Digging deeper: using Afrotropical dung beetles to better understand quality and coverage of biodiversity data

Bryony Blades
St. Hugh's College
Department of Biology
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Trinity 2025

Declaration

I declare that this thesis is entirely my own work, except where explicitly stated in the text. The work has not been submitted for any degree or professional qualification except as specified.

Funding

This work was supported by funding from the African Natural History Research Trust (ANHRT) Scholarship, through the University of Oxford Department of Biology. Without the backing of ANHRT, and especially that of Richard Smith, this research would not have been possible. I will be eternally grateful, thank you.

Acknowledgements

Firstly, I would like to thank my supervisors Tim Coulson, Mike Bonsall, and Lizzy Jeffers for their guidance, patience, and time spent helping me become a better scientist. I have always felt extremely lucky to have landed so firmly on my feet in having this team around me, and my respect and admiration of you all is boundless. I'm also grateful for your understanding when life outside work has been complicated, and especially to Lizzy for your guidance when I was learning more about how my brain works. Thank you, too, to Sonya Clegg for inviting me into your lab group to learn about genomics, and for your support completing these parts of my project.

I would also like to thank Richard Smith, as this DPhil, its fieldwork, and all of the other experiences I've had during this time, would not have been possible without the ANHRT Scholarship. More than that, though, I want to thank you for believing in me from the very beginning, for the care and concern you have consistently shown me, and for your excellently good company. Working with you has been such a pleasure.

None of my thesis could have been completed without the hard work of Hitoshi Takano, whose taxonomic revision of *Catharsius* was the basis of this project. Thank you for your belief in me, for continuing to identify dung beetles long after you would have liked to stop, and also for the stimulating discussions on collection digitisation, biodiversity modelling, riverine barriers, cryptic crosswords, and test cricket. Thank you, also, to everyone else at ANHRT for making my time there so enjoyable and welcoming me back whenever I have visited.

Thank you to Will for accompanying me on fieldwork, and sharing in the more challenging tasks, for helping to adjust plans when things weren't working out, and for sharing your love of mantids. The trip would have been even more difficult without William, Dom, Abel, and Cecil to whom I owe an eternal debt of gratitude for welcoming me to Zambia, helping with collecting, and keeping both of us happy, safe, and well-fed with delicious food.

One of the best parts of my DPhil was undoubtedly my visit to Joaquín Hortal's group at MNCN in Madrid. Thank you for hosting me, it was an immense privilege to learn from you, and to collaborate with you on my second chapter. I will not forget the kindness and compassion of your lab, and it is my sincere hope that we can work together again in the future. I am also thankful to everyone else in the group for embracing me as one of your own, especially Cristina. Critically, this stay wouldn't have been possible without the generosity of the Pedrosa family who invited me to stay in their home.

There are many friends who have supported me in different ways during this journey. Thank you to those that I knew before who have not forgotten me when I dropped off the face of the planet for a few years, and who have continued to invite me to their plans nonetheless. Thank you to new friends in my online community for your empathy, understanding, and kindness that comes from our shared experiences. Finally, thank you to the friends I have made at Oxford. Andrew, for being the first to welcome me into the E2D lab group, Tabby for sitting next to me these last couple of years, and sharing in many of the highs and lows (lab work …), letting me constantly interrupt you, and the hugs, and Liv for matching my oversharing (and also more hugs!). Also, everyone else in the E2D, Clegg, and Mathematical Ecology Labs that have made my time so much more enjoyable, especially Maisie, who continues to be one of the kindest people I have ever been lucky enough to meet, and Andrea, whose help with bioinformatics was instrumental in producing the latter half of this thesis. Thanks, also, to friends and colleagues in the wider Department of Biology for making it such a great place to work.

Most importantly, thank you to my family. Katie, I'm grateful to Oxford for two things: my DPhil, and you. I'm not sure I will ever be able to properly convey the immense impact that you have had on my life. It is largely down to meeting you that I started to feel more like myself again after a little while feeling lost. Without the utterly unconditional support, wholeheartedly judgement-free friendship, nights discussing everything from climate change and biodiversity to meaningless gossip on my sofa, hours doing the same on facetime when one or both of us were away, inappropriately loud laughs, and endless complaints about that one chapter we've each had that has hung over us, I know I wouldn't have been able to get through the last few years. Your room will always be made up in our home, and when you finally stop making me live with a long distance best friend, we can get to working on that paper together.

To my older siblings, your partners, my nieces and nephews, my grandparents, and my parents, thank you for your unwavering support and love throughout my DPhil, and before. Thank you for helping us to plan a wedding mid-doctorate, thank you for the time we spend together that reminds me of the importance of stepping away from work and of how lucky I am to have a family like ours, and thank you for having always inspired me with your own successes. You make my life immeasurably better. Mum and Dad, thank you for everything you have always done to provide me with the opportunities that have helped me get to this point, and for always looking after me, especially in this last month. All of my achievements are shared with you. To Archie, thank you for the cuddles.

Martyn. This is the most difficult section to write, as I would need the length of this thesis again to do justice to all the ways in which this huge undertaking has been made easier because of you. Thank you for agreeing to put our life plans on hold, for moving hours away from your work, for the things you have given up for me, for being there for a hug or a rant or as a shoulder to cry on when things were hard, for learning to love birds and beetles, and for always making me laugh. As I sit here, I can't think of what else I can say to express the depth of my gratitude. So, instead, this is for you...

This thesis is dedicated to my husband, Martyn, for our shared love of things that fly.

Publications and contributions

The following chapters consist of manuscripts that have been accepted for publication, or are in preparation. In chapters that are a collaborative effort, the majority of the work is my own.

Chapter 2: The impact of taxonomic revision on species distribution modelling
Bryony Blades (In Prep)

> BB conceived the project, prepared data for analysis, conducted analysis, and wrote the manuscript.

Chapter 3: Blades, B., Ronquillo, C., & Hortal, J. (2025). Mobilisation of Data From Natural History Collections Can Increase the Quality and Coverage of Biodiversity Information. *Ecology and Evolution*, *15*(4), e71139. https://doi.org/10.1002/ece3.71139

> BB led on data curation, formal analysis, methodology, project administration, writing the original draft, and writing review and editing, and had an equal role in conceptualisation, validation, and visualisation.
>
> JH led on supervision, had an equal role in conceptualisation, methodology and validation, and had a supporting role in data curation, formal analysis, visualisation, and writing review and editing.
>
> CR had an equal role in validation and visualisation, and had a supporting role in conceptualisation, data curation, formal analysis, methodology, and writing review and editing.

Chapter 4: Exploring the congruence of data types: do molecular and morphological trait data describe phylogenetic relationships in the same way? (In Prep)
Bryony Blades, Andrea Estandía

> BB conceived the project. BB did fieldwork. BB did the lab work. BB prepared morphological data. AE prepared raw sequencing data for analysis. BB carried out analysis. BB wrote the manuscript.

Chapter 5: Not the be all and environm-end all: landscape genomics hints that biotic interactions may override environmental adaptations even in strong bioindicator species (In Prep)
Bryony Blades, Andrea Estandía

> BB conceived the project. BB did fieldwork. BB did the lab work. AE prepared raw sequencing data for analysis. BB carried out analysis. BB wrote the manuscript.

Whilst completing my thesis, I have also contributed to the following publication.

Ladle, R J., Diniz-Filho, J A F., Lessa, T., Mott, T., Pertierra, L R., Sobral-Souza, T., Jiménez-Valverde, A., Castro-Souza, R., Tessarolo, G., Stropp, J., Blades, B., Moura, M R., Malhado, A C M., Efe, M A., Gouveia, S., Zarzo-Arias, A., Meyer, L., Leo, M., Azevedo Farias, A K., Ronquillo, C., Guedes, J., Frateles, L E F., Santos, A., Llorente-Culebras, S., Medina, N G., Valcárcel, V., Mestre, A., Mesquita-Joanes, F., Dubeux, M., Menegotto, A., Nakamura, G., Rangely, J., Hortal, J. (2025). The Linnean Shortfall's silent partner: how misidentification drives taxonomic uncertainty. Biological Reviews. (In Review)

Thesis advisors: Professor Tim Coulson, Professor Michael Bonsall, Dr. Elizabeth Jeffers

Digging deeper: using Afrotropical dung beetles to better understand quality and coverage of biodiversity data

Bryony Blades

Abstract

Rapid technological advance over the last century has fundamentally altered the way we understand the world, and the study of biodiversity is no exception. Developments in earth observation and Geographical Information Systems (GIS) tools have generated expansive maps of the environment, next-generation (NGS) sequencing has facilitated access to extensive genetic information, and mass digitisation and collation of records on online biodiversity databases have expanded the spatial, temporal, and taxonomic scope of research far beyond what has ever been possible before. However, despite the many opportunities this data proliferation has afforded us, many issues in their quality, coverage, and congruence have been identified. In the first half of this thesis, I use records from a recent taxonomic revision of *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidea), some of which have not been shared online, to evaluate the impact that up-to-date taxonomy has on species distribution models (SDMs) and whether the data to fill acknowledged coverage gaps exist in natural history collections that are not yet mobilised. In the second half of this thesis, whole genome sequencing of *Catharsius* specimens collected in Zambia is used to assess the degree to which different types of biodiversity data agree, and whether their integration improves understanding of species' relationships and distributions. Specifically, in **Chapter 1**, I provide a general introduction into biodiversity data. In **Chapter 2**, I used species occurrence records

identified according to outdated and up-to-date taxonomic understanding to create individual and ensemble SDMs and measure improvements to model performance as a consequence of taxonomic revision. These improvements were unanimous, and also refined identification of areas of high and low habitat suitability. Critically, ensemble models still performed excellently when based on outdated taxonomy, highlighting that they may obscure taxonomic data quality errors. In **Chapter 3**, I combined occurrence records from the aforementioned taxonomic revision that had not yet been shared with the world's biggest online biodiversity network, the Global Biodiversity Information Facility (GBIF), with all existing *Catharsius* records on GBIF. Quantifying the improvement to inventory completeness and well-sampled cells as a consequence of doing this, I showed that not-yet-mobilised natural history collections have the capacity to disproportionately enhance coverage in poorly sampled regions and climates. In **Chapter 4**, I used the morphological trait matrix from the taxonomic revision alongside whole genome sequencing data to compare the congruence of data types in their description of species' relationships. Using distance matrices and clustering analyses, I found that morphological and molecular data describe the same overall population structure, but differ in their measurements of inter- and intra-cluster relatedness, underlining the merits of an integrative approach to taxonomy. In **Chapter 5**, occurrence, genetic, morphological, and climatic data were integrated in a study of landscape genomics to assess the degree to which adaptation to the environment drives the parapatric distributions of two closely-related species of *Catharsius*. Despite strong links between dung beetles and their habitats, the explanatory power of environmental and geographic variables on their genetic diversity was weak. Instead, I found a genetic continuum along which both species are mixed, suggesting that a

possible hybrid zone in this area is responsible for maintaining their parapatry rather than the current landscape. In **Chapter 6**, I discuss the thesis findings in the broader context of inferring biodiversity knowledge. Overall, this body of work advances understanding of the ways in which data can be improved and utilised for the study of biodiversity.

Contents

CHAPTER 1

Introduction

The variety of life on earth has interested human beings since long before the term "biodiversity" was coined in the 1980s (Sarkar, 2021). Paintings of Sulawesi warty pigs in Maros-Pangkep show that observations of the natural world have captured our attention for as many as 45,500 years (Brumm et al., 2021). Since then, from cave art to the pioneering works of Aristotle, Charles Darwin and Alfred Russel Wallace, to now, how we record these observations has developed alongside technology. It is a sad irony, then, that the technological advance of modern times has undoubtedly contributed to the acute rate of biodiversity loss we are currently experiencing (Hald-Mortensen, 2023). As a silver lining, it has also provided materials and methods to combat this decline, such as the staggering collation of data on easily accessible online databases. Understanding how accurately and comprehensively these data represent Earth's diversity is a critical step in ensuring its protection.

The role of data in characterising biodiversity

Full many a flow'r is born to blush unseen,
And waste its sweetness on the desert air.

- *Thomas Gray, Elegy Written in a Country Churchyard, 1751*

In order to characterise what we do know about biodiversity, it is first important to describe what we don't and, as such, gaps in biodiversity knowledge have been categorised into a number of named shortfalls. These describe mismatches between the actual state of biodiversity, and our current knowledge of it. The best known characterise the discrepancy between the true number of species that exist or have existed and those that have been described (the Linnean shortfall) and understanding of their real, as opposed to known, distributions (Wallacean), abundance and population dynamics (Prestonian), evolution (Darwinian), functions and traits (Raunkiæran), abiotic tolerances (Hutchinsonian), and ecological interactions (Eltonian) (Hortal et al., 2015). Fortunately, as large-scale data collection and analysis has become increasingly embedded in science and broader society, its prominence offers a powerful opportunity to begin addressing these deficits.

The rise of 'Big Data' has been pervasive across industries and fields of research, bringing "datafication" into many aspects of life, from words, friendships, and human health, to vehicle performance, professional networks, fire risk, and even floors (Cukier & Mayer-Schönberger, 2013). However, 'big' does not just refer to quantity; as well as volume, it is defined by velocity, variety, veracity, value, and sometimes variability. These are qualities also reflected in the heterogeneity of biodiversity data, and, as a

result, its study has increasingly benefitted from big data approaches that can handle the complexity and scale. The rapid digitisation and accumulation of biodiversity data (its velocity) has been of particular note, resulting in extensive online databases of easily accessible information on species' distributions, genetics, traits, movement, and population demographics, as well as maps of the environment including soil and land cover, vegetation, climate, and terrain (Franklin et al., 2017; Nelson & Ellis, 2019; Wüest et al., 2020). One of the best known examples is the Global Biodiversity Information Facility (GBIF), which currently holds over three and a half billion occurrence records in a network of databases from 2,570 publishing institutions (GBIF, 20/10/2025). These have enabled analysis of many facets of biodiversity at previously unimaginable spatial, temporal, and taxonomic scales, and have been used across fields such as macroecology, biogeography, conservation, invasive species, functional ecology, population genetics, and taxonomy (Franklin et al., 2017; Heberling et al., 2021; Soberón & Peterson, 2004).

Data quality

Given the widespread use of these data, and their application across wide-ranging disciplines, research interest in their quality, and its effects on downstream analysis, has grown. Online databases that aggregate occurrence records from different sources, such as natural history collections, field surveys, and citizen science are prone to a number of inconsistencies. In the first instance, these can be between data fields such as coordinates and collecting location, or between data providers in the fields that they include, with other ambiguities including uncertain abbreviations of place names or

collectors (Feng et al., 2022; Soberón et al., 2002). A different type of error, in georeferencing, can see coordinates fall in the sea or in a location or habitat not expected for that species, have the latitude and longitude switched, or have one of these erroneously equal to 0. Some may also have coordinates that correspond with the centroid of the country they were observed or collected in, those of the institution they are kept in, or none at all (Ronquillo et al., 2024; Serra-Diaz et al., 2017; Soberón et al., 2002; Soberón & Peterson, 2004; Yesson et al., 2007). On top of this, georeferencing uncertainty is often not recorded, which may mislead biogeographical studies (Marcer et al., 2022). A number of tools and protocols have been developed to filter occurrence points prior to their inclusion in downstream analyses (Ronquillo et al., 2024; Serra-Diaz et al., 2024; Zizka et al., 2020), and geographic errors such as these are the most straightforward to identify and remove or fix. On the other hand, errors in the content of the records are much more difficult to identify after they have been digitised and uploaded. For example, misidentification of the organism, which has been shown to negatively impact biodiversity assessments and species distribution models (SDMs) (Coca-de-la-Iglesia et al., 2024; Costa et al., 2015; A. V. Rodrigues et al., 2022).

Related to misidentifications, taxonomic errors can refer to unstandardised taxonomies across datasets, misspelled scientific names, and synonyms, and recent developments seek to integrate fixing these problems with more basic filtering steps (Jin & Yang, 2020; Ronquillo et al., 2024). A perhaps more problematic taxonomic error, though, is that of outdated taxonomy (Ball-Damerow et al., 2019; Soberón & Peterson, 2004). Given the ever evolving state of life on earth, taxonomic understanding of species and their relationships to one another is also non-static. This dynamic nature is not reflected in

the fixed label of a presence record without taxonomic revision of the group involved, threatening our ability to characterise species relationships with changing environments, and is almost impossible to fix without re-inspection of the original material. What makes this ever more problematic is the declining support for the field of taxonomy over recent decades, resulting in drained resources and expertise, especially in hyper diverse groups such as insects (Hochkirch et al., 2022). With the unlikely chance of publishing in high impact factor journals, and the non-indexed nature of taxonomy-specific publications, difficulty getting funding has a knock-on effect on hiring for faculty positions, and therefore teaching (Hopkins & Freckleton, 2002; Hutchings, 2021; Anon. 1946; Lagomarsino & Frost, 2020; Wägele et al., 2011). Internally the field is not without trouble either, as disagreements continue on the validity of varying species concepts, the integration of modern genetic methods with traditional morphological techniques, and the use of artificial intelligence (Bernard, 2025; Karbstein et al., 2024; Taylor et al., 2019; Valdecasas, 2024; Zachos et al., 2019). This strain means that many species are expected to go extinct before they are described, and revision of already described species is probably even less likely.

A common application of occurrence data is in correlative SDMs, exhibiting the biggest rise of any topic using GBIF data in recent years (Ball-Damerow et al., 2019; Heberling et al., 2021). These use presence records and environmental variables to derive a species-environment relationship that can be applied to describe distributions, assess vulnerability to climate change, predict future range shifts, evaluate risk from invasive species, model the spread of disease vectors, and inform conservation planning across terrestrial, freshwater, and marine environments (Elith et al., 2010; Elith & Franklin,

2013; Elith & Leathwick, 2009; Guisan & Zimmermann, 2000). Geographical errors in occurrence data have been shown to negatively impact the performance of SDMs (Graham et al., 2008; Osborne & Leitão, 2009) and, although the importance of taxonomic accuracy has been underlined (Lozier et al., 2009; Soley-Guardia et al., 2024), the impact of its revision is more difficult to test. In the face of decreasing expertise and funding for taxonomy, and the impossibility of easily filtering for these errors, it is a priority to measure the extent to which these problems damage our ability to describe biodiversity, and how it may be threatened in a changing climate.

Data coverage

Even if we were able to assume that all biodiversity occurrence records on online databases had no errors in quality, and were based on the most up-to-date taxonomic understanding, they have another pervasive flaw. It is now widely-accepted that record coverage is strongly taxonomically and geographically biased, resulting in records that are significantly skewed towards vertebrates, particularly birds, and Europe and North America (Amano et al., 2016; Hughes et al., 2021; Troudet et al., 2017). This unevenness in the geographic distribution of records is so extreme, with comparatively so few records in the tropics, that it appears inversely correlated with species richness (Collen et al., 2008; Yesson et al., 2007). This is not only a problem due to the inadequacy it underlines in our databases, but because these regions are amongst those most at risk from the negative impacts of climate change (Sinivasan, 2010). Mirroring this, taxonomic bias is such that the most diverse order of organisms on the planet, insects, is by far the worst represented in the online record (Girardello et al., 2019; Rocha-

Ortega et al., 2021). It is estimated that there are 200 million fewer records on GBIF than would be expected according to the number of known species (Troudet et al., 2017), with over a third of the world's terrestrial cells lacking any insect data at all (García-Rosello et al., 2023). Spatial biases are also problematic on a smaller scale, with record density seen to correlate with Protected Areas and proximity to roads, showing that coverage is notably affected by accessibility (García-Rosello et al., 2023; Girardello et al., 2019; Newbold, 2010; Petersen et al., 2021; Romo et al., 2006). As a central aggregator of many databases, GBIF's coverage is partly driven by data sharing agreements and its bias affected by the different data types in contributing sources (J. Beck et al., 2014). On the most basic level, there is a clear continental unevenness in the countries that have signed up to participate in the GBIF intergovernmental initiative, with notably poorer representation from Africa and Asia (GBIF, n.d.). More precisely, though, the different data types included in these collections are known to suffer from varying biases. Citizen science, which is quickly taking over as the primary contributor to GBIF (Waller, 2019), has shown its potential to rapidly fill sampling gaps, but still with strong biases towards developed areas, widespread species, and birds (Amano et al., 2016; Shirey et al., 2021; Speed et al., 2018). The distribution of citizen science schemes is also skewed towards regions that already have good coverage (Chandler et al., 2017). Although specimens from natural history collections are also known to be unevenly distributed in space, and this has affected their coverage of environmental conditions (Hortal et al., 2008), they are thought to contain more rare taxa (J. Beck et al., 2013), and contributions from small institutions in particular have been seen to lessen spatial bias (de Araujo et al., 2022; Glon et al., 2017). What remains to be seen is whether data pertaining to underrepresented Afrotropical insects could be found in

institutions that have not digitised and / or shared their data, or whether they do not exist at all. If the latter, we already have a clear picture of where further sampling and surveying should be directed, but if the former, there is potential for knowledge improvement via much less resource demanding methods.

Data integration

In theory, assuming both good data quality and coverage should mean our ability to characterise biodiversity is similarly reliable, and this is somewhat the case. However, some criticisms of online biodiversity databases underline the need to integrate or link databases containing different data types for further improvement (Ball-Damerow et al., 2019; Peterson et al., 2010). The potential benefit in doing so has been demonstrated through fields that have adopted integrative practices, such as taxonomy. Integration of different data types, though, is not always straightforward, as they have been found to convey different information at times. For example, the aforementioned disagreements in preferred taxonomic methods are made even more difficult by their sometimes describing contrasting species boundaries and relationships. One of the best cited examples is the superorder Afrotheria. For many years, its members, including animals such as sea cows, tenrecs, anteaters, armadillos, elephant shrews and elephants, were placed in a number of different groups based on their morphological characteristics. However, with the use of genetic data, their monophyly was later discovered (Lee & Palci, 2015; Oyston et al., 2022; Van Den Ende et al., 2023). Disagreements are not usually this stark, though, and it is now generally agreed that integration of both data types allows for the most comprehensive phylogenies and measurements of genetic

diversity (A. A. Alves et al., 2013; R. M. Alves et al., 2017; E. K. V. D. Andrade et al., 2017; Darkwa et al., 2020; Kadoić Balaško et al., 2021; Keating et al., 2023; Van Den Ende et al., 2023). Even with the agreed need for data integration in taxonomy focused studies, this does raise a concern in the context of the rise of citizen science and mass record digitalization. If genetic information is needed alongside visual cues for identification, which are often not accessible in these scenarios, this opens the door to even more errors of misidentification on databases. It has been suggested that disagreement between morphological classification and genetic data might be as a result of the use of molecular markers rather than whole genome single nucleotide polymorphisms (SNPs) (A. A. Alves et al., 2013; R. M. Alves et al., 2017), but studies since both support and dispute this hypothesis, (Darkwa et al., 2020; Kadoić Balaško et al., 2021). As such, it is important to test whether we can continue to rely on morphological methods of species identification given their returning importance.

Beyond taxonomy and phylogenetics, molecular data has also been integrated with environmental data and spatial methods in the interdisciplinary field of landscape genomics (Sgrò et al., 2011). This seeks to understand patterns of adaptation to the environment using at least thousands of genetic loci, but often whole genomes, facilitated by easier access to sequencing data via the invention and popularisation of next generation sequencing (Chaulk & Keyghobadi, 2022; Storfer et al., 2018). This has added a new dimension to our understanding of biodiversity and its conservation in the ability to measure the degree of genetic adaptation required of a species to adjust to projected changes in climate (Capblancq et al., 2020). In insects these techniques have been applied to pests, disease vectors, and species of conservation concern, with

studies on beetles primarily focused on agricultural and forest pests (Chaulk & Keyghobadi, 2022). One recent study examined the relationship between genetic diversity and habitat fragmentation in the Tropical Dry Forest in the Colombian Caribbean, and this appears to be the only landscape genomics study conducted on true dung beetles (Scarabaeinae) anywhere in the world to date. This is surprising as dung beetles are known not just to be strong indicators of their environment, but also of wider biodiversity, as well as the providers of globally important ecosystem services (Beynon et al., 2015; A. L. V. Davis et al., 2004; McGeoch et al., 2002; Nichols et al., 2008; Slade et al., 2016; Spector, 2006). Measuring their degree of adaptation to environmental conditions allows us to quantify the threat climate change poses to them and, by extension, both wider biodiversity and ecosystem health.

Study system: *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidea)

Catharsius is a genus of large-bodied true dung beetle that is distributed across Africa and Asia. They are strong fliers and, as paracoprids, bury beneath dung pats to construct underground nesting structures (Takano, 2018). Before now they were most recently revised by Ferreira (Ferreira, 1960a, 1960b, 1972), but a revision of the Afrotropical species has taken place in recent years, becoming the largest revision of a group of dung beetles anywhere in the world (Takano, 2018). As effective bioindicators, this taxon is ideal for integrating occurrence, genetic, and environmental data. On top of this, they are members of the most underrepresented taxa and regions in species occurrence data, as well as a group suffering a particular decline in taxonomic expertise. A dataset with reliable taxonomy and a mix of data points that both have, and

have not, been mobilised online presents a unique opportunity to test data quality, coverage, and integration for species and places that need them the most.

Thesis aims

In this thesis, I explore the quality, coverage, and congruence of biodiversity data in an underrepresented taxon in the sparsely sampled Afrotropical realm.

In **Chapter 2**, occurrence information collated for the recent taxonomic revision (Takano, 2018) is used to test the impact of revising taxonomy on the output of SDMs. Along with climatic variables, presence records identified according to outdated and up-to-date taxonomy are used to create ensemble models highlighting suitable habitat across the Afrotropical realm. Results reveal the unanimous improvement in performance from revising taxonomy across three commonly used SDM evaluation metrics. This is driven by an enhanced ability to distinguish high and low suitability habitats, and results in the characterisation of much more clearly distinguished climatic niches. Critically, unlike individual model replicates, the ensemble model informed by outdated taxonomy performed excellently, highlighting that this method may obscure taxonomic data quality issues.

In **Chapter 3**, the data coverage of Afrotropical insects on GBIF is evaluated using occurrence points for all species from the taxonomic revision. For this, *Catharsius* records from GBIF were merged with those from the taxonomic revision that had not been uploaded to measure the improvement in inventory completeness and the

generation of well-sampled cells. Not only did the additional information from the revision increase the quantity of usable data, but it also disproportionately improved coverage of under sampled regions, potential environmental conditions, and rare climates. This study shows that further digitisation of natural history collections would not just increase data quantity, but would lessen knowledge shortfalls in the places and groups for which they are known to be most severe.

In **Chapter 4**, whole genome sequencing and morphological trait data are used to compare the congruence of data types in the description of species' relationships. It finds that the use of single nucleotide polymorphisms (SNPs) from across the whole genome describe the same overall population structure as is suggested by species morphology, but they differ in the degree of intra-cluster relatedness. This underlines the merit of an integrative taxonomic approach, but also shows that taxonomic revision using morphological traits is still a valid methodology, and especially so given the importance of visual identification for biodiversity digitisation.

In **Chapter 5**, the occurrence, genetic, morphological, and climatic data types are integrated in a study of landscape genomics to ascertain the degree of local adaptation in beetles collected in northwest Zambia. Specifically, through the integration of various data types, it seeks to explain the adjacent but non-overlapping distributions of *Catharsius dux* and *Catharsius duciformis*. Unexpectedly for a strong bioindicator taxon that builds underground nests, climatic and soil variables could not explain genetic variation to any significant degree. Notably, this genetic variation could also not be explained by visual species identification, suggesting that the sampling for this study

took place in a genetic hybrid zone within which morphological identities have been conserved. Furthermore, it is this zone, rather than environmental adaptation, that is proposed as the barrier to expansion into suitable habitat exhibited by *C. duciformis* in Chapter 2. Not only does this exemplify a key criticism regarding the difficulty integrating biotic interactions into SDMs, but it also underlines that environment is not always the strongest driver of distributions even in taxa known to be effective bioindicators.

The thesis concludes with the general discussion in **Chapter 6**, where I discuss the results of Chapters 2-5 alongside the broader context in which they are situated. This closes with wider considerations of the trade-off between model complexity and understanding their process, and the difficulty of inferring new knowledge from biodiversity data in the absence of prior understanding.

References

Alves, A. A., Bhering, L. L., Rosado, T. B., Laviola, B. G., Formighieri, E. F., & Cruz, C. D. (2013). Joint analysis of phenotypic and molecular diversity provides new insights on the genetic variability of the Brazilian physic nut germplasm bank. *Genetics and Molecular Biology*, *36*(3), 371–381. https://doi.org/10.1590/S1415-47572013005000033

Alves, R. M., Silva, C. R. D. S., Albuquerque, P. S. B. D., & Santos, V. S. D. (2017). Phenotypic and genotypic characterization and compatibility among genotypes to select elite clones of cupuassu. *Acta Amazonica*, *47*(3), 175–184. https://doi.org/10.1590/1809-4392201602104

Amano, T., Lamming, J. D. L., & Sutherland, W. J. (2016). Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. *BioScience*, *66*(5), 393–400. https://doi.org/10.1093/biosci/biw022

Andrade, E. K. V. D., Andrade Júnior, V. C. D., Laia, M. L. D., Fernandes, J. S. C., Oliveira, A. J. M., & Azevedo, A. M. (2017). Genetic dissimilarity among sweet potato genotypes using morphological and molecular descriptors. *Acta Scientiarum. Agronomy*, *39*(4), 447. https://doi.org/10.4025/actasciagron.v39i4.32847

Anonymous (1946) Importance of Taxonomy. *Nature*, *158*, 105–106. https://doi.org/10.1038/158105b0

Ball-Damerow, J. E., Brenskelle, L., Barve, N., Soltis, P. S., Sierwald, P., Bieler, R., LaFrance, R., Ariño, A. H., & Guralnick, R. P. (2019). Research applications of primary biodiversity databases in the digital age. *PLOS ONE*, *14*(9), e0215794. https://doi.org/10.1371/journal.pone.0215794

Beck, J., Ballesteros-Mejia, L., Nagel, P., & Kitching, I. J. (2013). Online solutions and the "Wallacean shortfall': What does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions*, *19*(8), 1043–1050. https://doi.org/10.1111/ddi.12083

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bernard, J. (2025). Combining new technology with classic taxonomy to overcome hurdles to discovering dark taxa. *Systematics and Biodiversity*, *23*(1), 2454014. https://doi.org/10.1080/14772000.2025.2454014

Beynon, S. A., Wainwright, W. A., & Christie, M. (2015). The application of an ecosystem services framework to estimate the economic value of dung beetles to the U.K. cattle industry. *Ecological Entomology*, *40*(S1), 124–135. https://doi.org/10.1111/een.12240

Brumm, A., Oktaviana, A. A., Burhan, B., Hakim, B., Lebe, R., Zhao, J., Sulistyarto, P. H., Ririmasse, M., Adhityatama, S., Sumantri, I., & Aubert, M. (2021). Oldest cave art found in Sulawesi. *Science Advances*, *7*(3), eabd4648. https://doi.org/10.1126/sciadv.abd4648

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., & Keller, S. R. (2020). Genomic Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes. *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 245–269. https://doi.org/10.1146/annurev-ecolsys-020720-042553

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294. https://doi.org/10.1016/j.biocon.2016.09.004

Chaulk, A., & Keyghobadi, N. (2022). Insect Landscape Genomics. In J. Dupuis & O. P. Rajora (Eds), *Population Genomics: Insects*. Springer Nature.

Coca-de-la-Iglesia, M., Gallego-Narbón, A., Alonso, A., & Valcárcel, V. (2024). High rate of species misidentification reduces the taxonomic certainty of European biodiversity databases of ivies (*Hedera* L.). *Scientific Reports*, *14*(1), 4876. https://doi.org/10.1038/s41598-024-54735-0

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75–88. https://doi.org/10.1177/194008290800100202

Costa, H., Foody, G., Jiménez, S., & Silva, L. (2015). Impacts of Species Misidentification on Species Distribution Modeling with Presence-Only Data. *ISPRS International Journal of Geo-Information*, *4*(4), 2496–2518. https://doi.org/10.3390/ijgi4042496

Cukier, K., & Mayer-Schönberger, V. (2013). The Rise of Big Data: How It's Changing the Way We Think About the World. *Foreign Affairs*, *92*(3), 28–40.

Darkwa, K., Agre, P., Olasanmi, B., Iseki, K., Matsumoto, R., Powell, A., Bauchet, G., De Koeyer, D., Muranaka, S., Adebola, P., Asiedu, R., Terauchi, R., & Asfaw, A. (2020). Comparative assessment of genetic diversity matrices and clustering methods in white Guinea yam (*Dioscorea rotundata*) based on morphological and molecular markers. *Scientific Reports*, *10*(1), 13191. https://doi.org/10.1038/s41598-020-69925-9

Davis, A. L. V., Scholtz, C. H., Dooley, P. W., Bham, N., & Kryger, U. (2004). Scarabaeine dung beetles as indicators of biodiversity, habitat transformation and pest control chemicals in agro-ecosystems. *South African Journal of Science*. 100(9/10), 415–424.

de Araujo, M. L., Quaresma, A. C., & Ramos, F. N. (2022). GBIF information is not enough: National database improves the inventory completeness of Amazonian epiphytes. *Biodiversity and Conservation*, *31*(11), 2797–2815. https://doi.org/10.1007/s10531-022-02458-x

Elith, J., & Franklin, J. (2013). Species Distribution Modeling. In S. A. Levin (Ed.), *Encyclopedia of Biodiversity* (2nd edn, Vol. 6, pp. 692–705). Elsevier. https://doi.org/10.1016/B978-0-12-384719-5.00318-X

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Feng, X., Enquist, B. J., Park, D. S., Boyle, B., Breshears, D. D., Gallagher, R. V., Lien, A., Newman, E. A., Burger, J. R., Maitner, B. S., Merow, C., Li, Y., Huynh, K. M., Ernst, K., Baldwin, E., Foden, W., Hannah, L., Jørgensen, P. M., Kraft, N. J. B., … Hurlbert, A. (2022). A review of the heterogeneous landscape of biodiversity databases: Opportunities and challenges for a synthesized biodiversity

knowledge base. *Global Ecology and Biogeography*, geb.13497. https://doi.org/10.1111/geb.13497

Ferreira, M. C. (1960a). Descricao de especies novas de *Catharsius* s.str. *Novos Taxa Entomológicos*, *23*, 3–8.

Ferreira, M. C. (1960b). Revisao das esepcies Africanas de *Catharsius* s.str. Do Groupo Adamastor e descricao de especies novas. *Revista Entomologia Moçambique*, *3*(1), 1–73.

Ferreira, M. C. (1972). Os Escarabideos de Africa (sul do Saara). I. *Revista de Entomologia de Mocambique*, *11*, 5–1088.

Franklin, J., Serra-Diaz, J. M., Syphard, A. D., & Regan, H. M. (2017). Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography*, *26*(1), 6–17. https://doi.org/10.1111/geb.12501

García-Rosello, E., Gonzalez-Dacosta, J., Guisande, C., & Lobo, J. M. (2023). GBIF falls short of providing a representative picture of the global distribution of insects. *Systematic Entomology*, *48*(4), 489–497. https://doi.org/10.1111/syen.12589

GBIF. (2025, June 25). *GBIF*. https://www.gbif.org/

GBIF. (n.d.). *The GBIF Network*. https://www.gbif.org/the-gbif-network/africa

Girardello, M., Chapman, A., Dennis, R., Kaila, L., Borges, P. A. V., & Santangeli, A. (2019). Gaps in butterfly inventory data: A global analysis. *Biological Conservation*, *236*, 289–295. https://doi.org/10.1016/j.biocon.2019.05.053

Glon, H. E., Heumann, B. W., Carter, J. R., Bartek, J. M., & Monfils, A. K. (2017). The contribution of small collections to species distribution modelling: A case study from *Fuireneae* (Cyperaceae). *Ecological Informatics*, *42*, 67–78. https://doi.org/10.1016/j.ecoinf.2017.09.009

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A., & The Nceas Predicting Species Distributions Working Group. (2008). The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, *45*(1), 239–247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Hald-Mortensen, C. (2023). The Main Drivers of Biodiversity Loss: A Brief Overview. *Journal of Ecology & Natural Resources*, *7*(3). https://doi.org/10.23880/jenr-16000346

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, *118*(6), e2018093118. https://doi.org/10.1073/pnas.2018093118

Hochkirch, A., Casino, A., Lyubomir, P., Allen, D., Tilley, L., Georgiev, T., Gospodinov, K., & Barov, B. (2022). *European Red List of Insect Taxonomists*. Publication Office of the European Union.

Hopkins, G. W., & Freckleton, R. P. (2002). Declines in the numbers of amateur and professional taxonomists: Implications for conservation. *Animal Conservation*, *5*(3), 245–249. https://doi.org/10.1017/S1367943002002299

Hortal, J., De Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*(1), 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Hortal, J., Jiménez-Valverde, A., Gómez, J. F., Lobo, J. M., & Baselga, A. (2008). Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, *117*(6), 847–858. https://doi.org/10.1111/j.0030-1299.2008.16434.x

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., & Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, *44*(9), 1259–1269. https://doi.org/10.1111/ecog.05926

Hutchings, P. (2021). Potential loss of biodiversity and the critical importance of taxonomy—An Australian perspective. In *Advances in Marine Biology* (Vol. 88, pp. 3–16). Elsevier. https://doi.org/10.1016/S0065-2881(21)00015-8

Jin, J., & Yang, J. (2020). BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation*, *21*, e00852. https://doi.org/10.1016/j.gecco.2019.e00852

Kadoić Balaško, M., Mikac, K. M., Benítez, H. A., Bažok, R., & Lemic, D. (2021). Genetic and Morphological Approach for Western Corn Rootworm Resistance

Management. *Agriculture*, *11*(7), 585. https://doi.org/10.3390/agriculture11070585

Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., Wagner, N. D., Emerson, B. C., Albach, D. C., Scheu, S., Bradler, S., De Vries, J., Irisarri, I., Li, H., Soltis, P., Mäder, P., & Wäldchen, J. (2024). Species delimitation 4.0: Integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution*, *39*(8), 771–784. https://doi.org/10.1016/j.tree.2023.11.002

Keating, J. N., Garwood, R. J., & Sansom, R. S. (2023). Phylogenetic congruence, conflict and consilience between molecular and morphological data. *BMC Ecology and Evolution*, *23*(1), 30. https://doi.org/10.1186/s12862-023-02131-z

Lagomarsino, L. P., & Frost, L. A. (2020). The Central Role of Taxonomy in the Study of Neotropical Biodiversity. *Annals of the Missouri Botanical Garden*, *105*(3), 405–421. https://doi.org/10.3417/2020601

Lee, M. S. Y., & Palci, A. (2015). Morphological Phylogenetics in the Genomic Age. *Current Biology*, *25*(19), R922–R929. https://doi.org/10.1016/j.cub.2015.07.009

Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of Sasquatch in western North America: Anything goes with ecological niche modelling. *Journal of Biogeography*, *36*(9), 1623–1627. https://doi.org/10.1111/j.1365-2699.2009.02152.x

Marcer, A., Chapman, A. D., Wieczorek, J. R., Xavier Picó, F., Uribe, F., Waller, J., & Ariño, A. H. (2022). Uncertainty matters: Ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. *Ecography*, *2022*(9), e06025. https://doi.org/10.1111/ecog.06025

McGeoch, M. A., Van Rensburg, B. J., & Botes, A. (2002). The verification and application of bioindicators: A case study of dung beetles in a savanna ecosystem. In *Journal of Applied Ecology* (Vol. 39, pp. 661–672).

Nelson, G., & Ellis, S. (2019). The history and impact of digitization and digital data mobilization on biodiversity research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1763), 20170391. https://doi.org/10.1098/rstb.2017.0391

Newbold, T. (2010). Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in*

Physical Geography: Earth and Environment, *34*(1), 3–22. https://doi.org/10.1177/0309133309355630

Nichols, E., Spector, S., Louzada, J., Larsen, T., Amezquita, S., & Favila, M. E. (2008). Ecological functions and ecosystem services provided by Scarabaeinae dung beetles. *Biological Conservation*, *141*(6), 1461–1474. https://doi.org/10.1016/j.biocon.2008.04.011

Osborne, P. E., & Leitão, P. J. (2009). Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions*, *15*(4), 671–681. https://doi.org/10.1111/j.1472-4642.2009.00572.x

Oyston, J. W., Wilkinson, M., Ruta, M., & Wills, M. A. (2022). Molecular phylogenies map to biogeography better than morphological ones. *Communications Biology*, *5*(1), 521. https://doi.org/10.1038/s42003-022-03482-x

Petersen, T. K., Speed, J. D. M., Grøtan, V., & Austrheim, G. (2021). Species data for understanding biodiversity dynamics: The what, where and when of species occurrence data collection. *Ecological Solutions and Evidence*, *2*(1), e12048. https://doi.org/10.1002/2688-8319.12048

Peterson, A. T., Knapp, S., Guralnick, R., Soberón, J., & Holder, M. T. (2010). The big questions for biodiversity informatics. *Systematics and Biodiversity*, *8*(2), 159–168. https://doi.org/10.1080/14772001003739369

Rocha-Ortega, M., Rodriguez, P., & Córdoba-Aguilar, A. (2021). Geographical, temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. *Ecological Entomology*, *46*(4), 718–728. https://doi.org/10.1111/een.13027

Rodrigues, A. V., Nakamura, G., Staggemeier, V. G., & Duarte, L. (2022). Species misidentification affects biodiversity metrics: Dealing with this issue using the new R package naturaList. *Ecological Informatics*, *69*, 101625. https://doi.org/10.1016/j.ecoinf.2022.101625

Romo, H., García-Barros, E., & Lobo, J. M. (2006). Identifying recorder-induced geographic bias in an Iberian butterfly database. *Ecography*, *29*(6), 873–885. https://doi.org/10.1111/j.2006.0906-7590.04680.x

Ronquillo, C., Stropp, J., & Hortal, J. (2024). OCCUR Shiny application: A user-friendly guide for curating species occurrence records. *Methods in Ecology and Evolution*, *15*(5), 816–823. https://doi.org/10.1111/2041-210X.14271

Sarkar, S. (2021). Origin of the Term Biodiversity. *BioScience*, *71*(9), 893–893. https://doi.org/10.1093/biosci/biab071

Serra-Diaz, J. M., Borderieux, J., Maitner, B., Boonman, C. C. F., Park, D., Guo, W., Callebaut, A., Enquist, B. J., Svenning, J., & Merow, C. (2024). occTest: An integrated approach for quality control of species occurrence data. *Global Ecology and Biogeography*, *33*(7), e13847. https://doi.org/10.1111/geb.13847

Serra-Diaz, J. M., Enquist, B. J., Maitner, B., Merow, C., & Svenning, J.-C. (2017). Big data of tree species distributions: How big and how good? *Forest Ecosystems*, *4*(1), 30. https://doi.org/10.1186/s40663-017-0120-0

Sgrò, C. M., Lowe, A. J., & Hoffmann, A. A. (2011). Building evolutionary resilience for conserving biodiversity under climate change. *Evolutionary Applications*, *4*(2), 326–337. https://doi.org/10.1111/j.1752-4571.2010.00157.x

Shirey, V., Belitz, M. W., Barve, V., & Guralnick, R. (2021). A complete inventory of North American butterfly occurrence data: Narrowing data gaps, but increasing bias. *Ecography*, *44*(4), 537–547. https://doi.org/10.1111/ecog.05396

Sinivasan, U. T. (2010). Economics of climate change: Risk and responsibility by world region. *Climate Policy*, *10*(3), 298–316. https://doi.org/10.3763/cpol.2009.0652

Slade, E. M., Riutta, T., Roslin, T., & Tuomisto, H. L. (2016). The role of dung beetles in reducing greenhouse gas emissions from cattle farming. *Scientific Reports*, *6*(1), 18140. https://doi.org/10.1038/srep18140

Soberón, J., Arriaga, L., & Lara, L. (2002). Issues of quality control in large, mixed-origin entomological databases. In H. Saarenmaa & E. S. Nielsen (Eds), *Towards a global biological information infrastructure.pdf* (pp. 15–22). European Environment Agency. https://www.eea.europa.eu/publications/technical_report_2001_70

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 689–698. https://doi.org/10.1098/rstb.2003.1439

Soley-Guardia, M., Alvarado-Serrano, D. F., & Anderson, R. P. (2024). Top ten hazards to avoid when modeling species distributions: A didactic guide of assumptions,

problems, and recommendations. *Ecography*, *2024*(4), e06852. https://doi.org/10.1111/ecog.06852

Spector, S. (2006). Scarabaeine dung beetles (Coleoptera: Scarabaeidae: Scarabaeinae): An invertebrate focal taxon for biodiversity research and conservation. *The Coleopterists Bulletin*, *60*, 71–83.

Speed, J. D. M., Bendiksby, M., Finstad, A. G., Hassel, K., Kolstad, A. L., & Prestø, T. (2018). Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PLOS ONE*, *13*(4), e0196417. https://doi.org/10.1371/journal.pone.0196417

Storfer, A., Patton, A., & Fraik, A. K. (2018). Navigating the Interface Between Landscape Genetics and Landscape Genomics. *Frontiers in Genetics*, *9*, 68. https://doi.org/10.3389/fgene.2018.00068

Takano, H. (2018). *A systematic revision of the Afrotropical members of the dung beetle genus* Catharsius *Hope, 1837 (Coleoptera: Scarabaeidae)* [DPhil Thesis]. University of Oxford.

Taylor, P. J., Denys, C., & Cotterill, F. P. D. (Woody). (2019). Taxonomic anarchy or an inconvenient truth for conservation? Accelerated species discovery reveals evolutionary patterns and heightened extinction threat in Afro-Malagasy small mammals. *Mammalia*, *83*(4), 313–329. https://doi.org/10.1515/mammalia-2018-0031

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*(1), 9132. https://doi.org/10.1038/s41598-017-09084-6

Valdecasas, A. G. (2024). Can Taxonomists Think? Reversing the AI Equation. *Taxonomy*, *4*(4), 713–722. https://doi.org/10.3390/taxonomy4040037

Van Den Ende, C., Puttick, M. N., Urrutia, A. O., & Wills, M. A. (2023). Why should we compare morphological and molecular disparity? *Methods in Ecology and Evolution*, *14*(9), 2390–2410. https://doi.org/10.1111/2041-210X.14166

Wägele, H., Klussmann-Kolb, A., Kuhlmann, M., Haszprunar, G., Lindberg, D., Koch, A., & Wägele, J. W. (2011). The taxonomist—An endangered race. A practical proposal for its survival. *Frontiers in Zoology*, *8*(1), 25. https://doi.org/10.1186/1742-9994-8-25

Waller, J. (2019, January 21). *Will citizen science take over?* GBIF Data Blog. https://data-blog.gbif.org/post/gbif-citizen-science-data/

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M., & Karger, D. N. (2020). Macroecology in the age of Big Data – Where to go from here? *Journal of Biogeography*, *47*(1), 1–12. https://doi.org/10.1111/jbi.13633

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How Global Is the Global Biodiversity Information Facility? *PLoS ONE*, *2*(11), e1124. https://doi.org/10.1371/journal.pone.0001124

Zachos, F. E., Christidis, L., & Garnett, S. T. (2019). Mammalian species and the twofold nature of taxonomy: A comment on Taylor et al. 2019. *Mammalia*, *84*(1), 1–5. https://doi.org/10.1515/mammalia-2019-0009

Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J. F. R., Colli-Silva, M., Fantinati, M. R., Fernandes, M. F., Ferreira-Araújo, T., Gondim Lambert Moreira, F., Santos, N. M. C., Santos, T. A. B., Dos Santos-Costa, R. C., Serrano, F. C., Alves Da Silva, A. P., De Souza Soares, A., Cavalcante De Souza, P. G., Calisto Tomaz, E., … Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. *PeerJ*, *8*, e9916. https://doi.org/10.7717/peerj.9916

CHAPTER 2

The impact of taxonomic revision on species distribution modelling

In Preparation

Bryony Blades[1,2]
[1]Department of Biology, University of Oxford, Oxford, UK
[2]African Natural History Research Trust, Kingsland, Herefordshire, UK

Abstract

The rapid rise of species distribution models (SDMs), especially over the last decade, can be partly attributed to the equally expeditious collation of occurrence records on easily accessible online biodiversity databases. Despite their utility, these records are known to be negatively impacted by a number of data quality problems, including their taxonomic accuracy. This is problematic as, in an opposite trend to that of distribution modelling, the popularity of taxonomy has been in decline, especially for insects. The importance of taxonomy to SDMs has been demonstrated, but we have yet to evaluate the threat that this declining taxonomic capacity poses to the integrity of these databases, as well as the models that make use of them. In this study, I aim to measure the potential improvement to SDMs as a consequence of taxonomic revision, qualify in what way this impacts our ability to describe species' niches, and assess the ability of commonly-used modelling algorithms to flag taxonomic errors in

occurrence data. Using records for six species from a recent revision of *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidae), I compared individual and ensemble modelling outputs to those generated using occurrences identified according to outdated taxonomic understanding. I analysed how the revision influenced SDM performance according to three oft-used evaluation metrics and maps of projected habitat suitability across the Afrotropical realm. Findings show that changes to species identifications as a consequence of the taxonomic revision consistently improved model performance and, in doing so, enhanced recognition of areas of high and low habitat suitability. Critically, the ensemble model based on outdated taxonomy still performed above the threshold for excellence in all three metrics, underlining that individual replicates should be inspected carefully as an ensemble approach may obscure data problems.

Introduction

As the rate of technological advance has accelerated over recent decades, so too has our ability to describe global distributions of biodiversity. Improvements in earth observation technologies and production of environmental maps have provided access to myriad descriptions and predictions of global conditions through time (Franklin et al., 2017; He et al., 2015; Wüest et al., 2020), and increased digitisation of natural history collections and collation of data on online biodiversity networks such as the Global Biodiversity Information Facility (GBIF) has assembled billions of more easily accessible records of species occurrences (Heberling et al., 2021). Species distribution modelling, an often correlative tool designed to take advantage of these types of data to derive a species-environment relationship and assess habitat suitability, has consequently been at the forefront of studies in biogeography, conservation biology, ecology, paleoecology, and wildlife management (Araújo & Guisan, 2006). The capacity to identify biodiversity hotspots, track shifts driven by the changing climate, and predict the spread of invasive species and disease vectors are but some of the practical applications of this tool which are so critical at a time of unprecedented anthropogenic impact on the natural world (Elith & Franklin, 2013; Elith & Leathwick, 2009; Franklin, 2023; Guisan & Zimmermann, 2000)

It is unsurprising, then, that species distribution modelling is the most prevalent topic amongst papers published using GBIF data over the period 2013–2019 (Heberling et al., 2021), and many of the most highly cited papers in biogeography focus on developing methods (Dawson et al., 2023; Journal of Biogeography, 2022) and seeking to better understand elements such as modelling techniques (Segurado & Araújo, 2004),

parameterization and evaluation (Araújo & Guisan, 2006; C. Liu et al., 2019; Radosavljevic & Anderson, 2014), uncertainty (R. G. Pearson et al., 2006), transferability (Elith & Leathwick, 2009; Moreno-Amat et al., 2015; Randin et al., 2006), variable selection (Austin & Van Niel, 2011; Bradie & Leung, 2017), spatial bias in sampling (Veloz, 2009), and small sample sizes (R. G. Pearson et al., 2007; van Proosdij et al., 2016). Crucially, models are fundamentally limited by the quality of the data they use, and many criticisms have been levelled at occurrence data quality and transparency (Soberón et al., 2002). Users have been encouraged away from its automatised use without taking measures to lessen uncertainty, correct for spatial bias, or verify data with subject experts (J. Beck et al., 2014; Moudrý et al., 2024).

