machine learning methods improved cluster visualization and interpretation like standard multivariate analyses, revealing hidden genetic structure in morphologically cryptic taxa. These approaches represent powerful, time-efficient complements to conventional species delimitation tools.

Recent advances combining deep learning and ABC have improved inference of demographic history and hybridization from genomic data. Here we initiated CNN model training to explore demographic scenarios and potential hybrid zones in the *T. pallidipennis* species complex, extending applications of neural networks in systematics and phylogenomics.

This work emphasizes the complexity of species boundaries in the *Trimerotropis pallidipennis* and the value of integrative frameworks combining morphology, cytogenetics, and genomics.

These appoaches and future analyses may eventually help to incorporate CNNs trained with genomic, cytogenetic, and morphological data to achieve a deeper understanding of this fasinating species complex.

REFERENCES

1 - Cigliano, MM, Braun H, Eades DC et al. Orthoptera Species File. Version 2025 http://Orthoptera.SpeciesFile.org 2 - Confalonieri et al. 1998: Confalonieri VA, Sequeira AS, Todaro L et al. Mitochondrial DNA and phylogeography of the grasshopper *Trimerotropis pallidipennis* in relation to clinal distribution of chromosome polymorphisms. Heredity 1998;81:444–52. https://doi.org/10.1038/sj.hdy.6884050 3 - Husemann M, Guzman NV, Danley PD et al. Biogeography of *Trimerotropis pallidipennis* (Acrididae: Oedipodinae): deep divergence across the Americas. Journal of Biogeography 2013;40:261–73. https://doi.org/10.1111/jbi.12007 4- Guzman NV, Confalonieri VA. The evolution of South American populations of *Trimerotropis pallidipennis* (Oedipodinae: Acrididae) revisited: dispersion routes and origin of chromosomal inversion clines. Journal of Orthoptera Research 2010;19:253–60. https://doi.org/10.1665/034.019211 5 -Guzmán, N. V., Pietrokovsky, S. M., Cigliano, M. M., Confalonieri, V. A. (2017). Unraveling the diversification history of grasshoppers belonging to the *Trimerotropis pallidipennis* (Oedipodinae: Acrididae) species group: a hotspot of biodiversity in the Central Andes. PeerJ, 5, e3835. https://doi.org/10.7717/peerj.3835 6 - Confalonieri, V. A., & Colombo, P. C. (1989). Inversion polymorphisms in *Trimerotropis pallidipennis* (Orthoptera): clinal variation along an altitudinal gradient. Heredity, 62(1), 107–112. https://doi.org/10.1038/hdy.1989.14 7 -Confalonieri, V. A. (1994). Inversion Polymorphisms and Natural Selection in *Trimerotropis Pallidipennis* (Orthoptera). Correlationwith environmental variables: an overall view. Heredias, 125(2–3), 289–296. https://doi.org/10.1111/j.1601-5223.1994.00289.x 8 -Guzmán NV. Kemppainen P, Monti G et al. Stable inversion clines in a grasshopper species group despite complex geographical history.Molecular Ecology 2022;31:1196–215.10 -Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.11-Venables, W.N., Ripley, B.D., 2002. Statistics and Computing. Modern Applied Statistics with S. Springer.12- Donaldson, J., 2016. tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE). R package version 0.1-3 13-Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., 2018. cluster: cluster analysis basics and extensions. R package version 2.0.7-1. 14- Kassambara, A., Mundt, F., 2017. factoextra: extract and visualize the results of multivariate data analyses. R package version 1.0.5. 15-Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2017. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal 8 (1), 205–233. 16- Coombes K.R., Wang H. 2018. PCDimension: finding the number of significant principal components. R package version 1.1.9. 17-Kirschner, P., Perez, M.F., Zȁveská, E. et al. Congruent evolutionary responses of European steppe biota to late Quaternary climate change. Nat Commun 13, 1921 (2022). https://doi.org/10.1038/s41467-022-29267-8