Compared to the increase in species distribution model (SDM) studies, taxonomic studies using GBIF data decreased proportionately during the same time period (Heberling et al., 2021). This mirrors a wider decline in taxonomic research by both professionals and amateurs, which not only threatens our ability to discern evolutionary patterns and better understand rates of biodiversity decline, but jeopardizes the basis of other disciplines such as ecology and genetics (Bacher, 2012; Clarkson, 1998; Hochkirch et al., 2022; Hopkins & Freckleton, 2002; D. L. Pearson et al., 2011; Tahseen, 2014). This decline is particularly bad in insect taxonomy, with taxonomic capacity–the available knowledge, skills, and resources of experts to identify and classify species–threatened for over 40% insect orders at the European level (Hochkirch et al., 2022), despite their diversity and the integral role they play in ecology, agriculture, human health, and natural resources (Scudder, 2017). As an example, dung beetles (Coleoptera: Scarabaeidae: Scarabaeinae) are considered to be effective bioindicators

at a variety of scales (A. L. V. Davis et al., 2004) as they are widely distributed, sensitive to ecological and anthropogenic change, ecologically and economically important, and can be used effectively as a surrogate for overall diversity of an area (Spector, 2006). For this reason, they feature in many studies of biogeography across the globe (see, for example: Daniel et al., 2021; A. L. V. Davis & Dewhurst, 1993; A. L. V. Davis & Scholtz, 2001; Escobar et al., 2007; Halffter, 1991; Herzog et al., 2013; Hewavithana et al., 2016; Jos, 2012; McGeoch et al., 2002; Raine & Slade, 2019; Salomão et al., 2022; Shahabuddin et al., 2014; Villamarin-Cortez et al., 2022). Despite this proliferation of work, identification of dung beetles remains challenging (Takano, 2018) and, according to expert opinion, much of the work on dung beetle taxonomy can be outdated or unreliable (H. Takano, personal communication, 10 August, 2021).

With regards to practical applications, it is intuitive to say that taxonomic accuracy of occurrence data is important to species distribution modelling. Full sampling of environmental gradients in which a species is found, a key assumption of these methods (Elith & Leathwick, 2009), cannot be achieved without first understanding the species. The particular interplay has been explored in more detail, for example in invasive species management, in which the use of occurrences pertaining only to ecologically relevant subspecies improves prediction of invasive risk (Agarwal et al., 2021; Mori et al., 2019), and heterogenous niche properties described by taxonomically unreliable records generate a much larger potential invasion range than do models built upon reliable records (Ensing et al., 2013). The importance of classification level has also been tested outside this discipline, agreeing that incorporation of intra-specific variation improves SDM predictions (Barria et al., 2020; Goudarzi et al., 2021).

Convincing distribution predictions have also been shown to still be attainable despite taxonomic error in publicly available records (Lozier et al., 2009). However, the scale and nature of change to distribution modelling outcomes as a consequence of taxonomic revision has not been explicitly tested. The concurrent increase in SDM studies and decline of taxonomy make it even more important to quantify this now.

Given their sensitivity to environmental change (A. L. V. Davis et al., 2004; Spector, 2006), dung beetles are the ideal candidate for correlative species distribution modelling. Fortunately, the largest ever revision of a group of dung beetles using modern methods was recently undertaken (Takano, 2018), presenting an opportunity to assess the ongoing importance of such research. In this study, occurrence points for the Afrotropical members of the genus *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidea) identified according to both outdated and up-to-date taxonomy are used in SDMs to determine the effect of taxonomic revision on model outputs. Specifically, it investigates model performance according to commonly used evaluation metrics, and the ability to distinguish areas of high and low suitability.

Materials and Methods

Study Area and Taxon

The study area is the African mainland of the Afrotropical biogeographical realm spanning 17.5°W–51°E, 21°N–35°S. The study region exhibits a wide breadth of climatic conditions, including eight biomes and 91 ecoregions, encompassing tropical forest to

xeric shrublands. For use later in the analyses, a shapefile of the realm (Olson et al., 2001; World Wildlife Fund, 2012) was modified to include only the mainland of Africa using QGIS version 3.22.9-Białowieża (QGIS, 2021).

Catharsius is a genus of large copro- and necrophagous dung beetles with species distributed across both Africa and Asia. For the aforementioned revision, an extensive collection of distribution data pertaining to the Afrotropical members was compiled from natural history collections (Takano, 2018). As a first step in my analysis, I extracted a set of 4,998 records including species names, coordinates, years and counts from the text of the taxonomic revision. All occurrences were manually inspected to ensure they fell on land and in countries that were expected for that species, and a small number of errors from typing inaccuracies corrected with the expert help of the original taxonomist, Hitoshi Takano.

Occurrences

The previous revision of Afrotropical *Catharsius* species prior to Takano (2018) was conducted over fifty years ago (Ferreira, 1960a, 1960b, 1972). As a consequence of the most recent revision, changes in specimen identifications dramatically altered our understanding of some species' distributions. For example, specimens previously attributed to *Catharsius dux* (henceforth, *C. dux sensu* Ferreira) were re-identified as six different species (henceforth, the "true" species, including *C. dux sensu* Takano). To investigate the impact these changes have on SDMs, all occurrences of these six true species (*C. dux sensu* Takano, *Catharsius dominus, Catharsius duciformis, Catharsius*

gorilla, Catharsius satyrus, and *Catharsius ulysses*) were extracted from the aforementioned wider dataset of 4,998 records, and compiled alongside those for *C. dux sensu* Ferreira (Figure 2.1). For clarity moving forwards, all references to "species" include *C. dux sensu* Ferreira, but all references to "true species" do not.

As occurrence points have been extracted from natural history collections, it is possible that sampling is biased in environmental and geographic space, which is known to have a negative impact on model calibration (Veloz, 2009). To reduce this potential bias, points were thinned using the R package 'spThin' version 0.2.0 (Aiello-Lammens et al., 2015) with a minimum nearest neighbour distance of 2km. This resulted in 415 total points split between the seven species (Table 2.1), such that all seven have in excess of the minimum number of records required for building accurate SDMs according to three out of four different criteria investigated by van Proosdij et al. (2016).

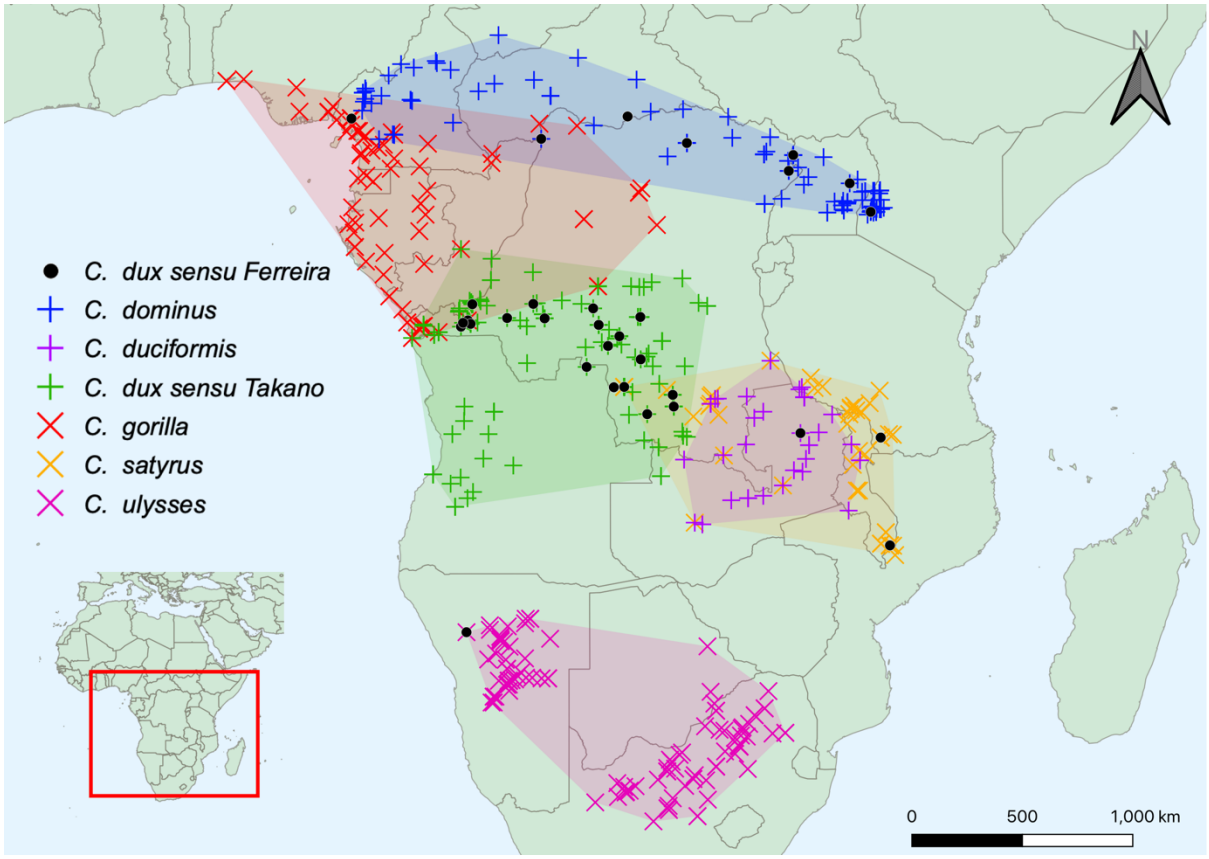


Figure 2.1: The taxonomic revision of Catharsius *re-identified occurrences previously identified as* C. dux sensu *Ferreira as six different species. Occurrences displayed on this map as black points pertain to the outdated taxonomy (*C. dux sensu *Ferreira) and coloured crosses are all records extracted from the taxonomic revision that pertain to these six species. Points shown are prior to spatial thinning.*

Table 2.1: Catharsius *occurrences extracted from the taxonomic revision were spatially thinned to reduce sampling bias for use in SDMs. The change in the number of records as a consequence of this process is displayed here by species.*

Species	***Number of occurrences prior to thinning***	***Number of occurrences after thinning***
C. dominus	156	78
C. duciformis	37	29
C. dux sensu Ferreira	33	33
C. dux sensu Takano	180	88
C. gorilla	93	62
C. satyrus	70	37
C. ulysses	121	88

Environmental variables

For use in the SDMs, the 19 bioclimatic variables from WorldClim were downloaded at a resolution of 30 s (approximately one km at the equator) and cropped to the study extent using the aforementioned shapefile. These variables represent trends, seasonality, and limiting environmental factors of temperature and precipitation derived from aggregating monthly data from the period 1970–2000 (Fick & Hijmans, 2017; WorldClim, 2020). As models that are fit with correlated variables are unreliable when projected to new times or places where those correlations may not hold true (Warren et al., 2014), they were subjected to a Pearson correlation test using the R package 'virtualspecies' version 1.5.1 (Leroy et al., 2016). This generated a list of layers wherein those with a correlation coefficient (|r|) of ≥ 0.7 are placed into groups. This threshold follows evidence that predictive models begin to lose their ability to distinguish between correlated predictors when pairwise correlations exceed approximately |0.7| (Smith & Santos, 2020). Uncorrelated layers were automatically included in the model, and a single variable chosen from each correlated group based on its ecological relevance and interpretability in relation to the study species. The layers included in the final modelling process were annual mean temperature, mean diurnal range, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, precipitation of wettest month, precipitation seasonality, precipitation of driest quarter, precipitation of warmest quarter, and precipitation of coldest quarter.

Species distribution models

First, absence points were generated for inclusion in the SDMs. A number of methods by which to generate these have been investigated, including random sampling, sampling geographically distant points, and sampling environmentally dissimilar locations (Barbet-Massin et al., 2012). Here, R package 'biomod2' version 4.2-5 (Thuiller et al., 2024) was used to generate a surface range envelope (SRE) for each species, a basic model that defines the potential suitable habitat based on the observed range of environmental values, excluding the 5% most extreme values at both ends to reduce the effect of outliers. Absences were randomly generated within unsuitable areas, totalling 3x the number of presence points for each species, as recommended by the creators of SDM R package 'biomod2' (Guéguen et al., n.d.). The SRE method has been shown to be effective with a higher number of presence points, likely due to better sampling completeness (Barbet-Massin et al., 2012). With fewer presence points it is comparatively unlikely that they sample the full extent of the species' climatic niche, and the chance of false absences is high, but with more presence points this risk decreases. In this case, randomly sampling absences in areas deemed unsuitable by an SRE are more likely to be informative than absences sampled according to geographic distance as, although the latter would still be more likely to be true absences, they may be too different from presences to be informative (Barbet-Massin et al., 2012). Here, although the number of points for each species is low, they are a subset of much wider sampling for this genus which includes almost 5000 records and, although not completely without bias, is comprehensive in its coverage (Takano, 2018). Furthermore, as dung beetles are known to be sensitive to environmental gradients (A. L. V. Davis et

al., 2004; Spector, 2006), and *Catharsius* is known to be a strong flier and consequently a good disperser (Takano, 2018), it was assumed that where suitable environmental conditions exist, the species is likely to have been recorded, and therefore their realised niches are expected to be well-sampled.

Given the varying outcomes of distribution modelling algorithms, amongst which there is no one method that universally outperforms the others (Elith et al., 2006), the use of ensemble models, which generate a consensus model from several individual model replicates, has been recommended (Buisson et al., 2010; Grenouillet et al., 2011; C. Liu et al., 2019; Marmion et al., 2009; Stohlgren et al., 2010), and is now widely implemented. Following this, 'biomod2' was used to create ensemble models for each of the species included here. Individual models were first created using two regression algorithms–Generalised Linear Model (GLM), and Generalised Additive Model (GAM)–and four machine learning algorithms–Generalised Boosting Model (also called Boosted Regression Trees, GBM), Artificial Neural Network (ANN), Random Forest (RF), and Maximum Entropy (MAXNET). Ten replicates of each were run, each with an 80/20 training/testing data split and equal weighting between presences and absences. The use of default settings for distribution modelling has been criticised (e.g. Morales et al. (2017) for Maxent), so parameters defined for each model type by the 'biomod2' developers (OPT.strategy = "bigboss") was used. Three frequently used evaluation metrics, the True Skill Statistic (TSS), Cohen's Kappa (KAPPA), and the Area-Under-Curve of Relative Operating Characteristic (ROC) were generated for validation. The first two metrics are threshold dependent, and results range from -1 to +1, with +1 indicating perfect agreement and values below zero indicating a performance no better than

random (Allouche et al., 2006; C. Liu et al., 2011). For presence–absence data, values for the latter (which is usually referred to as AUC and will be referred to as such henceforth) range from 0.5 (no better than random) to 1 (perfect discrimination) (Araujo et al., 2005). In presence–background models such as Maxent, however, the maximum achievable AUC depends on the species' true prevalence and the extent of the background sample (Phillips et al., 2006). There is much debate about which evaluation metric is best (see Discussion), but these have been chosen to establish the impact poor taxonomy has in combination with commonly used methods. An AUC threshold of 0.8 for 'good' and 0.9 for 'excellence' is recommend to judge SDM performance (Araujo et al., 2005) and, once the individual replicates had run, those with a validation AUC < 0.8 were filtered out. A single ensemble model for each species was then created using the mean of probabilities over the remaining replicates, weighted according to their validation AUC score. Lastly, the seven ensemble models were projected to the extent of the Afrotropical realm, and the raster images saved.

All SDMs were run on the University of Oxford Advanced Research Computing (ARC) facility (Richards, 2015).

Model performance

To test whether the model validation scores for *C. dux sensu* Ferreira were significantly different from those of the true species, i.e. whether there is a relationship between model performance and taxonomic accuracy, a Friedman test was conducted in R base package 'stats' for each modelling algorithm and each evaluation metric. The Friedman

test is a non-parametric test used for analysing repeated measures data (Friedman, 1937), allowing for all ten replicates of each algorithm for each species to be included individually. Variances between groups were unequal, as assessed using a Levene test from R package 'car' version 3.1-3 (Fox & Weisberg, 2019), and a Conover's post hoc test therefore used to highlight which groups were driving significance.

Habitat suitability projections

Ensemble model projections generated a value between zero (low) and 1000 (high) for each raster cell. To assess whether poor taxonomy affects the ability of distribution models to identify areas of high and low habitat suitability, QGIS 'Białowieża' was used to reclassify each cell as low (0–250), low–average (251–500), average–high (501–750), or high suitability (751–1000), and to count the total number of cells across each bin for each species. The R base package 'stats' was used to generate a contingency table of these values and run a Chi-squared test to establish whether there were significant differences in the distribution of cells across each bin between species. However, this test is sensitive to large sample sizes and the cell count in the projection rasters was extremely high. As such, a Cramér's V was run in R package 'vcd' version 1.4-13 (Meyer et al., 2006, 2024) to estimate the effect size, i.e. the strength of the relationship between species and the ability of amount of high and low suitability area generated by its model. As a measure of association between zero (no association) and one (perfect association), interpretation varies according to degrees of freedom (df*), defined as the smallest of (total rows - 1) or (total columns -1) from the contingency table (Gravetter & Wallnau, 2012). Here, with df*=3, effect size thresholds are 0.06 (small), 0.17 (medium),

and 0.29 (large) (Cohen, 1988). The standardised residuals, a measure of how far the observed count in each bin differs from the expected count, were used to highlight which species were driving the overall difference.

Results

Model performance

The ensemble model for *C. dux sensu* Ferreira generated the lowest performance scores across all three evaluation metrics (Table 2.2), but still above the threshold for the top performance category in each (0.8 for TSS "excellent" and KAPPA "almost perfect", and 0.9 for AUC "excellent" (Araujo et al., 2005; Landis & Koch, 1977)). The mean validation score of all individual replicates, including those that would have been filtered out for the ensemble model, was also lowest for this species across all three evaluation metrics (Supplementary Figure 2.1), so the difference between this mean and the performance of the ensemble model was much greater for *C. dux sensu* Ferreira than any true species (Figure 2.2). Although there were some individual replicates for true species that performed worse than the best of *C. dux sensu* Ferreira, its overall performance across all individual and ensemble models was lower than all true species in all other algorithms and across all metrics, with the clearest consistent difference using RF and the least clear using GLM (Supplementary Figure 2.2).

Table 2.2: Catharsius *ensemble model calibration scores across all three evaluation metrics*

Species	***KAPPA***	***AUC***	***TSS***
C. dominus	0.898	0.992	0.923
C. duciformis	0.933	0.997	0.966
C. dux sensu Takano	0.922	0.993	0.943
***C. dux sensu* Ferreira**	**0.874**	**0.975**	**0.838**
C. gorilla	0.954	0.997	0.934
C. satyrus	0.964	0.996	0.964
C. ulysses	0.939	0.998	0.951

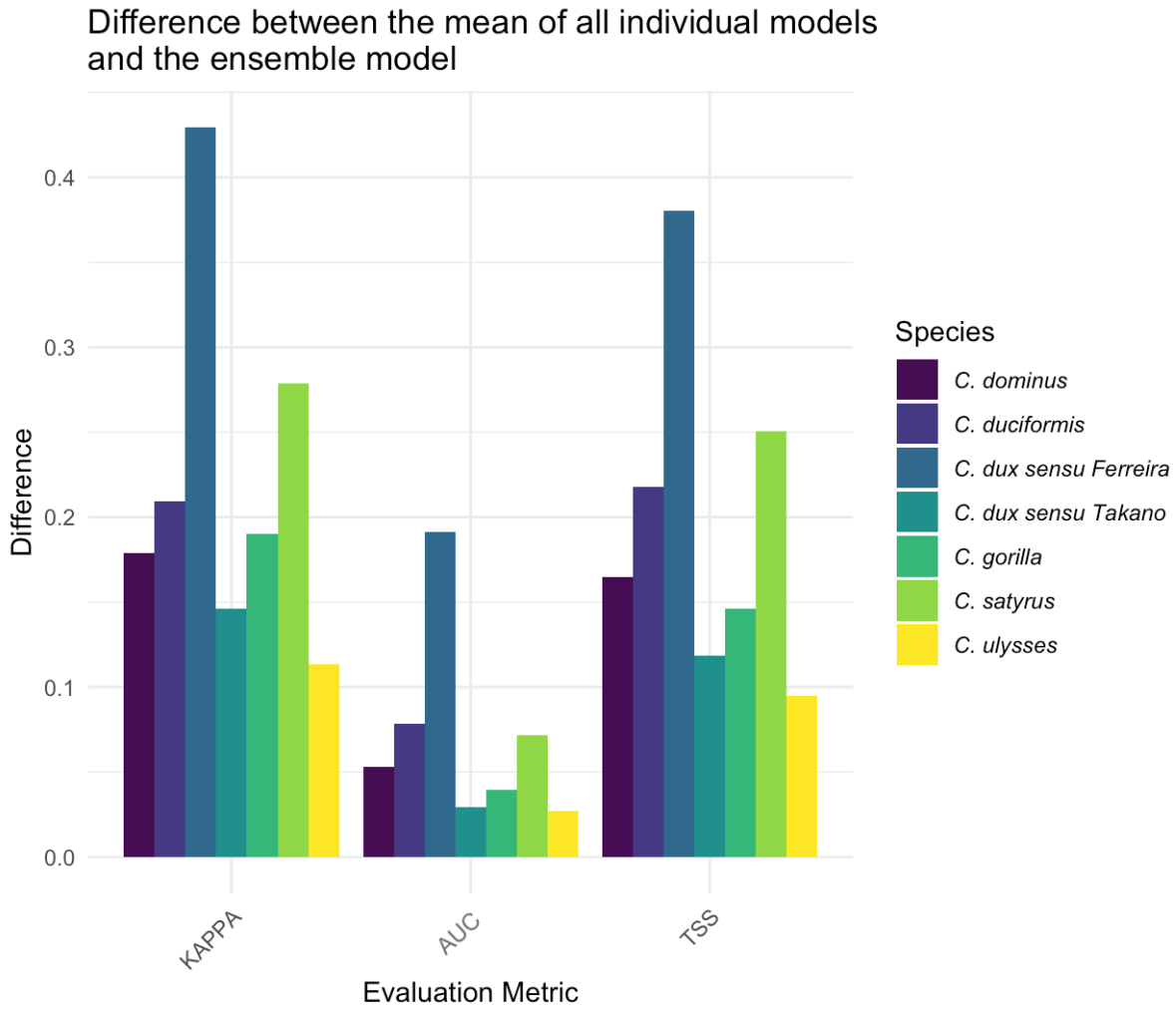


Figure 2.2: Difference between the mean validation scores of all 10 Catharsius *individual model replicates, for all modelling algorithms, and the ensemble model calibration scores. Note that the scales for metrics are not comparable to one another.*

Differences between species were significant for all evaluation metrics in all algorithms (all 18 Friedman tests were significant, see Supplementary Table 2.1). Out of 18 possible algorithm-metric combinations, half found differences between *C. dux sensu* Ferreira and all true species. There were significant differences between the AUC of *C. dux sensu* Ferreira and all true species for five out of six algorithms, excepting only the GLM. With TSS, this was the case in three out of six algorithms, and with KAPPA in just one. Machine learning algorithms consistently found more instances of significant differences between model performance for *C. dux sensu* Ferreira and the true species, and less significant differences between the true species, than the regression methods (Table 2.3). No true species had significant differences in model performance with all other true species with any modelling algorithm or evaluation metric (Supplementary Table 2.2).

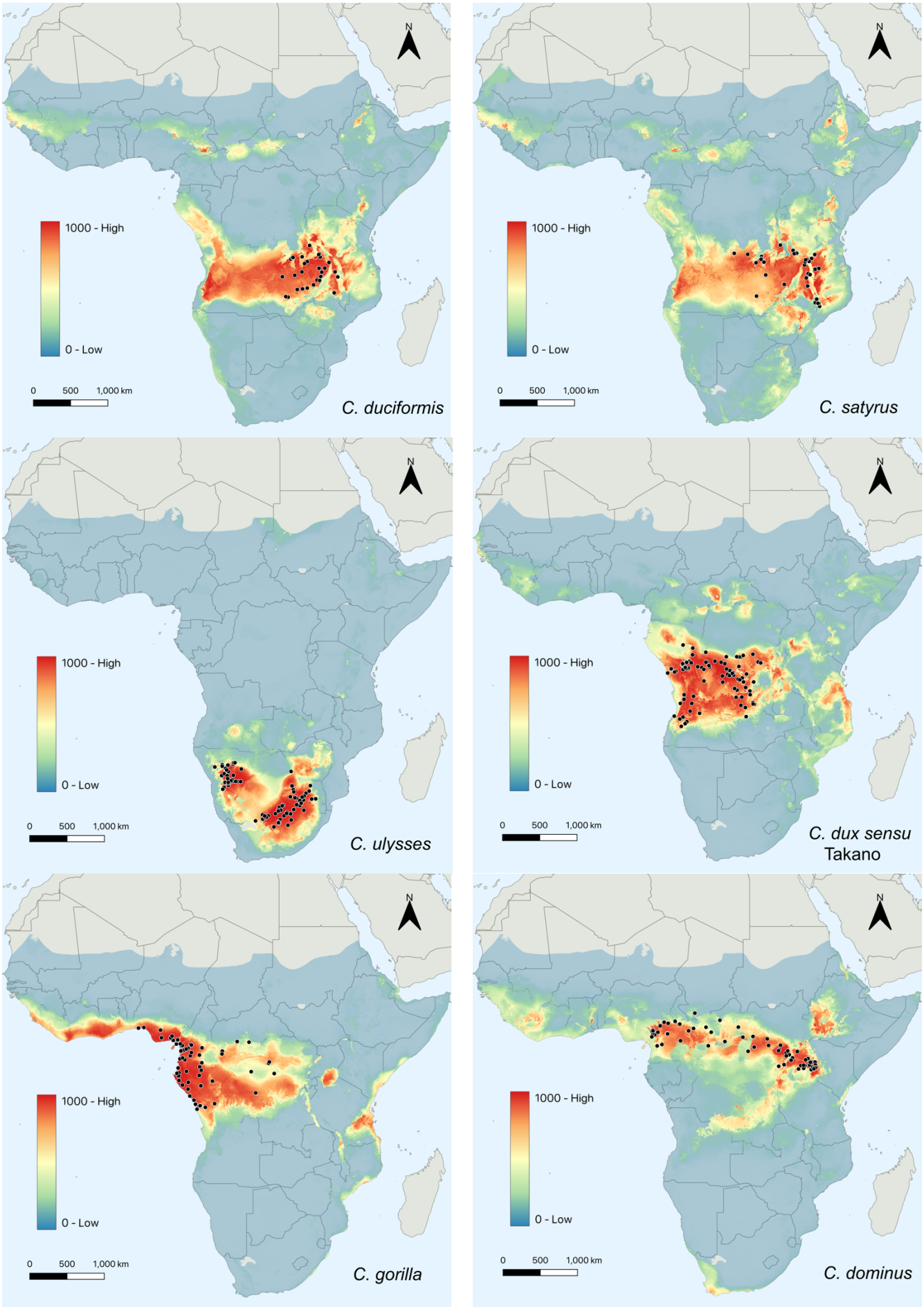
Ensemble model habitat suitability projections

Areas of habitat suitability across the Afrotropical realm vary notably between species, highlighting distinctive adaptations to their environments (Figures 2.3a–f). Generally, *C. ulysses* is adapted to the dry desert and grasslands of Southern Africa, *C. duciformis* and *C. satyrus* to tropical and subtropical savannah, *C. dux* similarly to tropical savannah but with higher tolerance of wetter conditions, *C. dominus* to forest-savannah mosaic areas, and C. *gorilla* to tropical forest. The projection for *C. dux sensu* Ferreira (Figure 2.4) shows an intermediate tolerance of these conditions, with the exception of Southern African desert and the most tropical rainforest, with large extents of averagely suitable habitat across them all.

Table 2.3: Summary of Conover test results detailing significant differences between Catharsius *model validation scores between species*

Evaluation metric	SDM algorithm	Type of SDM algorithm	Proportion of six true species significantly different from *C. dux sensu* Ferreira	Proportion of 21 true species comparisons that were significant
KAPPA	ANN	ML	1	0
KAPPA	GBM	ML	0.83	0.048
KAPPA	MAXNET	ML	0.67	0.048
KAPPA	RF	ML	0.67	0.095
KAPPA	GAM	Regression	0.83	0.19
KAPPA	GLM	Regression	0.5	0.38
AUC	ANN	ML	1	0.14
AUC	GBM	ML	1	0.048
AUC	MAXNET	ML	1	0.14
AUC	RF	ML	1	0.095
AUC	GAM	Regression	1	0.19
AUC	GLM	Regression	0.67	0.43
TSS	ANN	ML	1	0.095
TSS	GBM	ML	0.83	0
TSS	MAXNET	ML	1	0
TSS	RF	ML	1	0
TSS	GAM	Regression	0.67	0.14
TSS	GLM	Regression	0.67	0.29

In total, each raster projection of habitat suitability had 24,663,627 cells containing a suitability value between zero (low) and 1000 (high) suitability, and for all species, the majority were classified as low suitability when binned. Differences in the distributions of cells across suitability bins were significant between species ($\chi^2(18) = 8{,}984{,}232$, $p < 0.001$), and the strength of the association was small-to-moderate ($V = 0.131705$). Significance was driven by fewer low suitability cells and more average-low and average-high suitability cells than expected for *C. dux sensu* Ferreira (Figure 2.5). Whilst it did not have the fewest cells in the very high suitability category, with marginally more than *C. dominus* and *C. satyrus,* it did not generate any cells with values above 872, the lowest top value of all species (Supplementary Tables 2.3 and 2.4).



Figures 2.3a-f: projections of Catharsius *ensemble models* for C. duciformis, C. satyrus, C. ulysses, C. dux sensu *Takano,* C. gorilla, and C. dominus *in order of ensemble model TSS score. Areas of high suitability are shown in red, and areas of low suitability are shown in blue.*

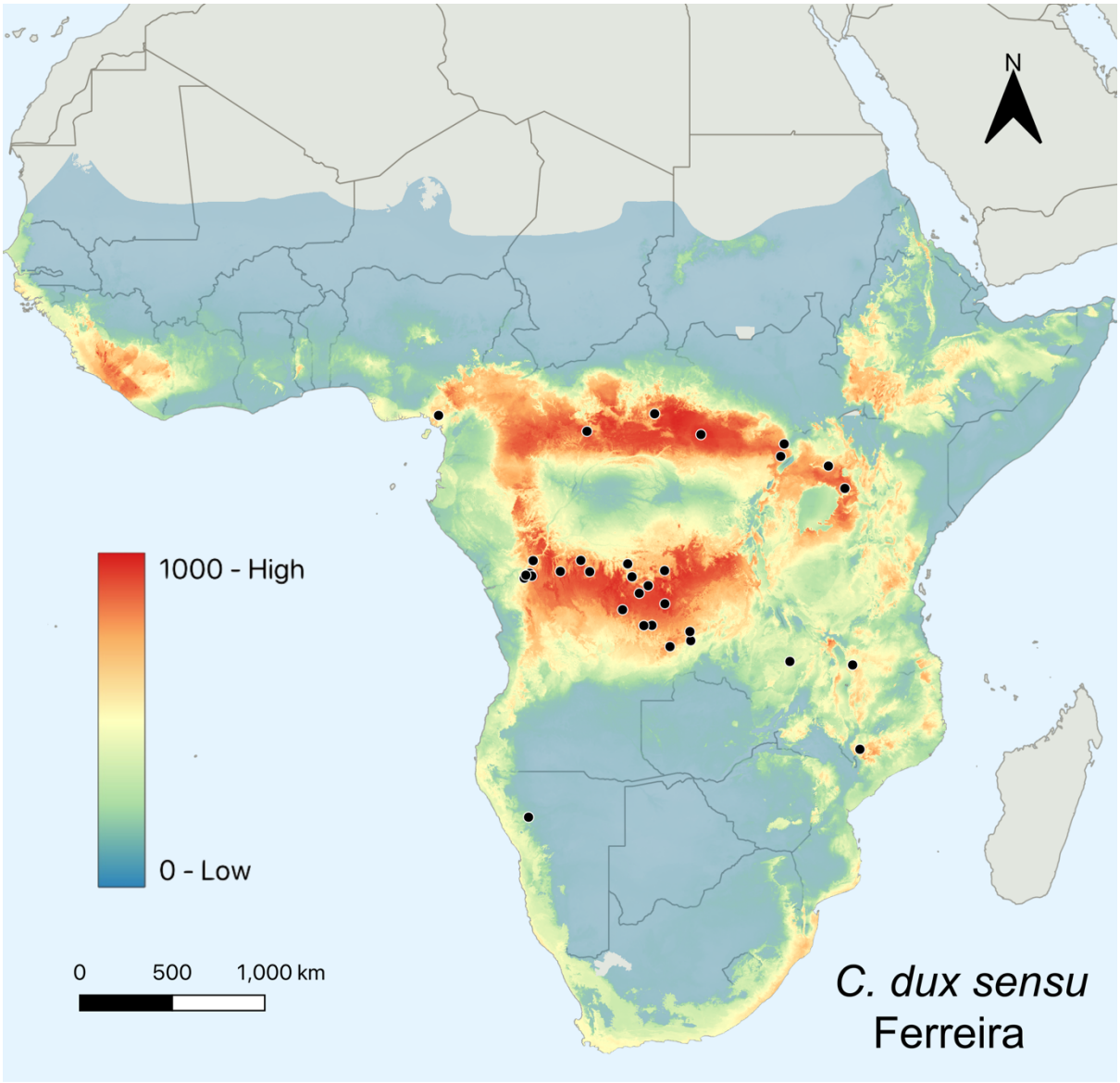


Figure 2.4: Catharsius *ensemble model projection for* C. dux sensu *Ferreira. Areas of high suitability are shown in red, and areas of low suitability are shown in blue.*

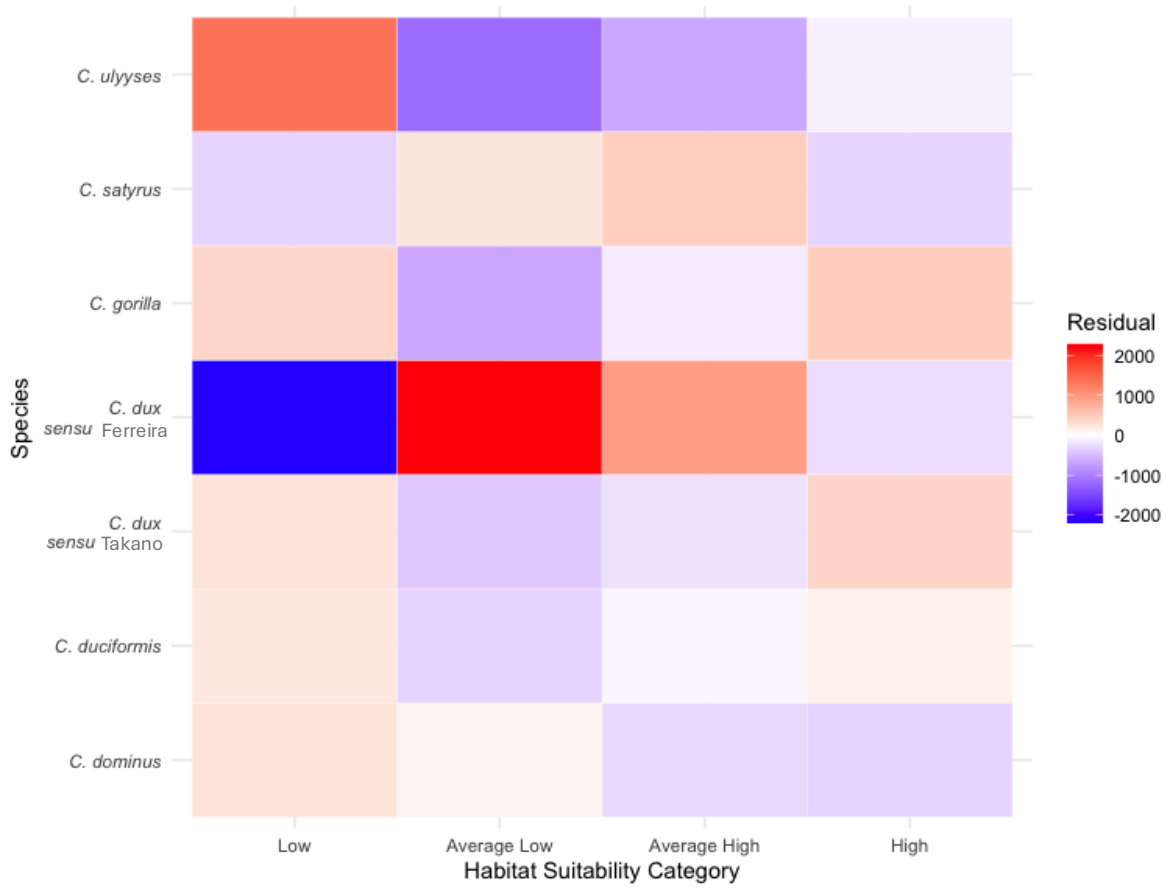


Figure 2.5: Standardised residuals from Chi-squared test comparing distribution of cells across suitability bins per Catharsius *species. Positive values indicate more cells in that bin than expected, and negative values indicate less cells in that bin than expected. Colour intensity indicates magnitude of difference.*

Discussion

This study underlines the importance of up-to-date taxonomy in biodiversity data. By quantifying the improvement to species distribution models achieved by revising the taxonomy of an Afrotropical genus of dung beetles, this approach assesses not only the extent to which taxonomic data quality issues pose a risk to modelling interpretation, but also cautions against unreserved use of ensemble modelling techniques without careful inspection of individual replicate results. In particular, it has shown that SDMs that use occurrence points identified according to unrevised taxonomy are less capable of identifying areas of high and low habitat suitability, which leads to poor understanding of the ecological niche of these species. Use of ensemble models improve model performance but, in doing so, may obscure data quality issues.

The positive impact of revising taxonomy is strongly supported throughout the results. The overall performance of *C. dux sensu* Ferreira models, which used occurrence points identified according to outdated taxonomic understanding, was consistently worse than that of the models informed by up-to-date taxonomy. According to both the ensemble models and the mean value of all individual model replicates, improved taxonomic data quality results in improved SDM fits. This is further supported by significant differences between *C. dux sensu* Ferreira models and those of every other species in half the evaluation metric-algorithm combinations, whilst no true species was significantly different from all others in any. Some of the worst individual model replicates had lower scores than the best of *C. dux sensu* Ferreira, but this is likely a result of chance combinations of occurrence points generated by the data partitioning

and does not negate other findings. Instead, it underscores the importance of running multiple model replicates (Sillero & Barbosa, 2021).

Projected habitat suitability across the Afrotropical realm demonstrated that poor model performance linked to taxonomic data quality issues affects our ability to accurately describe species climatic niches. The ensemble model using *C. dux sensu* Ferreira occurrence points was unable to distinguish areas of high and, in particular, low suitability, instead finding many more averagely suitable areas than expected, underlining its difficulty pinpointing where this species could survive. This is particularly notable in a taxon well-known for strong links to its environment across a large and climatically diverse landscape. Significance of the Chi-squared test used to compare the distribution of cells across suitability bins was noticeably inflated due to their huge number, but the small to moderate effect, as judged by the Cramér's V, and the clear pattern in the residuals support that the interpretation of poor taxonomy negatively impacting our ability to characterise species niches is a meaningful one. On top of this, the large number and extent of cells with average suitability across diverse climates means that interpretation of niche breadth using these occurrence points would be vastly overestimated. As tropical species with narrower niche widths are more vulnerable to climate change (Grinder & Wiens, 2023), outdated taxonomy then threatens not only our ability to describe a species' niche, but also to accurately assess its vulnerability. Niche breadth has also been linked to the likelihood of volant species shifting their ranges to track suitable climate as it changes (Hällfors et al., 2024), so its misinterpretation for a highly mobile group like *Catharsius* would likely also impact our ability to predict future range shifts.

Preference for vastly different environmental conditions is evident amongst the species as they are now known, and it is thus clear to see why the *C. dux sensu* Ferreira model, functioning much like an ensemble model of all species, is unable to identify any particularly strong relationship with climate. Models highlighted a wide variety of habitat specialisations amongst the true species, including suitability in desert, savannah, and tropical rainforest conditions. In prior studies, dung beetle distributions have been found to correlate with soil texture, water table level, elevation, and precipitation in South America (Salomão et al., 2022; Villamarin-Cortez et al., 2022), and annual rainfall, annual temperature, rainfall seasonality, and altitude in the most southerly countries of Africa (A. L. V. Davis et al., 1999; A. L. V. Davis & Scholtz, 2020). Generic richness of dung beetles on a global scale is most strongly correlated to tropical and warm summer rainfall climate types (A. L. V. Davis & Scholtz, 2001), demonstrated regionally by higher species richness in the mid-summer rainfall areas in the north east of South Africa (A. Davis, 1997). It is clear, then, that abiotic variables are globally important drivers of dung beetle distributions, and their sensitivity to environment is well understood. With this prior understanding, it is more straightforward to flag SDMs that find no particularly favoured climate as poorly resolved, but what of lesser known taxa? This study agrees with established findings that without establishing the quality of our taxonomic understanding first, we cannot be sure of any further biological interpretations (Agarwal et al., 2021; Barria et al., 2020; Ensing et al., 2013; Goudarzi et al., 2021; Mori et al., 2019).

The use of ensemble modelling approaches has been shown to improve SDM accuracy, likely due to balancing strengths and weaknesses of different algorithms within a single model (Buisson et al., 2010; Grenouillet et al., 2011; C. Liu et al., 2019; Marmion et al., 2009; Stohlgren et al., 2010), although this is not always the case (Hao et al., 2020). Here, ensemble model scores were higher than the mean of the individual replicate scores for all species because replicates that had performed poorly were filtered out and the remaining weighted according to performance during the ensemble model process. On the face of it, this agrees with prior findings. However, although the ensemble model for *C. dux sensu* Ferreira performed worse than those of the true species, it would still be classified in the top performance category for all three evaluation metrics. Despite using occurrence points that are misidentified according to current taxonomic understanding, this would not flag any data quality concerns if produced as a standalone model and, in fact, would obscure them. The difference between the mean of all individual replicates and the ensemble model was much greater for *C. dux sensu* Ferreira than for the true species, underlining that more replicates are being removed during filtering, but this is not made clear by the final ensemble model output. Furthermore, among the replicates that are retained, we cannot be sure whether this is because data partitioning for training and testing is selecting occurrences that align with up-to-date taxonomic understanding, or whether the occurrences selected align climatically by chance. Ensemble modelling, then, may improve external performance scores by masking data errors. It is of paramount importance that the performance of individual model replicates is manually inspected prior to the creation of an ensemble model to avoid missing potential data quality red flags.

The choice of modelling algorithm for studies of species distributions is both important and difficult. It has been repeatedly shown that there is no clear and consistent "best" method and instead should be decided according to the specific aims of the study and characteristics of the data (Aguirre-Gutiérrez et al., 2013; X. Li & Wang, 2013; Qiao et al., 2015; Wisz et al., 2008). Here, the GLM generated the least clear differences between the fit of *C. dux sensu* Ferreira models and those of the other species, whilst RF generated the clearest. Machine learning algorithms were generally more efficient at pinpointing significant differences between the fit of *C. dux sensu* Ferreira models and those of the true species whilst simultaneously not finding these between the true species. Degree of desired complexity is a key element of choosing an SDM algorithm, with overly simple methods more easily interpreted but potentially not capturing the intricacy of natural phenomena, and complex methods–like machine learning options–improving accuracy at the cost of interpretability, whilst also being prone to overfitting (Chollet Ramampiandra et al., 2023; Merow et al., 2014). This study suggests that the latter's ability to fit such complex relationships may aid in detection of taxonomic data quality issues, and with the integration of tools that explain what goes on in machine learning's 'black box' into species distribution modelling (Ryo et al., 2021), this may be a decreasingly costly choice for species with some degree of taxonomic uncertainty.

Much like modelling algorithm, choice of evaluation metric is not straightforward. Despite the frequency with which it is used, existing research has judged AUC to be limited in its capacity to evaluate models (J. Beck et al., 2014; Jiménez & Soberón, 2020; Jiménez-Valverde, 2012, 2014; Lobo et al., 2008). Here, it distinguishes between *C. dux*

sensu Ferreira and all other species in every algorithm but the GLM, but also found the most significance in comparisons between the true species, suggesting that its focus on ranking rather than ecological realism, allowing it to remain high when models fail to predict large areas of suitable habitat, may prevent it from helping to identify taxonomic errors in the occurrence data. Whilst the KAPPA statistic did not identify significant differences between *C. dux sensu* Ferreira and all other species in as many cases, it did find fewer significant differences between other species. However, its sensitivity to data prevalence has led to criticisms of its use in model validation (Jiménez-Valverde, 2012), and these results indicate that it cannot efficiently identify underlying taxonomic data quality issues. The TSS was most reliable when identifying significant differences according to taxonomic accuracy, especially with machine learning algorithms. This may be because, although relying on delimitation of a threshold, it is a prevalence-independent metric that balances sensitivity and specificity, allowing appropriate penalty when poor taxonomic data quality obscures true ecological signal (Allouche et al., 2006). Critically, these metrics were included to test where taxonomic data quality issues have a unanimously negative impact on model performance across commonly used assessments. Their specific ability to identify these underlying errors was not explicitly tested and should be explored in more detail in future work. Although the model for *C. dux sensu* Ferreira generated marginally more high suitability cells than two of the true species, it had the lowest maximum suitability value. As such, exploration of evaluation metrics should also investigate whether the range of suitability values generated in model projections could function as a less time and computationally intensive indicator of model quality.

Performance of models using up-to-date taxonomy still varied between species, and there may be a number of reasons for this. Firstly, occurrence points are well understood to represent the realised niche of a species – a subset of environmental conditions that facilitate existence, often restricted by biotic interactions, dispersal limitations, and disturbances (Elith & Franklin, 2013). Climate variables, as used here, are therefore not exhaustive in their explanatory power, and other drivers such as fire disturbance (R. B. de Andrade et al., 2011), human disturbance (A. J. Davis et al., 2001), and mammal community composition (Raine & Slade, 2019) have been shown to influence dung beetle dynamics, as well as the texture and humidity of the soil that may govern underground nest structures (Takano, 2018). It is also possible that the 1km resolution of the climate variables themselves masked local effects for this relatively small organism. However, as climate is often the dominant predictor of species distributions at extents of above 200km (R. G. Pearson & Dawson, 2003), this resolution was judged to be appropriate given the overall scale of the study. In either case, these have not impacted any one species more than another to such a degree that the effect of taxonomic accuracy on model outputs is obscured. However, further work on this genus may wish to investigate adaptation on a local scale through the use of landscape genomics. In particular, the ensemble model projection for *C. duciformis* indicates suitable habitat beyond the western edge of its current observed range, and the integration of genetic data into the study of its distribution may illuminate whether this is due to sampling, local adaptation, or something else entirely.

Over time, a lack of recognition of the contributions of taxonomists to downstream analyses has meant that taxonomy has declined in popularity. That said, there have

been calls for increased recognition, drawing attention to the difficulty publishing taxonomic works in high impact journals, and impact on citation counts from most taxonomy specific journals not being ISI indexed. They also underline the impact this has on grant applications and hiring chances, ultimately lessening students' exposure to the discipline (Hopkins & Freckleton, 2002; Hutchings, 2021; Anon. 1946; Lagomarsino & Frost, 2020; Wägele et al., 2011). In their support, the importance of up-to-date taxonomy in species distribution modelling is made clear throughout this study. In improving model accuracy and our ability to determine high and low suitability habitat, it enhances our ability to describe species niches and therefore predict future range shifts and assess vulnerability to climate change. Taxonomists have been encouraged to incorporate modern methods into their work and better integrate with other disciplines (Orr et al., 2020), but it is crucial that, in turn, ecologists and biodiversity scientists recognise their contribution. Of course, it is a field that is demanding of both time and financial resources–the data used in this study, for example, was the product of a multi-year project to revise a single genus– so further research into how taxonomic expertise can included efficiently in modelling protocols should be explored.

References

Agarwal, I., Ceríaco, L. M. P., Metallinou, M., Jackman, T. R., & Bauer, A. M. (2021). How the African house gecko (*Hemidactylus mabouia*) conquered the world. *Royal Society Open Science*, *8*(8), 210749. https://doi.org/10.1098/rsos.210749

Aguirre-Gutiérrez, J., Carvalheiro, L. G., Polce, C., van Loon, E. E., Raes, N., Reemer, M., & Biesmeijer, J. C. (2013). Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria – Dutch Hoverflies as a Case Study. *PLoS ONE*, *8*(5), e63708. https://doi.org/10.1371/journal.pone.0063708

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*(5), 541–545. https://doi.org/10.1111/ecog.01132

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*(6), 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Andrade, R. B. de, Barlow, J., Louzada, J., Vaz-de-Mello, F. Z., Souza, M., Silveira, J. M., & Cochrane, M. A. (2011). Quantifying Responses of Dung Beetles to Fire Disturbance in Tropical Forests: The Importance of Trapping Method and Seasonality. *PLoS ONE*, *6*(10), e26208. https://doi.org/10.1371/journal.pone.0026208

Anonymous (1946) Importance of Taxonomy. *Nature*, *158*, 105–106. https://doi.org/10.1038/158105b0

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Araujo, M. B., Pearson, R. G., Thuiller, W., & Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, *11*(9), 1504–1513. https://doi.org/10.1111/j.1365-2486.2005.01000.x

Austin, M. P., & Van Niel, K. P. (2011). Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography*, *38*(1), 1–8. https://doi.org/10.1111/j.1365-2699.2010.02416.x

Bacher, S. (2012). Still not enough taxonomists: Reply to Joppa et al. *Trends in Ecology and Evolution*, *27*(2), 65–66.

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in Ecology and Evolution*, *3*(2), 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Barria, A. M., Zamorano, D., Parada, A., Labra, F. A., Estay, S. A., & Bacigalupe, L. D. (2020). The Importance of Intraspecific Variation for Niche Differentiation and Species Distribution Models: The Ecologically Diverse Frog *Pleurodema thaul* as Study Case. *Evolutionary Biology*, *47*(3), 206–219. https://doi.org/10.1007/s11692-020-09510-0

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bradie, J., & Leung, B. (2017). A quantitative synthesis of the importance of variables used in MaxEnt species distribution models. *Journal of Biogeography*, *44*(6), 1344–1361. https://doi.org/10.1111/jbi.12894

Buisson, L., Thuiller, W., Casajus, N., Lek, S., & Grenouillet, G. (2010). Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, *16*(4), 1145–1157. https://doi.org/10.1111/j.1365-2486.2009.02000.x

Chollet Ramampiandra, E., Scheidegger, A., Wydler, J., & Schuwirth, N. (2023). A comparison of machine learning and statistical species distribution models: Quantifying overfitting supports model interpretation. *Ecological Modelling*, *481*, 110353. https://doi.org/10.1016/j.ecolmodel.2023.110353

Clarkson, E. N. K. (1998). *Invertebrate Palaeontology and Evolution* (Fourth Edition). Blackwell Science.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Lawrence Erlbaum Associates.

Daniel, G. M., Sole, C. L., Scholtz, C. H., & Davis, A. L. V. (2021). Historical diversification and biogeography of the endemic southern African dung beetle genus, *Epirinus* (Scarabaeidae: Scarabaeinae). *Biological Journal of the Linnean Society*, *133*(3), 751–765. https://doi.org/10.1093/biolinnean/blab051

Davis, A. (1997). Climatic and biogeographical associations of southern African dung beetles (Coleoptera: Scarabaeidae s. str.). *African Journal of Ecology*, *35*(1), 10–38. https://doi.org/10.1111/j.1365-2028.1997.051-89051.x

Davis, A. J., Holloway, J. D., Huijbregts, H., Krikken, J., Kirk-Spriggs, A. H., & Sutton, S. L. (2001). Dung beetles as indicators of change in the forests of northern Borneo. *Journal of Applied Ecology*, *38*(3), 593–616. https://doi.org/10.1046/j.1365-2664.2001.00619.x

Davis, A. L. V., & Dewhurst, C. F. (1993). Climatic and biogeographical associations of Kenyan and northern Tanzanian dung beetles (Coleoptera: Scarabaeidae). *African Journal of Ecology*, *31*(4), 290–305. https://doi.org/10.1111/j.1365-2028.1993.tb00543.x

Davis, A. L. V., & Scholtz, C. H. (2001). Historical vs. ecological factors influencing global patterns of Scarabaeine dung beetle diversity. *Diversity and Distributions*, *7*(4), 161–174. https://doi.org/10.1111/j.1472-4642.2001.00102.x

Davis, A. L. V., & Scholtz, C. H. (2020). Dung beetle conservation biogeography in southern Africa: Current challenges and potential effects of climatic change. *Biodiversity and Conservation*, *29*(3), 667–693. https://doi.org/10.1007/s10531-019-01904-7

Davis, A. L. V., Scholtz, C. H., & Chown, S. L. (1999). Species turnover, community boundaries and biogeographical composition of dung beetle assemblages across an altitudinal gradient in South Africa. *Journal of Biogeography*, *26*(5), 1039–1055. https://doi.org/10.1046/j.1365-2699.1999.00335.x

Davis, A. L. V., Scholtz, C. H., Dooley, P. W., Bham, N., & Kryger, U. (2004). Scarabaeine dung beetles as indicators of biodiversity, habitat transformation and pest control chemicals in agro-ecosystems. *South African Journal of Science*.

Dawson, M. N., Correia, R. A., & Ladle, R. J. (2023). Five decades of biogeography: A view from the *Journal of Biogeography*. *Journal of Biogeography*, *50*(1), 1–7. https://doi.org/10.1111/jbi.14559

Elith, J., & Franklin, J. (2013). Species Distribution Modeling. In S. A. Levin (Ed.), *Encyclopedia of Biodiversity* (2nd edn, Vol. 6, pp. 692–705). Elsevier. https://doi.org/10.1016/B978-0-12-384719-5.00318-X

Elith, J., H. Graham*, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., Li, J., G. Lohmann, L., A. Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. M. Overton, J., Townsend Peterson, A., … E. Zimmermann, N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Ensing, D. J., Moffat, C. E., & Pither, J. (2013). Taxonomic identification errors generate misleading ecological niche model predictions of an invasive hawkweed. *Botany*, *91*(3), 137–147. https://doi.org/10.1139/cjb-2012-0205

Escobar, F., Halffter, G., & Arellano, L. (2007). From forest to pasture: An evaluation of the influence of environment and biogeography on the structure of beetle (Scarabaeinae) assemblages along three altitudinal gradients in the Neotropical region. *Ecography*, *30*(2), 193–208. https://doi.org/10.1111/j.0906-7590.2007.04818.x

Ferreira, M. C. (1960a). Descricao de especies novas de *Catharsius* s.str. *Novos Taxa Entomológicos*, *23*, 3–8.

Ferreira, M. C. (1960b). Revisao das esepcies Africanas de *Catharsius* s.str. Do Groupo Adamastor e descricao de especies novas. *Revista Entomologia Moçambique*, *3*(1), 1–73.

Ferreira, M. C. (1972). Os Escarabideos de Africa (sul do Saara). I. *Revista de Entomologia de Mocambique*, *11*, 5–1088.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (Third). Sage. https://www.john-fox.ca/Companion/

Franklin, J. (2023). Species distribution modelling supports the study of past, present and future biogeographies. *Journal of Biogeography*, *50*, 1533–1545. https://doi.org/10.1111/jbi.14617

Franklin, J., Serra-Diaz, J. M., Syphard, A. D., & Regan, H. M. (2017). Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography*, *26*(1), 6–17. https://doi.org/10.1111/geb.12501

Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Goudarzi, F., Hemami, M.-R., Malekian, M., Fakheran, S., & Martínez-Freiría, F. (2021). Species versus within-species niches: A multi-modelling approach to assess range size of a spring-dwelling amphibian. *Scientific Reports*, *11*(1), 597. https://doi.org/10.1038/s41598-020-79783-0

Gravetter, F. J., & Wallnau, L. B. (2012). *Essentials of Statistics for the Behavioral Sciences, 8th ed*. (8th edn). Cengage Learning.

Grenouillet, G., Buisson, L., Casajus, N., & Lek, S. (2011). Ensemble modelling of species distribution: The effects of geographical and environmental ranges. *Ecography*, *34*(1), 9–17. https://doi.org/10.1111/j.1600-0587.2010.06152.x

Grinder, R. M., & Wiens, J. J. (2023). Niche width predicts extinction from climate change and vulnerability of tropical species. *Global Change Biology*, *29*(3), 618–630. https://doi.org/10.1111/gcb.16486

Guéguen, M., Blancheteau, H., & Thuiller, W. (n.d.). *Pseudo-absences*. Biomod2. Retrieved 14 June 2024, from https://biomodhub.github.io/biomod2/articles/vignette_pseudoAbsences.html

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9

Halffter, G. (1991). Historical and ecological factors determining the geographical distribution of beetles (Coleoptera: Scarabaeidae: Scarabaeinae). *Biogeographia*

– *The Journal of Integrative Biogeography*, *15*. https://doi.org/10.21426/B615110376

Hällfors, M. H., Heikkinen, R. K., Kuussaari, M., Lehikoinen, A., Luoto, M., Pöyry, J., Virkkala, R., Saastamoinen, M., & Kujala, H. (2024). Recent range shifts of moths, butterflies, and birds are driven by the breadth of their climatic niche. *Evolution Letters*, *8*(1), 89–100. https://doi.org/10.1093/evlett/qrad004

Hao, T., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2020). Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. *Ecography*, *43*(4), 549–558. https://doi.org/10.1111/ecog.04890

He, K. S., Bradley, B. A., Cord, A. F., Rocchini, D., Tuanmu, M., Schmidtlein, S., Turner, W., Wegmann, M., & Pettorelli, N. (2015). Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, *1*(1), 4–18. https://doi.org/10.1002/rse2.7

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, *118*(6), e2018093118. https://doi.org/10.1073/pnas.2018093118

Herzog, S. K., Hamel-Leigue, A. C., Larsen, T. H., Mann, D. J., Soria-Auza, R. W., Gill, B. D., Edmonds, W. D., & Spector, S. (2013). Elevational Distribution and Conservation Biogeography of Phanaeine Dung Beetles (Coleoptera: Scarabaeinae) in Bolivia. *PLoS ONE*, *8*(5), e64963. https://doi.org/10.1371/journal.pone.0064963

Hewavithana, D. K., Wijesinghe, M. R., Dangalle, C. D., & Dharmarathne, H. A. S. G. (2016). Habitat and dung preferences of scarab beetles of the subfamily Scarabaeinae: A case study in a tropical monsoon forest in Sri Lanka. *International Journal of Tropical Insect Science*, *36*(02), 97–105. https://doi.org/10.1017/S1742758416000023

Hochkirch, A., Casino, A., Lyubomir, P., Allen, D., Tilley, L., Georgiev, T., Gospodinov, K., & Barov, B. (2022). *European Red List of Insect Taxonomists*. Publication Office of the European Union.

Hopkins, G. W., & Freckleton, R. P. (2002). Declines in the numbers of amateur and professional taxonomists: Implications for conservation. *Animal Conservation*, *5*(3), 245–249. https://doi.org/10.1017/S1367943002002299

Hutchings, P. (2021). Potential loss of biodiversity and the critical importance of taxonomy—An Australian perspective. In *Advances in Marine Biology* (Vol. 88, pp. 3–16). Elsevier. https://doi.org/10.1016/S0065-2881(21)00015-8

Jiménez, L., & Soberón, J. (2020). Leaving the area under the receiving operating characteristic curve behind: An evaluation method for species distribution modelling applications based on presence-only data. *Methods in Ecology and Evolution*, *11*(12), 1571–1586. https://doi.org/10.1111/2041-210X.13479

Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling: Insights into the AUC. *Global Ecology and Biogeography*, *21*(4), 498–507. https://doi.org/10.1111/j.1466-8238.2011.00683.x

Jiménez-Valverde, A. (2014). Threshold-dependence as a desirable attribute for discrimination assessment: Implications for the evaluation of species distribution models. *Biodiversity and Conservation*, *23*(2), 369–385. https://doi.org/10.1007/s10531-013-0606-1

Jos, F. (2012). Composition and Distribution Patterns of Species at a Global Biogeographic Region Scale: Biogeography of Aphodiini Dung Beetles (Coleoptera, Scarabaeidae) Based on Species Geographic and Taxonomic Data. In L. Stevens (Ed.), *Global Advances in Biogeography*. InTech. https://doi.org/10.5772/31314

Journal of Biogeography. (2022). *Journal of Biogeography 50th Anniversary Top 50 Cited Papers*. https://onlinelibrary.wiley.com/doi/toc/10.1111/(ISSN)1365-2699.top-50-cited-papers

Lagomarsino, L. P., & Frost, L. A. (2020). The Central Role of Taxonomy in the Study of Neotropical Biodiversity. *Annals of the Missouri Botanical Garden*, *105*(3), 405–421. https://doi.org/10.3417/2020601

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). Virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*(6), 599–607. https://doi.org/10.1111/ecog.01388

Li, X., & Wang, Y. (2013). Applying various algorithms for species distribution modelling. *Integrative Zoology*, *8*(2), 124–135. https://doi.org/10.1111/1749-4877.12000

Liu, C., Newell, G., & White, M. (2019). The effect of sample size on the accuracy of species distribution models: Considering both presences and pseudo-absences or background sites. *Ecography*, *42*(3), 535–548. https://doi.org/10.1111/ecog.03188

Liu, C., White, M., & Newell, G. (2011). Measuring and comparing the accuracy of species distribution models with presence-absence data. *Ecography*, *34*(2), 232–243. https://doi.org/10.1111/j.1600-0587.2010.06354.x

Lobo, J. M., Jiménez-valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lozier, J. D., Aniello, P., & Hickerson, M. J. (2009). Predicting the distribution of Sasquatch in western North America: Anything goes with ecological niche modelling. *Journal of Biogeography*, *36*(9), 1623–1627. https://doi.org/10.1111/j.1365-2699.2009.02152.x

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, *15*(1), 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x

McGeoch, M. A., Van Rensburg, B. J., & Botes, A. (2002). The verification and application of bioindicators: A case study of dung beetles in a savanna ecosystem. In *Journal of Applied Ecology* (Vol. 39, pp. 661–672).

Merow, C., Smith, M. J., Edwards, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., & Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography*, *37*(12), 1267–1281. https://doi.org/10.1111/ecog.00845

Meyer, D., Zeileis, A., & Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-Way Contingency Tables with vcd. *Journal of Statistical Software*, *17*(3), 1–48. https://doi.org/10.18637/jss.v017.i03

Meyer, D., Zeileis, A., Hornik, K., & Friendly, M. (2024). *vcd: Visualizing Categorical Data*. https://doi.org/10.32614/CRAN.package.vcd

Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: Are we paying attention to recommendations? A systematic review. *PeerJ*, *5*, e3093. https://doi.org/10.7717/peerj.3093

Moreno-Amat, E., Mateo, R. G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J.-C., & García-Amorena, I. (2015). Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, *312*, 308–317. https://doi.org/10.1016/j.ecolmodel.2015.05.035

Mori, E., Menchetti, M., Zozzoli, R., & Milanesi, P. (2019). The importance of taxonomy in species distribution models at a global scale: The case of an overlooked alien squirrel facing taxonomic revision. *Journal of Zoology*, *307*(1), 43–52. https://doi.org/10.1111/jzo.12616

Moudrý, V., Bazzichetto, M., Remelgado, R., Devillers, R., Lenoir, J., Mateo, R. G., Lembrechts, J. J., Sillero, N., Lecours, V., Cord, A. F., Barták, V., Balej, P., Rocchini, D., Torresani, M., Arenas-Castro, S., Man, M., Prajzlerová, D., Gdulová, K., Prošek, J., … Šímová, P. (2024). Optimising occurrence data in species distribution models: Sample size, positional uncertainty, and sampling bias matter. *Ecography*, *2024*(12), e07294. https://doi.org/10.1111/ecog.07294

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth. *BioScience*, *51*(11), 933. https://doi.org/10.1641/0006-3568(2001)051%255B0933:TEOTWA%255D2.0.CO;2

Orr, M. C. C., Ascher, J. S., Bai, M., Chesters, D., & Zhu, C.-D. (2020). Three questions: How can taxonomists survive and thrive worldwide? *Megataxa*, *1*(1). https://doi.org/10.11646/megataxa.1.1.4

Pearson, D. L., Hamilton, A. L., & Erwin, T. L. (2011). Recovery Plan for the Endangered Taxonomy Profession. *BioScience*, *61*(1), 58–63. https://doi.org/10.1525/bio.2011.61.1.11

Pearson, R. G., & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? *Global Ecology and Biogeography*, *12*(5), 361–371. https://doi.org/10.1046/j.1466-822X.2003.00042.x

Pearson, R. G., Raxworthy, C. J., Nakamura, M., & Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *Journal of Biogeography*, *34*(1), 102–117. https://doi.org/10.1111/j.1365-2699.2006.01594.x

Pearson, R. G., Thuiller, W., Araújo, M. B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P., & Lees, D. C. (2006). Model-based uncertainty in species range prediction. *Journal of Biogeography*, *33*(10), 1704–1711. https://doi.org/10.1111/j.1365-2699.2006.01460.x

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

QGIS. (2021). *QGIS 3.22.9-Białowieża* (Version 3.22.9-Białowieża) [Computer software].

Qiao, H., Soberón, J., & Peterson, A. T. (2015). No silver bullets in correlative ecological niche modelling: Insights from testing among many potential algorithms for niche estimation. *Methods in Ecology and Evolution*, *6*(10), 1126–1136. https://doi.org/10.1111/2041-210X.12397

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643. https://doi.org/10.1111/jbi.12227

Raine, E. H., & Slade, E. M. (2019). Dung beetle-mammal associations: Methods, research trends and future directions. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1897). https://doi.org/10.1098/rspb.2018.2002

Randin, C. F., Dirnböck, T., Dullinger, S., Zimmermann, N. E., Zappa, M., & Guisan, A. (2006). Are niche-based species distribution models transferable in space?

Journal of Biogeography, *33*(10), 1689–1703. https://doi.org/10.1111/j.1365-2699.2006.01466.x

Richards, A. (2015). *University of Oxford Advanced Research Computing*. http://dx.doi.org/10.5281/zenodo.22558

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199–205. https://doi.org/10.1111/ecog.05360

Salomão, R. P., Pires, D. de A., Baccaro, F. B., Schietti, J., Vaz-de-Mello, F. Z., Lima, A. P., & Magnusson, W. E. (2022). Water table level and soil texture are important drivers of dung beetle diversity in Amazonian lowland forests. *Applied Soil Ecology*, *170*, 104260. https://doi.org/10.1016/j.apsoil.2021.104260

Scudder, G. G. E. (2017). The Importance of Insects. In *Insect Biodiversity: Science and Society* (Second Edition, Vol. 1, pp. 9–46). John Wiley & Sons.

Segurado, P., & Araújo, M. B. (2004). An evaluation of methods for modelling species distributions: Methods for modelling species distributions. *Journal of Biogeography*, *31*(10), 1555–1568. https://doi.org/10.1111/j.1365-2699.2004.01076.x

Shahabuddin, Hasanah, U., & Elijonnahdi. (2014). Effectiveness of dung beetles as bioindicators of environmental changes in land-use gradient in Sulawesi, Indonesia. *Biotropia*, *21*(1), 53–63. https://doi.org/10.11598/btb.2014.21.1.5

Sillero, N., & Barbosa, A. M. (2021). Common mistakes in ecological niche models. *International Journal of Geographical Information Science*, *35*(2), 213–226. https://doi.org/10.1080/13658816.2020.1798968

Smith, A. B., & Santos, M. J. (2020). Testing the ability of species distribution models to infer variable importance. *Ecography*, *43*(12), 1801–1813. https://doi.org/10.1111/ecog.05317

Soberón, J., Arriaga, L., & Lara, L. (2002). Issues of quality control in large, mixed-origin entomological databases. In H. Saarenmaa & E. S. Nielsen (Eds), *Towards a global biological information infrastructure.pdf* (pp. 15–22). European Environment Agency. https://www.eea.europa.eu/publications/technical_report_2001_70

Spector, S. (2006). Scarabaeine dung beetles (Coleoptera: Scarabaeidae: Scarabaeinae): An invertebrate focal taxon for biodiversity research and conservation. *The Coleopterists Bulletin*, *60*, 71–83.

Stohlgren, T. J., Ma, P., Kumar, S., Rocca, M., Morisette, J. T., Jarnevich, C. S., & Benson, N. (2010). Ensemble Habitat Mapping of Invasive Plant Species. *Risk Analysis*, *30*(2), 224–235. https://doi.org/10.1111/j.1539-6924.2009.01343.x

Tahseen, Q. (2014). Taxonomy-The Crucial yet Misunderstood and Disregarded Tool for Studying Biodiversity. *Journal of Biodiversity & Endangered Species*, *02*(03). https://doi.org/10.4172/2332-2543.1000128

Takano, H. (2018). *A systematic revision of the Afrotropical members of the dung beetle genus* Catharsius *Hope, 1837 (Coleoptera: Scarabaeidae)* [DPhil Thesis]. University of Oxford.

Thuiller, W., Georges, D., Gueguen, M., Engler, R., Breiner, F., Lafourcade, B., Patin, R., & Blancheteau, H. (2024). *biomod2: Ensemble Platform for Species Distribution Modeling*. https://doi.org/10.32614/CRAN.package.biomod2

van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, *39*(6), 542–552. https://doi.org/10.1111/ecog.01509

Veloz, S. D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *Journal of Biogeography*, *36*(12), 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x

Villamarin-Cortez, S., Hankin, L., Coronado, S., Macdonald, J., & Noriega, J. A. (2022). Diversity and distribution patterns of Ecuador's dung beetles (Coleoptera: Scarabaeinae). *Frontiers in Ecology and Evolution*, *10*, 1008477. https://doi.org/10.3389/fevo.2022.1008477

Wägele, H., Klussmann-Kolb, A., Kuhlmann, M., Haszprunar, G., Lindberg, D., Koch, A., & Wägele, J. W. (2011). The taxonomist—An endangered race. A practical proposal for its survival. *Frontiers in Zoology*, *8*(1), 25. https://doi.org/10.1186/1742-9994-8-25

Warren, D. L., Wright, A. N., Seifert, S. N., & Shaffer, H. B. (2014). Incorporating model complexity and spatial sampling bias into ecological niche models of climate

change risks faced by 90 California vertebrate species of concern. *Diversity and Distributions*, *20*(3), 334–343. https://doi.org/10.1111/ddi.12160

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., Elith, J., Dudík, M., Ferrier, S., Huettmann, F., Leathwick, J. R., Lehmann, A., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. C., … Zimmermann, N. E. (2008). Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, *14*(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

World Wildlife Fund. (2012). *Publications: Terrestrial Ecoregions of the World*. World Wildlife Fund. https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world

WorldClim. (2020, 2022). *Bioclimatic variables*. https://www.worldclim.org/data/bioclim.html

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M., & Karger, D. N. (2020). Macroecology in the age of Big Data – Where to go from here? *Journal of Biogeography*, *47*(1), 1–12. https://doi.org/10.1111/jbi.13633

CHAPTER 3

Mobilisation of Data From Natural History Collections Can Increase the Quality and Coverage of Biodiversity Information

***Ecology and Evolution* 15(4), e71139**

Bryony Blades[1,2], Cristina Ronquillo[3], Joaquín Hortal[3]

[1]Department of Biology, University of Oxford, Oxford, UK
[2]African Natural History Research Trust, Kingsland, Herefordshire, UK
[3]Department of Biogeography and Global Change, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain

Ecology and Evolution

WILEY

Ecology and Evolution Open Access

RESEARCH ARTICLE OPEN ACCESS

Mobilisation of Data From Natural History Collections Can Increase the Quality and Coverage of Biodiversity Information

Bryony Blades[1,2] | Cristina Ronquillo[3] | Joaquín Hortal[3]

[1]Department of Biology, University of Oxford, Oxford, UK | [2]African Natural History Research Trust (ANHRT), Kingsland, Herefordshire, UK | [3]Department of Biogeography & Global Change, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain

Correspondence: Bryony Blades (bryony.blades@biology.ox.ac.uk)

Received: 5 February 2025 | **Revised:** 4 March 2025 | **Accepted:** 5 March 2025

Funding: This work was supported by the Department of Biology, University of Oxford; the African Natural History Research Trust Scholarship; MCIN/AEI/10.13039/501100011033/FEDER, EU; Spanish AEI projects NICED (PID2022-140985NB-C21) and SCENIC (PID2019-106840GB-C21).

Keywords: biodiversity | data coverage | digitisation | GBIF | mobilisation | natural history collections

ABSTRACT

The surge of biodiversity data availability in recent decades has allowed researchers to ask questions on previously unthinkable scales, but knowledge gaps still remain. In this study, we aim to quantify potential gains to insect data on the Global Biodiversity Information Facility (GBIF) through further digitisation of natural history collections, assess to what degree this would fill biases in spatial and environmental record coverage, and deepen understanding of environmental bias with regard to climate rarity. Using mainland Afrotropical records for *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidae), we compared inventory completeness of GBIF data to a dataset which combined these with records from a recent taxonomic revision. We analysed how this improved dataset reduced regional and environmental bias in the distribution of occurrence records using an approach that identifies well-surveyed spatial units of 100 × 100km as well as emerging techniques to classify rarity of climates. We found that the number of cells for which inventory completeness could be calculated, as well as coverage of climate types by 'well-sampled' cells, increased threefold when using the combined set compared to the GBIF set. Improvements to sampling in Central and Western Africa were particularly striking, and coverage of rare climates was similarly improved, as not a single well-sampled cell from the GBIF data alone occurred in the rarest climate types. These findings support existing literature that suggests data gaps on GBIF are still pervasive, especially for insects and in the tropics, and so, is not yet ready to serve as a standalone data source for all taxa. However, we show that natural history collections hold the necessary information to fill many of these gaps, and their further digitisation should be a priority.

1 | Introduction

In recent decades, the concurrent intensification of technological advancement and the coupled climate and biodiversity crisis has given rise to massive mobilisation of biodiversity 'big data' (Newbold 2010; Wüest et al. 2020). This surge in data availability has allowed researchers to ask questions on previously impossible scales in fields such as conservation, biodiversity informatics, macroecology, disease biology and taxonomy (Heberling et al. 2021). Understanding of species' distributions and methods with which to model them have particularly benefitted, in turn advancing knowledge of evolutionary processes and conservation management (Acevedo et al. 2016; Elith and Franklin 2013; Guisan and Thuiller 2005; Soberón and Peterson 2004).

Despite the scale of these advancements, our understanding of species' distributions is still incomplete. Labelled the 'Wallacean shortfall' (Lomolino 2004), knowledge gaps arising from uneven sampling effort result in biases in spatial, temporal, climatic and taxonomic coverage of records (Collen et al. 2008; Hortal et al. 2015; Oliver et al. 2021; Sporbert et al. 2019; Troudet et al. 2017). Despite their richness, data deficits are particularly pervasive in the tropics, driven by tough environmental conditions, limited accessibility, poor infrastructure or security, and a lack of local capacity such as academic institutions or funds (Amano and Sutherland 2013; Araujo et al. 2022; Rocha-Ortega et al. 2021; Siddig 2019; Yesson et al. 2007). Such record unevenness in space can also result in their biased distribution across environmental gradients, affecting our ability to characterise the real breadth of species' fundamental niches (Hortal et al. 2008)—although this is not always the case (Newbold 2010). This hampers the reliability of model predictions, potentially misinforming applications in ecology and conservation (Troia and McManamay 2016).

Deficits and spatial biases are especially critical for invertebrates (Rocha-Ortega et al. 2021; Troudet et al. 2017), the importance of which to global biodiversity and ecosystem services cannot be overstated (Noriega et al. 2018; Wagner et al. 2021). This has led to particularly low inventory completeness—how comprehensively biodiversity has been surveyed and recorded—even in scenarios where records are numerous (García-Rosello et al. 2023; Iannella et al. 2019; Sánchez-Fernández et al. 2021), and that an unknown number may have already gone extinct (García-Rosello et al. 2023) paints a concerning picture for biodiversity moving forward. Although the strength of the Global Biodiversity Information Facility (GBIF)—the world's largest online biodiversity information network—lies in its compilation of data from myriad sources, these gaps remain, and not yet mobilised (digitised and made widely available through an online database) data sources such as taxonomic bibliographies and natural history collections can still provide novel insights (Beck et al. 2013; Shirey et al. 2019). As such, efforts to investigate spatial and environmental bias in insect sampling generally compile exhaustive databases from a number of sources (Ballesteros-Mejia et al. 2013; Romo et al. 2006; Sánchez-Fernández et al. 2008, 2022; Shirey et al. 2021), and less is known about insect inventories derived from GBIF data alone (see García-Rosello et al. 2023; Girardello et al. 2019; Rocha-Ortega et al. 2021; Troia and McManamay 2016). Additionally, whether inventory completeness is biased towards climate conditions that occur frequently is seldom analysed (but, see Ronquillo et al. 2020; Sobral-Souza et al. 2021), despite evidence that GBIF data are lacking in rare, locally restricted taxa (Beck et al. 2013).

Given the prevalence of correlative species distribution modelling, a technique that assumes comprehensive sampling of a species' fundamental niche using GBIF data (Heberling et al. 2021), it is problematic that research to date has not yet fully determined how significantly the completeness of its insect inventories is biased in climatic space. The strength of GBIF, though, lies in its role as a centralised aggregator, and its capacity as a collaborative platform can be leveraged to reduce this bias. Intensified digitisation and mobilisation of natural history collections have begun to help, especially with contributions from smaller herbaria and institutions, as well as private collections (Araujo et al. 2022; Beck et al. 2013; Yesson et al. 2007), but much remains to be done (Hardy et al. 2023; Popov et al. 2021). Fortunately, data compilation for purposes such as taxonomic revision often utilises collections that have heretofore not been mobilised, allowing an examination of how additional data could improve inventory completeness across space and climate and, here, for a severely data-deficient clade.

In this study, spatial and environmental bias in GBIF insect inventories is evaluated using data on dung beetles, which are known to act as a proxy for general biodiversity (Spector 2006). We determine how significantly bias could be reduced through the integration into GBIF of a dataset independently compiled for a taxonomic revision of the Afrotropical members of the genus *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidae) (Takano 2025). In particular, we investigate inventory completeness and how evenly this is distributed across climatic conditions and rarity.

2 | Materials and Methods

2.1 | Study Area and Taxon

The study area is the African mainland of the Afrotropical biogeographical realm spanning 17.5°W–51°E, 21°N–35°S. The study region exhibits a wide breadth of climatic conditions, including eight biomes and 91 ecoregions, encompassing tropical forest to xeric shrublands. Using QGIS version 3.22.9-Białowieża (QGIS 2021), a shapefile of the realm (Olson et al. 2001; World Wildlife Fund 2012) was modified to include only the mainland of Africa, categorising countries by region using the UN M49 standard (United Nations Statistics Division n.d.) as a reference (see Appendix S1). This shapefile was then used to create a grid of 100 km × 100 km cells for the statistical analyses, a resolution chosen given the scale of the study region and general sparsity of GBIF *Catharsius* records.

Catharsius is a genus of large copro- and necrophagous dung beetles with species distributed across both Africa and Asia. The Afrotropical members, including *Catharsius dux* (Figure 1), are currently being revised in the largest ever revision of any group of dung beetles in the world, for which an extensive collection of distributional data has been compiled from natural history collections (Takano 2025). It is these records that have been collated (Blades and Takano 2024) and are described below.

2.2 | Occurrence Data

To demonstrate the potential value of integrating unmobilised natural history collection data to GBIF, two datasets were used. First, all occurrences for *Catharsius* were downloaded from GBIF (2023) and subjected to a preprocessing procedure to ensure they did not fall foul of known GBIF data quality issues, as follows. Occurrences were filtered to include only those identified to the species level, and with precise and accurate coordinate information that did not suggest specimens were located in biodiversity institutions, the sea, in the centroid of a country, or in countries other than those listed on the physical label. The taxonomy of the

FIGURE 1 | An Afrotropical member of *Catharsius, C. dux*, in NW Zambia (2022).

resulting occurrences was standardised according to the most recent revision (Takano 2025) by removing occurrences that were identified as a non-African species and those that had been placed into *incertae sedis*, and correcting identifications to reflect new synonyms and homonyms. Further specifics of these processes can be found in Appendix S2. The geographic distribution of the resulting occurrences was cropped to the same extent as the bioclimatic variables, and this is henceforth referred to as the 'GBIF set'.

Second, a combined dataset was created by joining the GBIF set and a set of species names, coordinates, years and counts that had previously been extracted from the text of the taxonomic revision—the 'revision set' (Blades and Takano 2024). As part of the extraction process, all occurrences had been manually inspected to ensure they fell on land and in countries that were expected for that species. A small number of errors from typing inaccuracies were corrected with the expert help of the original taxonomist, and all records were cropped to the study extent before merging with GBIF data. Some of the records had been mobilised on GBIF, likely due to varying data sharing agreements between institutions holding the specimens, so duplicate entries were then manually removed. To be considered a duplicate, records must have shared the same species name and year, and in the circumstance that a more precise date was listed in either set, this must be identical. Records must also have had the same collector and location names, even if these were written in different formats, and neither their latitude nor longitude could be more than one geographic degree apart. Further specifics can be found in Appendix S2. This is henceforth referred to as the 'combined set'.

2.3 | Statistical Analyses

To assess survey completeness in all grid cells of the study area, the package 'KnowBR' version 2.2 (Guisande and Lobo 2023) was used. By generating species accumulation curves—representing the cumulative number of species observed as a function of the cumulative number of samples collected—it compares the number of observed species to the number of predicted species per spatial unit (here, the 100 km × 100 km cells). This determines a percentage of inventory completeness, that is, how many of the species that are likely to be present in each cell have been recorded as such (Lobo et al. 2018). Analysis was carried out separately for both the GBIF and combined sets, and then cells with > 20 records, a completeness score > 75% and a ratio of occurrences to the number of species > 5 were identified as 'well-sampled' (WS) in each set. These criteria used the approximate plateau of the cumulative distribution of records within each measure as a reference (see Ronquillo et al. 2023). The regional shapefile was used to quantify the bias of inventory completeness and WS cells for both sets in terms of how evenly they were distributed in each region.

As unevenness in the spatial coverage of occurrence data has been shown to result in biased sampling of environmental conditions (Hortal et al. 2008), the WS cells for each set were then used to evaluate if comprehensive sampling has been conducted across the full spectrum of potential environmental conditions in the study region. To describe environmental variations in the study area, we used the 19 bioclimatic variables from WorldClim, which represent trends, seasonality and limiting environmental factors of temperature and precipitation derived from aggregating monthly data from the period 1970–2000 (Fick and Hijmans 2017; WorldClim 2020; see Appendix S3). These were downloaded at a resolution of 30 s (approximately 1 km at the equator) and aggregated to 0.83° (approximately 100 km at the equator), the resolution of the grid used for spatial coverage, and cropped to the study extent.

A principal components analysis (PCA) of the study area was carried out in R package 'psych' version 2.4.1 (Revelle 2024) to reduce the dimensionality of the climate data to two axes (PC1 and PC2) that captured 73% of the variation. PC1 predominantly characterised levels of precipitation and variability in temperature, displaying a gradient from the high rainfall and more stable temperatures of tropical forests to the drier desert areas with greater annual and daily temperature ranges. PC2 broadly described a temperature gradient from the warmer north and low elevations to the cooler south and high elevations, as well as high to low precipitation seasonality to a lesser degree. Then, these axes were converted into classes by binning the PCA environmental space into equal-area cells (see Sobral-Souza et al. (2021) and Ronquillo et al. (2020) for a similar approach). The resulting bins correspond to distinct 'climate types' of which 33 were identified in this study area (Figure 2a). Although other approaches may provide a more continuous definition of climate rarity (see Fournier et al. 2020), the method used here has the advantage of ensuring that the analysis of WS cells uses the exact same description of climate frequency in both coverage of general climate conditions and climatic rarity.

The classes—or 'climate types'—and the frequency at which each was observed in the study region were quantified. The

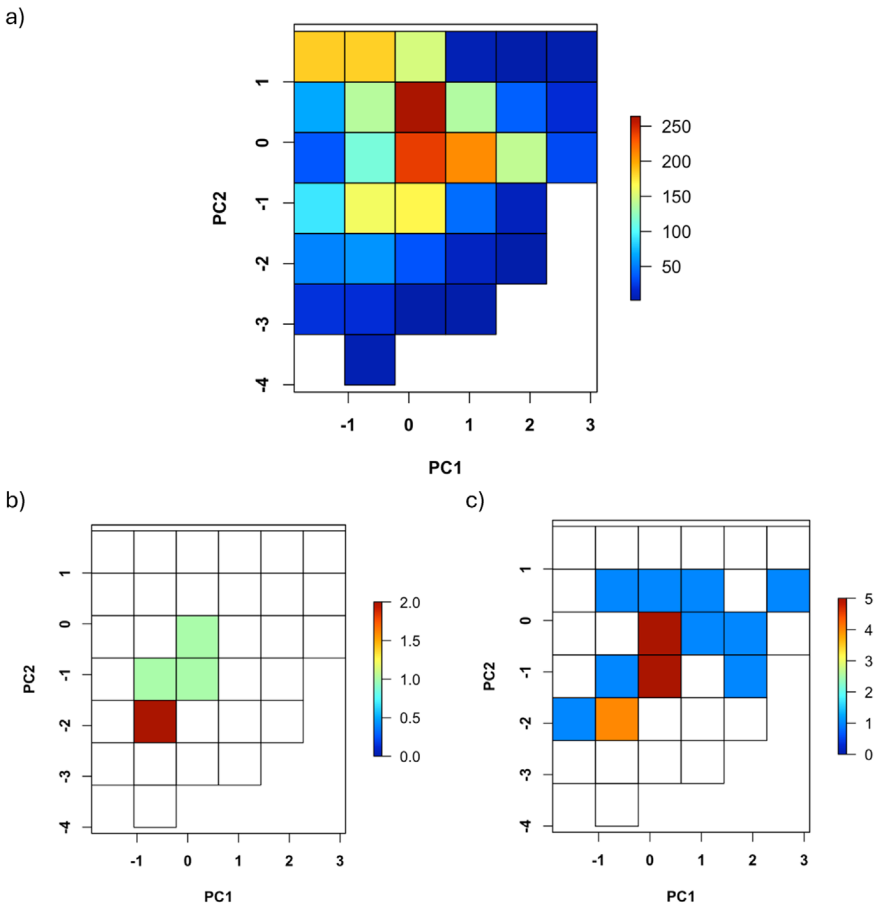


FIGURE 2 | Thirty-three climate types were identified in the Afrotropical mainland when principal components analysis values were converted into classes (a). Climate types found in well-sampled *Catharsius* cells for the GBIF set (b) and combined set (c) are displayed with colour indicating the number of cells that particular climate type was found in.

climate types that the WS cells fell into were identified, and the niche overlap between conditions in these locations and the study region as a whole was calculated using Schoener's *D*. This returns a value of between 0, for no overlap, and 1, for a complete overlap (Schoener 1970). Here, as WS cells are found in an increasing number of climate conditions, Schoener's *D* would be expected to rise, indicating good sampling over a more comprehensive coverage of potential environmental conditions in the Afrotropical realm. The significance of *D* values was tested by comparing them against a null distribution generated by randomly sampling 1000 sets of five (for the GBIF set) or 23 (for the combined set) occurrences (i.e., the same number of occurrences as WS cells for each set) and calculating the niche overlap of each sample with the study area.

To determine whether environmental conditions found in WS cells were an unbiased subset of those described by each PCA axis—that is, whether better sampling is correlated with particular conditions—two-sample Kolmogorov–Smirnov tests were run using the R base package 'stats'. Used to evaluate whether two groups are sampled from the same distribution (Massey 1951), the null hypothesis will be rejected if the WS cells sample certain climatic conditions at a rate that does not reflect their overall frequency in the study area. For example, if WS cells are found to disproportionately favour cooler climates, despite the presence of warmer climates in the realm.

Finally, min-max scaling was used to evaluate whether WS cells were found in climates that occur infrequently in the study region, or are 'rare'. For this, a rarity value between 0 (common) and 1 (rare) was assigned to each climate type, depending on its relative frequency in the study region. The rate at which each type was sampled by WS cells was then compared to its overall density in the study region, as such determining whether good sampling was biased towards common or rare climates.

All analyses were conducted in R version 4.2.1 (R Core Team 2022), using RStudio version 2023.12.1.402 (Posit Team 2024), and code adapted from Ronquillo (2023).

3 | Results

3.1 | Preprocessing of Occurrence Data

Downloading all records for *Catharsius* from GBIF returned a dataset of 4270 entries (GBIF 2023), pertaining to 72 species; more results of the quality assurance filtering and taxonomic standardisation procedures can be found in Appendix S2. After preprocessing, the GBIF set totalled 1686 entries, corresponding to 3915 specimens belonging to 50 species. The revision dataset totalled 4979 entries, corresponding to 15,943 specimens belonging to 146 species. Upon combination with the GBIF dataset, 489

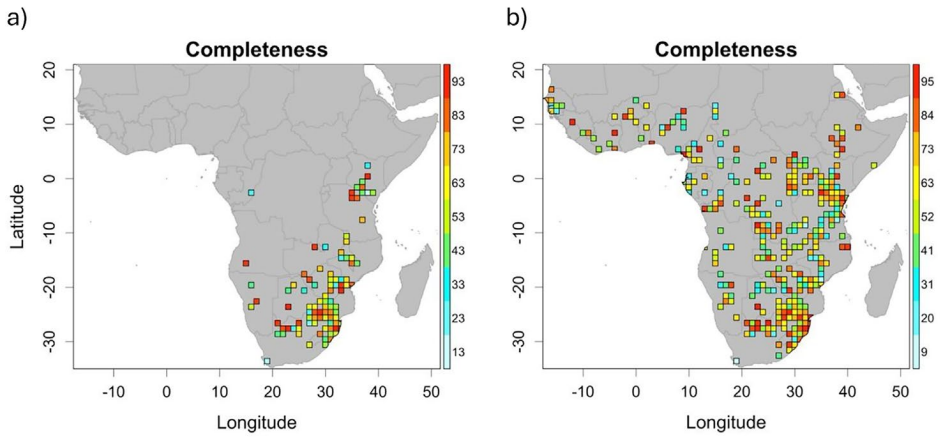


FIGURE 3 | Completeness of *Catharsius* inventories in the mainland Afrotropical realm using the (a) GBIF set and (b) combined set. Coloured cells are those for which data were sufficient to calculate completeness, with warmer colours indicating higher completeness percentage.

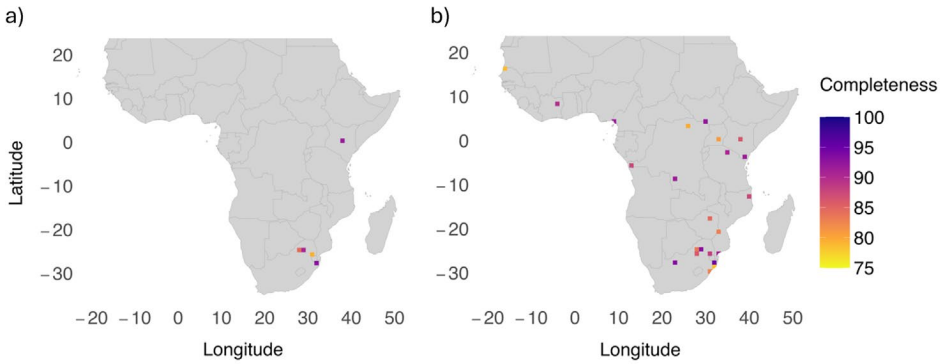


FIGURE 4 | Cells in the mainland Afrotropical realm in which *Catharsius* is well-sampled using (a) the GBIF set and (b) the combined set, where well-sampled is defined as containing > 20 records, completeness > 75% and a ratio of occurrences to species > 5. Colour indicates completeness percentage, with darker colours identifying higher completeness.

duplicate entries were removed from the former, resulting in a combined dataset of 6174 entries, corresponding to 18,043 specimens belonging to 146 species. As such, the revision set contributed the overwhelming majority of information to the combined set, with GBIF having contributed only 27.3% of entries, 21.7% of specimens and no new species.

3.2 | Statistical Analyses

A total of 94 cells, out of a potential 1867 (5%), contained sufficient information to compute inventory completeness for the GBIF set, which is predominantly concentrated in the northeast of South Africa (Figure 3). Regionally, 51 of these cells are found in Southern Africa, 41 in Eastern Africa, two in Central Africa and none in Western Africa. Contrastingly, inventory completeness could be computed for 314 (16.82%) with the combined set, an increase of 220. Of these, 75 are found in Southern Africa, 128 in Eastern Africa, 76 in Central Africa and 35 in Western Africa, illustrating a reduced regional sampling bias. Some cells for the combined set lie across the border between two regions, in which case the centre point of the grid square was the decider.

Cells with high completeness values were also concentrated in South Africa, with some further representation in Southern and Eastern Africa for the GBIF set, but a cross-realm spread for the combined set. The single grid cell in Central Africa with a high completeness value in the GBIF set was no longer considered to be so well completed in the context of a more extensive dataset.

Only five cells fulfilled the criteria to be considered well-sampled using the GBIF set, with four in the northeast of South Africa and one in Kenya; just 0.27% of 1867 potential cells and two countries. Using the combined set, 23 WS cells (1.23%) were identified across 11 countries (8 in South Africa, 3 in the Democratic Republic of Congo, 3 in Mozambique, 2 in Kenya and 1 in each of Senegal, Cote d'Ivoire, Cameroon, South Sudan, Zimbabwe, Uganda and Tanzania). Although South Africa is also comparatively overrepresented, and seven out of these 11 countries only returned a single WS cell, the combined set generated 4.5 times more WS cells in 5.5 times more countries than the GBIF set (Figure 4). Only four out of a potential 33 distinct climate types were found in WS cells from the GBIF set (12.12%; Figure 2b), as opposed to 12 for the combined set (36.36%; Figure 2c). The niche overlap (D) between WS cells and the study region as a whole was 0.238 ($p = 1$) and 0.468 ($p = 1$) for the GBIF and combined set respectively.

Kolmogorov–Smirnov tests show WS cells for both sets as unbiased in PCA1 (GBIF: $D = 0.30$, $p = 0.75$; combined: $D = 0.21$,

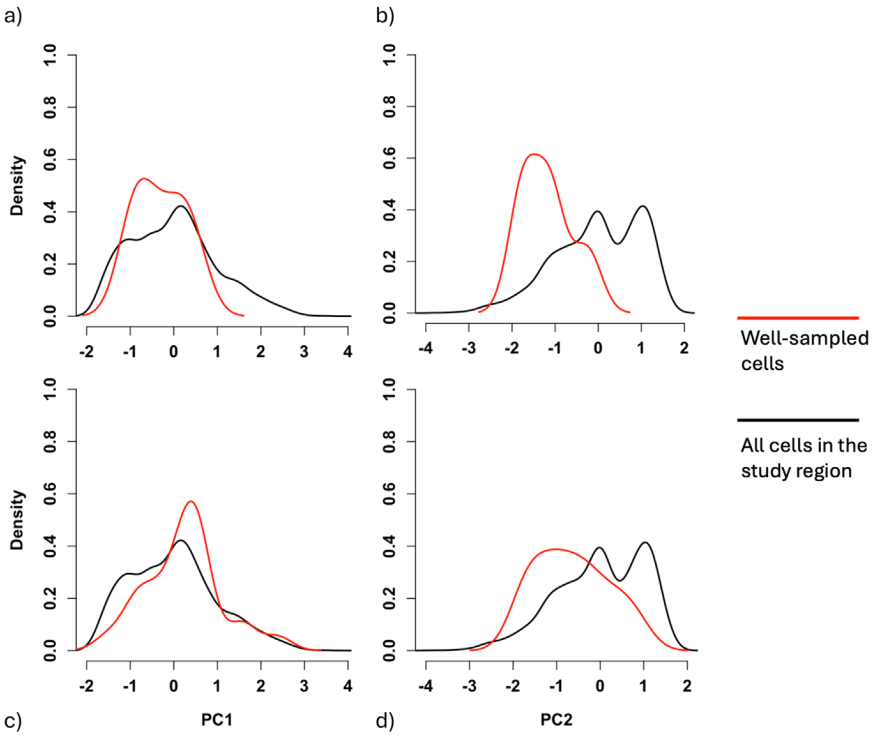


FIGURE 5 | Smoothed kernel density estimate of the distribution of principal component axes (PC1 and PC2) scores for the GBIF set (a and b) and the combined set (c and d), with the y-axis representing relative density. The red and black lines illustrate the continuous distribution of PCA values found in the *Catharsius* well-sampled cells and all cells in the mainland Afrotropical realm respectively.

$p = 0.26$) but biased in PCA2 (GBIF: $D = 0.63$, $p = 0.04$; combined: $D = 0.35$, $p = 0.0083$). As such, whilst both are representative samples of potential levels of precipitation and temperature variability, they oversample moderate and cooler temperatures, with comparatively little representation of warmer climates in the north (Figure 5).

No WS cells from the GBIF set were found in the rarest of climates (rarity > 0.8), and both very common and moderately common climate types were overrepresented compared to their relative density in the study region. Whilst WS cells from the combined set also undersampled the rarest of climates, it was to a lesser degree, and common and moderately common climate types were sampled at a rate more representative of their relative density in the study region (Figure 6a,c). Improved sampling in Central and Eastern Africa was responsible for coverage of the rarest climates by combined set WS cells, and the failure of GBIF WS cells to sample any of these was despite them both being most prevalent in South Africa (Figure 6b,d).

4 | Discussion

This study provides novel insights into the inventory completeness of the world's biggest online biodiversity data network. By quantifying the coverage improvement achieved with the addition of further natural history collection records, this comparative approach assesses not just how much value is still missing for insects on GBIF, but also the scale of potential gains from further record digitisation. In particular, it has shown that GBIF occurrence records are biased towards the southern and eastern areas of the Afrotropical realm and, consequently, fail to sample across the full range of potential environmental conditions. This leads to poor coverage of warmer climates and rare climate types. Whilst the inclusion of further natural history collection data does not entirely remove sampling bias, it greatly reduces unevenness in spatial and environmental sampling.

Inventory completeness for GBIF insect data is well documented as being poor worldwide, particularly so outside of Europe (García-Rosello et al. 2023; Rocha-Ortega et al. 2021), and this is reflected very clearly in these results. Here, GBIF data are only sufficient to compute completeness for 5% of grid cells and, even then, many of these return poor values and are strongly regionally biased, the pattern of which is broadly comparable with the general inventory completeness of insects on GBIF (Figure 2c, García-Rosello et al. 2023, 493). The combined set, though, allows completeness to be computed for over three times as many cells, and regional gaps are filled. It provides enough data to compute values in 74 and 35 more cells in Central and Western Africa, respectively, than the GBIF set, and also generated highly completed cells more evenly across the realm.

Notably, the two highly completed GBIF cells in Central Africa intersect Mupa and Bicuari National Parks in Angola, which is consistent with Girardello et al. (2019) who found gaps in GBIF butterfly inventories to correlate with the low density of protected areas. In fact, these results support that geographic biases in GBIF insect inventory completeness are comparable to those in GBIF raw data (García-Rosello et al. 2023), as they identify deficiencies driven by survey area attractiveness and socioeconomic factors. The GBIF set generated 4.5 times less WS cells

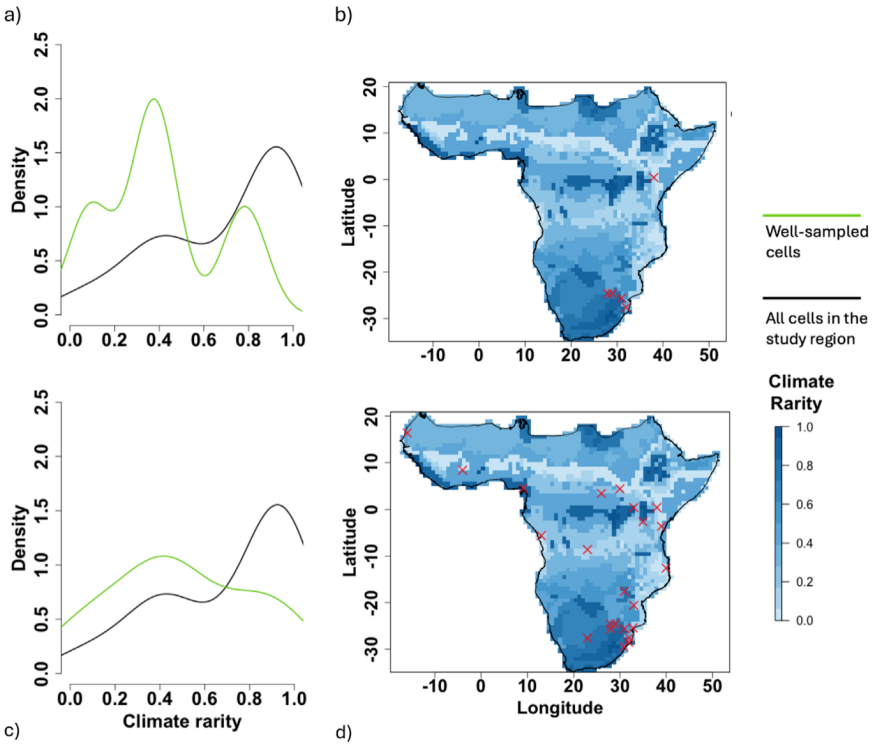


FIGURE 6 | Smoothed kernel density estimate of the distribution of climate rarity scores, between 0 (common) and rare (1), for the GBIF set (a) and the combined set (c), with the y-axis representing relative density. The green and black lines illustrate the continuous distribution of climate rarity values found in the *Catharsius* well-sampled cells and all cells in the mainland Afrotropical realm respectively. Maps illustrate the geographic distribution of climate rarity, with dark colours representing rarer climates, and well-sampled *Catharsius* cells for the GBIF set (b) and the combined set (d) illustrated by red crosses.

than the combined set, and their exclusive location in South Africa is consistent with findings that GBIF sampling coverage is strongly related to GDP per capita (Amano and Sutherland 2013; Hughes et al. 2021; IMF 2024), and shows similarities with the work of Stropp et al. (2016) on seed plants, evidencing that many biases in biodiversity knowledge are pervasive across groups. Contrastingly, almost a third of the combined set WS cells were generated in countries ranked in the bottom 10 for 2024 GDP per capita (Democratic Republic of Congo, Mozambique and South Sudan) (IMF 2024). Species data used in this study are not from 2024, but this does pose the question of whether bias in GBIF data driven by economic inequality could be reduced by further digitisation of natural history collections. Critically, records are made available on GBIF through data-sharing agreements between the platform and contributing institutions, and this study indicates that further mobilisation should leverage GBIF's strength as a centralised aggregator and aim to expand its pool of contributors rather than just the volume of information from existing participants.

In light of the stark regional biases in GBIF WS cells, it is not surprising that coverage of climatic conditions also falls short. Well-sampled cells from the combined set covered more climate types and had a higher niche overlap with the overall study region, demonstrating the potential value added to ecological inferences through increased record digitisation. That said, nonsignificant D values for both the GBIF set and the combined set indicate that WS cells are predominantly found in common climate types for both, despite general coverage improvement through the integration of further data. These results align with studies in which systematic surveys and natural history collections outperformed GBIF insect data in terms of survey coverage, but also derivation of ranges and climatic niches (Beck et al. 2013; Troia and McManamay 2016). It is concerning that the scale of bias in GBIF insect data is such that it compromises its usefulness despite the sheer number of records now accessible. This has also been observed in other taxa (Araujo et al. 2022; Shirey et al. 2019), and whilst its severity is likely not consistent across groups, it is probable that GBIF is not yet comprehensive enough to be used as an exclusive data source in modelling biodiversity patterns and should instead be used as a resource within a wider suite. This highlights the importance of investing in collection, digitisation and mobilisation of further datasets through the platform, as improving its coverage would reduce the need to source data from multiple, sometimes inaccessible, or nonstandardised datasets. Others have also called for the integration of biodiversity databases in one place to this same end (Araujo et al. 2022).

A further limitation of GBIF data here was that, whilst most climate types in this study occurred infrequently, this was not reflected in sampling; the GBIF WS cells did not occur in any of the rarest climate types, despite both being most prevalent in South Africa. This risks missing species that may be specialised to infrequent climatic conditions and, in the case of such habitat specificity, at risk on multiple fronts: smaller geographic range, few populations and specialised niche conditions (Işik 2011). It is intuitive that mobilising more records

on GBIF increases coverage and combats these weaknesses. Whilst not necessarily a given, as unmobilised specimens in natural history collections may well have been collected in the same places as existing online records, multiplying the quantity of data available is likely to provide new information, especially for a data-deficient group, as seen above. However, the relative density of the combined set WS cells across climate types and rarities (i.e., the likelihood of their being found in any given place) more closely resembles that of all cells in the study region, making it clear that this is not simply a case of increasing available data. The changing shapes of the smoother kernel density estimate curves in Figures 5 and 6 are a visual representation of closing climatic gaps in sampling. Explicitly, the data from the taxonomic revision are distributed differently across space and environment from those already available on GBIF, and aggregating these fills knowledge gaps.

Usefulness of digitised records, though, hinges on the quality of the data that is generated. Here, over half of the original GBIF data were removed during preprocessing as they did not meet the necessary standards for analysis (see Appendix S2), and concerns with geospatial errors, insufficient metadata and taxonomic inaccuracies are well documented (Ferro and Flick 2015; Prudic et al. 2023; Rocha-Ortega et al. 2021; Ronquillo et al. 2020; Yesson et al. 2007). Misidentifications and poor taxonomy are particularly difficult to pinpoint and correct after mobilisation (Soberón et al. 2002), and this is especially so for little-known or cryptic invertebrate species, such as those in *Catharsius*. The revision set data were extracted from a rigorous, multiyear project to revise *Catharsius*' taxonomy, so there were very few georeferencing errors and no taxonomic errors, according to this most recent understanding of the genus. The time-consuming nature of this work and declining taxonomic expertise (Hopkins and Freckleton 2002; Hutchings 2021; 'Importance of Taxonomy' 1946; Lagomarsino and Frost 2020; Wägele et al. 2011) precludes this as an option for many studies, especially those that encompass thousands or even millions of occurrence records. Methods to reduce GBIF taxonomic misidentifications have been tested with some success (Smith et al. 2016), but these still have distinct time and data demands of their own. Other tools to validate not just the taxonomic, but also geographic, temporal and metadata accuracy of records are also promising (Ronquillo et al. 2024), but the quality of GBIF data is such that these processes often greatly reduce the number of usable data points, as seen in this study. It seems clear that the best way to avoid compounding existing flaws in data quality is with meticulous digitisation in the first place, including the allocation of resources to taxonomic verification or revision as part of the data preparation process.

Efforts to understand the limitations of our knowledge are paramount. Research that seeks to better understand these flaws and suggest solutions not only improves theoretical knowledge of species distributions but can potentially better inform practical applications, such as directing fieldwork or prompting new data-sharing agreements with institutions whose data will maximise biogeographical inference. This study underlines that much value still stands to be gained by further digitisation of natural history collections, emphasising that GBIF is not yet ready to function as a standalone data source, but care must be taken to not compound existing data quality issues.

Author Contributions

Bryony Blades: conceptualization (equal), data curation (lead), formal analysis (lead), methodology (equal), project administration (lead), validation (equal), visualization (equal), writing – original draft (lead), writing – review and editing (equal). **Cristina Ronquillo:** conceptualization (supporting), data curation (supporting), formal analysis (supporting), methodology (supporting), validation (equal), visualization (equal), writing – review and editing (supporting). **Joaquín Hortal:** conceptualization (equal), data curation (supporting), formal analysis (supporting), methodology (equal), supervision (lead), validation (equal), visualization (supporting), writing – review and editing (supporting).

Acknowledgements

We are particularly thankful to Hitoshi Takano for his unselfish release of unpublished data, and his guidance with taxonomic disambiguation of the classification of *Catharsius* dung beetles, without which this work would not have been possible. The authors also thank Tim Coulson, Elizabeth Jeffers, Michael Bonsall and Robert Whittaker who commented on versions of this manuscript, and Richard Smith for his donation to the African Natural History Research Trust Scholarship.

Conflicts of Interest

The authors declare no Conflicts of Interest.

Data Availability Statement

Data and code supporting this paper are cited in this manuscript as (Blades & Takano, 2024), and are available at: 10.25446/oxford.27195816.

References

Acevedo, P., A. Jiménez-Valverde, P. Aragón, and A. Niamir. 2016. "New Developments in the Study of Species Distribution." In *Current Trends in Wildlife Research*, edited by R. Mateo, B. Arroyo, and J. T. Garcia, 151–175. Springer International Publishing. https://doi.org/10.1007/978-3-319-27912-1_7.

Amano, T., and W. J. Sutherland. 2013. "Four Barriers to the Global Understanding of Biodiversity Conservation: Wealth, Language, Geographical Location and Security." *Proceedings of the Royal Society B: Biological Sciences* 280, no. 1756: 2649. https://doi.org/10.1098/rspb.2012.2649.

Araujo, M. L., A. C. Quaresma, and F. N. Ramos. 2022. "GBIF Information Is Not Enough: National Database Improves the Inventory Completeness of Amazonian Epiphytes." *Biodiversity and Conservation* 31, no. 11: 2797–2815. https://doi.org/10.1007/s10531-022-02458-x.

Ballesteros-Mejia, L., I. J. Kitching, W. Jetz, P. Nagel, and J. Beck. 2013. "Mapping the Biodiversity of Tropical Insects: Species Richness and Inventory Completeness of A Frican Sphingid Moths." *Global Ecology and Biogeography* 22, no. 5: 586–595. https://doi.org/10.1111/geb.12039.

Beck, J., L. Ballesteros-Mejia, P. Nagel, and I. J. Kitching. 2013. "Online solutions and the 'Wallacean shortfall': What does GBIF contribute to our knowledge of species' ranges?" *Diversity and Distributions* 19, no. 8: 1043–1050. https://doi.org/10.1111/ddi.12083.

Blades, B., and H. Takano. 2024. *Catharsius Inventory Completeness [Dataset]*. 195816. Oxford. https://doi.org/10.25446/oxford.27195816.

Collen, B., M. Ram, T. Zamin, and L. McRae. 2008. "The Tropical Biodiversity Data Gap: Addressing Disparity in Global Monitoring." *Tropical Conservation Science* 1, no. 2: 75–88. https://doi.org/10.1177/194008290800100202.

Elith, J., and J. Franklin. 2013. "Species Distribution Modeling." In *Encyclopedia of Biodiversity*, 692–705. Elsevier. https://doi.org/10.1016/B978-0-12-384719-5.00318-X.

Ferro, M. L., and A. J. Flick. 2015. "Collection Bias and the Importance of Natural History Collections in Species Habitat Modeling: A Case Study Using *Thoracophorus Costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), with a Critique of GBIF.Org." *Coleopterists Bulletin* 69, no. 3: 415–425. https://doi.org/10.1649/0010-065X-69.3.415.

Fick, S. E., and R. J. Hijmans. 2017. "WorldClim 2: New 1-Km Spatial Resolution Climate Surfaces for Global Land Areas." *International Journal of Climatology* 37, no. 12: 4302–4315. https://doi.org/10.1002/joc.5086.

Fournier, B., H. Vázquez-Rivera, S. Clappe, L. Donelle, P. H. P. Braga, and P. R. Peres-Neto. 2020. "The Spatial Frequency of Climatic Conditions Affects Niche Composition and Functional Diversity of Species Assemblages: The Case of Angiosperms." *Ecology Letters* 23, no. 2: 254–264. https://doi.org/10.1111/ele.13425.

García-Rosello, E., J. Gonzalez-Dacosta, C. Guisande, and J. M. Lobo. 2023. "GBIF Falls Short of Providing a Representative Picture of the Global Distribution of Insects." *Systematic Entomology* 48, no. 4: 489–497. https://doi.org/10.1111/syen.12589.

GBIF. 2023. "Occurrence Download [Dataset]." https://doi.org/10.15468/DL.73MEZG.Global Biodiversity Information Facility.

Girardello, M., A. Chapman, R. Dennis, L. Kaila, P. A. V. Borges, and A. Santangeli. 2019. "Gaps in Butterfly Inventory Data: A Global Analysis." *Biological Conservation* 236: 289–295. https://doi.org/10.1016/j.biocon.2019.05.053.

Guisan, A., and W. Thuiller. 2005. "Predicting Species Distribution: Offering More Than Simple Habitat Models." *Ecology Letters* 8, no. 9: 993–1009. https://doi.org/10.1111/j.1461-0248.2005.00792.x.

Guisande, C., and J. Lobo. 2023. "*KnowBR: Discriminating Well Surveyed Spatial Units From Exhaustive Biodiversity Databases*." *R package version 2.2* [Computer software]. https://CRAN.R-project.org/package=KnowBR.

Hardy, H., L. Livermore, P. Kersey, K. Norris, and V. Smith. 2023. "Understanding the Users and Uses of UK Natural History Collections." *Research Ideas & Outcomes* 9: e113378. https://doi.org/10.3897/rio.9.e113378.

Heberling, J. M., J. T. Miller, D. Noesgaard, S. B. Weingart, and D. Schigel. 2021. "Data Integration Enables Global Biodiversity Synthesis." *Proceedings of the National Academy of Sciences* 118, no. 6: e2018093118. https://doi.org/10.1073/pnas.2018093118.

Hopkins, G. W., and R. P. Freckleton. 2002. "Declines in the Numbers of Amateur and Professional Taxonomists: Implications for Conservation." *Animal Conservation* 5, no. 3: 245–249. https://doi.org/10.1017/S1367943002002299.

Hortal, J., F. De Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle. 2015. "Seven Shortfalls That Beset Large-Scale Knowledge of Biodiversity." *Annual Review of Ecology, Evolution, and Systematics* 46, no. 1: 523–549. https://doi.org/10.1146/annurev-ecolsys-112414-054400.

Hortal, J., A. Jiménez-Valverde, J. F. Gómez, J. M. Lobo, and A. Baselga. 2008. "Historical Bias in Biodiversity Inventories Affects the Observed Environmental Niche of the Species." *Oikos* 117, no. 6: 847–858. https://doi.org/10.1111/j.0030-1299.2008.16434.x.

Hughes, A. C., M. C. Orr, K. Ma, et al. 2021. "Sampling Biases Shape Our View of the Natural World." *Ecography* 44, no. 9: 1259–1269. https://doi.org/10.1111/ecog.05926.

Hutchings, P. 2021. "Potential Loss of Biodiversity and the Critical Importance of Taxonomy—An Australian Perspective." In *Advances in Marine Biology*, vol. 88, 3–16. Elsevier. https://doi.org/10.1016/S0065-2881(21)00015-8.

Iannella, M., P. D'Alessandro, and M. Biondi. 2019. "Entomological Knowledge in Madagascar by GBIF Datasets: Estimates on the Coverage and Possible Biases (Insecta)." *Fragmenta Entomologica* 51, no. 1: 1–10. https://doi.org/10.4081/fe.2019.329.

Importance of Taxonomy. 1946. "Importance of Taxonomy." *Nature* 158: 105–106.

International Monetary Fund. 2024. "Report for Selected Countries and Subjects." World Economic Outlook Databas. https://www.imf.org/en/Publications/WEO/weo-database/2024/April/weo-report.

Işik, K. 2011. "Rare and Endemic Species: Why Are They Prone to Extinction?" *Turkish Journal of Botany* 35: 411–417. https://doi.org/10.3906/bot-1012-90.

Lagomarsino, L. P., and L. A. Frost. 2020. "The Central Role of Taxonomy in the Study of Neotropical Biodiversity." *Annals of the Missouri Botanical Garden* 105, no. 3: 405–421. https://doi.org/10.3417/2020601.

Lobo, J. M., J. Hortal, J. L. Yela, et al. 2018. "KnowBR: An Application to Map the Geographical Variation of Survey Effort and Identify Well-Surveyed Areas From Biodiversity Databases." *Ecological Indicators* 91: 241–248. https://doi.org/10.1016/j.ecolind.2018.03.077.

Lomolino, M. V. 2004. "Conservation Biogeography." In *Frontiers of Biogeography: New Directions in the Geography of Nature*, edited by L. R. Heaney and M. V. Lomolino, 293–296. Sinauer Associates.

Massey, F. J. 1951. "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association* 46, no. 253: 68–78. https://doi.org/10.2307/2280095.

Newbold, T. 2010. "Applications and Limitations of Museum Data for Conservation and Ecology, With Particular Attention to Species Distribution Models." *Progress in Physical Geography: Earth and Environment* 34, no. 1: 3–22. https://doi.org/10.1177/0309133309355630.

Noriega, J. A., J. Hortal, F. M. Azcárate, et al. 2018. "Research Trends in Ecosystem Services Provided by Insects." *Basic and Applied Ecology* 26: 8–23. https://doi.org/10.1016/j.baae.2017.09.006.

Oliver, R. Y., C. Meyer, A. Ranipeta, K. Winner, and W. Jetz. 2021. "Global and National Trends, Gaps, and Opportunities in Documenting and Monitoring Species Distributions." *PLoS Biology* 19, no. 8: e3001336. https://doi.org/10.1371/journal.pbio.3001336.

Olson, D. M., E. Dinerstein, E. D. Wikramanayake, et al. 2001. "Terrestrial Ecoregions of the World: A New Map of Life on Earth." *Bioscience* 51, no. 11: 933. https://doi.org/10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2.

Popov, D., P. Roychoudhury, H. Hardy, L. Livermore, and K. Norris. 2021. "The Value of Digitising Natural History Collections." *Research Ideas & Outcomes* 7: e78844. https://doi.org/10.3897/rio.7.e78844.

Posit Team. 2024. "*RStudio: Integrated Development Environment for R* [Computer Software]." Posit Software, PBC. http://www.posit.co/.

Prudic, K., E. Zylstra, N. Melkonoff, R. Laura, and R. Hutchinson. 2023. "Community Scientists Produce Open Data for Understanding Insects and Climate Change." *Current Opinion in Insect Science* 59: 101081. https://doi.org/10.1016/j.cois.2023.101081.

QGIS. 2021. "*QGIS 3.22.9-Białowieża* (Version 3.22.9-Białowieża) [Computer Software]."

R Core Team. 2022. "R: A Language and Environment for Statistical Computing." *R Foundation for Statistical Computing* [Computer software]. https://www.R-project.org/.

Revelle, W. 2024. "*psych: Procedures for Psychological, Psychometric, and Personality Research* [Computer software]." Northwestern University. https://CRAN.R-project.org/package=psych.

Rocha-Ortega, M., P. Rodriguez, and A. Córdoba-Aguilar. 2021. "Geographical, Temporal and Taxonomic Biases in Insect GBIF Data on Biodiversity and Extinction." *Ecological Entomology* 46, no. 4: 718–728. https://doi.org/10.1111/een.13027.

Romo, H., E. García-Barros, and J. M. Lobo. 2006. "Identifying Recorder-Induced Geographic Bias in an Iberian Butterfly Database." *Ecography* 29, no. 6: 873–885. https://doi.org/10.1111/j.2006.0906-7590.04680.x.

Ronquillo, C. 2023. "*cRonFer/Inv.Completeness-Env.Space* [R]." https://github.com/cRonFer/Inv.Completeness-Env.Space.

Ronquillo, C., F. Alves-Martins, V. Mazimpaka, et al. 2020. "Assessing Spatial and Temporal Biases and Gaps in the Publicly Available Distributional Information of Iberian Mosses." *Biodiversity Data Journal* 8: e53474. https://doi.org/10.3897/BDJ.8.e53474.

Ronquillo, C., J. Stropp, and J. Hortal. 2024. "OCCUR Shiny Application: A User-Friendly Guide for Curating Species Occurrence Records." *Methods in Ecology and Evolution* 15, no. 5: 816–823. https://doi.org/10.1111/2041-210X.14271.

Ronquillo, C., J. Stropp, N. G. Medina, and J. Hortal. 2023. "Exploring the Impact of Data Curation Criteria on the Observed Geographical Distribution of Mosses." *Ecology and Evolution* 13, no. 12: e10786. https://doi.org/10.1002/ece3.10786.

Sánchez-Fernández, D., R. Fox, R. L. H. Dennis, and J. M. Lobo. 2021. "How Complete Are Insect Inventories? An Assessment of the British Butterfly Database Highlighting the Influence of Dynamic Distribution Shifts on Sampling Completeness." *Biodiversity and Conservation* 30, no. 3: 889–902. https://doi.org/10.1007/s10531-021-02122-w.

Sánchez-Fernández, D., J. M. Lobo, P. Abellán, I. Ribera, and A. Millán. 2008. "Bias in Freshwater Biodiversity Sampling: The Case of Iberian Water Beetles." *Diversity and Distributions* 14, no. 5: 754–762. https://doi.org/10.1111/j.1472-4642.2008.00474.x.

Sánchez-Fernández, D., J. L. Yela, R. Acosta, et al. 2022. "Are Patterns of Sampling Effort and Completeness of Inventories Congruent? A Test Using Databases for Five Insect Taxa in the Iberian Peninsula." *Insect Conservation and Diversity* 15, no. 4: 406–415. https://doi.org/10.1111/icad.12566.

Schoener, T. W. 1970. "Nonsynchronous Spatial Overlap of Lizards in Patchy Habitats." *Ecology* 51, no. 3: 408–418. https://doi.org/10.2307/1935376.

Shirey, V., M. W. Belitz, V. Barve, and R. Guralnick. 2021. "A Complete Inventory of North American Butterfly Occurrence Data: Narrowing Data Gaps, but Increasing Bias." *Ecography* 44, no. 4: 537–547. https://doi.org/10.1111/ecog.05396.

Shirey, V., S. Seppälä, V. Branco, and P. Cardoso. 2019. "Current GBIF Occurrence Data Demonstrates Both Promise and Limitations for Potential Red Listing of Spiders." *Biodiversity Data Journal* 7: e47369. https://doi.org/10.3897/BDJ.7.e47369.

Siddig, A. A. H. 2019. "Why Is Biodiversity Data-Deficiency an Ongoing Conservation Dilemma in Africa?" *Journal for Nature Conservation* 50: 125719. https://doi.org/10.1016/j.jnc.2019.125719.

Smith, B. E., M. K. Johnston, and R. Lücking. 2016. "From GenBank to GBIF: Phylogeny-Based Predictive Niche Modeling Tests Accuracy of Taxonomic Identifications in Large Occurrence Data Repositories." *PLoS One* 11, no. 3: e0151232. https://doi.org/10.1371/journal.pone.0151232.

Soberón, J., L. Arriaga, and L. Lara. 2002. "Issues of Quality Control in Large, Mixed-Origin Entomological Databases." In *Towards a Global Biological Information Infrastructure.Pdf*, edited by H. Saarenmaa and E. S. Nielsen, 15–22. European Environment Agency. https://www.eea.europa.eu/publications/technical_report_2001_70.

Soberón, J., and T. Peterson. 2004. "Biodiversity Informatics: Managing and Applying Primary Biodiversity Data." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 359, no. 1444: 689–698. https://doi.org/10.1098/rstb.2003.1439.

Sobral-Souza, T., J. Stropp, J. P. Santos, et al. 2021. "Knowledge Gaps Hamper Understanding the Relationship Between Fragmentation and Biodiversity Loss: The Case of Atlantic Forest Fruit-Feeding Butterflies." *PeerJ* 9: e11673. https://doi.org/10.7717/peerj.11673.

Spector, S. 2006. "Scarabaeine Dung Beetles (Coleoptera: Scarabaeidae: Scarabaeinae): An Invertebrate Focal Taxon for Biodiversity Research and Conservation." *Coleopterists Bulletin* 60: 71–83.

Sporbert, M., H. Bruelheide, G. Seidler, et al. 2019. "Assessing Sampling Coverage of Species Distribution in Biodiversity Databases." *Journal of Vegetation Science* 30, no. 4: 620–632. https://doi.org/10.1111/jvs.12763.

Stropp, J., R. J. Ladle, A. C. Malhado, et al. 2016. "Mapping Ignorance: 300 Years of Collecting Flowering Plants in Africa." *Global Ecology and Biogeography* 25, no. 9: 1085–1096. https://doi.org/10.1111/geb.12468.

Takano, H. 2025. "A systematic revision of the Afrotropical members of the dung beetle genus Catharsius Hope, 1837 (Coleoptera: Scarabaeidae)." [Manuscript in Preparation]. African Natural History Research Trust.

Troia, M. J., and R. A. McManamay. 2016. "Filling in the GAPS: Evaluating Completeness and Coverage of Open-Access Biodiversity Databases in the United States." *Ecology and Evolution* 6, no. 14: 4654–4669. https://doi.org/10.1002/ece3.2225.

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. "Taxonomic Bias in Biodiversity Data and Societal Preferences." *Scientific Reports* 7, no. 1: 9132. https://doi.org/10.1038/s41598-017-09084-6.

United Nations Statistics Division. n.d. "UNSD—Methodology. Standard Country or Area Codes for Statistical Use (M49)." https://unstats.un.org/unsd/methodology/m49/.

Wägele, H., A. Klussmann-Kolb, M. Kuhlmann, et al. 2011. "The Taxonomist—An Endangered Race. A Practical Proposal for Its Survival." *Frontiers in Zoology* 8, no. 1: 25. https://doi.org/10.1186/1742-9994-8-25.

Wagner, D. L., E. M. Grames, M. L. Forister, M. R. Berenbaum, and D. Stopak. 2021. "Insect Decline in the Anthropocene: Death by a Thousand Cuts." *Proceedings of the National Academy of Sciences* 118, no. 2: e2023989118. https://doi.org/10.1073/pnas.2023989118.

World Wildlife Fund. 2012. "Publications: Terrestrial Ecoregions of the World." *World Wildlife Fund*. https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world.

WorldClim. 2020, 2022. "*Bioclimatic variables.*" https://www.worldclim.org/data/bioclim.html.

Wüest, R. O., N. E. Zimmermann, D. Zurell, et al. 2020. "Macroecology in the Age of Big Data—Where to Go From Here?" *Journal of Biogeography* 47, no. 1: 1–12. https://doi.org/10.1111/jbi.13633.

Yesson, C., P. W. Brewer, T. Sutton, et al. 2007. "How Global Is the Global Biodiversity Information Facility?" *PLoS One* 2, no. 11: e1124. https://doi.org/10.1371/journal.pone.0001124.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

CHAPTER 4

Exploring the congruence of data types: do molecular and morphological trait data describe phylogenetic relationships in the same way?

In Preparation

Bryony Blades[1,2], Andrea Estandía[1]
[1]Department of Biology, University of Oxford, Oxford, UK
[2]African Natural History Research Trust, Kingsland, Herefordshire, UK

Abstract

Descriptions of species and their relationships to one another are critical in understanding the evolution of life. However, disagreements are ongoing as to the correct methodology with which to define these, made more difficult by the diverging results sometimes produced by morphological and genetic data. This is better understood for the use of molecular markers, but there is no consensus on the degree to which inferences agree when using whole genome single nucleotide polymorphisms (SNPs). With the renewed importance of visual species identification due to mass record digitisation and the rise of citizen science, clarifying this is key. Using a morphological trait matrix created by an expert taxonomist for the revision of *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidae), this study compares how it portrays species relatedness to that described by whole genome SNPs. We analyse how the datasets cluster

species and depict both inter-species and inter-cluster variation using distance matrices, hierarchical clustering dendrograms, and multidimensional scaling. Results show that variation captured by whole genome SNPs agrees with morphological traits in overall population structure, but somewhat differs in the degree of relatedness within and between clusters, supporting existing literature that advocates for a “total evidence” approach in taxonomy. However, it underlines that where this is not possible, morphological traits remain a valid representation of diversity, and recommends that mapping them onto total evidence trees be adopted as standard procedure to assist visual identification in settings without access to genetic resources.

Introduction

Phylogenetics, the study of evolutionary history and relationships between organisms, has long been acknowledged as critical in comparative and evolutionary biology, in myriad fields such as developmental biology, functional morphology, endocrinology, ecology, behavioural science, genomics, epidemiology, conservation biology, and population dynamics (Losos, 1996; Lyubetsky et al., 2014). Description of species, in the related field of taxonomy, and derivation of their relationships, though, is not straightforward, and contentious debate continues about the best data and methods to use (e.g. Zamani et al., 2021). These have evolved alongside technological advance, with traditional techniques focusing on the grouping of individuals according to phenotypic similarity, and modern methods favouring the use of molecular data, from single- and multi-locus datasets, through to analysis of whole genomes made possible by next-generation sequencing (Karbstein et al., 2024). There has been a push for integrated studies using both types of data, but even a decade ago morphological traits already made up less than 2% of characters in combined analyses (Lee & Palci, 2015). This move away from morphology may prove problematic. If studies in taxonomy and phylogenetics only use molecular methods, visually diagnostic traits may not be captured, rendering morphological identification difficult. In the midst of methodological disagreement and decline of traditional methods, this study assesses whether taxonomic studies using morphological traits are able to capture species relatedness in a way that's supported by genetics.

Although still requiring taxonomic expertise, the strengths of morphological data lie in their logistical accessibility: they are cheaper and easier to generate and process, requiring less specialised lab work or bioinformatics training. As such, they can act as a financially accessible preliminary assessment of population structure (Kadoić Balaško et al., 2021). Most critically, though, morphological data are all that we have for fossils and are thus essential in generating time-calibrated phylogenies to track evolutionary change over time. To achieve this, fossils must be accurately positioned in the tree of life, which can only be done with description of their morphology, and that of the other taxa within the tree (Lee & Palci, 2015). Not only does doing this improve the trees (Keating et al., 2023; Oyston et al., 2022), it is only through such time-calibrated phylogenies that we can investigate further questions such as the role of dispersal and vicariance in distribution patterns (Oyston et al., 2022).

Despite this critical role, these data are criticised predominantly because choosing, coding, and scoring morphological and other phenotypic traits necessarily entails a number of subjective decisions. That only a subset of informative phenotypic characters are generally measured, which itself relies on which taxa are included in the study, means that definitions of homology across taxa are fundamentally inconsistent (Van Den Ende et al., 2023). This is compounded by the inclusion of absences as a trait state equal to others, despite the inherent inability to compare it in its degree of similarity, and susceptibility to sampling artefacts (Van Den Ende et al., 2023). It is also difficult to identify cryptic species, or traits which have evolved as a result of convergent evolution (Sales et al., 2018; Taylor et al., 2019).

With molecular data, possible states (i.e. nucleotides or amino acids) are known a priori and identified biochemically, and alignments optimised algorithmically. Furthermore, absences (missing data) are not considered an equal state, susceptibility to sampling artefacts is lower, and variation is not sampled at different resolutions as sequences are independent from taxon sampling density or degree of difference between taxa (Van Den Ende et al., 2023). This improved consistency across sampling events, combined with the decreased need for taxonomic expertise and difficulty in findings characters, means larger and more standardised datasets can be compiled (Oyston et al., 2022). However, whilst molecular models of evolution are more refined than morphological models (Oyston et al., 2022), these sophisticated algorithms are not always used, their efficacy is affected by the nature of the sample, and decisions on which sequences, algorithms, and modelling parameters to use are still subjective (Van Den Ende et al., 2023).

A number of studies have compared inferences from phenotypic and molecular data, many finding that resulting trees or relatedness measurements do not correlate, but are complementary, and combining them in a "total evidence" approach paints a more complete evolutionary picture than using one or the other (A. A. Alves et al., 2013; R. M. Alves et al., 2017; E. K. V. D. Andrade et al., 2017; Darkwa et al., 2020; Kadoić Balaško et al., 2021; Keating et al., 2023; Van Den Ende et al., 2023). Oft cited evidence of this comes from Afrotheria, a superorder of placental mammals whose monophyly was only uncovered with molecular data, as their vastly different morphologies had long-informed their placement in different groups (Lee & Palci, 2015; Oyston et al., 2022; Van Den Ende et al., 2023). It has been suggested that this low correlation is to be expected

when using molecular markers due to their limited ability to predict phenotype by virtue of predominantly representing limited, non-coding parts of the genome and consequently different selection pressures, but that using single nucleotide polymorphisms (SNPs) from the whole genome may overcome this (A. A. Alves et al., 2013; R. M. Alves et al., 2017). This has since been both supported (Kadoić Balaško et al., 2021) and disputed (Darkwa et al., 2020).

Comparative studies have not only evaluated accuracy of phylogenetic trees, resolved deep relationships, and sought to better understand the evolution of life (Oyston et al., 2022; Van Den Ende et al., 2023), but also assessed genetic diversity of important food and biofuel crops for breeding programs (A. A. Alves et al., 2013; R. M. Alves et al., 2017; E. K. V. D. Andrade et al., 2017; Darkwa et al., 2020), and identified resistance in pest populations (Kadoić Balaško et al., 2021). In these applications, with good access to expertise and funding, it is intuitive that a total evidence approach is the way forward. However, in a time when biological recording, by experts and amateurs alike, and mass digitisation of these records is at the forefront of theoretical and applied assessments of biodiversity, biogeography, and macroecology (Heberling et al., 2021), there is a disconnect between this ideal and what is feasible. Field surveying, and to a certain degree identification of museum specimens, fundamentally relies on visual identification, and if a species can only be distinguished genetically, it becomes invisible in these records, risking underreporting of taxa, biased biodiversity models based on incomplete data, and challenges for conservation efforts that depend on public monitoring.

A recent taxonomic revision of the Afrotropical members of the genus *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidae) (Takano, 2018) generated an extensive morphological trait matrix to characterise species, including some close relatives with very few visual dissimilarities. Fortunately, these closely-related species, along with others from the wider genus, have also been collected and their DNA sequenced. This presents an ideal opportunity to investigate the lack of consensus on whether inferences from whole genome SNPs correspond with phenotypic variation using expertly chosen morphological traits. If so, such expertise could be used to guide improved field- and museum-based identification when genetic analysis is not feasible, which is critical at a time when so much of this information is being uploaded to online databases. Specifically, we evaluate the congruence in how this morphological trait matrix and whole genome SNPs describe species clustering, as well as inter- and intra-cluster relatedness, using distance matrices, hierarchical clustering dendrograms, and multidimensional scaling.

Methods

Specimens

A total of 66 *Catharsius* dung beetles were collected from locations in the North-Western Province of Zambia (Figure 4.1) in November–December 2022. Immediately after euthanising specimens, a posterior leg was removed, split into three parts and the femur slightly crushed, and immersed in 96% ethanol in a 1.5ml Eppendorf tube. The body was then split in two between the pronotum and mesonotum and immersed in

96% ethanol in a 50ml centrifuge tube, along with the corresponding Eppendorf tube. Centrifuge tubes were kept in Ziplock bags that contained a label (pencil on waterproof paper) with a unique code and date for each collection event, each corresponding with a GPS coordinate of the collecting location, and these bags stored in boxes that were kept at ambient temperature in the shade. Remote field conditions precluded the use of a freezer to store specimens. Upon return to the United Kingdom, these boxes were transferred to a freezer at the University of Oxford. To begin the laboratory work, specimens were identified to species-level by Dr. Hitoshi Takano, the expert taxonomist responsible for the revision of the genus (Takano, 2018). Individuals were mixed and identified in a random order with no knowledge of the collecting location to prevent identification being informed in any way by where, or with which other species, each beetle was found. It is difficult to distinguish between species with females and minor males of *Catharsius dux* and *Catharsius duciformis* without dissection of their genitalia and, in these cases, 29 specimens were identified as *C. dux / C. duciformis*. As there were 28 other specimens (18 and 10, respectively) identified to just one of each of these species, a negative impact on downstream analyses was not anticipated, and identification accuracy over precision was thus prioritised. DNA was extracted from the femur of the hind leg using Qiagen DNeasy Blood and Tissue Kits and samples sent to Novogene. At Novogene, 49 of the samples required a DNA purification process to remove RNA contamination, and whole genomes were sequenced at 10X depth using Illumina NovaSeq X Plus. The specimens are now stored at the African Natural History Research Trust in Herefordshire, UK.

Bioinformatics

To begin the bioinformatics pipeline (Figure 4.2), the most closely related reference genome available, *Copris fidius* (JAUIMS000000000; Gustafson et al., 2023), was indexed with BWA module v0.7.17 (H. Li & Durbin, 2009). Then, the first ten bases were trimmed from raw sequencing reads using FastP (S. Chen, 2023; S. Chen et al., 2018) and mapped to the reference genome using SAMtools v1.14 (Danecek et al., 2021). Duplicate reads in the resulting BAM files were removed using Picard v3 (Broad Institute, 2019), and SNP calling performed using BCFtools v1.14 (Danecek et al., 2021). Indels were then excluded and only biallelic SNPs retained, and variants further filtered removing sites with quality ≤ 10, combined depth of coverage > 2000 and low read support <500 using BCFtools v1.14 (Danecek et al., 2021). As SNPs in linkage disequilibrium convey correlated genetic information and affect modelling outcomes, these were pruned using PLINK v2.00a2.3 (Chang et al., 2015; Purcell & Chang, n.d.), using the following parameters: a sliding window size of 150, a step size of 5, and an R^2 threshold of 0.5, as in Chen (2023).

Statistical analysis

To store and analyse the SNPs in R, a 'genlight' object was created in R package 'adegenet' (Jombart, 2008; Jombart & Ahmed, 2011), and a Euclidean distance matrix created to measure the genetic distance between pairs of specimens using R base package 'stats'. Using specimen identifications, the average pairwise distance between species was calculated to create a matrix of genetic distance between species. Some

of the following comparative analyses require identical matrix structures and, as the trait matrix from the taxonomic revision did not include C. *dux / C. duciformis* (it only includes single species) or *Catharsius merrettorum* (which was added as a later addition to the taxonomic revision (Takano, 2025)), two versions of this final distance matrix were created: a first with all specimens (henceforth, the "full" genetic distance matrix), and a second removing specimens identified as either *C. dux / C. duciformis* or *C. merrettorum* (henceforth, the "reduced" genetic distance matrix).

Figure 4.1: Catharsius *specimens were collected in November-December 2022 from eight locations, depicted by the red points, in the North-Western province of Zambia, shown in both the main and inset maps in dark blue. The inset map shows the fieldwork area in the wider geographical context of Zambia.*

Trait information was extracted from the character matrix of the taxonomic revision, comprising 130 traits from both external and internal morphological characters (Takano, 2018). These were a mix of binary and multistate characters that describe the shape and positioning of features on the head, forelegs, pronotum, elytrum, ventrite, mid/hind legs, aedeagus, and endophallus (detailed description of these features can be found in Appendix II of Takano (2018). Missing data was included when traits were absent in some species. As such, to calculate the trait distance matrix, pairwise distances

between species were computed using Gower's distance, in R package 'cluster', v2.1.6 (Maechler et al., 2023). This is known to handle missing and different types of data well, and is useful in clustering analysis (Gower, 1971).

Several multivariate statistical methods were used to analyse to what degree the morphological trait and genetic data describe the same relatedness between species. First, a Mantel test was performed with R package 'vegan' v2.6-8 (Oksanen et al., 2024) to assess the correlation between the morphological and reduced genetic distance matrices. The test computes Pearson's correlation coefficient (*r*), with values close to zero indicating no correlation and those close to one or minus one indicating a strong positive or negative correlation, respectively (K. Pearson, 1895). If the trait and genetic data describe the same degree of relatedness between species, as differences between species increase in one distance matrix, so should they in the other, indicated by a value of *r* close to one. Significance was derived using 999 permutations.

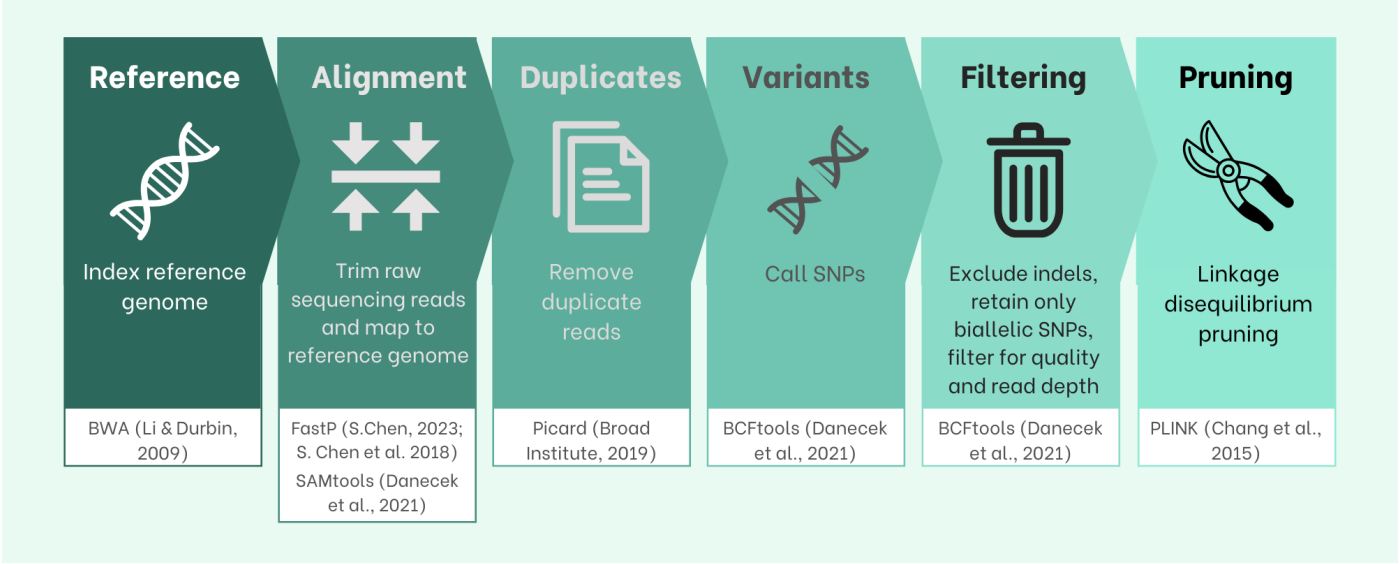


Figure 4.2: Bioinformatics pipeline outlining the processing steps between receiving raw sequencing data from Novogene and data prepared for downstream analysis.

Absolute values in each distance matrix varied drastically due to their different units of measurement, so these were scaled to facilitate more informative comparison in the following tests. As the findings of this study to assess the compatibility of trait and genetic data has implications for taxonomy and phylogenetics, hierarchical clustering dendrograms were constructed in R base package 'stats' for the trait and reduced genetic distance matrices using the complete linkage method. Although these are not true phylogenetic trees, they visualise relationships between species, and can be compared to assess whether the hierarchical clustering patterns from the datasets align. The cophenetic distance, a measure of the similarity required to cluster in the tree (Sneath & Sokal, 1973), was computed for each dendrogram, and the Pearson's correlation coefficient *(r*) between these distances calculated. This approach determines whether species that are genetically similar also cluster in the same way based on traits, in which case, *r* would tend towards one.

To visualise relationships between species in an alternate way, metric Multidimensional Scaling (MDS) was performed in R base package 'stats'. This reduces the dimensionality of the data whilst conserving the pairwise distances between species, and provides clear graphical representation of relatedness (Kruskal & Wish, 1978). Clustering of points indicates similarity between objects, and similar clustering of species within each dataset would be expected if the same relationships are described in both. This method can also highlight which species are driving dissimilarities in the case of differing relationships, and was completed with the trait, reduced genetic, and full genetic distance matrices.

To further assess whether the structure of the clusters persists across datasets, and also highlight species driving potential differences, the MDS results for the trait and reduced distance matrices were compared with a statistical shape analysis. A Procrustes analysis measures the level of spatial transformation required to minimise differences between datasets (Mardia et al., 1979), with values between zero and 0.2 for the Procrustes sum of squares (m^2), and close to one for the correlation in a symmetric Procrustes rotation (r), indicating perfect to good similarity. Significance was derived using 999 permutations. After transformation, the residual distance between each species and its counterpart in the other dataset highlighted whether the structure of the clusters remains unchanged; if the datasets portray the same relationships, movement of species within this space should be intra- rather than inter-cluster. This was conducted using R package 'vegan' v2.6-8 (Oksanen et al., 2024).

Finally, a further comparison of the overall similarity between the datasets was conducted by computing the RV coefficient, R_V (Robert & Escoufier, 1976), between the MDS results, using R package 'FactoMineR' v2.11 (Le et al., 2008). Designed to measure the degree of dissimilarity between multivariate datasets, coefficient values range between zero (no similarity between datasets) and one (perfect similarity). Much like the Mantel and cophenetic correlation tests, values close to one would indicate structural similarity between the genetic and trait data, indicating that we can infer the same relatedness between species from each.

All analyses were conducted in R version 4.4.1 (R Core Team, 2024), using RStudio version 2024.12.0.467 (Posit Team, 2024).

Results

The bioinformatics pipeline generated 66 genotypes and 208,765 binary SNPs, with 11.84% overall missing data. As there was more than one individual for some of the species, the full genetic distance matrix (Supplementary Table 4.1) returned values for some species compared to themselves (and NA for those with only one individual), which can be interpreted as the intra-species variation. As such, the top four most similar pairwise comparisons were intra-specific comparisons for *Catharsius biconifer*, *Catharsius machadoi*, *Catharsius luluensis*, and *Catharsius peregrinus*. However, *Catharsius satyrus* exhibited higher intra-specific variation than *C. luluensis* + *Catharsius orami* and *C. machadoi* + *Catharsius smithi* did with each other, and the comparisons of *C. dux*, *C. duciformis*, and *C. dux / C. duciformis* with themselves did not feature in the top ten most similar combinations. The genetic distance matrices used in all further calculations replaced intra-species variation values with zero for improved comparison with the trait distance matrix (Supplementary Table 4.2). When considering comparisons between two different species only, six of the top ten most similar pairings were shared between the genetic and trait matrices, albeit in a different order (Table 4.1). However, whilst the trait set showed the most similarity between the grouping of *C. dux*, *C. duciformis*, and *C. satyrus*, these did not feature in the top ten most similar pairings for the genetic set. Two of the top ten most dissimilar pairings were shared between the genetic and trait matrices (Table 4.2). Every entry on this list for the genetic dataset included at least one of *C. dux* or *C. duciformis*, whereas differences in the trait dataset found stronger dissimilarities between other species. The

Mantel test suggested a moderate to strong positive correlation between the trait and reduced genetic distance matrices (r = 0.5956, p = 0.001).

The dendrograms derived from the trait and genetic data were also positively correlated (r = 0.6658), both generating three clusters, each containing the same species (Figure 4.3). The same clusters were also identifiable in the results of the multidimensional scaling, not including *C. merrettorum* which is only found in the full genetic dataset (Figure 4.4). Differences between the MDS results of each dataset are minimal, and pertain to the distance of some species from their clusters: *C. satyrus* is more different from *C. dux* and *C. duciformis* in the genetic data than in the trait data, likewise *Catharsius peregrinus* from *C. luluensis*, *Catharsius pallas*, and *C. orami,* as well as *C. biconifer* from *Catharsius smithi* and *C. machadoi*. Overall, the structure of the clusters is consistent across both datasets.

There is a strong similarity between the shapes of the trait and reduced genetic datasets after MDS (m^2 = 0.07988, r = 0.9592, p = 0.001; R_V = 0.9018, p < 0.001), and residual distances in the Procrustes analysis between each species and its counterpart in the other dataset show only intra-cluster movement and, so, support a consistent overall structure across both datasets (Figure 4.5).

Table 4.1: The ten most similar pairwise comparisons between Catharsius *species as described, respectively, by the genetic (scaled Euclidean distance) and morphological trait (scaled Gower's distance) data. Fill colours link a species pairing with the same pairing in the other dataset to highlight relative position in the top ten. Numbers in parentheses and greyed out species names are species pairings that include at least one species from the full genetic dataset that is not present in the trait dataset.*

(Most similar)	**Genetic**		**Traits**
1	*C. luluensis + C. orami*	**1**	*C. dux + C. duciformis*
2	*C. smithi + C. machadoi*	**2**	*C. dux + C. satyrus*
3	*C. orami + C. pallas*	**3**	*C. duciformis + C. satyrus*
4	*C. orami + C. peregrinus*		*= C. luluensis + C. pallas*
5	*C. luluensis + C. pallas*	**5**	*C. orami + C. pallas*
6	*C. luluensis + C. peregrinus*	**6**	*C. luluensis + C. orami*
7	*C. peregrinus + C. pallas*	**7**	*C. smithi + C. machadoi*
(8)	*C. merrettorum + C. pallas*	**8**	*C. smithi + C. biconifer*
(9)	*C. merrettorum + C. orami*	**9**	*C. peregrinus + C. pallas*
(10)	*C. merrettorum + C. luluensis*	**10**	*C. orami + C. peregrinus*
8	*C. pallas + C. smithi*		
9	*C. orami + C. smithi*		
10	*C. pallas + C. machadoi*		

Table 4.2: The ten most dissimilar pairwise comparisons between Catharsius *species as described, respectively, by the genetic (scaled Euclidean distance) and morphological trait (scaled Gower's distance) data. Fill colours link a species pairing with the same pairing in the other dataset to highlight relative position in the top ten. Numbers in parentheses and greyed out species names are species pairings that include at least one species from the full genetic dataset that is not present in the trait dataset.*

(Most dissimilar)	**Genetic**		**Traits**
1	*C. machadoi + C. dux*	**1**	*C. peregrinus + C. satyrus*
2	*C. machadoi + C. duciformis*	**2**	*C. peregrinus + C. dux*
3	*C. biconifer + C. dux*		*= C. peregrinus + C. duciformis*
4	*C. biconifer + C. duciformis*	**4**	*C. satyrus + C. luluensis*
(5)	*C. machadoi + C. dux / duciformis*	**5**	*C. orami + C. machadoi*
(6)	*C. biconifer + C. dux / duciformis*	**6**	*C. satyrus + C. pallas*
5	*C. smithi + C. dux*		*= C. luluensis + C. machadoi*
6	*C. smithi + C. duciformis*		*= C. luluensis + C. duciformis*
(7)	*C. smithi + C. dux / duciformis*		*= C. luluensis + C. dux*
7	*C. peregrinus + C. duciformis*		*= C. peregrinus + C. biconifer*
8	*C. peregrinus + C. dux*		*= C. pallas + C. machadoi*
(8)	*C. peregrinus + C. dux / duciformis*		
(9)	*C. merrettorum + C. dux*		
(10)	*C. merrettorum + C. duciformis*		
9	*C. pallas + C. dux*		
10	*C. pallas + C. duciformis*		

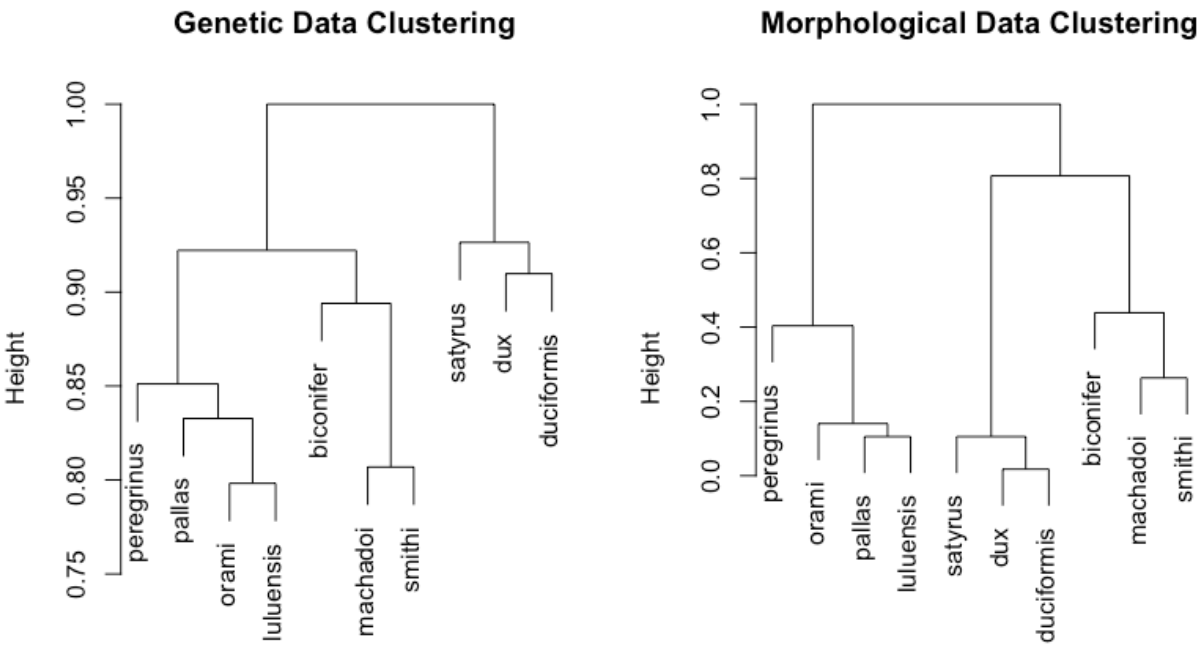


Figure 4.3: Hierarchical clustering dendrograms derived from genome-wide Catharsius *SNP data (left) and morphological traits (right) representing the inferred relationships between individuals or populations based on similarity. Congruence between the two trees highlights the extent to which variations in morphology reflect underlying genetic structure. On the y-axis, "Height" refers to the maximum distance between two clusters, as the complete linkage method was used. Dissimilarity in the genetic data was measured using Euclidean distance, and dissimilarity in the morphological data was measured using Gower's distance, both scaled to facilitate comparison.*

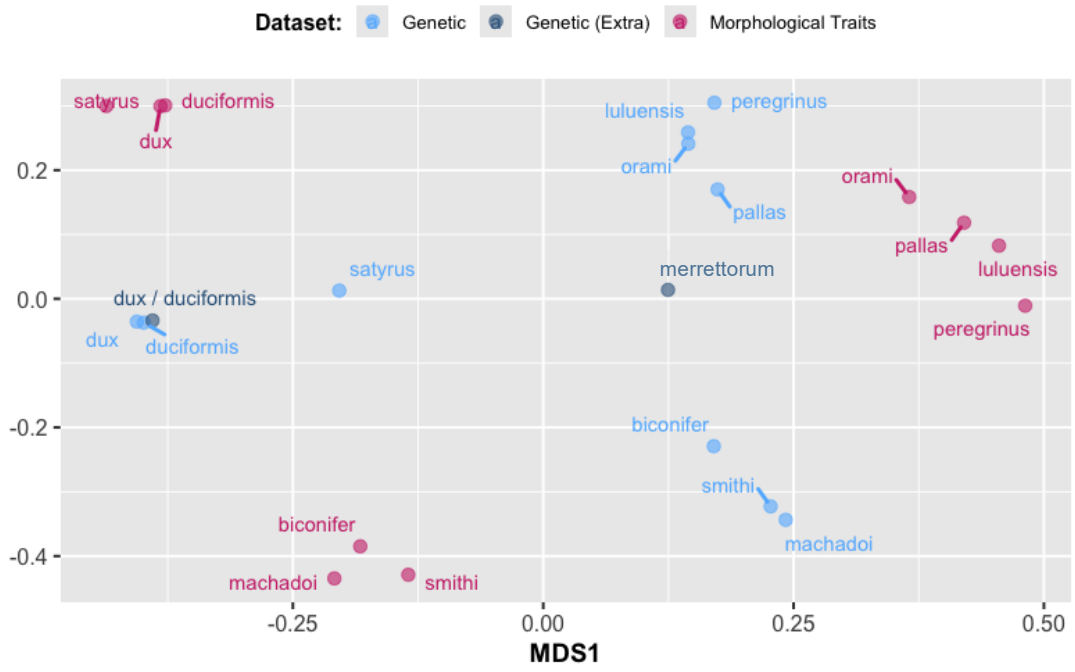


Figure 4.4: Metric multidimensional scaling (MDS) plot comparing patterns of similarity based on Catharsius *genetic and morphological trait data. Each point represents a species, positioned according to pairwise dissimilarities, and their relative positions within each dataset illustrate how relatedness between individuals and clusters is described by each data type.*

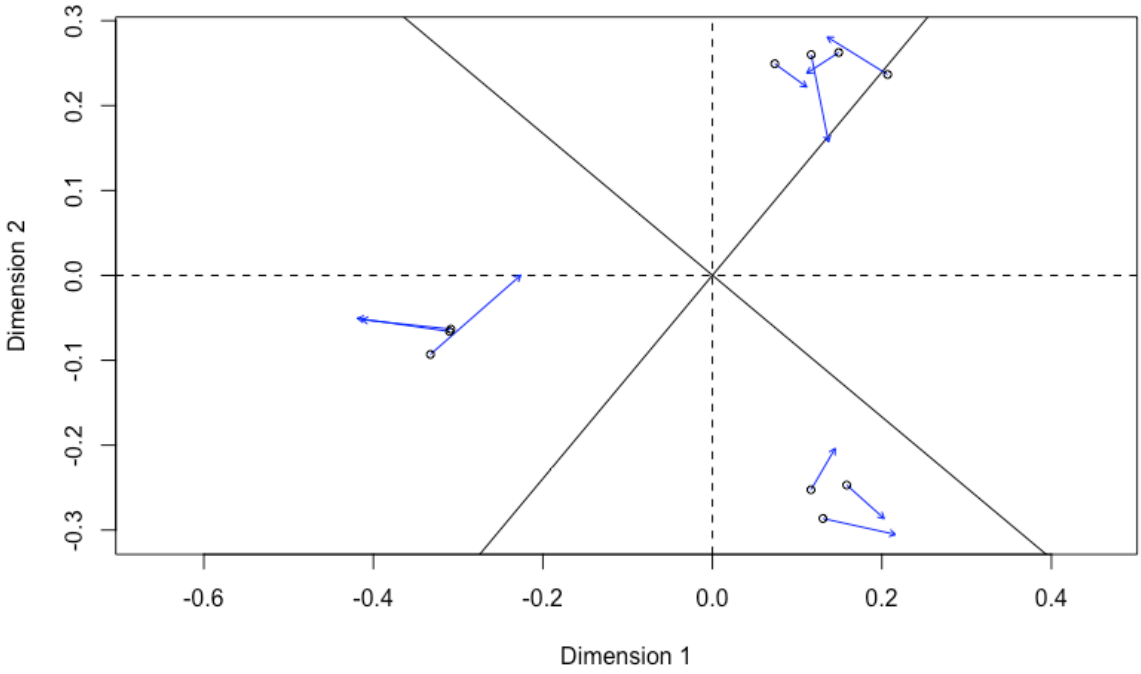


Figure 4.5 Errors plot showing the residual distances between corresponding points in Catharsius *genetic and morphological datasets after Procrustes analysis. Each line connects the position of an individual in the two datasets, with longer lines indicating greater mismatch. The grouping in the lower left quadrant is* C. dux, C. duciformis*, and* C. satyrus*, in the upper right quadrant is* C. orami, C. peregrinus, C. luluensis*, and* C. pallas*, and in the lower right is* C. biconifer, C. smithi*, and* C. machadoi.

Discussion

This study contributes to the ongoing debate surrounding the merits of morphological and molecular data in studies of taxonomy, phylogenetics, and population structure. By comparing distance matrices and species clustering using both morphological and molecular datasets, this comparative approach assesses the degree to which these different data types agree in the depiction of species boundaries and relationships. In particular, it has shown that the use of whole genome SNPs can capture phenotypic variation in such a way that molecular markers have been found not to (Kadoić Balaško et al., 2021). The overall structure and species clustering was consistent between datasets, but relatedness within and between clusters varied in its magnitude,

supporting existing literature that maintains these types of data are complementary, and a “total evidence” approach encompassing them both is the best course of action.

Molecular and morphological data have often been found to correlate very little in studies of intra- and inter-specific diversity (A. A. Alves et al., 2013; R. M. Alves et al., 2017; E. K. V. D. Andrade et al., 2017; Darkwa et al., 2020; Kadoić Balaško et al., 2021; Keating et al., 2023; Van Den Ende et al., 2023). Our results instead find that relatedness and species clustering derived from SNPs agree with those captured by a morphological trait matrix, shown in the strong similarities between the distance matrices, dendrograms, and shape of relationships within the MDS. The overall structure of the population is preserved between the two datasets, driven by similarities rather than dissimilarities: the top ten most similar species pairings from the distance matrices found more commonality than the top ten most dissimilar, which only shared two pairings, and even then in notably different places. This is corroborated by the dendrograms which recovered the same groupings, the MDS comparison—upon which the same clusters of species can be identified—and the Procrustes errors which showed only intra-cluster movement.

As speculated by Alves et al. (2013) and Alves et al. (2017), it is possible that the correlations described above can be explained by the use of whole genome SNPs rather than molecular markers, such as random DNA polymorphisms and microsatellites, which are poor at predicting phenotype. Although refuted by Darkwa et al. (2020) who found groupings of white Guinea yam accessions to differ between data types, this agrees with Kadoić Balaško et al. (2021), who found a high correlation between wing

shape and genetic structure in western corn rootworm. This disagreement with Darkwa et al. (2020) could be explained by their retaining 136,429 SNPs for analysis, whereas this study uses 208,765, whilst ability to predict phenotypic traits is stable only down to 150,000 (Ober et al., 2012). That said Kadoić Balaško et al. (2021) used a much smaller set of 7125 and still found good correlation, so this may not entirely explain the incongruity in results.

Despite the notable correlations between the datasets, this study does agree with the consensus that they do not confer identical information. Whilst the overall structure is maintained across the datasets, i.e. they both find the same clusters, the degree of inter- and intra-cluster relatedness differs. Although they are correlated overall, the distance matrices generally do not agree on the most dissimilar species pairings, and the dendrograms – despite grouping the same species on each main branch – differ in the positioning of these branches. In the genetic data, the *C. biconifer* clade is more closely related to that of *C. peregrinus*, whereas in the morphological data, it is closer to that of *C. satyrus*. This difference in degree of relatedness is also seen within clades, such as *C. biconifer* and *C. peregrinus* from their clusters on the dendrograms, and *C. satyrus* from its cluster on the MDS comparison. The Procrustes errors depict only intra-cluster movement, underlining that where the datasets can't be matched after transformation is in how related the species within clusters are to each other. Notably, the morphological dataset finds a much greater degree of difference between species and clusters than does the genetic dataset, as made clear by the differing heights of the dendrograms in Figure 4.3. Even after scaling, the former delimits much stronger boundaries between species and especially between their groupings.

Changes in genotype are thought to take longer to become established than those in phenotype, which has led some to argue that morphological traits are a better, as well as more affordable, way to track short term population changes (Kadoić Balaško et al., 2021). However, it is this asynchronous divergence that is thought to be partly responsible for the differences found between molecular and morphological clustering, even when using SNPs (Darkwa et al., 2020). This could explain both the differences in branch positioning in the dendrograms here, as well as the notably larger distances between clusters and species in the morphological dataset, and agrees with previous research in which more close relatedness was found between genotypic rather than phenotypic pairs (A. A. Alves et al., 2013). In fact, this phenomenon neatly encapsulates both the similarities and differences found in this study; although both types of data are capturing the same overall variation, that it is happening at different rates explains why it is measured at different magnitudes.

Of course, it is likely that the ratio of 130 morphological traits to >208,000 SNPs used here is also partly responsible for the larger and more defined absolute differences between species described by the former. The sheer number of SNPs enables them to capture much finer-scale variation. Whether or not this is an advantage depends on the overall goal; perhaps morphological data draw boundaries that are unrealistically defined, and are not able to capture the spectrum of genetic diversity on a precise enough level? On the other hand, that human desire to neatly categorise everything has meant conservation initiatives are largely based on “species”, perhaps means that the more difficult nature of defining boundaries along a spectrum makes it

counterproductive. This would especially be the case if reproductive isolation is not complete and continued gene flow complicates distinguishing ecological and / or behavioural groups whose boundaries may be better reflected by phenotype.

It is well-supported, then, that a total evidence approach integrating both phenotypic and molecular data provides a more comprehensive description of evolutionary history, better capturing the distinctive selective pressures and timescales of morphological and molecular evolution. However, if genetic information is critical to the accurate description of a species, and its placement within a phylogenetic tree, a challenge arises. The ability to visually identify a species is critical in many branches of biology, including in field- and museum-based work, but if genetic resources are needed and not available, these species become functionally invisible in those settings. In a parallel line of thinking, Turton-Hughes et al. (2024) introduce the term “shadow diversity”, highlighting the importance of unknown and undescribed species, underlining the complex web of socio-cultural and technological barriers that prevent them being recorded. Furthermore, they describe the self-perpetuating cycle in which these unknown unknowns remain unrecognised or under-recorded, and are therefore unable to draw research interest and conservation action, thus remaining unrecognised, and so forth. Although they underline that technological advancement alone is not sufficient to fully illuminate shadow diversity, eDNA is suggested as an example to combat one element of the problem: “If we cannot sense it, we cannot measure it” (Turton-Hughes et al., 2024, p. 14). In a sense, this is aligned with the challenge presented above but, whereas the innate human limitation they describe is our struggle to reach the absolute boundaries of knowledge, the limitation described here is our inability to perceive

information within these boundaries in environments without access to the technological tools with which to do so. We propose that another dimension to shadow diversity is those species that remain “dark” in situations where the only tool with which to perceive them is visual identification. As the importance of both digitising natural history collections and field observations, as well as mobilising them on online databases for use in biodiversity modelling, are increasingly well understood (Blades et al., 2025; Heberling et al., 2021; Popov et al., 2021), these records similarly contribute to the cycle of under-recording described by Turton-Hughes et al. (2024).

It is possible, though, to bridge this disconnect between the pursuit of perfection and what is achievable by extending the scope of the integrated approach in systematics. By mapping morphological traits onto a resolved total evidence tree, combinations of diagnostic characters can be identified to assist future visual identification (Hernández-Lara et al., 2018). This capitalises on the benefit of using both phenotypic and genotypic data to describe species and their relationships, whilst also acknowledging the importance of visual identification. It is my recommendation that this step be added as standard to any study employing a total evidence methodology.

It would be remiss to not acknowledge that a major barrier to achieving this is declining support for taxonomy, and the increasingly short supply of resources and expertise with which to carry out these studies in the first place (Engel et al., 2021). This, combined with conservative estimates of almost seven million undescribed species requiring study before they are lost, is known as the ‘taxonomic impediment’ (Bernard, 2025). Advances in artificial intelligence (AI) are being explored as a possible way to overcome

these challenges, and the potential of using machine learning algorithms to identify diagnostic features in species clustering from specimen images has been highlighted (Bernard, 2025; Karbstein et al., 2024). If this can be implemented successfully, it would help to both combat the taxonomic impediment more quickly, and bridge the gap between what can be identified genetically and visually. However, there are a number of unresolved hurdles in the application of AI in taxonomy, and there is disagreement as to the extent to which it should be employed. Some argue that automating data preparation processes such as taking morphometric measurements and translating existing revisions should be the limit for now, and that taxonomists can make use of the time this frees up by continuing their work as is (Valdecasas, 2024). Others say that traditional taxonomy will be required for many years to train AI to a point where it is useful, and also to validate outputs (Bernard, 2025), whilst there are also those that are exploring ways for artificial intelligence to conduct the taxonomic work itself, but agree that there are still some unanswered problems and again that taxonomists must be involved to validate outputs (Karbstein et al., 2024). What underlies this debate is that a "species" is fundamentally a human concept, so some degree of subjectivity is probably a necessary evil, and the importance of a taxonomist's intuition for their study taxa should not be understated. Whether this is used for generating training datasets for AI and validating its outputs, or continued work done by hand, we must acknowledge the importance of this discipline and recognise the need to reinstate funding and training (D. L. Pearson et al., 2011; Wägele et al., 2011).

The unprecedented rate of biodiversity loss is now thought to rival better acknowledged drivers of environmental change, and negatively impact both ecosystem functioning

and services (Cardinale et al., 2012). What is perhaps even more concerning, are those losses that we are not yet able to measure, especially as these undescribed species are likely at an even higher risk of extinction (J. Liu et al., 2022). Furthermore, species with common names and IUCN Red List assessments are more likely to receive scientific and societal interest (Mammola et al., 2023), which cannot be achieved without first being described. Technological advances in DNA sequencing have helped to combat this, improving our understanding of the evolution of life (Oyston et al., 2022; Van Den Ende et al., 2023), and opening doors to identifying cryptic species (Sales et al., 2018) and previously unrecognised genetic diversity (A. A. Alves et al., 2013). That said, with concurrent improvements in technology such as record digitisation and growth of online databases, reliance on molecular data threatens our ability to accurately record biodiversity in situations without access to these resources. This study supports findings that morphological traits are able to capture the overall genetic structure described by SNPs, although also that their information is complementary rather than identical. It underlines that the integrated approach to studies of taxonomy, phylogenetics, and diversity should be extended to highlight combinations of diagnostic traits to bridge the gap between the ideal and the accessible. Whether or not developments in AI and machine learning are used to facilitate a faster work rate, the importance of funding taxonomy cannot be understated. Even optimistic outlooks suggesting that the world's undescribed species are more likely to be described than go extinct underline the importance of investing in taxonomy to achieve this (Costello et al., 2013).

References

Alves, A. A., Bhering, L. L., Rosado, T. B., Laviola, B. G., Formighieri, E. F., & Cruz, C. D. (2013). Joint analysis of phenotypic and molecular diversity provides new insights on the genetic variability of the Brazilian physic nut germplasm bank. *Genetics and Molecular Biology*, *36*(3), 371–381. https://doi.org/10.1590/S1415-47572013005000033

Alves, R. M., Silva, C. R. D. S., Albuquerque, P. S. B. D., & Santos, V. S. D. (2017). Phenotypic and genotypic characterization and compatibility among genotypes to select elite clones of cupuassu. *Acta Amazonica*, *47*(3), 175–184. https://doi.org/10.1590/1809-4392201602104

Andrade, E. K. V. D., Andrade Júnior, V. C. D., Laia, M. L. D., Fernandes, J. S. C., Oliveira, A. J. M., & Azevedo, A. M. (2017). Genetic dissimilarity among sweet potato genotypes using morphological and molecular descriptors. *Acta Scientiarum. Agronomy*, *39*(4), 447. https://doi.org/10.4025/actasciagron.v39i4.32847

Bernard, J. (2025). Combining new technology with classic taxonomy to overcome hurdles to discovering dark taxa. *Systematics and Biodiversity*, *23*(1), 2454014. https://doi.org/10.1080/14772000.2025.2454014

Blades, B., Ronquillo, C., & Hortal, J. (2025). Mobilisation of Data From Natural History Collections Can Increase the Quality and Coverage of Biodiversity Information. *Ecology and Evolution*, *15*(4), e71139. https://doi.org/10.1002/ece3.71139

Broad Institute. (2019). *Picard Toolkit* [Computer software]. https://broadinstitute.github.io/picard/

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., Wardle, D. A., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S., & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), 59–67. https://doi.org/10.1038/nature11148

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, T., Xu, J., Wang, L., Wang, H., You, E., Deng, C., Bian, H., & Shen, Y. (2023). Landscape genomics reveals adaptive genetic differentiation driven by multiple environmental variables in naked barley on the Qinghai-Tibetan Plateau. *Heredity*, *131*(5–6), 316–326. https://doi.org/10.1038/s41437-023-00647-0

Costello, M. J., May, R. M., & Stork, N. E. (2013). Can We Name Earth's Species Before They Go Extinct? *Science*, *339*(6118), 413–416. https://doi.org/10.1126/science.1230318

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Darkwa, K., Agre, P., Olasanmi, B., Iseki, K., Matsumoto, R., Powell, A., Bauchet, G., De Koeyer, D., Muranaka, S., Adebola, P., Asiedu, R., Terauchi, R., & Asfaw, A. (2020). Comparative assessment of genetic diversity matrices and clustering methods in white Guinea yam (*Dioscorea rotundata*) based on morphological and molecular markers. *Scientific Reports*, *10*(1), 13191. https://doi.org/10.1038/s41598-020-69925-9

Engel, M. S., Ceríaco, L. M. P., Daniel, G. M., Dellapé, P. M., Löbl, I., Marinov, M., Reis, R. E., Young, M. T., Dubois, A., Agarwal, I., Lehmann A., P., Alvarado, M., Alvarez, N., Andreone, F., Araujo-Vieira, K., Ascher, J. S., Baêta, D., Baldo, D., Bandeira, S. A., … Zacharie, C. K. (2021). The taxonomic impediment: A shortage of taxonomists, not the lack of technical approaches. *Zoological Journal of the Linnean Society*, *193*(2), 381–387. https://doi.org/10.1093/zoolinnean/zlab072

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857. https://doi.org/10.2307/2528823

Gustafson, G. T., Glynn, R. D., Short, A. E. Z., Tarasov, S., & Gunter, N. L. (2023). To design, or not to design? Comparison of beetle ultraconserved element probe

set utility based on phylogenetic distance, breadth, and method of probe design. *Insect Systematics and Diversity*, *7*(4), 4. https://doi.org/10.1093/isd/ixad014

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, *118*(6), e2018093118. https://doi.org/10.1073/pnas.2018093118

Hernández-Lara, C., Espinosa De Los Monteros, A., Ibarra-Cerdeña, C. N., García-Feria, L., & Santiago-Alarcon, D. (2018). Combining morphological and molecular data to reconstruct the phylogeny of avian Haemosporida. *International Journal for Parasitology*, *48*(14), 1137–1148. https://doi.org/10.1016/j.ijpara.2018.10.002

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr521

Kadoić Balaško, M., Mikac, K. M., Benítez, H. A., Bažok, R., & Lemic, D. (2021). Genetic and Morphological Approach for Western Corn Rootworm Resistance Management. *Agriculture*, *11*(7), 585. https://doi.org/10.3390/agriculture11070585

Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., Wagner, N. D., Emerson, B. C., Albach, D. C., Scheu, S., Bradler, S., De Vries, J., Irisarri, I., Li, H., Soltis, P., Mäder, P., & Wäldchen, J. (2024). Species delimitation 4.0: Integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution*, *39*(8), 771–784. https://doi.org/10.1016/j.tree.2023.11.002

Keating, J. N., Garwood, R. J., & Sansom, R. S. (2023). Phylogenetic congruence, conflict and consilience between molecular and morphological data. *BMC Ecology and Evolution*, *23*(1), 30. https://doi.org/10.1186/s12862-023-02131-z

Kruskal, J., & Wish, M. (1978). *Multidimensional Scaling*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985130

Le, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Lee, M. S. Y., & Palci, A. (2015). Morphological Phylogenetics in the Genomic Age. *Current Biology*, *25*(19), R922–R929. https://doi.org/10.1016/j.cub.2015.07.009

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liu, J., Slik, F., Zheng, S., & Lindenmayer, D. B. (2022). Undescribed species have higher extinction risk than known species. *Conservation Letters*, *15*(3), e12876. https://doi.org/10.1111/conl.12876

Losos, J. B. (1996). Phylogenies and comparative biology, Stage II: Testing causal hypotheses derived from phylogenies with data from extant taxa. *Systematic Biology*, *45*(3), 259–260.

Lyubetsky, V., Piel, W. H., & Quandt, D. (2014). Current Advances in Molecular Phylogenetics. *BioMed Research International*, *2014*, 1–2. https://doi.org/10.1155/2014/596746

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2023). *Cluster: Cluster Analysis Basics and Extensions. R package version 2.1.6.* [Computer software].

Mammola, S., Adamo, M., Antić, D., Calevo, J., Cancellario, T., Cardoso, P., Chamberlain, D., Chialva, M., Durucan, F., Fontaneto, D., Goncalves, D., Martínez, A., Santini, L., Rubio-Lopez, I., Sousa, R., Villegas-Rios, D., Verdes, A., & Correia, R. A. (2023). Drivers of species knowledge across the tree of life. *eLife*, *12*, RP88251. https://doi.org/10.7554/eLife.88251

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

Ober, U., Ayroles, J. F., Stone, E. A., Richards, S., Zhu, D., Gibbs, R. A., Stricker, C., Gianola, D., Schlather, M., Mackay, T. F. C., & Simianer, H. (2012). Using Whole-Genome Sequence Data to Predict Quantitative Trait Phenotypes in *Drosophila melanogaster*. *PLoS Genetics*, *8*(5), e1002685. https://doi.org/10.1371/journal.pgen.1002685

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M. D., Durand, S., … Weedon, J. (2024). *vegan: Community Ecology R package version 2.6-8* [Computer software]. https://CRAN.R-project.org/package=vegan

Oyston, J. W., Wilkinson, M., Ruta, M., & Wills, M. A. (2022). Molecular phylogenies map to biogeography better than morphological ones. *Communications Biology*, *5*(1), 521. https://doi.org/10.1038/s42003-022-03482-x

Pearson, D. L., Hamilton, A. L., & Erwin, T. L. (2011). Recovery Plan for the Endangered Taxonomy Profession. *BioScience*, *61*(1), 58–63. https://doi.org/10.1525/bio.2011.61.1.11

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347–352), 240–242. https://doi.org/10.1098/rspl.1895.0041

Popov, D., Roychoudhury, P., Hardy, H., Livermore, L., & Norris, K. (2021). The Value of Digitising Natural History Collections. *Research Ideas and Outcomes*, *7*, e78844. https://doi.org/10.3897/rio.7.e78844

Posit Team. (2024). *RStudio: Integrated Development Environment for R* [Computer software]. Posit Software, PBC. http://www.posit.co/

Purcell, S., & Chang, C. (n.d.). *PLINK 2.0* [Computer software]. www.cog-genomics.org/plink/2.0/

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* [Computer software]. https://www.R-project.org/

Robert, P., & Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods: The RV- Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *25*(3), 257–265.

Sales, N. G., Mariani, S., Salvador, G. N., Pessali, T. C., & Carvalho, D. C. (2018). Hidden Diversity Hampers Conservation Efforts in a Highly Impacted Neotropical River System. *Frontiers in Genetics*, *9*, 271. https://doi.org/10.3389/fgene.2018.00271

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W. H. Freeman and Co.

Takano, H. (2018). *A systematic revision of the Afrotropical members of the dung beetle genus* Catharsius *Hope, 1837 (Coleoptera: Scarabaeidae)* [DPhil Thesis]. University of Oxford.

Takano, H. (2025). A systematic revision of the Afrotropical members of the dung beetle genus *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidae). [Manuscript in Preparation]. *African Natural History Research Trust*.

Taylor, P. J., Denys, C., & Cotterill, F. P. D. (Woody). (2019). Taxonomic anarchy or an inconvenient truth for conservation? Accelerated species discovery reveals evolutionary patterns and heightened extinction threat in Afro-Malagasy small mammals. *Mammalia*, *83*(4), 313–329. https://doi.org/10.1515/mammalia-2018-0031

Turton-Hughes, S., Holmes, G., & Hassall, C. (2024). The diversity of ignorance and the ignorance of diversity: Origins and implications of "shadow diversity" for conservation biology and extinction. *Cambridge Prisms: Extinction*, *2*, e18. https://doi.org/10.1017/ext.2024.21

Valdecasas, A. G. (2024). Can Taxonomists Think? Reversing the AI Equation. *Taxonomy*, *4*(4), 713–722. https://doi.org/10.3390/taxonomy4040037

Van Den Ende, C., Puttick, M. N., Urrutia, A. O., & Wills, M. A. (2023). Why should we compare morphological and molecular disparity? *Methods in Ecology and Evolution*, *14*(9), 2390–2410. https://doi.org/10.1111/2041-210X.14166

Wägele, H., Klussmann-Kolb, A., Kuhlmann, M., Haszprunar, G., Lindberg, D., Koch, A., & Wägele, J. W. (2011). The taxonomist—An endangered race. A practical proposal for its survival. *Frontiers in Zoology*, *8*(1), 25. https://doi.org/10.1186/1742-9994-8-25

Zamani, A., Vahtera, V., Sääksjärvi, I. E., & Scherz, M. D. (2021). The omission of critical data in the pursuit of 'revolutionary' methods to accelerate the description of species. *Systematic Entomology*, *46*(1), 1–4. https://doi.org/10.1111/syen.12444

CHAPTER 5

Not the be all and environm-end all: landscape genomics hints that biotic interactions may override environmental adaptations even in strong bioindicator species

In Preparation

Bryony Blades[1,2], Andrea Estandía[1]
[1]Department of Biology, University of Oxford, Oxford, UK
[2]African Natural History Research Trust, Kingsland, Herefordshire, UK

Abstract

Uncovering the spatial distributions of genetic diversity is key in elucidating evolutionary processes, predicting how life on Earth will respond to environmental change, and informing targeted conservation strategies. Facilitated by next-generation sequencing (NGS), researchers are now able to capitalise on more accessible sets of genetic loci, that are orders of magnitude more extensive than before, to analyse patterns of adaptation and selection across the landscape. The study of Coleoptera using these methods has hitherto focused on agricultural and forest pests and, despite their importance to ecology and ecosystem services, no landscape genomics studies on true dung beetles (Scarabaeinae) have sought to evaluate the extent to which climatic

adaptation drives their distributions. With whole genome sequences of *Catharsius* Hope, 1837 (Coleoptera: Scarabaeidae) specimens collected in Zambia, this study aims to assess whether adaptation to environmental conditions is responsible for maintaining the parapatric distributions of two closely-related species. Single nucleotide polymorphisms (SNPs) are used in partial redundancy analyses and a latent factor mixed model to test the hypothesis that genetic diversity across *Catharsius dux* and *Catharsius duciformis* can be explained by geographic and environmental variation. To contextualise their genetic diversity in the context of their genus, SNPs from other *Catharsius* species are integrated in analyses of current and ancestral population structure. Ultimately, a genetic continuum including all *C. dux* and *C. duciformis* individuals is uncovered that cannot be well explained by environment, collecting location, or species identification, suggesting that a possible hybrid zone in this area is responsible for maintaining their parapatry rather than the current landscape.

Introduction

Dung beetles are effective bioindicators of biodiversity and ecosystem health at a variety of spatial scales, and provide globally important ecosystem services (Beynon et al., 2015; A. L. V. Davis et al., 2004; Nichols et al., 2008; Spector, 2006). As such, they have been studied extensively, covering broad topics of agriculture and biology, ecological functions, taxonomy, seed dispersal, and sexual selection and traits (Hemmings et al., 2025). Given both their sensitivity to environmental change and ecological importance, they are ideal models for studying how environmental variation shapes diversity and adaptation, but no studies have yet assessed the climatic drivers of genomic adaptation in any true dung beetles (Scarabaeinae). This is surprising given the disproportionate amount of research relative to the size of the group, and their importance to agriculture and ecology, and it is hoped that analysing to what degree climate drives variation will add an important dimension to our understanding of patterns in biodiversity.

The parapatric distributions of two Scarabaeinae species *Catharsius dux* and *Catharsius duciformis* were recently highlighted in a revision of the Afrotropical members of their genus (Takano, 2018). It is not clear what maintains the barrier between the distributions of these close relatives but, given their somewhat distinctive climatic niches (Blades, 2025) and dung beetles' sensitivity to their environment, we can hypothesise that adaptation to local climate plays a role. Genetic sequencing has provided tools to investigate such relationships between species and their environments, as well as how genetic diversity is distributed in space, elucidating the mechanisms behind observed distributions. Using these tools, we can also predict how

populations may respond to future environmental changes and inform targeted conservation strategies that integrate species' evolutionary potential (Capblancq et al., 2020; Funk et al., 2019; Sgrò et al., 2011).

When sequencing data was first incorporated in spatial studies, 'landscape genetics' methods typically used a handful of loci to understand the impact of landscape features on gene flow and population structure (Storfer et al., 2018). However, the advent of next generation sequencing in 2005 improved access to large quantities of financially accessible sequencing data (Mardis, 2017) and, combined with technological advances in remote sensing, climate modelling, and ecosystem observation, studies have moved towards 'landscape genomics' (Dauphin et al., 2023). By comparison, this uses thousands or even millions of loci, and often whole genomes, to understand spatial patterns of selection and adaptation (Storfer et al., 2018).

Insect landscape genomics studies have mirrored general trends, increasing since the early 2000s, and moving from the use of genetic microsatellites to single nucleotide polymorphisms (SNPs) since 2015 (Chaulk & Keyghobadi, 2022). Applied studies have focused on species of conservation concern, pests, and disease vectors. Coleoptera feature heavily in studies on both agricultural and forest pests, including on the mountain pine (*Dendroctonus ponderosae*), Colorado potato (*Leptinotarsa decemlineata*), and spruce bark (*Ips typographus*) beetles (Chaulk & Keyghobadi, 2022; Mykhailenko et al., 2024). Landscape genomics approaches have also recently been used to show winter-associated variation in the willow leaf beetle (*Chrysomela aeneicollis*) (Keller et al., 2023). It appears that just one landscape genomics study has been carried out so far on Scarabaeinae but, as this assesses the relationship between

habitat fragmentation and genetic diversity (González-Molina et al., 2024), to the best of my knowledge, the current study is the first to assess climatic drivers of genomic adaptation in any true dung beetles.

The parapatric distributions of *C. dux* and *C. duciformis* come closest in Zambia's North-Western Province. This region is characterised by a mosaic of tropical savannah, and temperate areas with both warm and hot summers (H. E. Beck et al., 2018). The forest-savannah mosaic that results from the confluence of these climate types fosters rich diversity, especially in the Ikelenge Pedicle in its north-west, which has been described as a regionally and globally important biodiversity hotspot (Cotterill, 2002). The climatic heterogeneity in this region provides an opportunity to improve understanding of the relationship between genomic diversity and environmental or geographical factors. Specifically, this study will use landscape genomics techniques to analyse SNPs from whole genomes of *C. dux*, *C. duciformis*, and other *Catharsius* species that we were able to collect on fieldwork to assess the extent to which climatic and edaphic adaptation shapes patterns of genetic diversity and drives their distributions. Ultimately, a genetic continuum containing all specimens of both *C. dux* and *C. duciformis* is uncovered which cannot be explained by environment or species identification, suggesting a possible hybrid zone.

Methods

Summary

This study hypothesises that adaptation to environmental conditions maintains the strict parapatry of *C. dux* and *C. duciformis*. The following steps are used to test this, and generate an alternative hypothesis upon its disproval.

After sample collection and DNA sequencing, bioinformatics steps prepared the raw sequence data for analysis by mapping them to a reference genome, calling SNPs, and filtering these for quality. The overall genetic structure of all sequenced species was summarised using a Principal Components Analysis (PCA) of the filtered SNPs. To determine if adaptation to the environment drives their distributions, the influence of landscape variables on the genetic diversity of *C. dux* and *C. duciformis* was evaluated using partial redundancy analyses (pRDAs) and a latent factor mixed model (LFMM). This influence was found to be weak, disproving the main hypothesis and motivating further analysis of their population structure. To compare how different *C. dux* and *C. duciformis* were to each other with how different they were to less closely-related species, sequences of other *Catharsius* species were then included in an analysis of wider population structure using *k*-means clustering, a Discriminant Analysis of Principal Components (DAPC), a calculation of pairwise similarity based on the proportion of shared alleles, and an ADMIXTURE analysis. Finally, to determine if the population structure of *C. dux* and *C. duciformis,* as described by these analyses, was

in any way driven by geographical space, the results of the DAPC were tested for a relationship with collecting location.

All analyses in R were conducted with R version 4.5.0 (R Core Team, 2025), using RStudio version 2024.12.1.563 (Posit team, 2025).

Whole genome sequencing

Sample collection, identification, and DNA sequencing

A total of 66 *Catharsius* dung beetles were collected from locations in the North-Western Province of Zambia (Figure 5.1, and denoted by the number between "Z" and "_" in specimen identifiers) as the study samples and identified to species-level by Dr. Hitoshi Takano, the expert taxonomist responsible for the revision of the genus (Takano, 2018). Individuals were mixed and identified in a random order with no knowledge of the collecting location to prevent identification being informed in any way by where, or with which other species, each beetle was found. It is difficult to distinguish between species with females and minor males of *C. dux* and *C. duciformis* without dissection of their genitalia and, in these cases, 29 specimens were identified as *C. dux / C. duciformis*. As there were 28 other specimens (18 and 10, respectively) identified to just one of each of these species, a negative impact on downstream analyses was not anticipated, and identification accuracy over precision thus prioritised. For each specimen, DNA was extracted from the femur of the hind leg using Qiagen DNeasy Blood and Tissue Kits and samples sent to Novogene. Some 49 of the samples were found to be contaminated with RNA, so were put through a DNA purification process at

Novogene, and whole genomes were sequenced at 10X depth using Illumina NovaSeq X Plus.

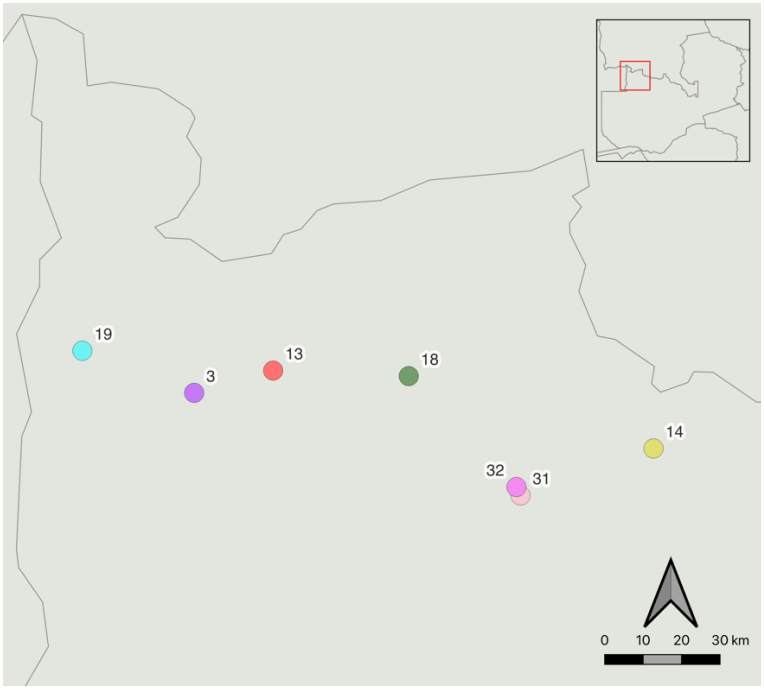


Figure 5.1: Catharsius *collecting locations in the North-Western Province of Zambia, labelled with their number (e.g. Collecting Location 14). Colours are to distinguish these from one another, and correspond with Table 5.2.*

Bioinformatics

Mapping raw sequencing data, calling SNPs, and filtering

Raw sequencing data must be first mapped to a reference genome and SNPs called and then filtered for quality before use in further analyses. As illustrated in the pipeline in Chapter 4, the most closely related reference genome available, *Copris fidius* (JAUIMS000000000; Gustafson et al., 2023), was indexed with BWA module v0.7.17 (H. Li & Durbin, 2009). Then, the first ten bases were trimmed from raw sequencing reads using FastP (S. Chen, 2023; S. Chen et al., 2018) and mapped to the reference genome using SAMtools v1.14 (Danecek et al., 2021). Duplicate reads in the resulting BAM files

were removed using Picard v3 (Broad Institute, 2019), and SNP calling performed using BCFtools v1.14 (Danecek et al., 2021). Indels were then excluded and only biallelic SNPs retained, and variants further filtered removing sites with quality ≤ 10, combined depth of coverage > 2000 and low read support <500 using BCFtools v1.14 (Danecek et al., 2021). As SNPs in linkage disequilibrium convey correlated genetic information and affect modelling outcomes, these were pruned using PLINK v2.00a2.3 (Chang et al., 2015; Purcell & Chang, n.d.) using the following parameters: a sliding window size of 150, a step size of 5, and an R^2 threshold of 0.5, as in Chen (2023).

The PLINK files were converted to a genlight file, a data format which allows analysis of SNPs in R, using package 'dartR' v1.0.5 (Gruber et al., 2018; Mijangos et al., 2022). Sex-linked loci were identified and removed using 'dartR.sexlinked' (Gruber et al., 2018; Mijangos et al., 2022; Robledo-Ruiz et al., 2023) to ensure unbiased assessment of population structure (Benestan et al., 2017). 'DartR' was then used to remove monomorphic loci, those with all missing values, impute any other missing values with those taken from the nearest neighbour, and run a PCA to reduce the dimensionality of the SNPs and generate axes that describe the genetic variation in the dataset.

Environmental data

Study area

A shapefile of the study area was created with QGIS version 3.22.9-Białowieża (QGIS, 2021) with a bounding extent of 23.7°E–26.4°E, 10.9°S–13.3°S, but confined entirely on the northern side by the border with the DRC, and almost entirely on the western side by the border with Angola (Supplementary Figure 5.1).

Climate

The 19 bioclimatic variables from WorldClim were chosen to describe environmental variation. These represent trends, seasonality, and limiting environmental factors of temperature and precipitation derived from aggregating monthly data from the period 1970–2000 (Fick & Hijmans, 2017; WorldClim, 2020). These were downloaded for the extent of Zambia at a resolution of 30s (approximately one km at the equator) using R package 'geodata' version 0.6-2 (Hijmans et al., 2024). These were cropped using the study area shapefile, and Precipitation of Driest Month (BIO14) removed as it had zero variation over this area. A PCA of the remaining 18 layers was completed with R package 'terra' version 1.8-42 (Hijmans, 2025).

Soil

As *Catharsius* is a genus of paracoprid beetles (Takano, 2018) – they create their nests by burying underneath dung pats – characteristics of the soil were hypothesised to be an important element of their environment. Soil variables from iSDA (Hengl et al., 2021), also at 30s resolution, representing the carbon, clay, sand, silt, and stone content, as well as its bulk density and pH of the water, were downloaded using the 'soil_af_isda' function of 'geodata'. As above, these were cropped using the study area shapefile and a PCA completed with 'terra'.

Partial redundancy analyses

To investigate the extent to which landscape is responsible for the genetic structure within *C. dux* and *C. duciformis*, five pRDAs were completed in R package 'vegan' version 2.6-10 (Oksanen et al., 2025). Whilst a traditional redundancy analysis is designed to explain variation in a multivariate response variable, a pRDA controls for any effect from covariates to isolate the influence of explanatory variables (Legendre & Legendre, 2012). To create the response variable, a matrix of SNP genotypes for individuals of *C. dux* and *C. duciformis* was first retrieved from the genlight object using R package 'adegenet' v2.1.10 (Jombart, 2008; Jombart & Ahmed, 2011). This records a "0", "1", or "2", for each individual at each locus, representing a homozygous reference allele, a heterozygous reference allele, or a homozygous alternate allele respectively. As individuals had been sampled at one of six (out of the seven) collecting locations, those from the same sites shared identical geographic coordinates and associated environment values, so could not be considered independent. To overcome this, SNP data were aggregated by collection site by taking the average allele frequency at each collecting location. As the original SNP matrix was very large, R package 'bigstatsr' version 1.6.1 (Privé et al., 2018) was used to convert it to a file-backed big matrix beforehand for more efficient computation. The first axis of the genetic PCA was used as the 'genetic structure' explanatory variable in the pRDAs. As SNP data were aggregated by collection site for the response variable, the values from the genetic PCA were also averaged by collection site. The 'geography' explanatory variable used the latitude of the collecting locations, as longitude was found to correlate notably with

climate and its inclusion led to overfitting. The remaining explanatory variables were values from the first principal component of both the climatic ('climate') and soil ('soil') PCAs that corresponded to each collecting location. The first model constrained all four variables to measure how much of the variation in the site-level SNP data they could explain overall. Models two to five each constrained a single explanatory variable whilst controlling for the remaining three, to isolate how much of the total variation each explained, following a similar methodology to Chen et al. (2023).

Latent factor mixed model

To test for adaptation to the local environment, a LFMM was run with the 'lfmm2' function from R package 'LEA' version 3.20.0 (Caye et al., 2019; Frichot & Francois, 2015; Gain & François, 2021). Latent factor mixed models infer and account for population structure and spatial autocorrelation in the dataset whilst testing for correlations between loci and the environment to reduce the rate of false positives (Frichot et al., 2013). Although it is not explicitly designed to handle repeated environmental values across individuals, it does not treat them as independent replicates. Instead, its correction for population structure helps mitigate spurious associations that could otherwise arise from pseudoreplication. As a result of fewer unique environmental observations, LFMMs tend to be conservative, potentially leading to fewer significant SNP–environment associations, but those that are detected are robust. Findings should therefore be interpreted as revealing only the strongest associations. For this model, the matrix of *C. dux* and *C. duciformis* SNP genotypes described above was used as the response variable. Explanatory variables were values from the first principal component of both the climatic and soil PCAs corresponding to

each individual specimen. These were considered together as 'environment' as models were set to compute a single significance value for all variables at each locus. Preliminary models tested for up to seven latent factors, finding two. This value was carried forward to the final model, and SNPs with significant associations with the environment were identified by correcting P values for false discovery rate (FDR) in 'stats'. This method was chosen as it is less stringent, and therefore more powerful, so less likely to remove true positives in an already conservative model (Benjamini & Hochberg, 2018). To illustrate significant SNPs in space, allele frequency was normalised per individual and aggregated by collecting location.

Using R base package 'stats', Pearson's correlation coefficient was computed for comparisons between the first latent factor and the first principal component of the climatic and soil PCAs, as well as both latitude and longitude. A strong positive correlation will return a value of 1, and strong negative a value of –1, indicating that environment or geography explains the variation captured in the first latent factor, whereas zero would indicate no correlation (K. Pearson, 1895).

Population structure across all species

k-means clustering

To contextualise the diversity of *C. dux* and *C. duciformis* within the wider genus, 'adegenet' was used to infer clusters of individuals using the *k*-means algorithm and five axes of the genetic PCA, including all specimens of all species. This finds clusters by maximising variation between groups (*k*) by running the algorithm for an increasing

number of *k,* and generating a Bayesian Information Criterion (BIC) statistic to measure goodness-of-fit. In theory, the lowest BIC indicates the optimal number of clusters, but this is not always clear in practice, where an elbow in the curve of BIC values can be used instead (Jombart & Collins, 2022). Two values of k were chosen for further analysis: *k*=5 as there was an uptick in BIC at *k*=6 and this also corresponds with number of clusters found in the genetic PCA, and *k*=11 as BIC levels off at this point, which also corresponded with the number of species identified in the dataset (Supplementary Figure 5.2). The significance of these data partitions was tested with a Permutational Multivariate Analysis of Variance (PERMANOVA) in 'vegan', which assessed whether the clustering explains a significant portion of the variance in pairwise distance between individuals (described here with a Euclidean distance matrix). The genetic differentiation between clusters (Fst) was calculated using 'dartR'.

Discriminant analysis of principal components

A DAPC was run in 'adegenet' with the first five axes of the genetic PCA for both *k*=5 and *k*=11. Using the *k*-means clusters and the genetic PCA axes, this derives linear combinations of variables that best separate the pre-defined groups, as well as a visual representation of between-population structure. This visualisation is useful as it can help to determine if clusters are being artificially forced onto a cline (Jombart et al., 2010). Posterior probabilities of cluster membership highlight individuals whose group identity is less well-defined, and those which did not have a probability of >0.9 for any group were labelled as uncertain.

Similarity based on shared alleles

To examine genetic variation on a finer scale, pairwise genetic similarity between all individuals was calculated based on the proportion of shared alleles using 'dartR'. By producing a similarity matrix from SNP genotype data, relationships between specimens can be visualised in such a way that highlights subtle patterns of relatedness that may be obscured when only considering population-level or cluster-based analyses. As well as fine-scale structure, this individual-based approach allows for the detection of potential migrants or hybrids, and within-population diversity.

ADMIXTURE

ADMIXTURE version 1.3.0 was used to estimate the ancestry proportions of individuals in an approach similar to STRUCTURE, but optimised for large SNP datasets (Alexander et al., 2009). Understanding the proportions of different ancestral populations from which an individual's genetic makeup is derived is useful for identifying cryptic population structure and recent hybridisation. Cautions have been made about interpreting historical demography with this method due to its sensitivity to relatives, isolation by distance, hierarchical and subtle population structure, and fluidity in populations over time, but it is still informative in explaining the most prominent variation in the dataset (Lawson et al., 2018). Here, the number of ancestral populations *K* – note that these are not related to and do not correspond with the *k*-means clusters described above – was identified by running ADMIXTURE with its default 5-fold cross-validation and identifying the number of ancestral populations with the lowest cross-validation error. This was completed on the University of Oxford Advanced Research Computing (ARC) facility (Richards, 2015).

Population structure across *C. dux* and *C. duciformis*

Spatial drivers of population structure

To compare the influence of environmental and geographical variables to that of population structure on variation in *C. dux* and *C. duciformis*, Pearson's correlation coefficient was calculated in 'stats' between the first latent factor of the LFMM and the first principal component of the genetic PCA. A PERMANOVA from 'vegan' was then used to test for a spatial driver of structure in *C. dux* and *C. duciformis* by assessing the degree to which collecting location influenced the DAPC posterior probabilities of cluster membership. In essence, are beetles from the same collecting location more similar to each other than those from other places? This was run for both *k*-means *k*=5 and *k*=11 and, for significant results, the R package 'pairwiseAdonis' version 0.4.1 (Arbizu, 2017) was used to test between pairs of locations to uncover which were driving the observed overall difference.

Results

Mapping raw sequencing data, calling SNPs, and filtering

The bioinformatics pipeline generated 66 genotypes and 208,765 binary SNPs with 11.84% overall missing data. After filtering for sex-linked loci, 74 were removed including 14 Y-linked loci, 32 sex-biased loci, one X-linked locus, and 27 gametologs, leaving 208,691 autosomal loci. After removing monomorphic loci, as well as those with all missing values, and imputing any remaining missing values, 208,201 loci remained.

The analysis of population structure among the 66 *Catharsius* specimens with PCA identified five groups using the first two principal components (Figure 5.2), which explained 11.77% and 2.89% of the variation in the SNPS, respectively. Nearly all of the specimens of *C. dux* and *C. duciformis* group closely together, forming a single large green cluster. The exceptions to this are the two green points close to the origin, which are specimens of *C. satyrus*, and the grey point at approximately (10, -5) which is specimen Z3_18, identified as *C. dux/duciformis*. The location of the *C. satyrus* specimens is reflective of their position in the phylogeny, but Z3_18 is an anomaly. When zoomed in to the *C. dux* and *C. duciformis* cluster (not including Z3_18), there was no apparent pattern related to collecting location or identification (Supplementary Figures 5.3 and 5.4).

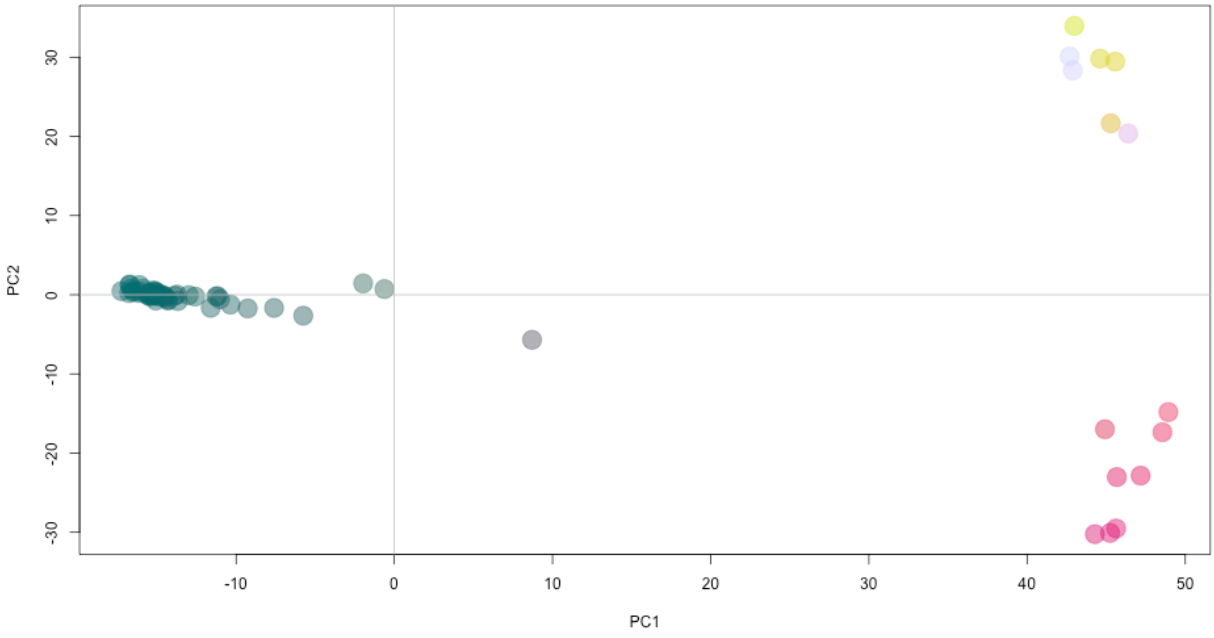


Figure 5.2: The first and second principal components of the PCA of Catharsius *SNPs showing the main axes of genetic variation amongst study samples. The colours of the five main groups correspond with Table 5.2 and show* C. dux, C. duciformis*, and* C. satyrus *in green, a single specimen of* C. dux / duciformis *(Z3_18) in grey,* C. machadoi *and* C. smithi *in yellow,* C. biconifer *in white, and* C. merrettorum, C. peregrinus, C. orami, C. luluensis*, and* C. pallas *in pink. The eigenvalues for the first three axes are 0.1177, 0.02894, 0.02184, respectively, and ~0 for the final, 66th, axis.*

Adaptation to the environment in *C. dux* and *C. duciformis*

Partial redundancy analyses

Five individual partial redundancy analyses were used to assess the degree to which genetic structure, climate, geography, and soil drove variation in the average allele frequency at each location. Model 1 showed that 66.05% of the variation could be explained by the combination of genetic structure, climate, soil, and geography (Table 5.1). Climate was responsible for almost a third of this explained variation. Together, geography and soil were reportedly responsible for 44.11% of the explained variation (19.55 and 24.56 in % explained variation in Table 5.1), but some degree of correlation (Supplementary Figure 5.5) means that this may be slightly inflated, which explains why the total of the '% explained variation' column for rows two to five is greater than 100%. Genetic structure was reportedly responsible for 27.52% of the explained variation. However, all models were non-significant, possibly due to the reduced power from aggregating SNP data across only six locations.

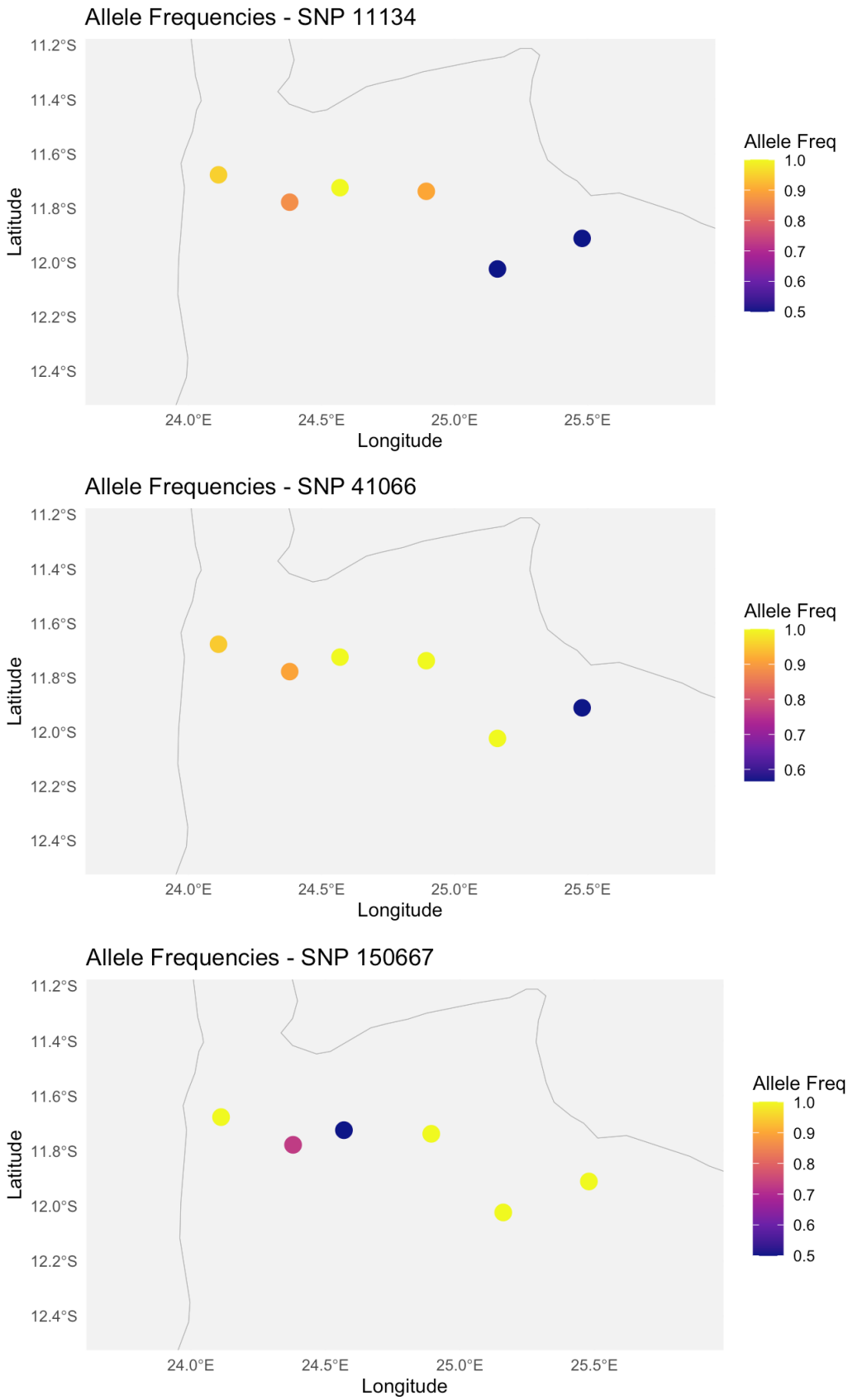
Table 5.1: Results of the five Catharsius *pRDAs. Percent of total variation refers to how much of the total variation in the SNPs that the variable(s) in bold explain. Percent of explained variation is how much of the total explained variation (66.05%) that the variable in bold is responsible for. The pRDA analyses include four constrained axes for the full model (all explanatory variables), and one constrained axis for each partial model assessing individual variable contributions while conditioning on the others.*

No.	Model	% total variation	% explained variation	*p*
1	**Genetic structure, climate, soil, geography** constrained	66.05	100	0.9028
2	**Genetic structure** constrained; climate, soil, geography conditioned	18.18	27.52	0.6861
3	**Climate** constrained; soil, genetic structure, geography conditioned	20.56	31.13	0.5681
4	**Geography** constrained; climate, soil, genetic structure conditioned	12.91	19.55	0.7222
5	**Soil** constrained; genetic structure, geography, climate conditioned	16.22	24.56	0.6764

Latent factor mixed model

Filtering the SNP dataset containing only specimens of *C. dux* and / or *C. duciformis* removed 19,467 monomorphic loci (leaving 188,724 remaining). The LFMM run on the resulting subset found three SNPs that were significantly correlated with the environment (climate and / or soil) when corrected for FDR. When allele frequencies for these significant SNPs were aggregated across collecting locations, all three showed visible geographical differences (Figures 5.3a-c). Specifically, SNP 41066 was less frequent in collecting location 14 (the most easterly), SNP 11134 less frequent in locations 14 and 31 (the two most easterly), and SNP 150667 less frequent in locations 13 and 3.

Two latent factors were identified, with latent factor 2 clearly separating male and female specimens (Supplementary Figure 5.6). Although sex-linked loci had previously been filtered, it is probable that some sex-linked loci specific to *C. dux* and *C. duciformis* were not identified during this process as it included all species. An explanation for latent factor 1 was not visually identifiable, and it was also not correlated with either climate ($r = 0.0063$), soil ($r = 0.016$), latitude ($r = 0.033$) or longitude ($r = 0.064$). There was no pattern according to species identification, so it did not explain morphological identities.



Figures 5.3a-c: Aggregated allele frequency at each Catharsius *collecting location for the significant SNPs. Grey lines show the Zambian border.*

Population structure across all species

k-means clustering

Specimens identified as *C. dux* and / or *C. duciformis* do not ever cluster with any other species, including *C. satyrus,* which is the only member of its own cluster for both *k* = 5 and *k* = 11 (Table 5.2, note that here, and in subsequent figures, "duci" refers to "duciformis"). When *k* is increased from five to 11, it is the *C. dux* and / or *C. duciformis* specimens which split further rather than the other separate species in the dataset, suggesting more variation **within** this group than **between** other species. The only non-*dux/duciformis* species which splits into its own cluster when *k*=11 is *Catharsius biconifer*, which is grouped with *Catharsius machadoi* and *Catharsius smithi* when *k*=5. There is no obvious pattern according to species identification or collecting location when trying to explain how *C. dux* and / or *C. duciformis* specimens cluster, with the exception of all confident single-species identifications at location 14 being *C. duciformis* and *C. dux* at all other collecting locations. In accordance with its separation in genetic PCA space, specimen Z3_18 belongs to its own cluster when *k*=11, but does cluster with other *C.dux/duciformis* specimens when *k*=5.

Differences between the clusters were significant for both *k*=5 (adonis2: $F(4, 61) = 3.36$, $R^2 = 0.18$, $p = 0.001$) and *k*=11 (adonis2: $F(10, 55) = 2.10$, $R^2 = 0.28$, $p = 0.001$). When *k*=11, Fst values between clusters show a clear split between those that contain *C. dux, C. duciformis, and C. satyrus* (Group A) and those that don't (Group B). The lowest dissimilarity values are between clusters containing just *C. dux* and / or *C. duciformis,* ranging from 0.0003 to 0.0056. Intra-Group B comparisons returned intermediate

dissimilarity values ranging from 0.0657 to 0.0797, whilst comparisons between clusters from Group A and Group B returned the most notable dissimilarity values, ranging from 0.0763 to 0.1154. Comparisons between *C. satyrus* and Group B clusters were the most dissimilar. Within Group A, values ranging from 0.0207 to 0.0249 reflected *C. satyrus'* separation from *C. dux* and / or *C. duciformis* but membership of the same group. Specific values can be found in Supplementary Figures 5.7 and 5.8, along with those from when *k*=5, which show the same overall pattern.

Discriminant analysis of principal components

The first discriminant function described most of the variation for both values of *k*, but two were included to visualise scatterplots (Supplementary Figures 5.9 and 5.10). The DAPC posterior probabilities described each individual's probability of belonging to each prior assigned cluster, and pinpointed those whose membership is uncertain (here, those which did not have a probability of assignment >0.9 to any group). When *k*=5, there were 45 uncertain individuals, all of which were identified as *C. dux* and / or *C. duciformis* (Supplementary Figure 5.11). Only four individuals identified as such were assigned confidently, and these were Z3_18, Z19_17, Z18_16, and Z18_15. When *k*=11, there were 56 uncertain individuals, including all specimens identified as *C. dux* and / or *C. duciformis*, except Z3_18 (Supplementary Figure 5.12). Only ten individuals (of any species) were assigned confidently, and these were Z3_18, Z14_18, Z14_19, Z19_1, Z19_5, Z19_7, Z19_9, Z3_14, Z3_17, and Z32_1.

Table 5.2: Results of Catharsius *k-means clustering for both k=5 and k=11, ordered by k=5 cluster membership and then collecting location (the number between "Z" and "_" in the specimen name). The colours of the collecting locations correspond with Figure 5.1, the colours of the species identifications correspond with Figure 5.2, and the colours of the 5k and 11k cluster membership correspond with Supplementary Figures 5.9 and 5.10.*

Specimen	Species	5k	11k	Specimen	Species	5k	11k	Specimen	Species	5k	11k
Z14_12	*dux / duci*	1	4	Z13_4	*merrettorum*	3	8	Z18_16	*dux / duci*	5	9
Z14_13	*duci*	1	4	Z14_18	*peregrinus*	3	8	Z18_4	*dux*	5	9
Z14_15	*dux / duci*	1	4	Z14_19	*peregrinus*	3	8	Z18_5	*dux / duci*	5	9
Z14_22	*dux / duci*	1	4	Z14_21	*orami*	3	8	Z18_8	*dux*	5	9
Z14_23	*duci*	1	4	Z18_11	*luluensis*	3	8	Z18_6	*dux / duci*	5	10
Z14_4	*duci*	1	7	Z19_1	*luluensis*	3	8	Z18_7	*dux / duci*	5	10
Z14_5	*duci*	1	7	Z19_5	*peregrinus*	3	8	Z19_11	*dux*	5	9
Z18_12	*dux / duci*	1	5	Z32_1	*pallas*	3	8	Z19_17	*dux*	5	9
Z18_13	*dux / duci*	1	5	Z13_5	*satyrus*	4	1	Z19_18	*dux / duci*	5	2
Z19_13	*dux*	1	5	Z13_6	*satyrus*	4	1	Z19_12	*dux*	5	10
Z19_8	*dux / duci*	1	5	Z13_2	*dux / duci*	5	9	Z19_14	*dux*	5	10
Z3_12	*dux / duci*	1	7	Z13_3	*dux*	5	9	Z19_15	*dux / duci*	5	10
Z3_13	*dux / duci*	1	7	Z14_11	*duci*	5	2	Z19_16	*dux / duci*	5	10
Z3_16	*dux / duci*	1	7	Z14_16	*duci*	5	2	Z19_19	*dux*	5	10
Z3_18	*dux / duci*	1	6	Z14_17	*dux / duci*	5	2	Z3_8	*dux / duci*	5	9
Z18_9	*machadoi*	2	3	Z14_14	*dux / duci*	5	10	Z3_9	*dux / duci*	5	2
Z19_4	*machadoi*	2	3	Z14_6	*duci*	5	10	Z3_11	*dux*	5	10
Z19_9	*machadoi*	2	3	Z14_7	*dux / duci*	5	10	Z3_15	*dux*	5	10
Z19_6	*smithi*	2	3	Z14_8	*duci*	5	10	Z3_19	*dux*	5	10
Z19_7	*biconifer*	2	11	Z14_9	*duci*	5	10	Z3_21	*dux*	5	10
Z3_14	*biconifer*	2	11	Z18_14	*dux*	5	9	Z3_7	*dux / duci*	5	10
Z3_17	*biconifer*	2	11	Z18_15	*dux*	5	9	Z31_1	*dux*	5	2

When *k* is increased from five to 11, the confidence with which *C. satyrus* individuals are assigned drops, and their similarity to the *C. dux* and / or *C. duciformis* specimens is reflected in the increased probability of clustering together, especially Z13_6. Specimen Z14_4 also stands out in its increased probability of clustering with the *C. satyrus* specimens when *k*=11, and consequent similarity with Z13_5.

Similarity based on shared alleles

Genetic distance between individuals based on shared alleles showed the two main groups in the dataset as Fst between clusters (Group A: *C. dux* and / or *C. duciformis*,

and *C. satyrus*; Group B: all other species) (Supplementary Figure 5.13). Supporting its slight separation from Group A in other results, Z3_18 showed some similarity with Group B specimens. Similarly, *C. satyrus* specimens grouped as expected and showed some dissimilarity with *C. dux* and / or C. *duciformis*. However, contrary to the notable cluster dissimilarity, there were some individuals of Group B that appear to be more similar to *C. satyrus* than *C. satyrus* was to *C. dux* and / or *C. duciformis*. Most notably, similarity is higher within Group B than within Group A despite often describing inter-species diversity between accepted and easily distinguishable distinct species.

ADMIXTURE

Both one and two ADMIXTURE clusters (*K*) produced low cross-validation errors (0.28773 and 0.28615, respectively) (Supplementary Figure 5.14). The slightly lower error at *K* = 2 suggests that although the variation could mostly be captured within a single cluster, there is enough genetic structure to say that the specimens hail from two closely-related ancestral populations. These populations almost exactly capture the split between Group A and Group B seen in other results, with the former in ancestral population 1, and the latter in population 2 (Figure 5.4). Once again, this is with the exception of *C. satyrus* and specimen Z3_18 which are notably admixed. Two further specimens identified as *C. dux / C. duciformis* (Z3_13, Z3_16), and two as *C. duciformis* (Z14_4 and Z14_5) displayed low levels of admixture, and these were also the specimens that were most similar to Group B species according to shared alleles. Specimen Z14_4 also stood out in the DAPC posteriors for its increased probability of clustering with the *C. satyrus* specimens.

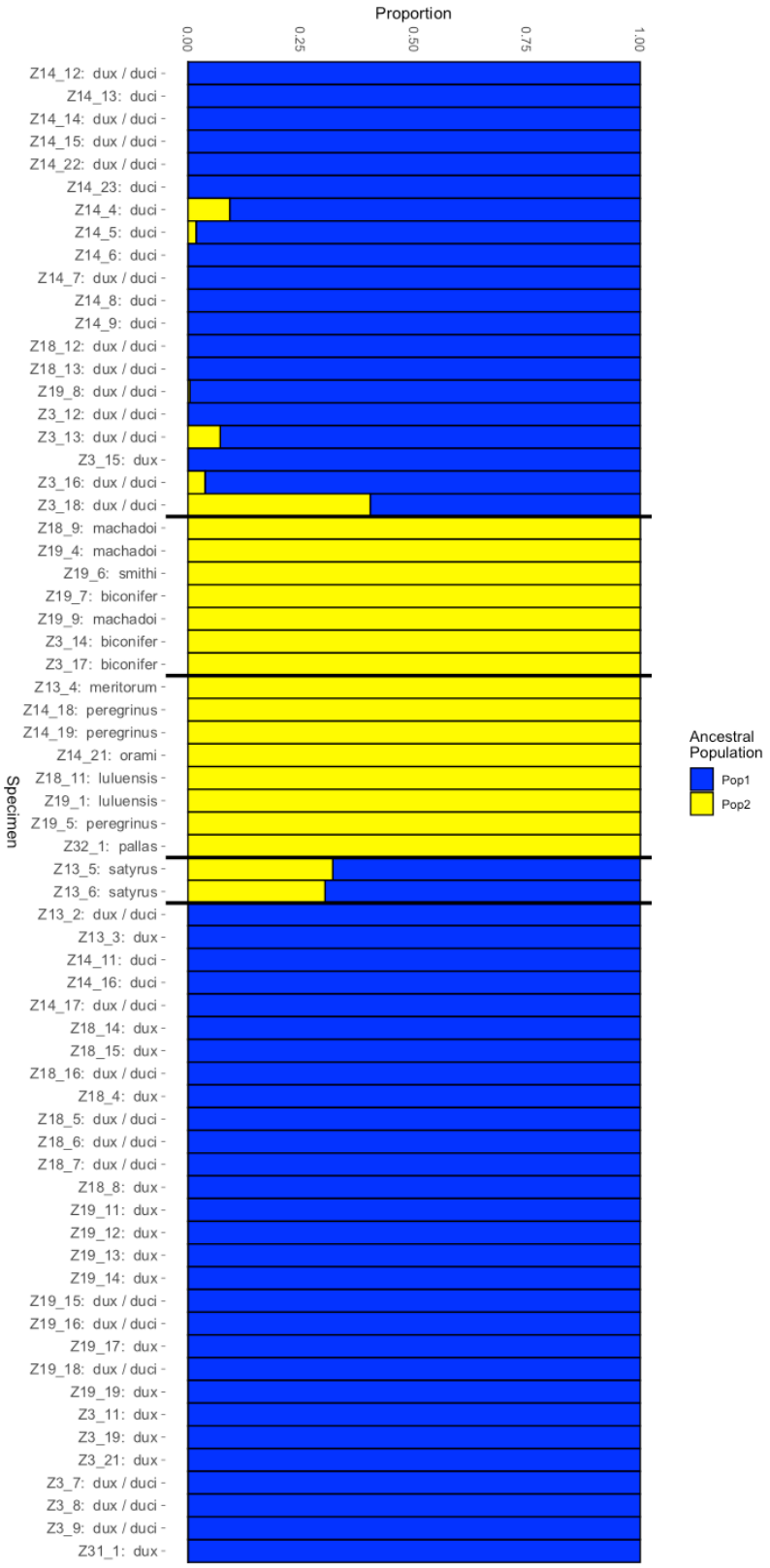


Figure 5.4: Results of the Catharsius *ADMIXTURE analysis showing the proportion of membership of two ancestral populations for each individual specimen. Individuals are in the same order as Table 5.2 (according to membership of k-means clusters) and the bold lines separate clusters 1 to 5, and also correspond with those in Table 5.2. There is a clear split between specimens of* C. dux / C. duciformis *and all other species, with the exception of* C. satyrus *and a small number of* C. dux *and/or* C. duciformis specimens that show a small amount of mixture between the two ancestral populations, and specimen Z3_18 which shows much more mixture, reflecting its position as an outlier in other analyses and likely hybrid nature.

Population structure across *C. dux* and *C. duciformis*

Spatial drivers of population structure

Latent factor 1 from the LFMM, which did not split specimens according to their species identification, and did not correlate with climate, soil, or latitude was strongly correlated with the first axis of the genetic PCA ($r = -0.81$, $p < 0.001$), which also clearly illustrated the result of *k*-means clustering, including the notable separation of Z3_18 (Figure 5.5).

Combining spatial and population results, collection location significantly influenced the DAPC posterior probability of group assignment for specimens of *C. dux* and / or *C. duciformis* when *k*=5 (adonis2: $F(5, 43) = 2.68$, $R^2 = 0.24$, $p = 0.036$), but not when *k*=11 (adonis2: $F(5, 43) = 1.18$, $R^2 = 0.12$, $p = 0.234$). For *k*=5, pairwise PERMANOVA comparisons for locations revealed significant comparisons between locations 14 and 13, 14 and 18, and 14 and 19, however these comparisons did not remain significant after FDR correction (Table 5.3).

Table 5.3: Significant pairwise PERMANOVA comparisons of Catharsius *DAPC cluster membership probability at each collecting location when individuals are split into five clusters. All other comparisons between collecting locations were non-significant. This demonstrates that collecting specimens from location 14, the most easterly, is likely the only location that may have an impact on the probability of membership of a particular genetic cluster.*

Pairs	Df	Sum of Squares	F	R^2	p	FDR adjusted p
13 vs 14	1	0.203	5.187	0.257	0.038	0.19
14 vs 18	1	0.292	7.256	0.240	0.02	0.15
14 vs 19	1	0.260	8.205	0.263	0.016	0.15

Key takeaways for *C. dux* and *C. duciformis*

- Despite morphological differences, individuals are randomly mixed along a genetic cline which cannot be significantly explained by the environmental or geographical variables included in this study.

- There is evidence of hybridisation in specimen Z3_18.

- The most easterly collecting location is loosely linked to identification, clustering, and allele frequency.

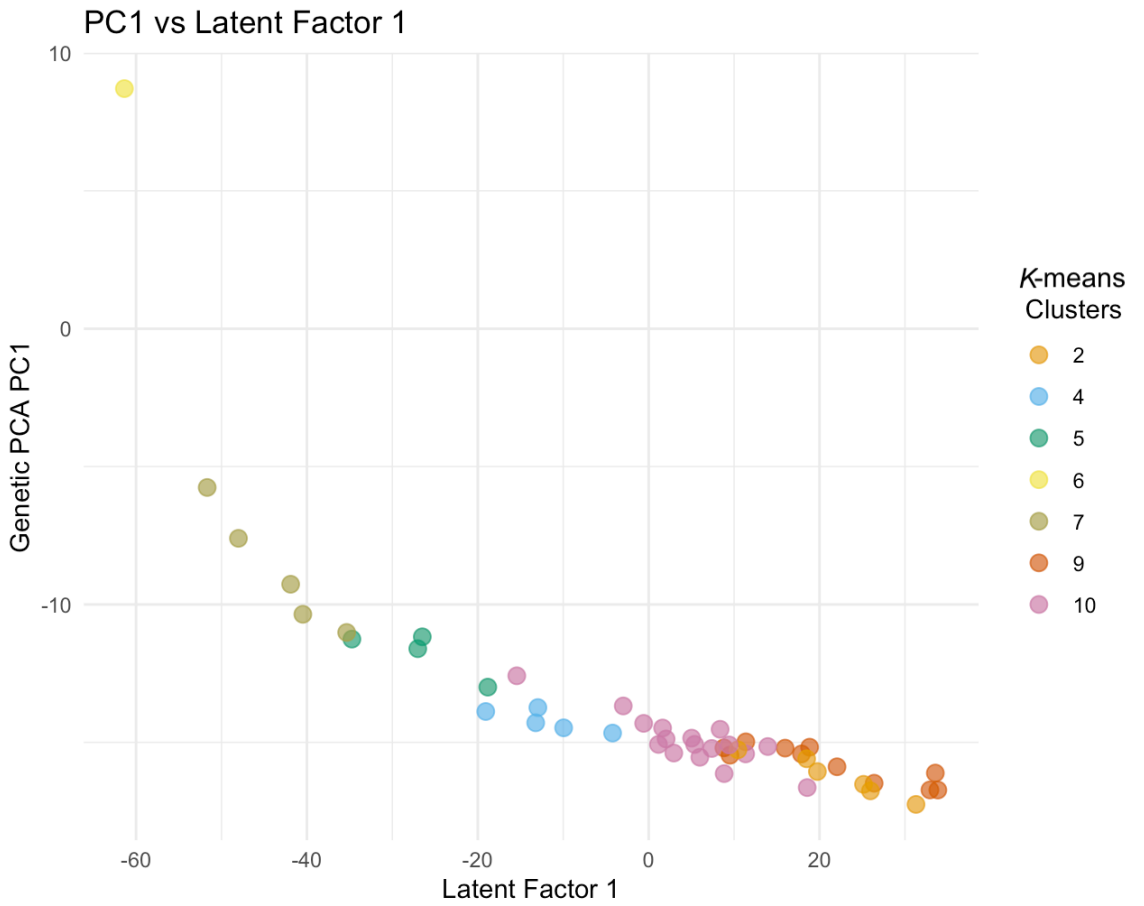


Figure 5.5: The correlation between latent factor 1 from the LFMM and the first axis of the genetic PCA of SNPs. Each point corresponds to an individual Catharsius *specimen, and is coloured according to its membership of a k-means cluster when k=11. Colours correspond with Table 5.2. This demonstrates a strong correlation between the first latent factor and the first axis of the genetic PCA, as well as the fact that genetic clusters as established by the K-means clustering lie clearly along this continuous axis of genetic variation.*

Discussion

Due to evidence in the literature of strong links between dung beetles and their environments, it was hypothesised that adaptation to local conditions was maintaining parapatry between *C. dux* and *C. duciformis*. Although some adaptation to environmental conditions was found, the results presented here suggest that population structure and demographic processes, including hybridisation, may play a more significant role in shaping patterns of genetic diversity and therefore driving distributions. Using SNPs from whole genomes, specimens of these two species were

found to be indistinguishable despite morphological differences. They exist along a genetic continuum that cannot be explained by our environmental data or space, with the exception of a subtle link between genetic identity and collecting location at the most easterly site. Together with evidence of hybridisation in this dataset, this suggests the possibility of a hybrid zone in the study area with an eastern edge captured at its periphery.

Adaptation to the environment in *C. dux* and *C. duciformis*

With the exception of specimen Z3_18, individuals of *C. dux* and / or *C. duciformis* consistently clustered together throughout the results, but there was no obvious structure within this grouping. Although climate was responsible for driving 21%, and soil 16%, of the total variation in average allele frequency at collecting locations, no significant results mean we cannot reliably conclude that these beetles were strongly adapted to the environment in which they were found. Similarly, the correlation between climate and latitude makes it difficult to attribute this variation to one or the other, and it could be that the variation attributed to climate could be all, or in part, actually driven by geographic distance. It is possible that the limitations introduced to the models by aggregating allele frequency to just six locations are responsible for the lack of significance, but if adaptations to the environment were strong enough to be perpetuating parapatry between these two species, we would still expect to see more notable differences between collecting locations given that they traverse the boundaries of their ranges.

The LFMM identified just three SNPs that were significantly correlated with the environment, and whilst this must be interpreted in the conservative context of the model, T. Chen et al. (2023) found markedly more candidate SNPs from a much smaller pool of loci using LFMMs. A weak influence of climate is in line with the results of the pRDAs, and together they suggest that high gene flow throughout the sampling area may be swamping any effect of local adaptation. That said, allele frequency of the significant SNPs hints that there may be an exception to this as, for two out of three, frequency is notably lower at the easterly collecting locations. Although environment has little explanatory power throughout the sampling area, it may be that it influences some structure at the eastern edge, but care should be taken with this interpretation given the low number of significant SNPs in the first place.

In line with the generally weak influence of environment, it was not surprising that latent factor 1 from the LFMM did not correlate with any environmental or spatial factors. However, its correlation with the first principal component of the genetic PCA and the *k*-means clusters, but simultaneous lack of pattern according to species identification was more unexpected. What this suggested was a continuous axis of genetic diversity that not only could not be explained by environmental adaptation, but also on which all individuals were mixed and experiencing gene flow, regardless of morphological differences.

Population structure across all species

Analyses of population structure including all species provided important context for further understanding the dynamics between *C. dux* and / or *C. duciformis*. Their

distinction from all other species, with the exception of *C. satyrus*, was clear and unanimous, except for specimen Z3_18. Their lower inter-individual but higher inter-cluster similarities when compared to other species underlines their simultaneously high overall genetic diversity and low level of differentiation between groups, supporting the presence of a continuous axis of diversity with ongoing gene flow. Clusters of non-*dux/duciformis* individuals, with their more pronounced separation, provide a reference for the degree of dissimilarity which would signpost more complete reproductive isolation and speciation. Whilst it is possible that the overall genetic diversity of *C. dux* and / or *C. duciformis* individuals is amplified by more specimens, this cannot account for the extremely low inter-cluster differentiation, and ongoing gene flow amongst them is further supported by the uncertain DAPC posterior probabilities. This is especially the case when there are eleven clusters, which underlines the difficulty in delineating meaningful boundaries between individuals regardless of identification.

As diagnostic features for identification, it was thought that the genitalia of *C. dux* and *C. duciformis* had diverged sufficiently to constitute a complete prezygotic barrier to reproduction through mechanical incompatibility (H. Takano, personal communication, October 2021), but it is clear that this barrier is imperfect. Studies have shown that the shape of genitalia in beetles evolves rapidly and in parallel between males and females of the same species (Macagno et al., 2011), but, whilst this means interspecific copulation comes at a high fitness cost, hybridisation and gene flow is still possible (Sota & Kubota, 1998). Here, specimen Z3_18's notable admixture, intermediate position in genetic PCA space, and similarities with other species based on shared alleles are strong evidence of hybridisation between *C. dux* and / or *C. duciformis* and

another species. It is reasonable to suggest, then, that *C. dux* and *C. duciformis* have also been able to hybridise amongst themselves on rare occasions. Such events, though infrequent, would be sufficient over thousands of years to homogenise their genomes in the study area. Furthermore, that they are still morphologically distinguishable does not preclude the possibility of hybridisation and gene flow, as decoupled differentiation in which morphological distinctiveness is preserved despite minimal genomic differentiation has been previously described. For example, gene flow between all-black Western European carrions crows and grey-coated Northern and Southern European hooded crows has resulted in genomic homogenisation, with the exception of a single colour locus responsible for their continued morphological difference (Gwee et al., 2025).

A possible hybrid zone could also explain why the most easterly collecting location (14) emerged as an exception to the weak influence of environment or geography seen across the dataset as a whole. Although pairwise comparisons of cluster membership probability (when $k = 5$) at each location did not remain significant after FDR correction, all unadjusted p-values below 0.05 involved Location 14, suggesting it as the primary driver of the observed global difference. The lack of overall significance when $k = 11$ does not negate this, but instead conforms with the continuity of genetic diversity; fewer clusters improve detection of broad-scale differences in population structure which may coincide with collecting locations, but more clusters accommodates the continuity of the genetic variation, resulting in less distinctive assignments and diminished associations with geographic or environmental variables. In line with these differences in genetic cluster membership, certain IDs at Location 14 were all *C.*

duciformis, but *C. dux* elsewhere. Although there are too few samples to draw confident conclusions from these identifications alone, together these results indicate that clustering at this collecting location may be somewhat influenced by a distinctive *C. duciformis* genetic identity and, as such, could signpost the eastern edge of a hybrid zone, beyond which beetles are both morphologically and genetically differentiated. The first axis of the climatic PCA that was used in the landscape analyses shows a sharp change at the eastern edge of the sampling area (Supplementary Figure 5.15). Coinciding with low allele frequencies for two out of three significant SNPs, and broadly describing precipitation, this transition could signal associations with a drier habitat at Location 14, although this has not been explicitly tested.

Although it seems unlikely that the influence of current climatic conditions is sufficiently strong enough to be maintaining the parapatric distributions of *C. dux* and *C. duciformis*, paleoclimatic changes may be able to account for their observed distribution. The effect of Pleistocene glaciation on the expansion and contraction of tropical rainforest in Africa is well documented (Leal, 2004; Maley, 1996). In addition, during cooler and drier periods, an arid corridor on a NE-SW orientation formed, joining Somalia with the Kalahari sands of Namibia and South Africa, to which present-day distributions of many animals are attributed (Balinsky, 1962; Perkins, 2020; Werger, 1978), including insects (Kirk-Spriggs & McGregor, 2009). It is plausible that the historic distribution of *C. dux* was split by this corridor, triggering allopatric divergence of *C. dux* and *C. duciformis* in climatic refugia. Range expansion driven by post-glaciation humidification and subsequent secondary contact in the sampling area where continued gene flow has eroded much of the previous genomic differentiation would be

consistent with the formation of other accepted hybrid zones (Barton & Hewitt, 1985; Hewitt, 1999). Furthermore, the projection for the ensemble model of *C. dux* clearly illustrated suitable habitat for *C. dux* to be centred on tropical savannah areas with a boundary at the approximate edge of the arid corridor and, whilst *C. duciformis* appears to tolerate these conditions to some degree too, it favours the temperate zone to the south (Figure 2.3 in Blades (2025)). Their overlapping but nonetheless distinct realised climatic niches support evolution in slightly different habitats, consistent with paleoclimatic changes (Figure 5.6).

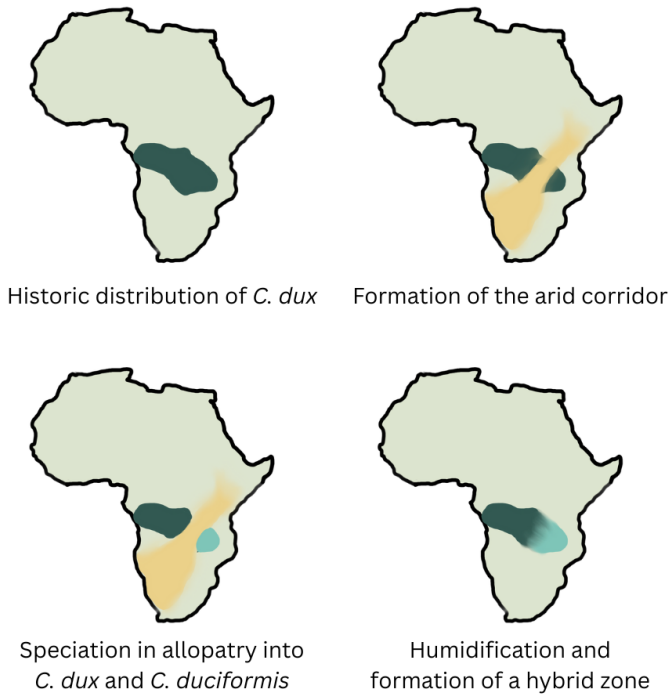


Figure 5.6: This Catharsius *study propose an evolutionary for* C. dux *and* C. duciformis *that is governed by paleoclimate changes. The placement and orientation of the arid corridor during drier periods of the Pleistocene glaciation proposed by Balinsky (1962) aligns with the boundary between the orientations of these two species.*

Informed by the strong links between dung beetles and their environment, as well as the divergent climatic niches of *C. dux* and *C. duciformis*, this study hypothesised that adaptation to environmental conditions is responsible for maintaining their parapatric

distributions. However, in the absence of strong links between genomic identity and climate, soil, or morphological differences, as well as the suggestion of an exception to this at a peripheral collection site, it instead suggests the possibility of a *C. dux* and *C. duciformis* hybrid zone resulting from post-glacial secondary contact. However, more specific population genomics analyses designed to test for the dynamics of hybridisation and relevant selective pressures are required to test this hypothesis. Critically, this study was motivated by observations of local range edges, so the current scale of sampling is not sufficient to answer these questions. Study area extent in landscape genomics studies is critical and should account for the dispersal ability of the study species (Storfer et al., 2018), and a distance of 150km between external sampling sites was deemed sufficient for the aims of this study. However, more extensive sampling across the entirety of their ranges would be required to adequately test the hybridisation hypothesis, with the aim of capturing the full genetic continuum and parental genotypes of *C. dux* and *C. duciformis*. Nonetheless, what is underlined here is that even in taxa known to be strong bioindicators (A. L. V. Davis et al., 2004; Spector, 2006), and even those with different realised climatic niches, the impact of abiotic conditions on their distributions can be outweighed by biotic interactions. The difficulty of their integration into spatial biodiversity modelling is well-documented (Wisz et al., 2013), but developing solutions for this should remain a priority if we are to reliably fill knowledge gaps.

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Arbizu, P. M. (2017). *pairwiseAdonis: Pairwise Multilevel Comparison using Adonis*. https://github.com/pmartinezarbizu/pairwiseAdonis

Balinsky, B. I. (1962). Patterns of animal distribution on the African Continent. *Annals of the Cape Provincial Museums*, *2*, 299–310.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, *16*, 113–148.

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, *5*(1), 180214. https://doi.org/10.1038/sdata.2018.214

Benestan, L., Moore, J., Sutherland, B. J. G., Le Luyer, J., Maaroufi, H., Rougeux, C., Normandeau, E., Rycroft, N., Atema, J., Harris, L. N., Tallman, R. F., Greenwood, S. J., Clark, F. K., & Bernatchez, L. (2017). Sex matters in massive parallel sequencing: Evidence for biases in genetic parameter estimation and investigation of sex determination systems. *Molecular Ecology*, *26*(24), 6767–6783. https://doi.org/10.1111/mec.14217

Benjamini, Y., & Hochberg, Y. (2018). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Beynon, S. A., Wainwright, W. A., & Christie, M. (2015). The application of an ecosystem services framework to estimate the economic value of dung beetles to the U.K. cattle industry. *Ecological Entomology*, *40*(S1), 124–135. https://doi.org/10.1111/een.12240

Blades, B. (2025). Chapter Two: The impact of taxonomic revision on species distribution modelling. In *Digging deeper: Using Afrotropical dung beetles to better understand quality and coverage of biodiversity data*. Doctoral Thesis.

Broad Institute. (2019). *Picard Toolkit* [Computer software]. https://broadinstitute.github.io/picard/

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., & Keller, S. R. (2020). Genomic Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes. *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 245–269. https://doi.org/10.1146/annurev-ecolsys-020720-042553

Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. *Molecular Biology and Evolution*, *36*(4), 852–860. https://doi.org/10.1093/molbev/msz008

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8

Chaulk, A., & Keyghobadi, N. (2022). Insect Landscape Genomics. In J. Dupuis & O. P. Rajora (Eds), *Population Genomics: Insects*. Springer Nature.

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, *2*(2), e107. https://doi.org/10.1002/imt2.107

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, T., Xu, J., Wang, L., Wang, H., You, E., Deng, C., Bian, H., & Shen, Y. (2023). Landscape genomics reveals adaptive genetic differentiation driven by multiple environmental variables in naked barley on the Qinghai-Tibetan Plateau. *Heredity*, *131*(5–6), 316–326. https://doi.org/10.1038/s41437-023-00647-0

Cotterill, F. (2002). *Mammal collections and biodiversity conservation in the Ikelenge Pedicle, Mwinilunga District, Northwest Zambia* (Occasional Publications in Biodiversity No. 10). Biodiversity Foundation for Africa.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Dauphin, B., Rellstab, C., Wüest, R. O., Karger, D. N., Holderegger, R., Gugerli, F., & Manel, S. (2023). Re-thinking the environment in landscape genomics. *Trends in Ecology & Evolution*, *38*(3), 261–274. https://doi.org/10.1016/j.tree.2022.10.010

Davis, A. L. V., Scholtz, C. H., Dooley, P. W., Bham, N., & Kryger, U. (2004). Scarabaeine dung beetles as indicators of biodiversity, habitat transformation and pest control chemicals in agro-ecosystems. *South African Journal of Science*.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Frichot, E., & Francois, O. (2015). LEA: an R package for Landscape and Ecological Association studies. *Methods in Ecology and Evolution*. http://membres-timc.imag.fr/Olivier.Francois/lea.html

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, *30*(7), 1687–1699. https://doi.org/10.1093/molbev/mst063

Funk, W. C., Forester, B. R., Converse, S. J., Darst, C., & Morey, S. (2019). Improving conservation policy with genomics: A guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists. *Conservation Genetics*, *20*(1), 115–134. https://doi.org/10.1007/s10592-018-1096-1

Gain, C., & François, O. (2021). LEA 3: Factor models in population genetics and ecological genomics with R. *Molecular Ecology Resources*, *21*(8), 2738–2748. https://doi.org/10.1111/1755-0998.13366

González-Molina, M., Martínez-Hernández, N., & Rico, Y. (2024). Genetic structure and demographic history of the dung beetle *Deltochilum guildingii* (Scarabaeinae): Implications for conservation of the Tropical Dry Forest in the Colombian caribbean. *Journal of Insect Conservation*, *28*(6), 1211–1221. https://doi.org/10.1007/s10841-024-00618-8

Gruber, B., Unmack, P. J., Berry, O. F., & Georges, A. (2018). dartr: An R package to facilitate analysis of SNP data generated from reduced representation genome

sequencing. *Molecular Ecology Resources*, *18*, 691–699. https://doi.org/10.1111/1755-0998.12745

Gustafson, G. T., Glynn, R. D., Short, A. E. Z., Tarasov, S., & Gunter, N. L. (2023). To design, or not to design? Comparison of beetle ultraconserved element probe set utility based on phylogenetic distance, breadth, and method of probe design. *Insect Systematics and Diversity*, *7*(4), 4. https://doi.org/10.1093/isd/ixad014

Gwee, C. Y., Metzler, D., Fuchs, J., & Wolf, J. B. W. (2025). Reconciling Gene Tree Discordance and Biogeography in European Crows. *Molecular Ecology*, *34*(10), e17764. https://doi.org/10.1111/mec.17764

Hemmings, Z., Evans, M. J., & Andrew, N. R. (2025). Spatial and temporal trends in dung beetle research. *PeerJ*, *13*, e18907. https://doi.org/10.7717/peerj.18907

Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S. M., McGrath, S. P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F. B. T., … Crouch, J. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, *11*(1), 6130. https://doi.org/10.1038/s41598-021-85639-y

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, *68*(1–2), 87–112. https://doi.org/10.1111/j.1095-8312.1999.tb01160.x

Hijmans, R. J. (2025). *terra: Spatial Data Analysis*. https://doi.org/10.32614/CRAN.package.terra

Hijmans, R. J., Barbosa, M., Ghosh, A., & Mandel, A. (2024). *geodata: Download Geographic Data*. https://doi.org/10.32614/CRAN.package.geodata

Jombart, T. (2008). adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btr521

Jombart, T., & Collins, C. (2022). *A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.1.6*.

https://github.com/thibautjombart/adegenet/raw/master/tutorials/tutorial-dapc.pdf

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*(1), 94. https://doi.org/10.1186/1471-2156-11-94

Keller, A. G., Dahlhoff, E. P., Bracewell, R., Chatla, K., Bachtrog, D., Rank, N. E., & Williams, C. M. (2023). Multi-locus genomic signatures of local adaptation to snow across the landscape in California populations of a willow leaf beetle. *Proceedings of the Royal Society B: Biological Sciences*, *290*(2005), 20230630. https://doi.org/10.1098/rspb.2023.0630

Kirk-Spriggs, A. H., & McGregor, G. (2009). Disjunctions in the Diptera (Insecta) fauna of the Mediterranean Province and southern Africa and a discussion of biogeographical considerations. *Transactions of the Royal Society of South Africa*, *64*(1), 32–52. https://doi.org/10.1080/00359190909519236

Lawson, D. J., Van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*(1), 3258. https://doi.org/10.1038/s41467-018-05257-7

Leal, M. E. (2004). *The African rain forest during the Last Glacial Maximum, an archipelago of forests in a sea of grass* [PhD thesis]. Wageningen University.

Legendre, P., & Legendre, L. (2012). *Numerical ecology* (3rd English ed.). Elsevier.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Macagno, A. L. M., Pizzo, A., Parzer, H. F., Palestrini, C., Rolando, A., & Moczek, A. P. (2011). Shape—but Not Size—Codivergence between Male and Female Copulatory Structures in *Onthophagus* Beetles. *PLoS ONE*, *6*(12), e28893. https://doi.org/10.1371/journal.pone.0028893

Maley, J. (1996). The African rain forest – main characteristics of changes in vegetation and climate from the Upper Cretaceous to the Quaternary. *Proceedings of the Royal Society of Edinburgh. Section B. Biological Sciences*, *104*, 31–73. https://doi.org/10.1017/S0269727000006114

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nature Protocols*, *12*(2), 213–218. https://doi.org/10.1038/nprot.2016.182

Mijangos, J. L., Berry, O. F., Pacioni, C., & Georges, A. (2022). dartR v2: An accessible genetic analysis platform for conservation, ecology and agriculture. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13918

Mykhailenko, A., Zieliński, P., Bednarz, A., Schlyter, F., Andersson, M. N., Antunes, B., Borowski, Z., Krokene, P., Melin, M., Morales-García, J., Müller, J., Nowak, Z., Schebeck, M., Stauffer, C., Viiri, H., Zaborowska, J., Babik, W., & Nadachowska-Brzyska, K. (2024). Complex Genomic Landscape of Inversion Polymorphism in Europe's Most Destructive Forest Pest. *Genome Biology and Evolution*, *16*(22), 1–23. https://doi.org/10.1093/gbe/evae263

Nichols, E., Spector, S., Louzada, J., Larsen, T., Amezquita, S., & Favila, M. E. (2008). Ecological functions and ecosystem services provided by Scarabaeinae dung beetles. *Biological Conservation*, *141*(6), 1461–1474. https://doi.org/10.1016/j.biocon.2008.04.011

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M. D., Durand, S., … Borman, T. (2025). *vegan: Community Ecology Package*. https://doi.org/10.32614/CRAN.package.vegan

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347–352), 240–242. https://doi.org/10.1098/rspl.1895.0041

Perkins, J. S. (2020). Take me to the River along the African drought corridor: Adapting to climate change. *Botswana Journal of Agriculture and Applied Sciences*, *14*(1), 60–71. https://doi.org/10.37106/bojaas.2020.77

Posit team. (2025). *RStudio: Integrated Development Environment for R*. Posit Software, PBC. http://www.posit.co/

Privé, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2018). Efficient analysis of large-scale genome-wide data with two R packages: Bigstatsr and bigsnpr. *Bioinformatics*, *34*(16), 2781–2787. https://doi.org/10.1093/bioinformatics/bty185

Purcell, S., & Chang, C. (n.d.). *PLINK 2.0* [Computer software]. www.cog-genomics.org/plink/2.0/

QGIS. (2021). *QGIS 3.22.9-Białowieża* (Version 3.22.9-Białowieża) [Computer software].

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Richards, A. (2015). *University of Oxford Advanced Research Computing*. http://dx.doi.org/10.5281/zenodo.22558

Robledo-Ruiz, D. A., Austin, L., Amos, J. N., Castrejón-Figueroa, J., Harley, D. K. P., Magrath, M. J. L., Sunnucks, P., & Pavlova, A. (2023). Easy-to-use R functions to separate reduced-representation genomic datasets into sex-linked and autosomal loci, and conduct sex assignment. *Molecular Ecology Resources*, 1–21. https://doi.org/10.1111/1755-0998.13844

Sgrò, C. M., Lowe, A. J., & Hoffmann, A. A. (2011). Building evolutionary resilience for conserving biodiversity under climate change. *Evolutionary Applications*, *4*(2), 326–337. https://doi.org/10.1111/j.1752-4571.2010.00157.x

Sota, T., & Kubota, K. (1998). Genital lock-and-key as a selective agent against hybridization. *Evolution*, *52*(5), 1507–1513. https://doi.org/10.1111/j.1558-5646.1998.tb02033.x

Spector, S. (2006). Scarabaeine dung beetles (Coleoptera: Scarabaeidae: Scarabaeinae): An invertebrate focal taxon for biodiversity research and conservation. *The Coleopterists Bulletin*, *60*, 71–83.

Storfer, A., Patton, A., & Fraik, A. K. (2018). Navigating the Interface Between Landscape Genetics and Landscape Genomics. *Frontiers in Genetics*, *9*, 68. https://doi.org/10.3389/fgene.2018.00068

Takano, H. (2018). *A systematic revision of the Afrotropical members of the dung beetle genus* Catharsius *Hope, 1837 (Coleoptera: Scarabaeidae)* [DPhil Thesis]. University of Oxford.

Werger, M. J. A. (1978). *Biogeography and Ecology of Southern Africa*. Dr W. Junk bv Publishers.

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M., Normand, S., Öckinger, E.,

Schmidt, N. M., … Svenning, J. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, *88*(1), 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

WorldClim. (2020, 2022). *Bioclimatic variables*. https://www.worldclim.org/data/bioclim.html

CHAPTER 6

Discussion

Questions about data are fundamentally questions about knowledge. In a time characterised so distinctly by technological advance, descriptions of what we know are driven by objective, measurable, reproducible evidence. However, how reliable really is that information? In this thesis, I explored the quality, coverage, and congruence of data, and how this affects our understanding of distributions of biodiversity.

The impact of the rise of "Big Data" in recent decades has been pervasive across disciplines and even industries. Technological advance has facilitated data collection and storage on unprecedented scales, facilitating the analysis of patterns at much wider spatial and temporal extents. The study of biodiversity is no exception, with massive increases in data not just on species occurrences and the environment, but also genetics, traits, movement, and population demographics (Wüest et al., 2020). These have not only improved our understanding of species distributions but also improved predictions of the impacts of climate and land-use change, as well as disturbance regimes and invasive species (Franklin et al., 2017; Heberling et al., 2021; Soberón & Peterson, 2004). Alongside the opportunities it presents, there has been a concerted effort to describe its limitations, with criticisms including spatial and taxonomic bias in records, outdated or unstable taxonomy, erroneous or missing metadata, and inconsistency between sources in aggregated databases (Peterson et al., 2010; Soberón et al., 2002; Soberón & Peterson, 2004; Troudet et al., 2017). Of accepted pitfalls in data quality, biased records and unreliable taxonomy present a particular problem for insects. Much like taxonomy's general decline in popularity, insect taxonomic expertise has dropped dramatically (Hochkirch et al., 2022). On top of this, insects have been found to be the most under-represented class in occurrence data, with an estimated deficit of 200

million records on the Global Biodiversity Information Facility (GBIF) (García-Rosello et al., 2023; Rocha-Ortega et al., 2021; Troudet et al., 2017). These knowledge and data coverage deficits continue to present barriers to insect conservation at a time when they're suffering from high rates of extinction despite their importance for global ecosystem health (Cardoso et al., 2011). As if this wasn't bad enough, the coverage of biodiversity data is known to be biased away from the tropics to such a degree that it appears to negatively correlate with species richness (Collen et al., 2008). The outlook for tropical insects, then, is ominous.

Fortunately, this is not to say that there is nobody working to lessen these shortfalls. The African Natural History Research Trust is a museum and research organisation focused on the taxonomy and biogeography of African invertebrates. In collaboration with them, the biggest ever taxonomic revision of any group of dung beetles was recently completed, focusing on the Afrotropical members of the dung beetle genus *Catharsius* (Takano, 2018). The up-to-date understanding of their taxonomy and the extensive compilation of distributional data, much of which was not available online, presented an opportunity to assess the impact of these specific problems not just on our understanding of their distributions, but on the outcomes of models that frequently use these types of data.

One of the notable changes from the taxonomic revision of *Catharsius* was the distribution of *Catharsius dux*. Previous understanding of its taxonomy meant that individuals were allegedly dispersed across much of the Afrotropical realm and in vastly different habitats (Ferreira, 1960a, 1960b, 1972). Their re-identification to six separate species meant that *C. dux* is now known to be confined to the forest-savanna mosaics

south of the Congo Basin. Spatial bias in occurrence records and the inability of a commonly-used (albeit heavily criticised) evaluation metric to identify data quality issues have driven warnings against automated use of GBIF occurrences in species distribution models (SDMs) (J. Beck et al., 2014). The way in which Takano's (2018) revision updated species identification and also had a knock on effect on our understanding of their ecology inspired me to test whether taxonomic error is yet another reason why users should take care to vet data quality manually before use in modelling. Extracting the occurrence data from the taxonomic revision, I compared individual and ensemble SDMs for the aforementioned six species with those made using occurrences identified as *C. dux* by Ferreira. Results unanimously showed that SDMs using occurrences identified according to updated taxonomy perform better, specifically because those using outdated taxonomy cannot find areas of high or low suitability habitat, impacting our ability to characterise species' niches, assess their vulnerability to climate change, and predict range shifts. Critically, much like Beck et al. (2014) finding the frequently used Area Under the Receiver Operating Curve (AUC) to be ineffective in flagging data quality problems, ensemble modelling – a recommended protocol to overcome the varying strengths and weaknesses of individual modelling algorithms (Marmion et al., 2009) – was found to obscure data errors and perform above the threshold for excellence in three separate metrics. This is not to say that these errors were obscured beyond being found; ensemble modelling is not a 'black box', just a closed one. It was straightforward to ascertain that there was a bigger difference between the performance of individual replicates and the ensemble model for *C. dux sensu* Ferreira than current species, but the motivation to go looking was driven by prior knowledge of the taxonomic error. Easy access to millions of species occurrence records and the

development of graphical user interfaces for species distribution modelling, such as that of Maxent, have facilitated their widespread use, and despite advice to the contrary, default model settings are used in many scenarios (Morales et al., 2017). It stands to reason, then, that without prior knowledge of a taxonomic error, much clearer instructions to inspect individual model replicate performance are needed to prevent misinformed practical recommendations. This thesis is therefore clear in its agreement that automated use of species occurrence data in SDMs should be avoided, and clearer again in the assertion that the impact of improved taxonomic accuracy goes beyond species identification, and the declining support, funding, and training in taxonomy threatens our ability to properly understand distributions of biodiversity. Future research should prioritise ways in which taxonomic expertise can be more easily integrated in studies using occurrence data. There has been some success in improving SDMs by weighting presence records according to the number of times a species was recorded over time (Zhang et al., 2020), and perhaps weighting according to the likelihood of taxonomic accuracy would be similarly successful whilst not being as demanding on resources as a full taxonomic revision.

Another opportunity presented by the taxonomic revision was as a consequence of the provenance of its data. All the specimens included in the revision were inspected by hand, and many came from institutions whose collections had not yet been digitised and / or shared with GBIF. As described, the data deficit for insects and the tropics is well-understood, but data points from offline natural history collections allowed for an extension of the question—do data to fill those deficits exist, and they just haven't been shared, or are they under sampled full stop? Unsurprisingly, the study agreed that

inventory completeness on GBIF is not yet good enough for it to act as a standalone data source. However, it found that not-yet-mobilised collections may have the answers, as their records increase the quantity of available data, and also disproportionately fill spatial and environmental gaps in sampling, in particular in rare climates. Data gaps may not be gaps in knowledge, then, but in sharing. Fortunately, GBIF's strength lies in its being a central aggregator of databases, meaning that in theory this could be relatively easily overcome, and although this wouldn't solve coverage problems for underrepresented regions and taxa, it would lessen them. In reality, though, this is much more complicated, and forces us to recognise that scientific study does not happen in a vacuum, but is fundamentally entwined with socioeconomics and political stability. Whilst this chapter suggested that widened data sharing agreements could lessen bias related to GDP per capita, sampling coverage is also known to be related to political stability (Hilario-Husain et al., 2024). This is increasingly complicated given the drastic rise in global violence in recent years, and particularly for Africa, which has seen its number of state-based conflicts almost double between 2013 and 2023 (Rustad, 2024). It's possible that reduced sampling in the north in the *Catharsius* data could be as a consequence of the warmer climates and more difficult collecting conditions, but it also corresponds with the Sahel countries, which are experiencing a long and worsening period of extreme violent conflict (Center for Preventive Action, 2024; Nsaibia, 2024; Raleigh et al., 2021). Attempting to disentangle the sociopolitical and climatic drivers of knowledge gaps risks undermining their intersectional nature, but advice has been developed to make progress in this particular area (Hilario-Husain et al., 2024), and tools developed to explore this relationship (Zizka et al., 2021).

In the first half of this thesis, a dataset with both reliable taxonomy and improved coverage for an underrepresented taxon and region allowed me to analyse the impact of accepted gaps and biases in biodiversity data, what we would refer to as shortfalls in knowledge. However, at this point, it would be remiss to not also recognise that 'what we know', as reflected by the quality and completeness of biodiversity data, is not actually 'what is known'. Recognition of different ways of knowing, such as indigenous and local knowledge, is improving, but attempts to integrate different systems with our own have been criticised for power imbalances and the distortion of local knowledge that happens when documented in our frameworks (Ruheza & Kilugwe, 2012). Indeed, on fieldwork for this project, local communities demonstrated an in-depth understanding of when and where to find *Catharsius* beetles that far outweighed that of myself and our fieldwork team. The magnitude and delicacy of the topic precludes it from being addressed at the depth it warrants in this thesis, but for an introduction see Rodrigues et al. (2022) and the references therein.

The second half of this thesis focuses on the integration of genetic information with morphological and environmental data. When *Catharsius* was revised, and the distribution of *C. dux* was better understood, it also flagged the adjacency of its range with its closest relative *Catharsius duciformis*. Evidence suggested that they had never been caught together, and were divided by a line that ran along the orientation of the Kabompo River (Takano, 2018). The SDMs created as part of the second chapter explained this distribution fairly well for *C. dux,* but not for *C. duciformis*, for whom habitat was suitable beyond the western boundary of where it had been sampled. It was suggested there that local adaptation, obscured by the scale of the variables and the

model, could be responsible and recommended a landscape genomics approach. Interested specifically in whether environment could explain the dividing line between their observed ranges, beetles were collected along a transect perpendicular to this hypothesised barrier. However, having returned to the UK and sequenced their whole genomes, I found that there was no clear and consistent genetic distinction between the two species that could be inferred from these specimens, despite their (albeit low) morphological distinctiveness. Mismatches in the way that morphological and molecular data describe species relationships and diversity have been explored (Van Den Ende et al., 2023), but it was thought that this may be at least in part down to the use of molecular markers rather than whole genome single nucleotide polymorphisms (SNPs) (A. A. Alves et al., 2013; R. M. Alves et al., 2017). A lack of consensus on this (Darkwa et al., 2020; Kadoić Balaško et al., 2021) was the inspiration for Chapter 4.

Using the trait matrix from the taxonomic revision, which details the diagnostic features of each species, I created a distance matrix that described similarity between species based on their shared (or not) morphological characteristics. Comparing this with a second distance matrix based on species similarity as described by the SNPs, I found that morphological and whole genome molecular data somewhat agree on phylogenetic relationships. Describing the same overall structure, but disagreeing on the degree of intra-cluster relatedness, they are complimentary datasets and support the use of an integrated approach in taxonomy and phylogenetics, as argued elsewhere (Keating et al., 2023). Critically, though, this chapter also draws attention to the disparity between what is ideal and what is possible. Situated in the context of the wider thesis, it refers to the impossibility of employing the use of genetic information in all instances of species

identification at a time of mass record digitisation. This research supports that visual identification is still a valid approach, despite a push for DNA prominence (as described in Zamani et al. (2021), but also introduces a wider point of interest about biodiversity data usage, specifically its interpretation. Even with a reliable, good quality dataset – as that of *Catharsius* is understood to be – different types of data do not always agree. This highlights the way in which data plays to our need to describe and categorise in a way that may be too simple to reflect the reality of nature (Sandberg et al., 2025). Disagreement on species concepts, which aligns with the contrasts between molecular and morphological approaches to describing species relatedness, is one example of the way that epistemological plurality manifests in the study of biodiversity, and underlines that there cannot be an objective truth when interpreted through the lens of our subjectivity. Even perfect data would not constitute perfect understanding, it is too much affected by our humanness.

Nonetheless, imperfect knowledge is better than none, and the improved complexity achieved through the integration of different data types may be a way to move closer to more accurate descriptions of natural realities. Dung beetles are known to be effective bioindicators at a variety of spatial scales (A. L. V. Davis et al., 2004; McGeoch et al., 2002), so I hoped that combining geographical, environmental, morphological, and molecular data in Chapter 5's study of landscape genomics would elucidate the biogeography of *C. dux* and *C. duciformis* in a way that the SDMs weren't fully able to do. However, the variation in the SNPs couldn't be well explained by where individuals had been collected, or the environmental conditions in those places and, as described above, genetic identify was also not linked to morphological classifications. A clear

genetic gradient that was nevertheless apparent, suggesting that this study took place entirely within a hybrid zone, the location of which aligns with a known paleoclimatic feature (Balinsky, 1962; Perkins, 2020). In a fitting example of data disagreement, this cline was not reflected in their morphologies, which remained distinctive across the dividing line. The only exception was a small number of beetles from the easternmost sampling location, which had a weak link to a particular genetic cluster and identification as *C. duciformis*, so may have been the edge of this potential zone. Although its western limit was not captured by sampling, it is reasonable to hypothesise this secondary hybrid zone as the barrier to *C. duciformis*' expansion into suitable habitat to the west.

Aptly, this brings the data analysis chapters of this thesis full circle. Difficulty including biotic interactions is a key criticism of correlative SDMs, with tools somewhat incorporating these processes still requiring prior knowledge of which species to include in the first place (Wisz et al., 2013). Encapsulated by Chapters 2 and 5, this remains an important direction for further research. Furthermore, the cyclical nature of discovery within this body of work captures wider questions about knowledge acquisition. Namely, whether biodiversity data are good enough to infer new knowledge or only to support our pre-existing understanding, and how we can balance positive outcomes from model and data complexity with knowledge of process. As closing remarks, I will address these below.

How do we balance positive outcomes from model and data complexity with knowledge of process? Integration of the different data types improved understanding of the dynamics governing *C. dux* and *C. duciformis*, as well as species relationships within the

wider genus. Together with assertions in this discussion that simple categorisation is not sufficient to describe the complexity of the natural world, this thesis seems fairly straightforward in its support of increased complexity. However, it also underlines the importance of understanding process. For example, in the name of a model that generated an excellent validation score, the ensemble SDMs obscured key steps that would inform us about their (un)reliability. Balancing complexity and interpretability has been discussed before, and overly complex models criticised for their reduced transferability in time and space (Bell & Schlaepfer, 2016; Moreno-Amat et al., 2015; Warren et al., 2014). I add to this that an understanding of how data are being treated throughout modelling procedures is not just key in interpreting results, but also in judging whether they are biologically realistic. Earlier, I described these ensembles as 'closed boxes' as, although the overall pipeline can be opaque, it is not difficult to find decisions that have been made if you go looking. In contrast, the 'black box', oft-used to describe artificial intelligence (AI) tools, references the almost insurmountable challenge in picking apart what is happening within due to their overwhelming complexity. As machine learning has been widely adopted in the field of species distribution modelling, there are ongoing attempts to improve knowledge of their process in recognition of its importance (Ryo et al., 2021). The use of AI in taxonomy, however, is in its infancy and although the speed with which it could operate is theoretically promising in the face of so many undescribed species, the importance of continued human contribution to its development is underlined (Bernard, 2025; Karbstein et al., 2024; Valdecasas, 2024). As it advances, adopters should take pains to understand its 'thought' process in a field which is already so fraught with disagreement. Model complexity can be positive when it

mirrors the intricacies of nature, but should not come at a cost of unclear data handling, or else we risk missing errors in its quality.

Are biodiversity data good enough to infer new knowledge or only to support our pre-existing understanding? This thesis adds to the body of literature that shows imperfections in data from taxonomic errors and poor coverage threaten descriptions of biodiversity and our ability to predict future changes. In this sense, we should still be cautious in our inferences. However, it also goes further, illustrating that even with reliable taxonomy and improved sampling, conflicts in data interpretation, for example of what a species is or how they are related, can make progress complicated. Even if we were able to access 'perfect' data, we need pre-existing biological understanding to make sense of it. For example, reliable taxonomy as an important prerequisite of modelling, and the human input that is so critical in progressing taxonomic understanding described above. Furthermore, prior knowledge of species interactions are required to test them; although genomic analysis did generate new and unexpected knowledge of a potential hybrid zone, this study was informed in the first place by a query about why the distributions of *C. dux* and *C. duciformis* seemed not to overlap. Scientific advance is well understood to be an iterative process, but is it possible to use these data to start new cycles? Future research should explore how drastic increases in available records can be leveraged to infer interactions as part of biogeographical studies, rather than before them. In the spirit of this added complexity, and because this would necessitate records at a finer resolution in time, more complicated climatic phenomena could be included, rather than averages over extended time periods. In the region of this

study, this could include the movement of the Congo Air Boundary and how the timing of this changes year to year (Howard & Washington, 2019; Knight & Washington, 2024).

At the beginning of this discussion, I said that questions about data are fundamentally questions about knowledge, but this is an oversimplification. They represent some measures of what we know, but not what is known. They describe some ways of knowing but, as our creation, are fundamentally limited by our subjectivity. And, even if they were perfect, a lack of an objective truth in a field like evolutionary biology that, by its very nature, is constantly changing, necessitates that they would not stay that way over time. Descriptions of biodiversity are, of course, critical for its protection, but applications of data would perhaps be more useful if we try to bear these limitations in mind.

References

Alves, A. A., Bhering, L. L., Rosado, T. B., Laviola, B. G., Formighieri, E. F., & Cruz, C. D. (2013). Joint analysis of phenotypic and molecular diversity provides new insights on the genetic variability of the Brazilian physic nut germplasm bank. *Genetics and Molecular Biology*, *36*(3), 371–381. https://doi.org/10.1590/S1415-47572013005000033

Alves, R. M., Silva, C. R. D. S., Albuquerque, P. S. B. D., & Santos, V. S. D. (2017). Phenotypic and genotypic characterization and compatibility among genotypes to select elite clones of cupuassu. *Acta Amazonica*, *47*(3), 175–184. https://doi.org/10.1590/1809-4392201602104

Balinsky, B. I. (1962). Patterns of animal distribution on the African Continent. *Annals of the Cape Provincial Museums*, *2*, 299–310.

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Bell, D. M., & Schlaepfer, D. R. (2016). On the dangers of model complexity without ecological justification in species distribution modeling. *Ecological Modelling*, *330*, 50–59. https://doi.org/10.1016/j.ecolmodel.2016.03.012

Bernard, J. (2025). Combining new technology with classic taxonomy to overcome hurdles to discovering dark taxa. *Systematics and Biodiversity*, *23*(1), 2454014. https://doi.org/10.1080/14772000.2025.2454014

Cardoso, P., Erwin, T. L., Borges, P. A. V., & New, T. R. (2011). The seven impediments in invertebrate conservation and how to overcome them. *Biological Conservation*, *144*(11), 2647–2655. https://doi.org/10.1016/j.biocon.2011.07.024

Center for Preventive Action. (2024). *Violent Extremism in the Sahel*. Global Conflict Tracker. https://cfr.org/global-conflict-tracker/conflict/violent-extremism-sahel

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: Addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75–88. https://doi.org/10.1177/194008290800100202

Darkwa, K., Agre, P., Olasanmi, B., Iseki, K., Matsumoto, R., Powell, A., Bauchet, G., De Koeyer, D., Muranaka, S., Adebola, P., Asiedu, R., Terauchi, R., & Asfaw, A. (2020).

Comparative assessment of genetic diversity matrices and clustering methods in white Guinea yam (*Dioscorea rotundata*) based on morphological and molecular markers. *Scientific Reports*, *10*(1), 13191. https://doi.org/10.1038/s41598-020-69925-9

Davis, A. L. V., Scholtz, C. H., Dooley, P. W., Bham, N., & Kryger, U. (2004). Scarabaeine dung beetles as indicators of biodiversity, habitat transformation and pest control chemicals in agro-ecosystems. *South African Journal of Science*.

Ferreira, M. C. (1960a). Descricao de especies novas de *Catharsius* s.str. *Novos Taxa Entomológicos*, *23*, 3–8.

Ferreira, M. C. (1960b). Revisao das esepcies Africanas de *Catharsius* s.str. Do Groupo Adamastor e descricao de especies novas. *Revista Entomologia Moçambique*, *3*(1), 1–73.

Ferreira, M. C. (1972). Os Escarabideos de Africa (sul do Saara). I. *Revista de Entomologia de Mocambique*, *11*, 5–1088.

Franklin, J., Serra-Diaz, J. M., Syphard, A. D., & Regan, H. M. (2017). Big data for forecasting the impacts of global change on plant communities. *Global Ecology and Biogeography*, *26*(1), 6–17. https://doi.org/10.1111/geb.12501

García-Rosello, E., Gonzalez-Dacosta, J., Guisande, C., & Lobo, J. M. (2023). GBIF falls short of providing a representative picture of the global distribution of insects. *Systematic Entomology*, *48*(4), 489–497. https://doi.org/10.1111/syen.12589

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, *118*(6), e2018093118. https://doi.org/10.1073/pnas.2018093118

Hilario-Husain, B. A., Tanalgo, K. C., Guerrero, S. J. C., Garcia, F. G. N., Lerios, T. E., Garcia, M. E. Z., Alvaro-Ele, R. J., Manampan-Rubio, M., Murray, S. A., Casim, L. F., Delos Reyes, J. L., Dela Cruz, K. C., Abdullah, S. S., Balase, S. M. P., Respicio, J. M. V., Lidasan, A. K., Buday, Z. S., Cabasan, Ma. T. N., Pimentel, J. L., … Agduma, A. R. (2024). Caught in the crossfire: Biodiversity conservation paradox of sociopolitical conflict. *Npj Biodiversity*, *3*(1), 10. https://doi.org/10.1038/s44185-024-00044-8

Hochkirch, A., Casino, A., Lyubomir, P., Allen, D., Tilley, L., Georgiev, T., Gospodinov, K., & Barov, B. (2022). *European Red List of Insect Taxonomists*. Publication Office of the European Union.

Howard, E., & Washington, R. (2019). Drylines in Southern Africa: Rediscovering the Congo Air Boundary. *Journal of Climate*, *32*(23), 8223–8242. https://doi.org/10.1175/JCLI-D-19-0437.1

Kadoić Balaško, M., Mikac, K. M., Benítez, H. A., Bažok, R., & Lemic, D. (2021). Genetic and Morphological Approach for Western Corn Rootworm Resistance Management. *Agriculture*, *11*(7), 585. https://doi.org/10.3390/agriculture11070585

Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., Wagner, N. D., Emerson, B. C., Albach, D. C., Scheu, S., Bradler, S., De Vries, J., Irisarri, I., Li, H., Soltis, P., Mäder, P., & Wäldchen, J. (2024). Species delimitation 4.0: Integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution*, *39*(8), 771–784. https://doi.org/10.1016/j.tree.2023.11.002

Keating, J. N., Garwood, R. J., & Sansom, R. S. (2023). Phylogenetic congruence, conflict and consilience between molecular and morphological data. *BMC Ecology and Evolution*, *23*(1), 30. https://doi.org/10.1186/s12862-023-02131-z

Knight, C., & Washington, R. (2024). Remote Midlatitude Control of Rainfall Onset at the Southern African Tropical Edge. *Journal of Climate*, *37*(8), 2519–2539. https://doi.org/10.1175/JCLI-D-23-0446.1

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., & Thuiller, W. (2009). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, *15*(1), 59–69. https://doi.org/10.1111/j.1472-4642.2008.00491.x

McGeoch, M. A., Van Rensburg, B. J., & Botes, A. (2002). The verification and application of bioindicators: A case study of dung beetles in a savanna ecosystem. In *Journal of Applied Ecology* (Vol. 39, pp. 661–672).

Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: Are we paying attention to recommendations? A systematic review. *PeerJ*, *5*, e3093. https://doi.org/10.7717/peerj.3093

Moreno-Amat, E., Mateo, R. G., Nieto-Lugilde, D., Morueta-Holme, N., Svenning, J.-C., & García-Amorena, I. (2015). Impact of model complexity on cross-temporal transferability in Maxent species distribution models: An assessment using paleobotanical data. *Ecological Modelling*, *312*, 308–317. https://doi.org/10.1016/j.ecolmodel.2015.05.035

Nsaibia, H. (2024). *Conflict Watchlist 2024—The Sahel: A Deadly New Era in the Decades-long Conflict*. ACLED. https://acleddata.com/conflict-watchlist-2024/sahel/

Perkins, J. S. (2020). Take me to the River along the African drought corridor: Adapting to climate change. *Botswana Journal of Agriculture and Applied Sciences*, *14*(1), 60–71. https://doi.org/10.37106/bojaas.2020.77

Peterson, A. T., Knapp, S., Guralnick, R., Soberón, J., & Holder, M. T. (2010). The big questions for biodiversity informatics. *Systematics and Biodiversity*, *8*(2), 159–168. https://doi.org/10.1080/14772001003739369

Raleigh, C., Nsaibia, H., & Dowd, C. (2021). The Sahel crisis since 2012. *African Affairs*, *120*(478), 123–143. https://doi.org/10.1093/afraf/adaa022

Rocha-Ortega, M., Rodriguez, P., & Córdoba-Aguilar, A. (2021). Geographical, temporal and taxonomic biases in insect GBIF data on biodiversity and extinction. *Ecological Entomology*, *46*(4), 718–728. https://doi.org/10.1111/een.13027

Rodrigues, A., Bloom, D., Zermoglio, P., Guralnick, R., Hirsch, T., Campbell, J., Ali, N., Ferrier, S., Niamir, A., Londoño, M. C., & Sica, Y. (2022). *Primary biodiversity data and the Post-2020 Global Biodiversity Framework*. GBIF Secretariat. https://doi.org/10.15468/doc-pgg2-xn60

Ruheza, S., & Kilugwe, Z. (2012). Integration of the indigenous and the scientific knowledge systems for conservation of biodiversity: Significances of their different worldviews and their win-loss relationship. *Journal of Sustainable Development in Africa*, *14*(6), 160–174.

Rustad, S. A. (2024). *Conflict Trends: A Global Overview, 1946-2023*. Peace Research Institute Oslo.

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of

black-box species distribution models. *Ecography*, *44*(2), 199–205. https://doi.org/10.1111/ecog.05360

Sandberg, O. M., Schultz, A., Guðmundsdóttir, R., & Skúlason, S. (2025). 'Species' Is Not the (Only) Unit of Biodiversity: A Process-Philosophical Perspective on Conservation Concepts. *Marine Ecology*, *46*(1), e12857. https://doi.org/10.1111/maec.12857

Soberón, J., Arriaga, L., & Lara, L. (2002). Issues of quality control in large, mixed-origin entomological databases. In H. Saarenmaa & E. S. Nielsen (Eds), *Towards a global biological information infrastructure.pdf* (pp. 15–22). European Environment Agency. https://www.eea.europa.eu/publications/technical_report_2001_70

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 689–698. https://doi.org/10.1098/rstb.2003.1439

Takano, H. (2018). *A systematic revision of the Afrotropical members of the dung beetle genus* Catharsius *Hope, 1837 (Coleoptera: Scarabaeidae)* [DPhil Thesis]. University of Oxford.

Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports*, *7*(1), 9132. https://doi.org/10.1038/s41598-017-09084-6

Valdecasas, A. G. (2024). Can Taxonomists Think? Reversing the AI Equation. *Taxonomy*, *4*(4), 713–722. https://doi.org/10.3390/taxonomy4040037

Van Den Ende, C., Puttick, M. N., Urrutia, A. O., & Wills, M. A. (2023). Why should we compare morphological and molecular disparity? *Methods in Ecology and Evolution*, *14*(9), 2390–2410. https://doi.org/10.1111/2041-210X.14166

Warren, D. L., Wright, A. N., Seifert, S. N., & Shaffer, H. B. (2014). Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. *Diversity and Distributions*, *20*(3), 334–343. https://doi.org/10.1111/ddi.12160

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J., Guisan, A., Heikkinen, R. K., Høye, T. T.,

Kühn, I., Luoto, M., Maiorano, L., Nilsson, M., Normand, S., Öckinger, E., Schmidt, N. M., … Svenning, J. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, *88*(1), 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Wüest, R. O., Zimmermann, N. E., Zurell, D., Alexander, J. M., Fritz, S. A., Hof, C., Kreft, H., Normand, S., Cabral, J. S., Szekely, E., Thuiller, W., Wikelski, M., & Karger, D. N. (2020). Macroecology in the age of Big Data – Where to go from here? *Journal of Biogeography*, *47*(1), 1–12. https://doi.org/10.1111/jbi.13633

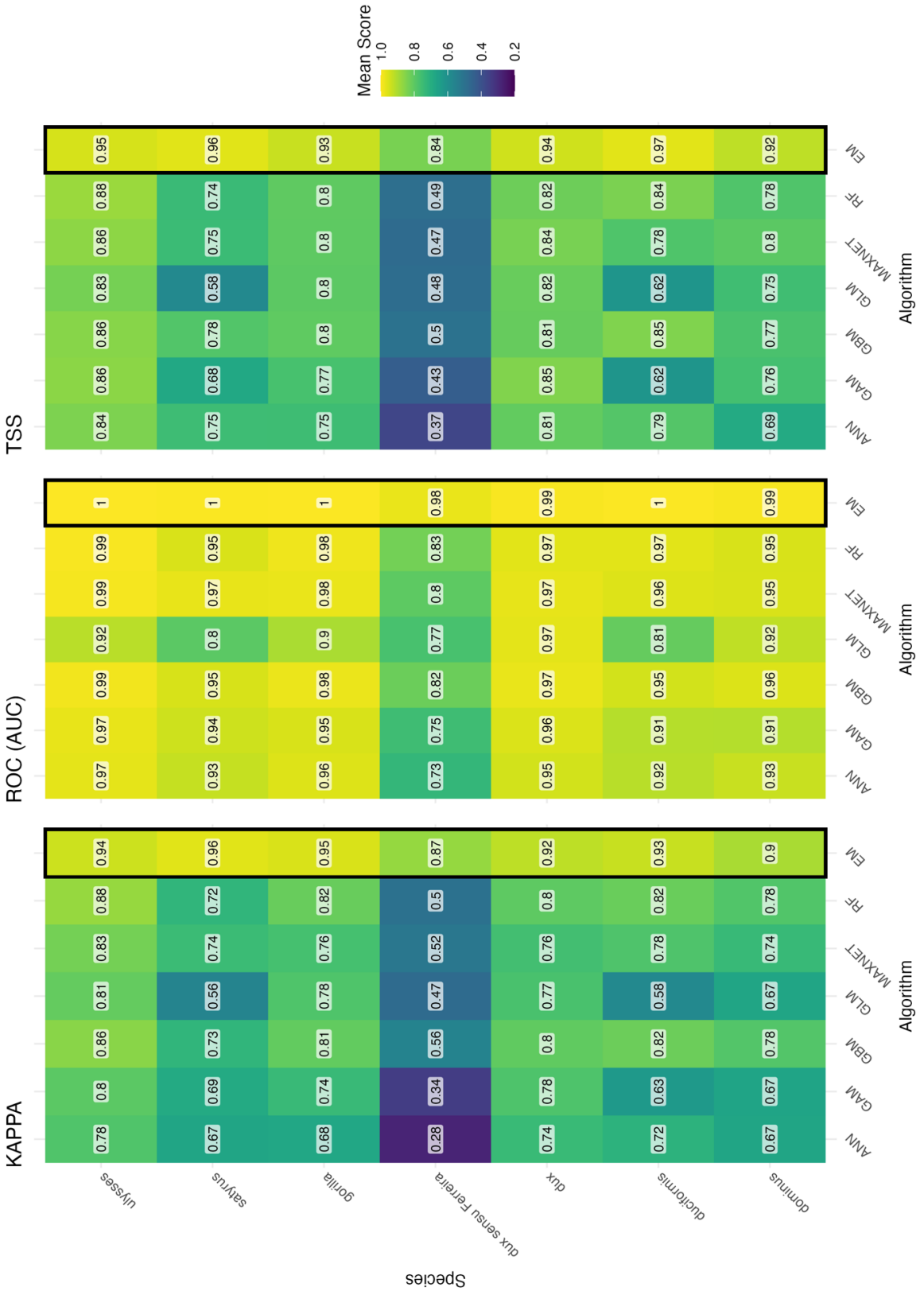
Zamani, A., Vahtera, V., Sääksjärvi, I. E., & Scherz, M. D. (2021). The omission of critical data in the pursuit of 'revolutionary' methods to accelerate the description of species. *Systematic Entomology*, *46*(1), 1–4. https://doi.org/10.1111/syen.12444

Zhang, Z., Mammola, S., & Zhang, H. (2020). Does weighting presence records improve the performance of species distribution models? A test using fish larval stages in the Yangtze Estuary. *Science of The Total Environment*, *741*, 140393. https://doi.org/10.1016/j.scitotenv.2020.140393

Zizka, A., Rydén, O., Edler, D., Klein, J., Perrigo, A., Silvestro, D., Jagers, S. C., Lindberg, S. I., & Antonelli, A. (2021). BIO-DEM, a tool to explore the relationship between biodiversity data availability and socio-political conditions in time and space. *Journal of Biogeography*, *48*(11), 2715–2726. https://doi.org/10.1111/jbi.14256

A

Supplementary Information Chapter 2



Supplementary Figure 2.1: Mean validation scores of all Catharsius *individual model replicates per evaluation metric, with ensemble model calibration score for comparison. Note that scales for metrics are not comparable to one another.*

Supplementary Table 2.1: Results of Friedman tests for all modelling algorithms and evaluation metrics tests that compare the performance of Catharsius SDMs *for all modelling algorithms and evaluation metrics. This demonstrates, for each algorithm and metric, whether there were significant differences between the model performance of each species.*

	Model	chi-sq	df	p	significant
1	KAPPA ANN	25.54	6	< 0.001	*
2	KAPPA GAM	31.27	6	< 0.001	*
3	KAPPA GBM	24.47	6	< 0.001	*
4	KAPPA GLM	31.16	6	< 0.001	*
5	KAPPA MAXNET	24.05	6	< 0.001	*
6	KAPPA RF	23.14	6	< 0.001	*
7	ROC ANN	33.50	6	< 0.001	*
8	ROC GAM	33.69	6	< 0.001	*
9	ROC GBM	31.32	6	< 0.001	*
10	ROC GLM	40.52	6	< 0.001	*
11	ROC MAXNET	35.03	6	< 0.001	*
12	ROC RF	34.26	6	< 0.001	*
13	TSS ANN	32.53	6	< 0.001	*
14	TSS GAM	30.84	6	< 0.001	*
15	TSS GBM	24.21	6	< 0.001	*
16	TSS GLM	30.60	6	< 0.001	*
17	TSS MAXNET	25.92	6	< 0.001	*
18	TSS RF	25.36	6	< 0.001	*

Supplementary Table 2.2: Significance (p values) of Conover tests that compare the performance of Catharsius SDMs *for all modelling algorithms and evaluation metrics. This demonstrates, for each algorithm and metric, which species' models performed significantly differently. The summary of these results is presented in Table 2.3.*

Model	*Species*	*dominus*	*duciformis*	*dux sensu Takano*	*dux sensu Ferreira*	*gorilla*	*satyrus*
KAPPA ANN	***duciformis***	0.9871	NA	NA	NA	NA	NA
KAPPA ANN	***dux***	0.8542	0.9986	NA	NA	NA	NA
KAPPA ANN	***dux sensu Ferreira***	0.0053	0.0002	0.0000	NA	NA	NA
KAPPA ANN	***gorilla***	0.9986	0.8542	0.5384	0.0309	NA	NA
KAPPA ANN	***satyrus***	1.0000	0.9871	0.8542	0.0053	0.9986	NA
KAPPA ANN	***ulysses***	0.3687	0.8542	0.9871	0.0000	0.1284	0.3687

KAPPA GAM	***duciformis***	0.8621	NA	NA	NA	NA	NA
KAPPA GAM	***dux***	0.2946	0.0091	NA	NA	NA	NA
KAPPA GAM	***dux sensu Ferreira***	0.0116	0.3346	0.0000	NA	NA	NA
KAPPA GAM	***gorilla***	0.8621	0.1396	0.9690	0.0001	NA	NA
KAPPA GAM	***satyrus***	1.0000	0.8621	0.2946	0.0116	0.8621	NA
KAPPA GAM	***ulysses***	0.0236	0.0002	0.9554	0.0000	0.4686	0.0236
KAPPA GBM	***duciformis***	0.9956	NA	NA	NA	NA	NA
KAPPA GBM	***dux***	0.9998	1.0000	NA	NA	NA	NA
KAPPA GBM	***dux sensu Ferreira***	0.0102	0.0009	0.0025	NA	NA	NA
KAPPA GBM	***gorilla***	0.9987	1.0000	1.0000	0.0015	NA	NA
KAPPA GBM	***satyrus***	0.9129	0.5572	0.7259	0.2453	0.6436	NA
KAPPA GBM	***ulysses***	0.3123	0.7259	0.5572	0.0000	0.6436	0.0158
KAPPA GLM	***duciformis***	0.9796	NA	NA	NA	NA	NA
KAPPA GLM	***dux***	0.1004	0.0073	NA	NA	NA	NA
KAPPA GLM	***dux sensu Ferreira***	0.6163	0.9796	0.0002	NA	NA	NA
KAPPA GLM	***gorilla***	0.0120	0.0004	0.9922	0.0000	NA	NA
KAPPA GLM	***satyrus***	1.0000	0.9977	0.0464	0.7932	0.0043	NA
KAPPA GLM	***ulysses***	0.0043	0.0001	0.9560	0.0000	1.0000	0.0014
KAPPA MAXNET	***duciformis***	0.9504	NA	NA	NA	NA	NA
KAPPA MAXNET	***dux***	0.9976	0.9994	NA	NA	NA	NA
KAPPA MAXNET	***dux sensu Ferreira***	0.0537	0.0016	0.0086	NA	NA	NA
KAPPA MAXNET	***gorilla***	1.0000	0.9340	0.9957	0.0642	NA	NA
KAPPA MAXNET	***satyrus***	1.0000	0.9884	0.9999	0.0250	0.9999	NA
KAPPA MAXNET	***ulysses***	0.0537	0.4759	0.2197	0.0000	0.0447	0.1062
KAPPA RF	***duciformis***	0.5794	NA	NA	NA	NA	NA

KAPPA RF	***dux***	0.8137	0.9998	NA	NA	NA	NA
KAPPA RF	***dux sensu Ferreira***	0.3348	0.0020	0.0082	NA	NA	NA
KAPPA RF	***gorilla***	0.8731	0.9988	1.0000	0.0127	NA	NA
KAPPA RF	***satyrus***	1.0000	0.4941	0.7429	0.4116	0.8137	NA
KAPPA RF	***ulysses***	0.0288	0.8137	0.5794	0.0000	0.4941	0.0193
ROC ANN	***duciformis***	0.9944	NA	NA	NA	NA	NA
ROC ANN	***dux***	0.5653	0.9256	NA	NA	NA	NA
ROC ANN	***dux sensu Ferreira***	0.0071	0.0005	0.0000	NA	NA	NA
ROC ANN	***gorilla***	0.0600	0.2854	0.9256	0.0000	NA	NA
ROC ANN	***satyrus***	0.9971	1.0000	0.9007	0.0007	0.2476	NA
ROC ANN	***ulysses***	0.0023	0.0248	0.3704	0.0000	0.9621	0.0196
ROC GAM	***duciformis***	0.8996	NA	NA	NA	NA	NA
ROC GAM	***dux***	0.2103	0.8996	NA	NA	NA	NA
ROC GAM	***dux sensu Ferreira***	0.0476	0.0007	0.0000	NA	NA	NA
ROC GAM	***gorilla***	0.0116	0.2823	0.9453	0.0000	NA	NA
ROC GAM	***satyrus***	0.9453	1.0000	0.8358	0.0012	0.2103	NA
ROC GAM	***ulysses***	0.0003	0.0305	0.4618	0.0000	0.9742	0.0190
ROC GBM	***duciformis***	0.9153	NA	NA	NA	NA	NA
ROC GBM	***dux***	0.8595	0.1878	NA	NA	NA	NA
ROC GBM	***dux sensu Ferreira***	0.0001	0.0110	0.0000	NA	NA	NA
ROC GBM	***gorilla***	0.9683	0.3715	0.9998	0.0000	NA	NA
ROC GBM	***satyrus***	0.9918	0.9995	0.4162	0.0023	0.6557	NA
ROC GBM	***ulysses***	0.2892	0.0141	0.9683	0.0000	0.8595	0.0535
ROC GLM	***duciformis***	0.0001	NA	NA	NA	NA	NA
ROC GLM	***dux***	0.0524	0.0000	NA	NA	NA	NA

ROC GLM	***dux sensu Ferreira***	0.0000	0.9932	0.0000	NA	NA	NA
ROC GLM	***gorilla***	0.9872	0.0039	0.0039	0.0002	NA	NA
ROC GLM	***satyrus***	0.0001	1.0000	0.0000	0.9967	0.0028	NA
ROC GLM	***ulysses***	1.0000	0.0001	0.0524	0.0000	0.9872	0.0001
ROC MAXNET	***duciformis***	0.8862	NA	NA	NA	NA	NA
ROC MAXNET	***dux***	0.1275	0.8160	NA	NA	NA	NA
ROC MAXNET	***dux sensu Ferreira***	0.0461	0.0005	0.0000	NA	NA	NA
ROC MAXNET	***gorilla***	0.1525	0.8535	1.0000	0.0000	NA	NA
ROC MAXNET	***satyrus***	0.0368	0.5265	0.9993	0.0000	0.9983	NA
ROC MAXNET	***ulysses***	0.0000	0.0062	0.2882	0.0000	0.2490	0.5785
ROC RF	***duciformis***	0.6945	NA	NA	NA	NA	NA
ROC RF	***dux***	0.8613	0.9999	NA	NA	NA	NA
ROC RF	***dux sensu Ferreira***	0.0073	0.0000	0.0000	NA	NA	NA
ROC RF	***gorilla***	0.0517	0.8254	0.6452	0.0000	NA	NA
ROC RF	***satyrus***	0.9994	0.9192	0.9821	0.0013	0.1647	NA
ROC RF	***ulysses***	0.0007	0.1384	0.0639	0.0000	0.8926	0.0042
TSS ANN	***duciformis***	0.2322	NA	NA	NA	NA	NA
TSS ANN	***dux***	0.0068	0.8807	NA	NA	NA	NA
TSS ANN	***dux sensu Ferreira***	0.0370	0.0000	0.0000	NA	NA	NA
TSS ANN	***gorilla***	0.6837	0.9910	0.4397	0.0001	NA	NA
TSS ANN	***satyrus***	0.9083	0.9083	0.1997	0.0005	0.9995	NA
TSS ANN	***ulysses***	0.0068	0.8807	1.0000	0.0000	0.4397	0.1997
TSS GAM	***duciformis***	0.3438	NA	NA	NA	NA	NA
TSS GAM	***dux***	0.3866	0.0006	NA	NA	NA	NA
TSS GAM	***dux sensu Ferreira***	0.0127	0.8663	0.0000	NA	NA	NA
TSS GAM	***gorilla***	0.9978	0.1036	0.7568	0.0015	NA	NA
TSS GAM	***satyrus***	0.9800	0.8663	0.0590	0.1459	0.7968	NA

TSS GAM	***ulysses***	0.0865	0.0000	0.9924	0.0000	0.3035	0.0060
TSS GBM	***duciformis***	0.6476	NA	NA	NA	NA	NA
TSS GBM	***dux***	0.7293	1.0000	NA	NA	NA	NA
TSS GBM	***dux sensu Ferreira***	0.0761	0.0002	0.0003	NA	NA	NA
TSS GBM	***gorilla***	0.8028	1.0000	1.0000	0.0005	NA	NA
TSS GBM	***satyrus***	0.9503	0.9956	0.9987	0.0026	0.9998	NA
TSS GBM	***ulysses***	0.1441	0.9742	0.9503	0.0000	0.9143	0.7293
TSS GLM	***duciformis***	0.0884	NA	NA	NA	NA	NA
TSS GLM	***dux***	0.7168	0.0004	NA	NA	NA	NA
TSS GLM	***dux sensu Ferreira***	0.0103	0.9925	0.0000	NA	NA	NA
TSS GLM	***gorilla***	0.9209	0.0021	0.9996	0.0001	NA	NA
TSS GLM	***satyrus***	0.2030	0.9999	0.0016	0.9412	0.0081	NA
TSS GLM	***ulysses***	0.7595	0.0005	1.0000	0.0000	0.9999	0.0021
TSS MAXNET	***duciformis***	1.0000	NA	NA	NA	NA	NA
TSS MAXNET	***dux***	0.8503	0.7831	NA	NA	NA	NA
TSS MAXNET	***dux sensu Ferreira***	0.0099	0.0154	0.0000	NA	NA	NA
TSS MAXNET	***gorilla***	0.9041	0.8503	1.0000	0.0001	NA	NA
TSS MAXNET	***satyrus***	0.9999	1.0000	0.6627	0.0290	0.7452	NA
TSS MAXNET	***ulysses***	0.2215	0.1673	0.9439	0.0000	0.9041	0.1050
TSS RF	***duciformis***	0.7524	NA	NA	NA	NA	NA
TSS RF	***dux***	0.5405	0.9999	NA	NA	NA	NA
TSS RF	***dux sensu Ferreira***	0.0462	0.0002	0.0000	NA	NA	NA
TSS RF	***gorilla***	0.8831	1.0000	0.9973	0.0005	NA	NA
TSS RF	***satyrus***	0.9997	0.9285	0.7896	0.0136	0.9807	NA
TSS RF	***ulysses***	0.1298	0.9285	0.9872	0.0000	0.8239	0.2958

Supplementary Table 2.3: The Catharsius *ensemble model projections produced rasters within which each cell has a suitability value. Each cell was classified as low (0-250), average-low (251-500), average-high (501-750), and high (751-100) suitability, and the total number of cells within each bin totalled for each species*

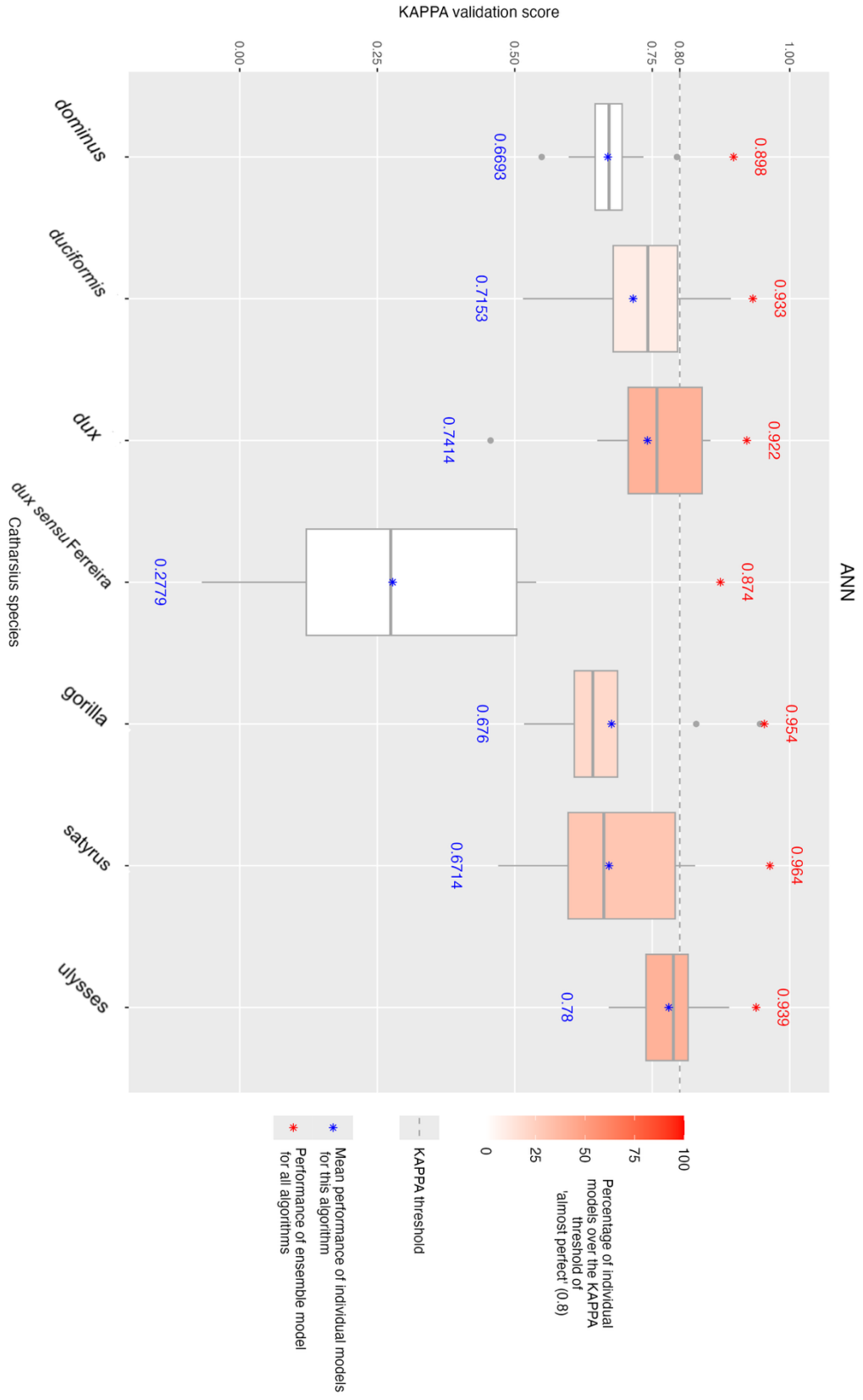
	Species	**0-250**	**251-500**	**501-750**	**751-100**
1	*dux sensu Ferreira*	15302773	5926600	2746834	687420
2	*dux sensu Takano*	19924424	2089837	1393741	1255625
3	*duciformis*	19864334	2198552	1574617	1026124
4	*dominus*	19936679	2807425	1306451	613072
5	*gorilla*	20168669	1684970	1484967	1325021
6	*satyrus*	18817900	3026640	2184095	634992
7	*ulysses*	22016645	951170	876915	818897

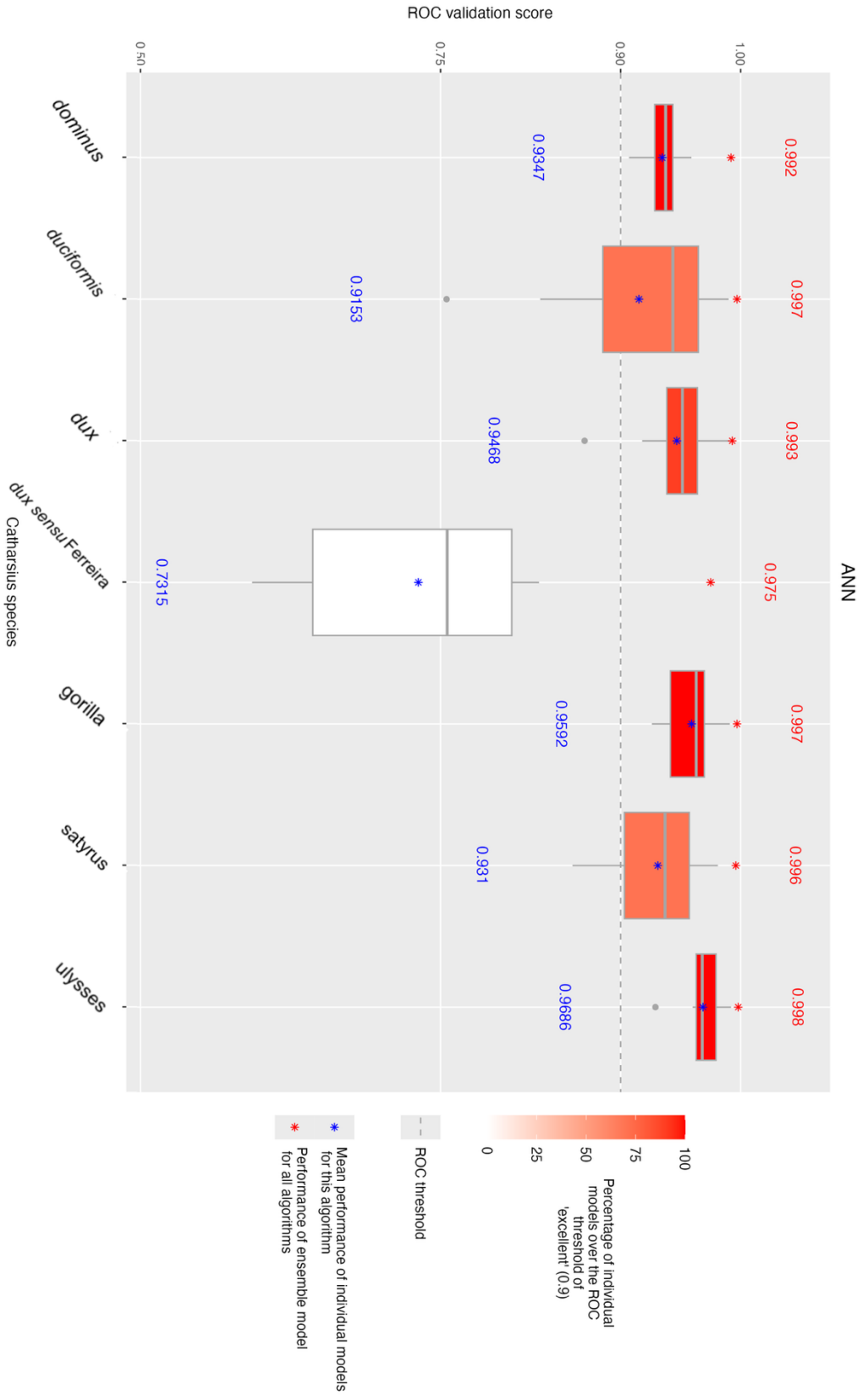
Supplementary Table 2.4: The Catharsius *ensemble model projections produced rasters within which each cell has a suitability value, and the minimum and maximum suitability values within each raster are displayed here*

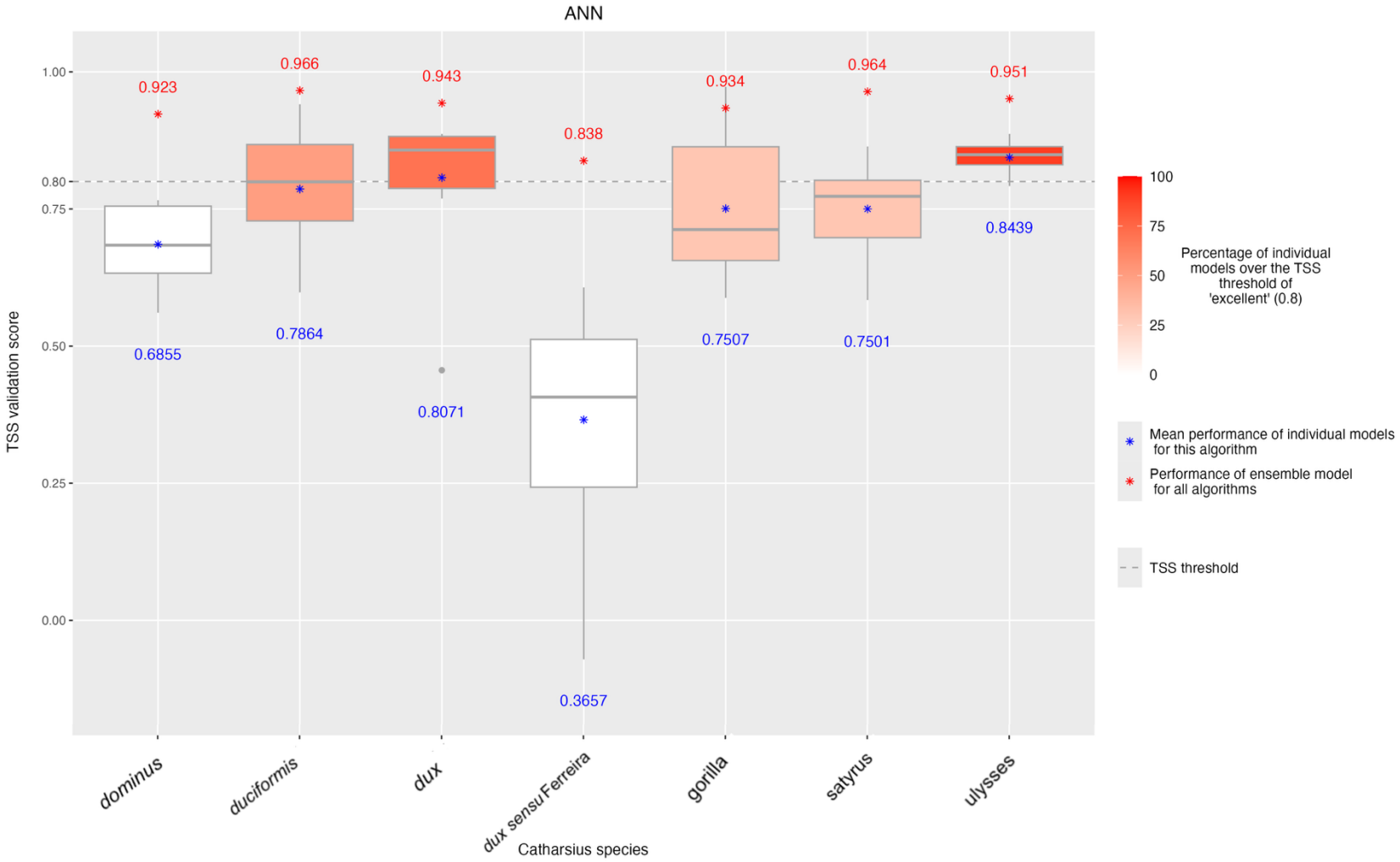
Species	**Min**	**Max**
dominus	10	922
duciformis	8	895
dux sensu Takano	9	912
dux sensu Ferreira	16	872
gorilla	8	926
satyrus	10	997
ulysses	7	926

Supplementary Figure 2.2 Full size images of individual Catharsius model replicates across all individual algorithms and evaluation metrics.

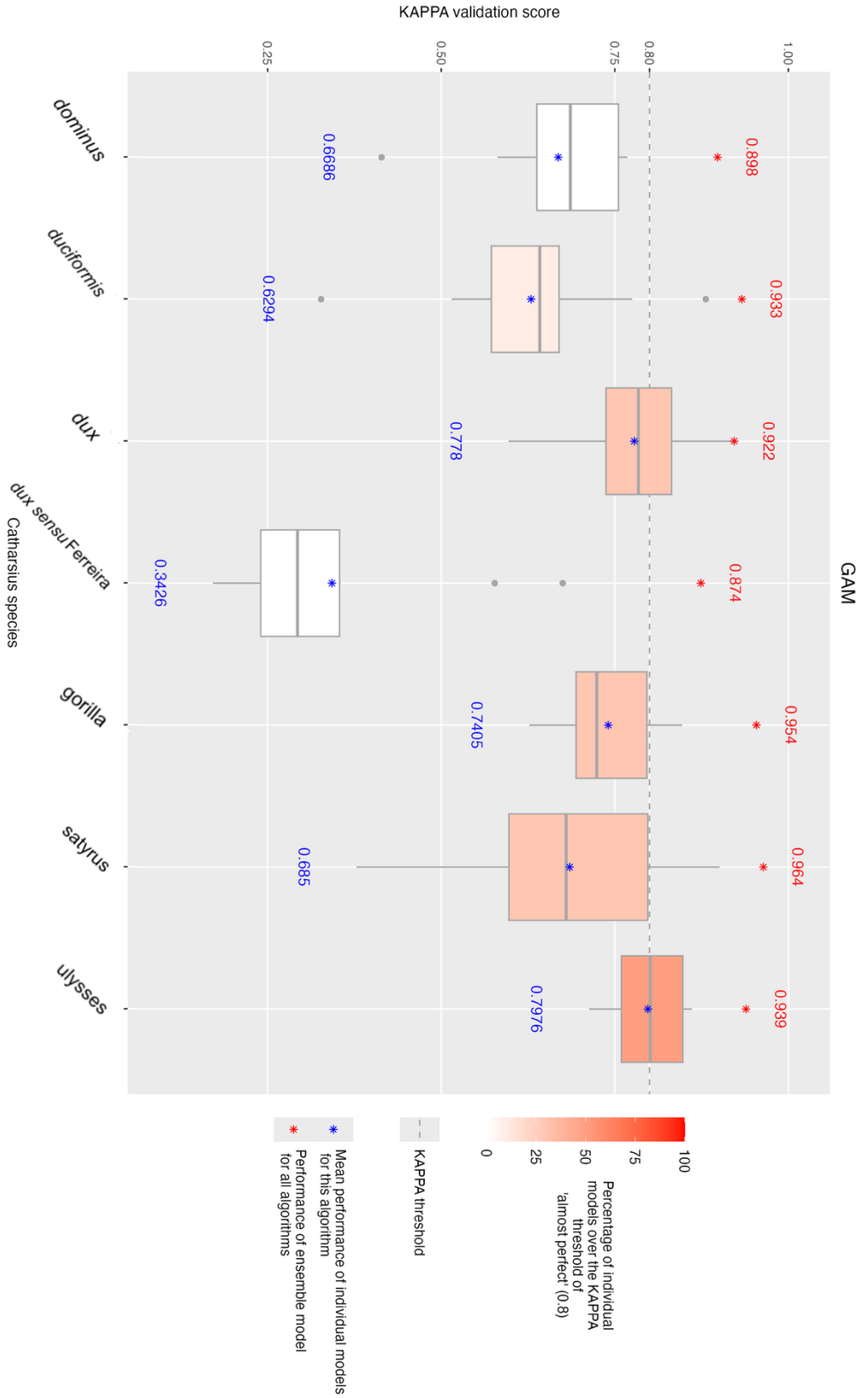
ANN

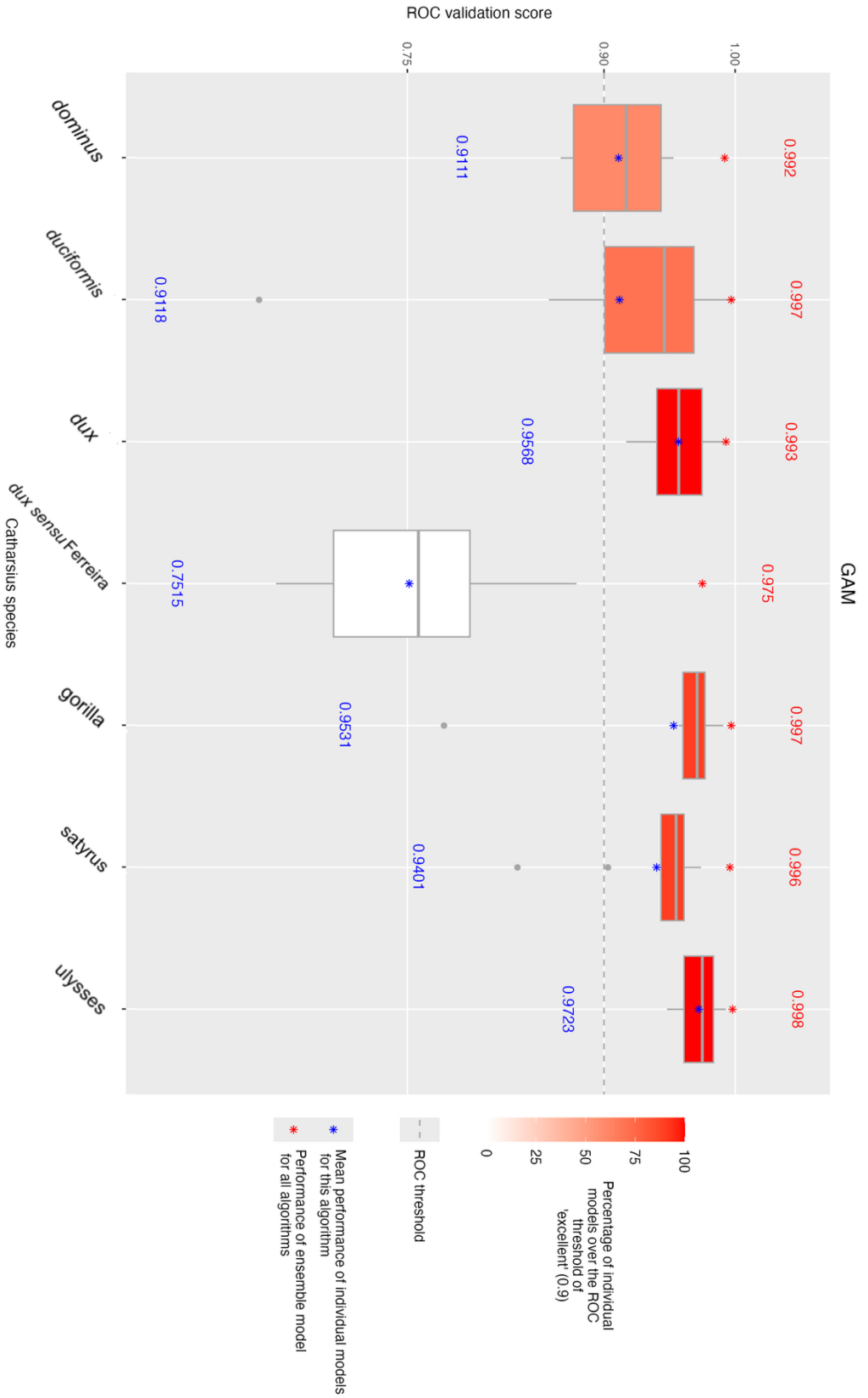


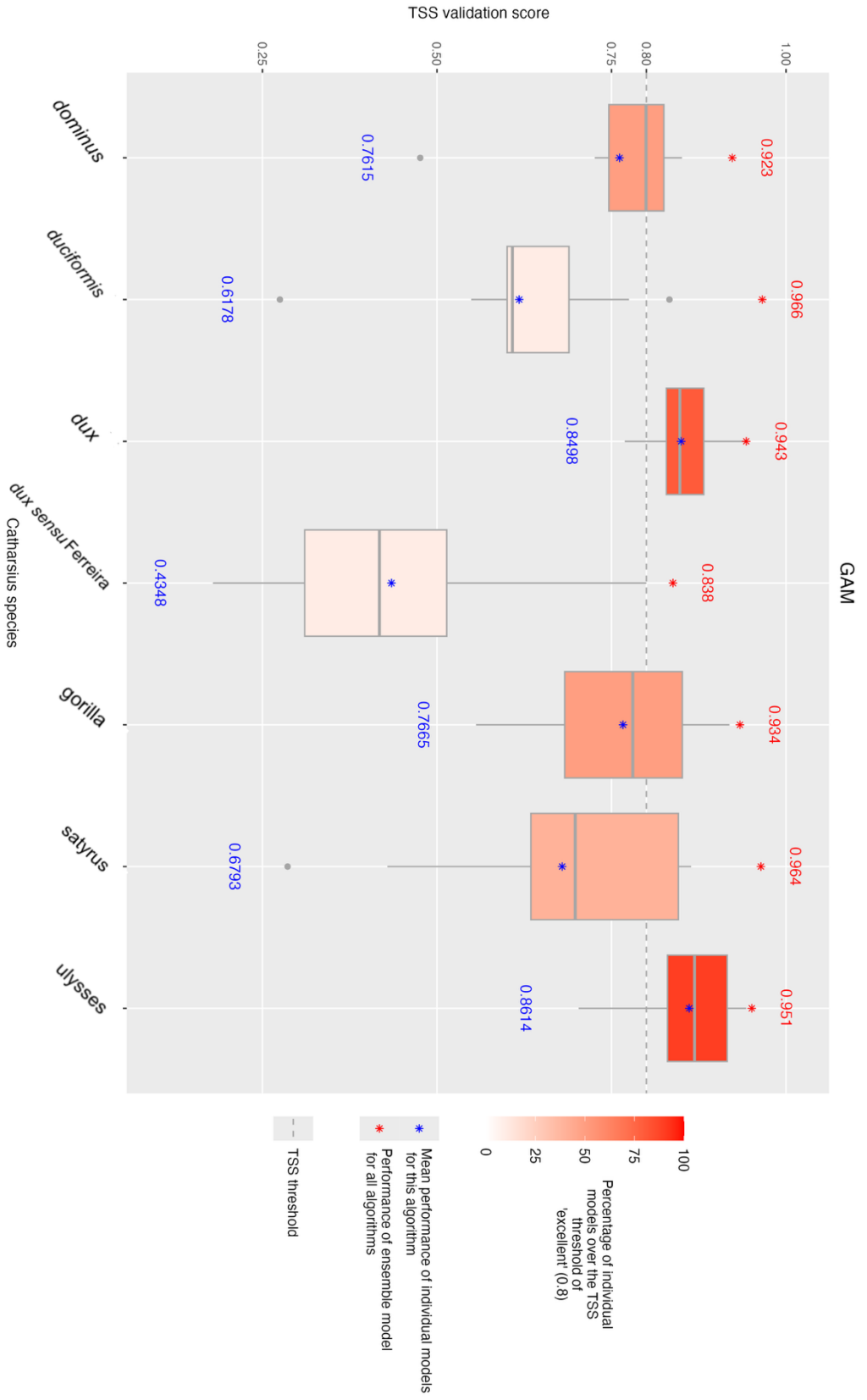

ROC validation score
0.50
0.75
0.90
1.00
ANN
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.9347
0.9153
0.9468
0.7315
0.9592
0.931
0.9686
0.992
0.997
0.993
0.975
0.997
0.996
0.998
Percentage of individual models over the ROC threshold of 'excellent' (0.9)
0
25
50
75
100
ROC threshold
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms


ANN
TSS validation score
1.00
0.80
0.75
0.50
0.25
0.00
0.923
0.966
0.943
0.838
0.934
0.964
0.951
0.6855
0.7864
0.8071
0.3657
0.7507
0.7501
0.8439
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
100
75
50
25
0
Percentage of individual models over the TSS threshold of 'excellent' (0.8)
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
TSS threshold

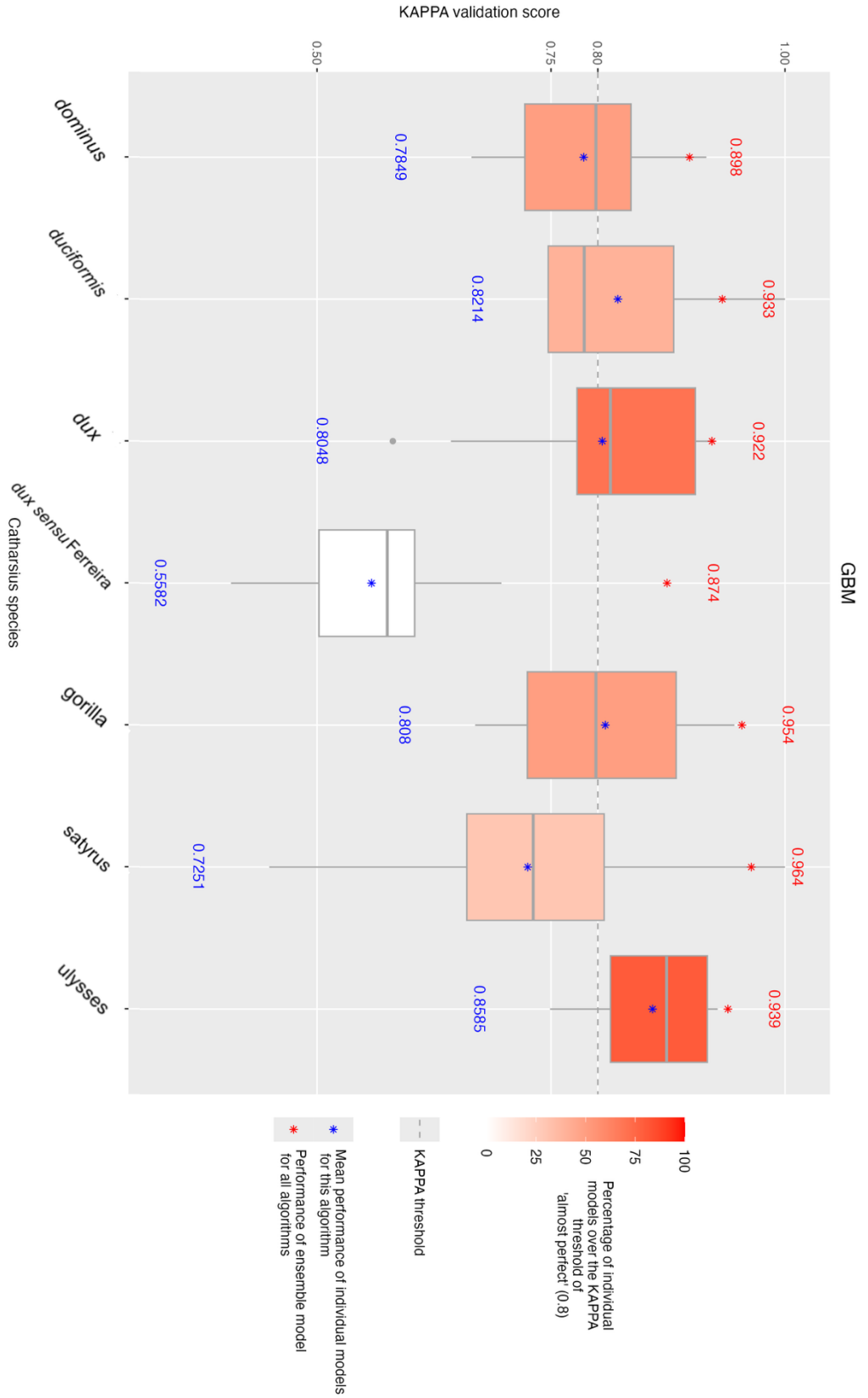
GAM

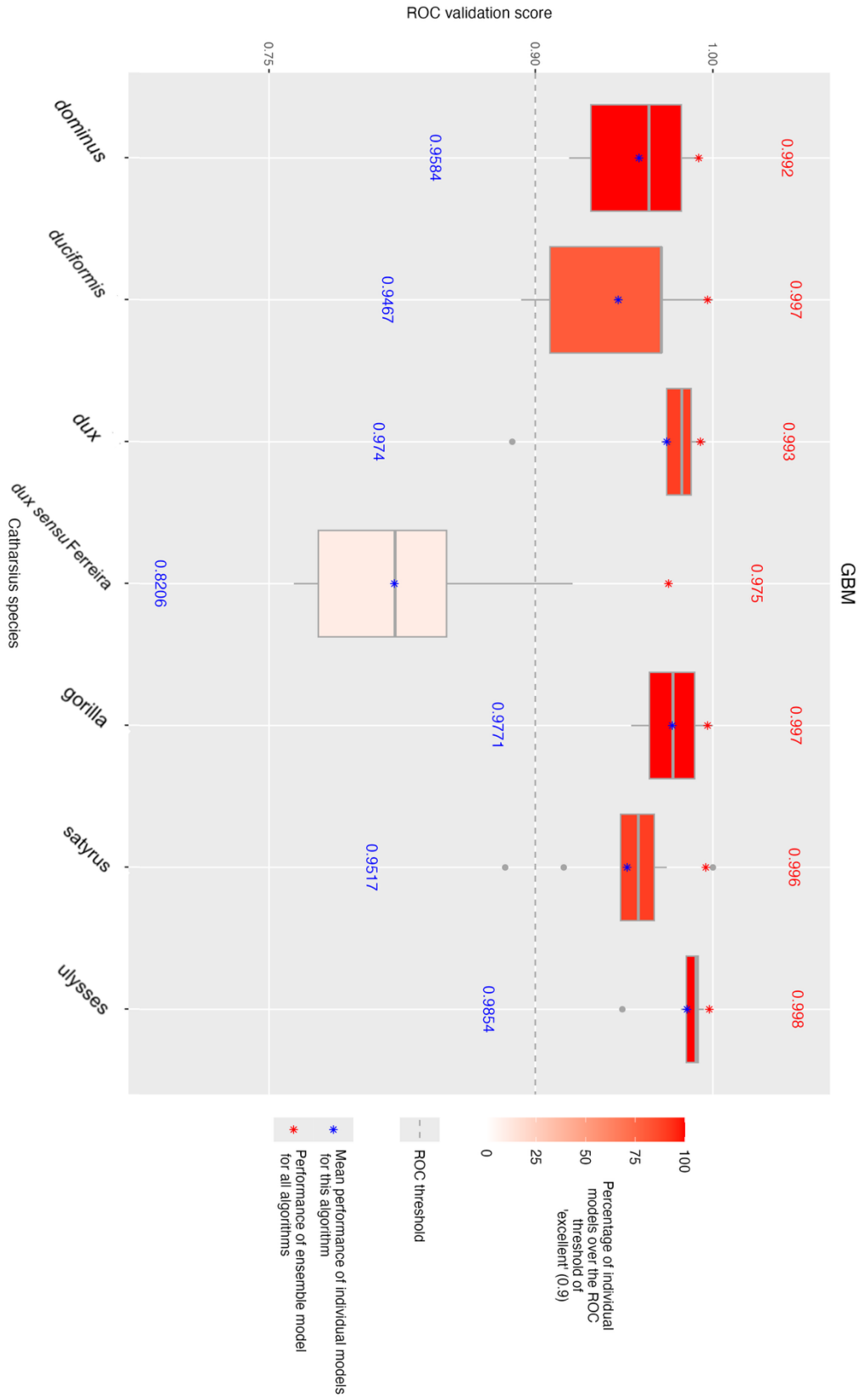

KAPPA validation score
0.25
0.50
0.75
0.80
1.00
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.6686
0.6294
0.778
0.3426
0.7405
0.685
0.7976
0.898
0.933
0.922
0.874
0.954
0.964
0.939
GAM
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
KAPPA threshold
0
25
50
75
100
Percentage of individual models over the KAPPA threshold of 'almost perfect' (0.8)

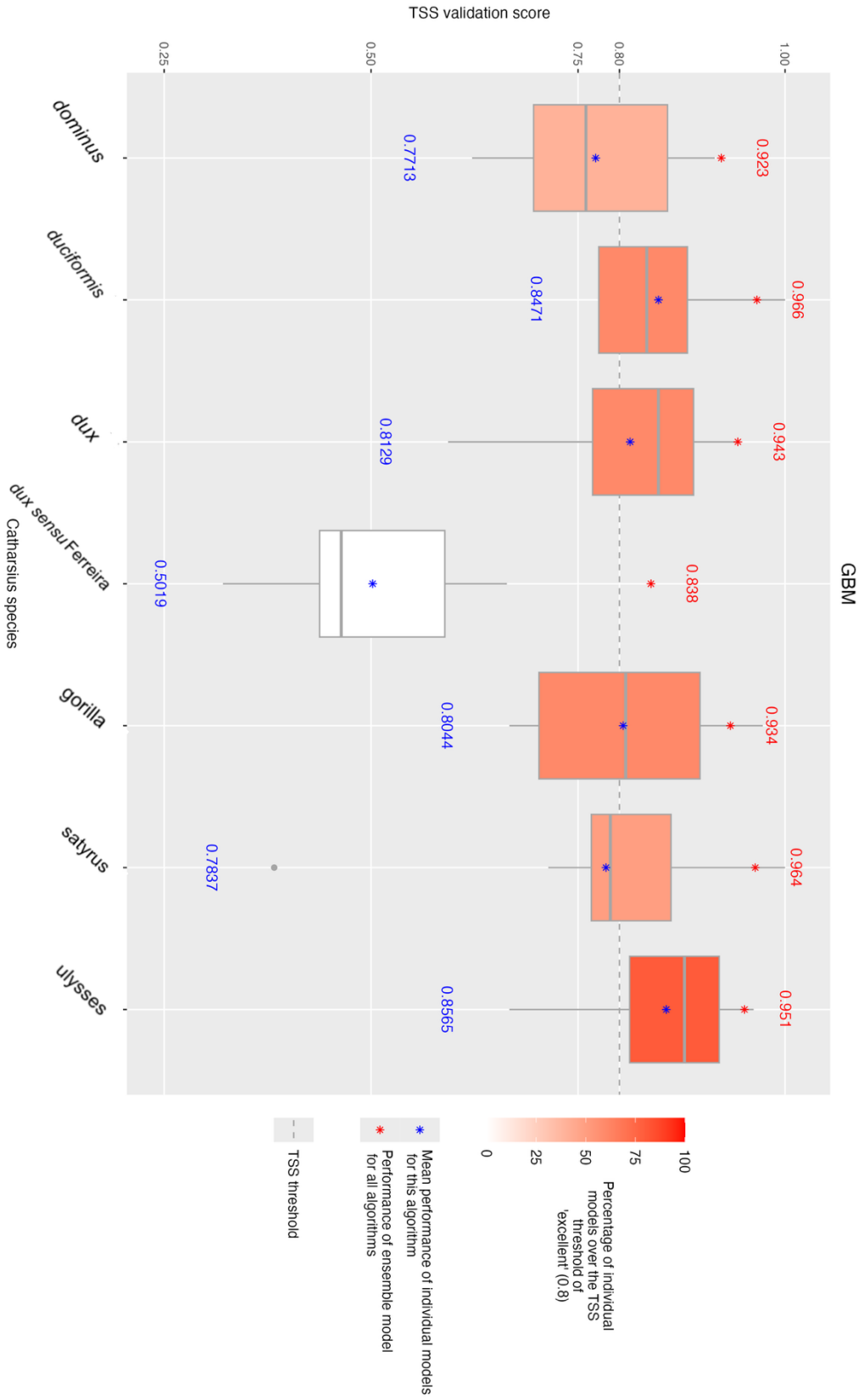

ROC validation score
0.75
0.90
1.00
GAM
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.9111
0.9118
0.9568
0.7515
0.9531
0.9401
0.9723
0.992
0.997
0.993
0.975
0.997
0.996
0.998
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
ROC threshold
0
25
50
75
100
Percentage of individual models over the ROC threshold of 'excellent' (0.9)


TSS validation score
0.25
0.50
0.75
0.80
1.00
GAM
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.923
0.966
0.943
0.838
0.934
0.964
0.951
0.7615
0.6178
0.8498
0.4348
0.7665
0.6793
0.8614
TSS threshold
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
0
25
50
75
100
Percentage of individual models over the TSS threshold of 'excellent' (0.8)

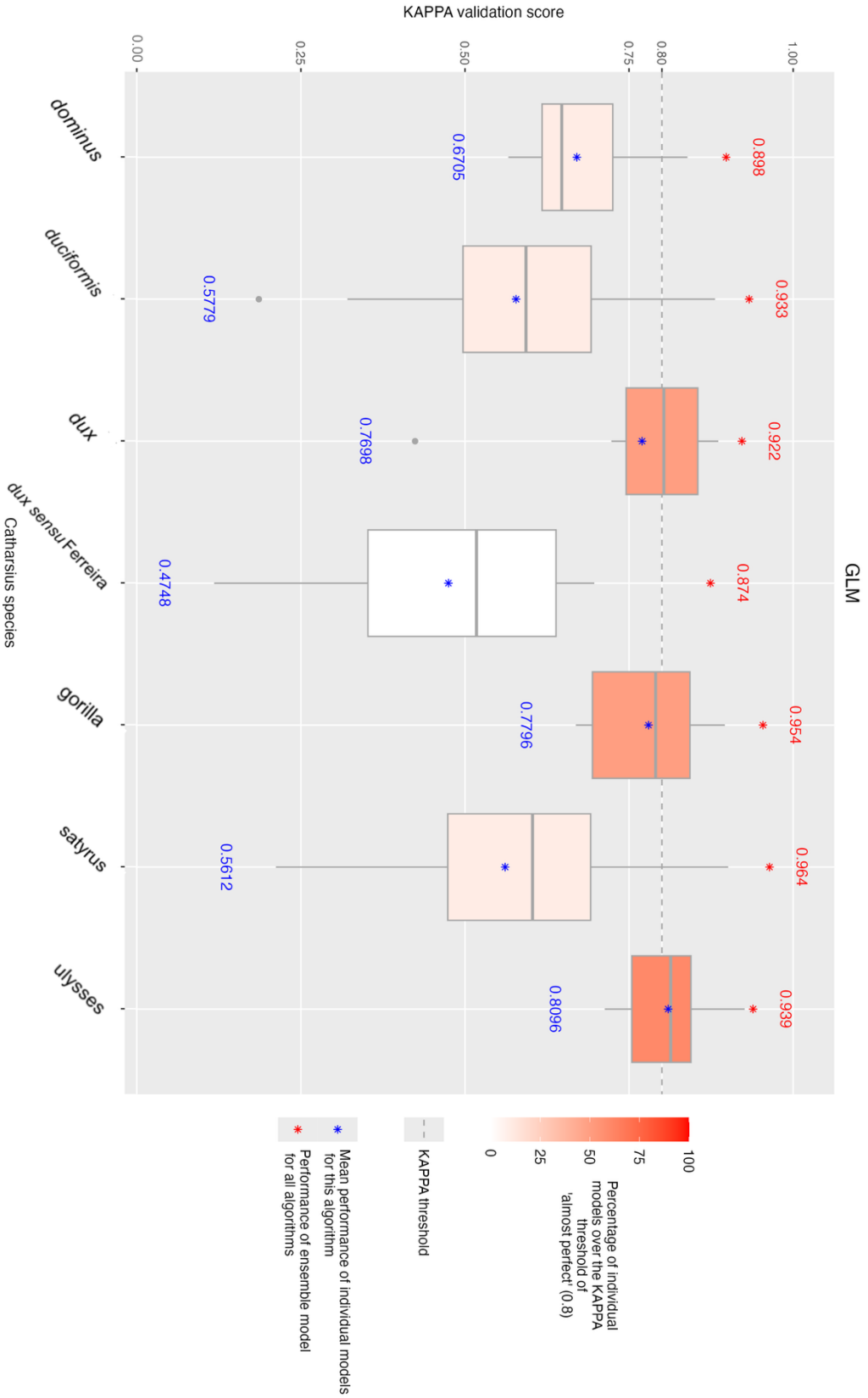
GBM

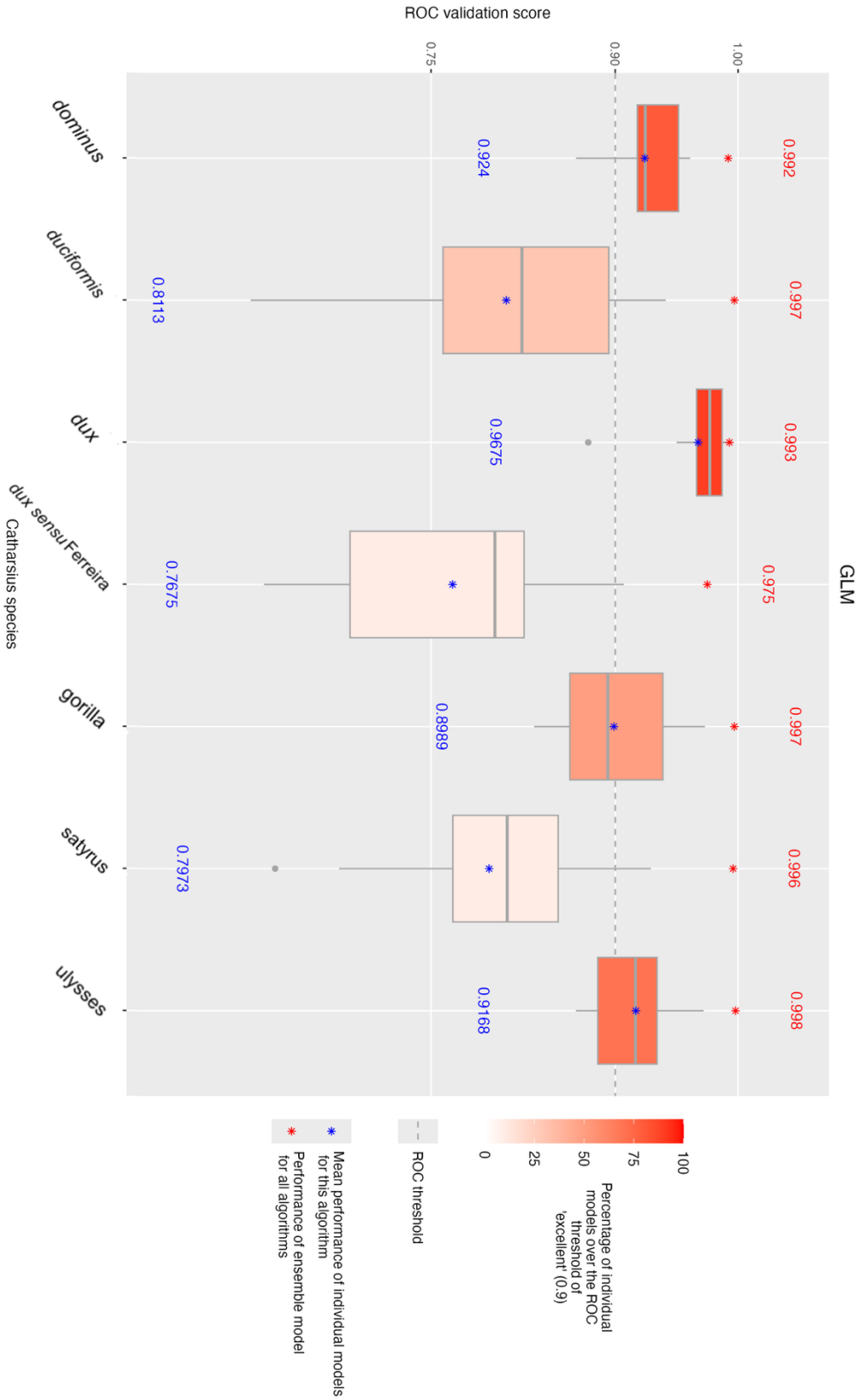

KAPPA validation score
0.50
0.75
0.80
1.00
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.7849
0.898
0.8214
0.933
0.8048
0.922
0.5582
0.874
0.808
0.954
0.7251
0.964
0.8585
0.939
GBM
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
KAPPA threshold
0
25
50
75
100
Percentage of individual models over the KAPPA threshold of 'almost perfect' (0.8)

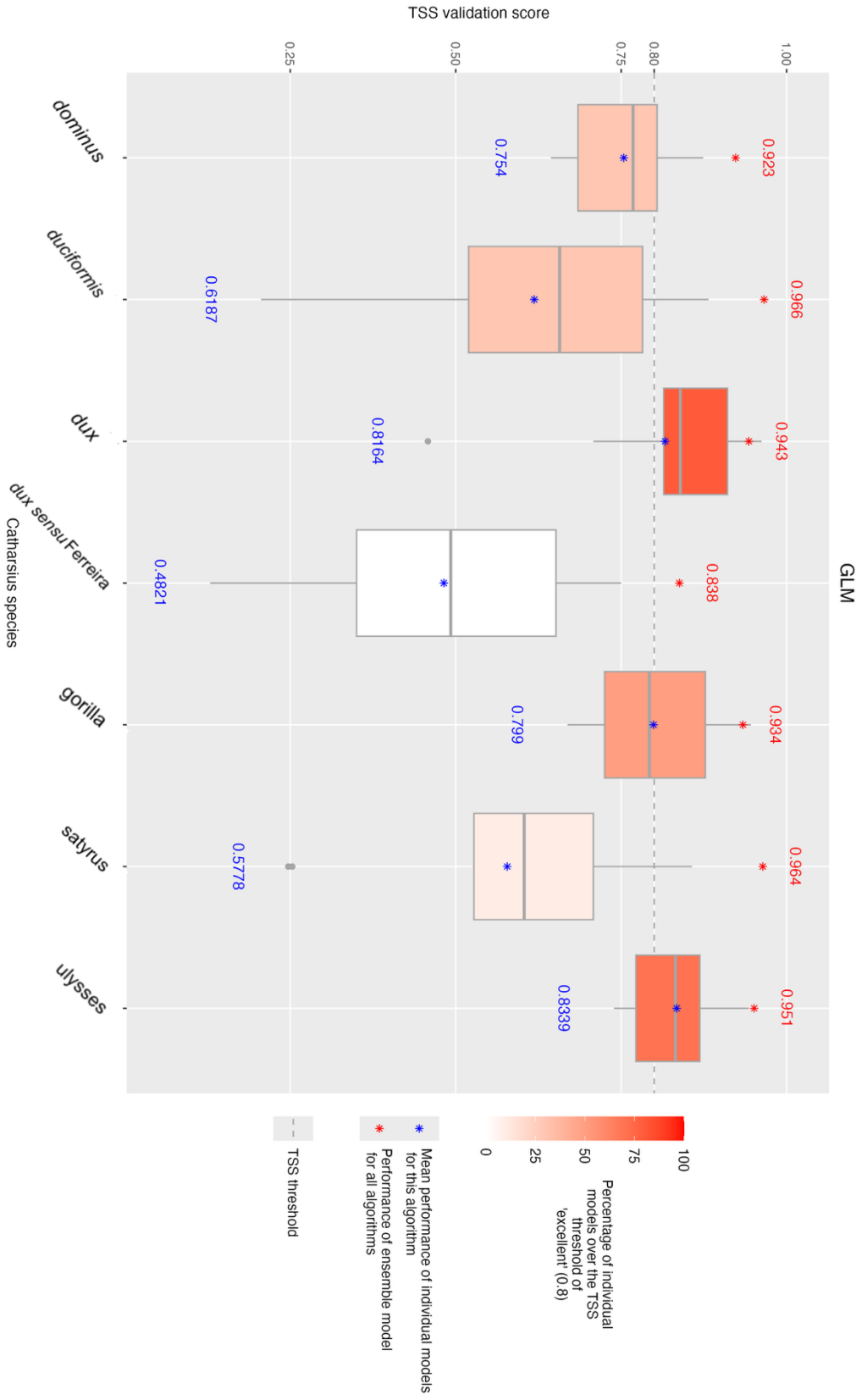

ROC validation score
0.75
0.90
1.00
GBM
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.9584
0.9467
0.974
0.8206
0.9771
0.9517
0.9854
0.992
0.997
0.993
0.975
0.997
0.996
0.998
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
ROC threshold
Percentage of individual models over the ROC threshold of 'excellent' (0.9)
0
25
50
75
100


TSS validation score
0.25
0.50
0.75
1.00
GBM
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.7713
0.8471
0.8129
0.5019
0.8044
0.7837
0.8565
0.923
0.966
0.943
0.838
0.934
0.964
0.951
TSS threshold
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
0
25
50
75
100
Percentage of individual models over the TSS threshold of 'excellent' (0.8)

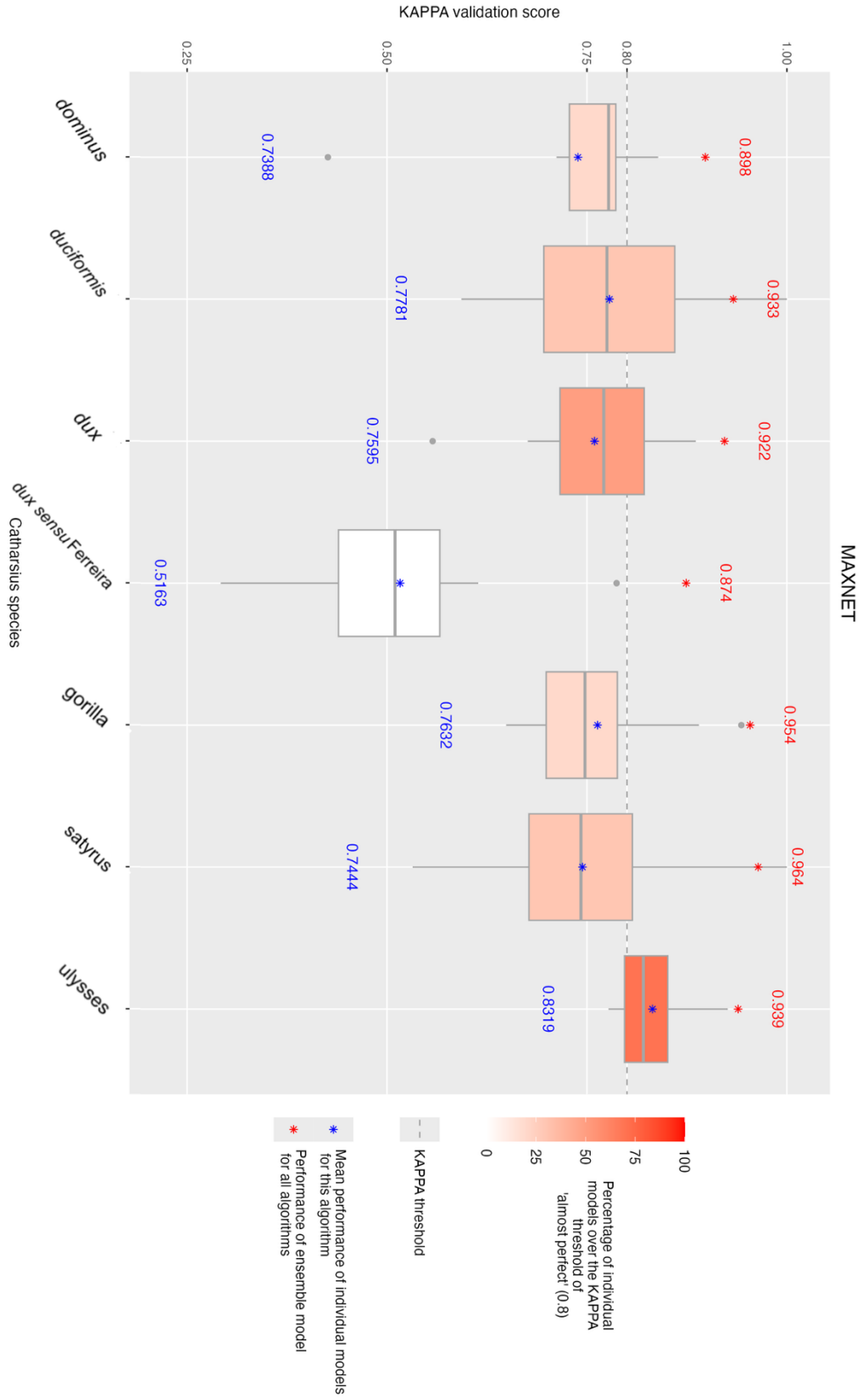
GLM

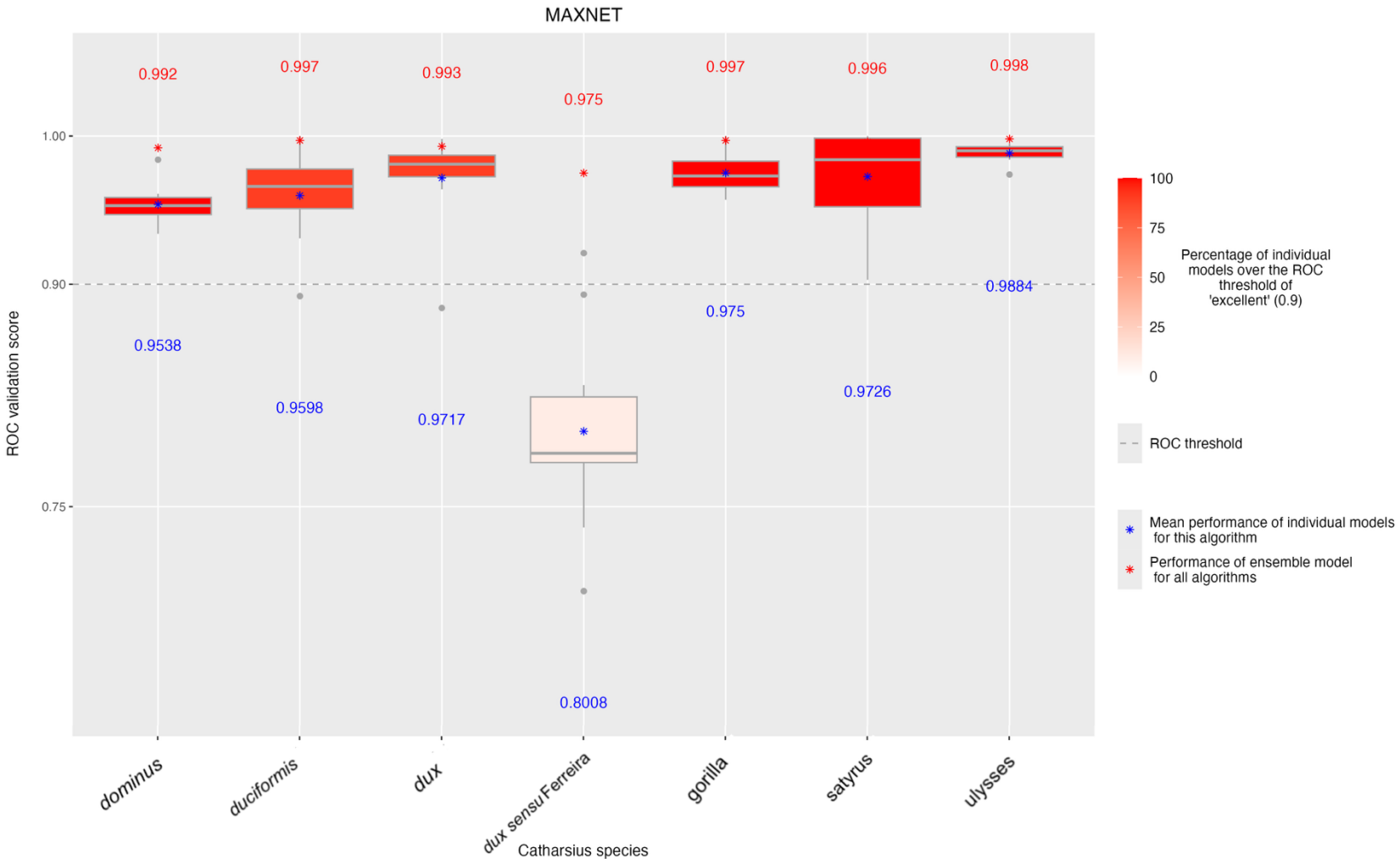

KAPPA validation score
0.00
0.25
0.50
0.75
0.80
1.00
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.6705
0.898
0.5779
0.933
0.7698
0.922
0.4748
0.874
0.7796
0.954
0.5612
0.964
0.8096
0.939
GLM
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
KAPPA threshold
0
25
50
75
100
Percentage of individual models over the KAPPA threshold of 'almost perfect' (0.8)

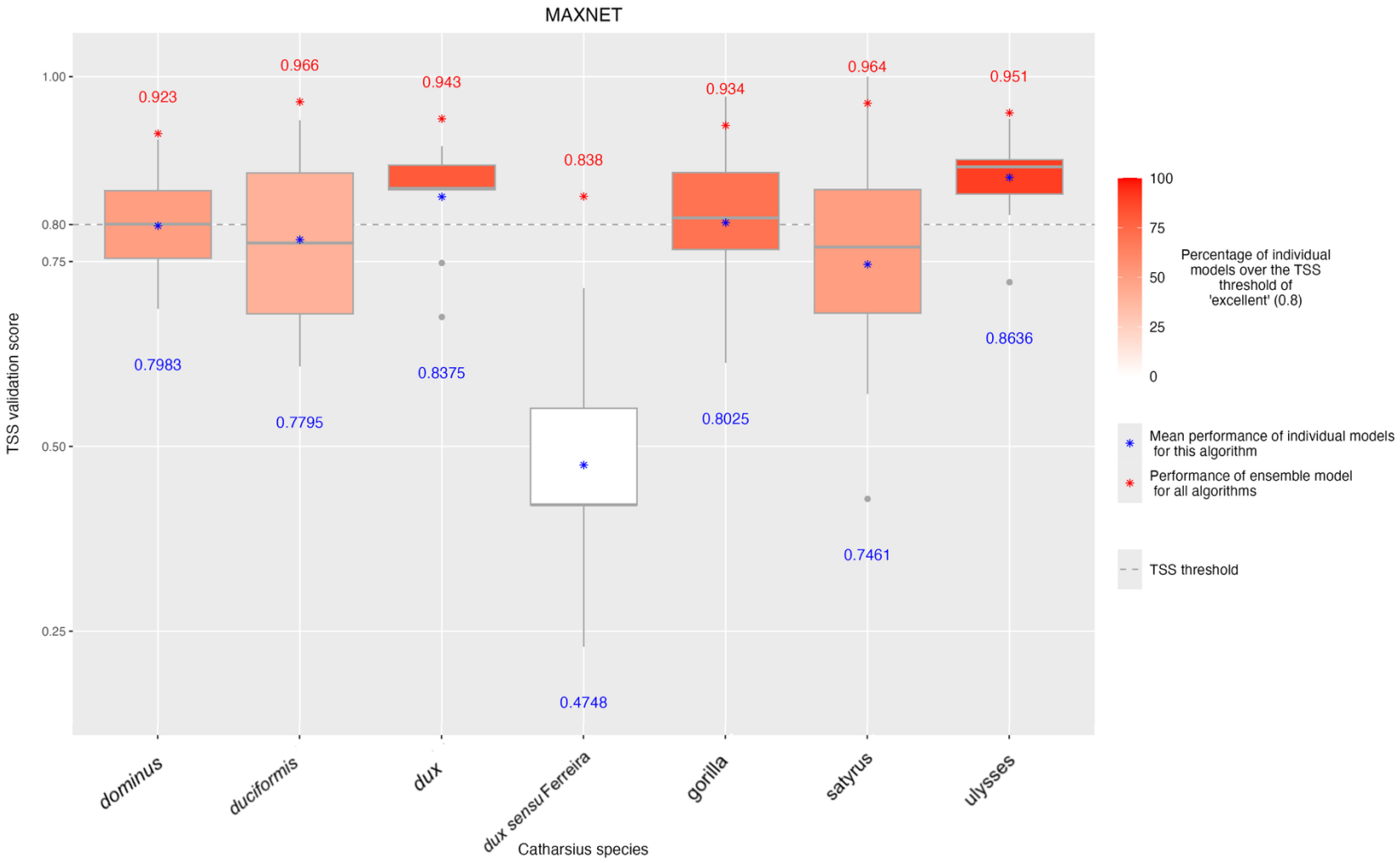

ROC validation score
0.75
0.90
1.00
GLM
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Cathartsius species
0.924
0.8113
0.9675
0.7675
0.8989
0.7973
0.9168
0.992
0.997
0.993
0.975
0.997
0.996
0.998
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
ROC threshold
0
25
50
75
100
Percentage of individual models over the ROC threshold of 'excellent' (0.9)


TSS validation score
0.25
0.50
0.75
0.80
1.00
GLM
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.754
0.923
0.6187
0.966
0.8164
0.943
0.4821
0.838
0.799
0.934
0.5778
0.964
0.8339
0.951
TSS threshold
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
0
25
50
75
100
Percentage of individual models over the TSS threshold of 'excellent' (0.8)

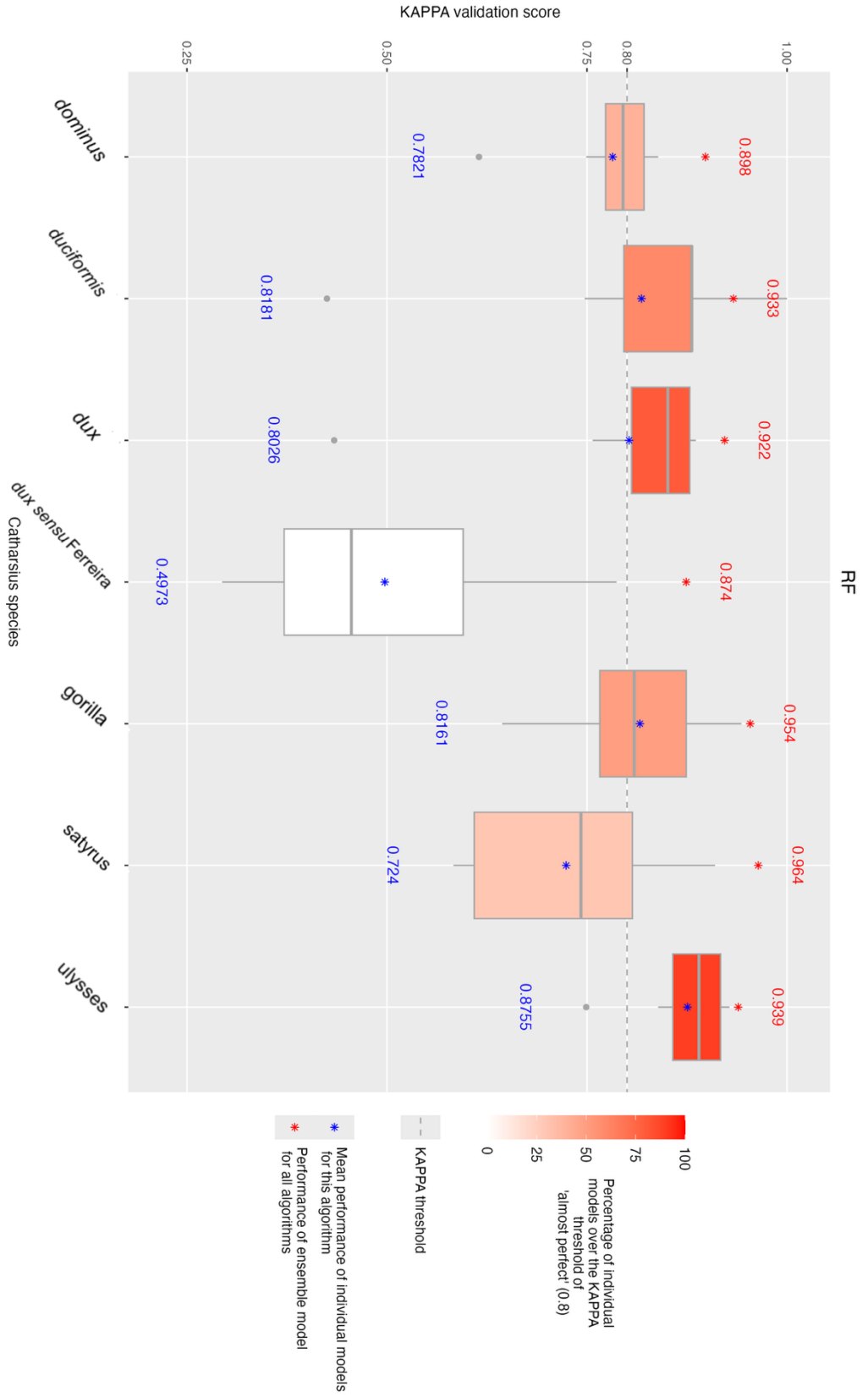
MAXNET

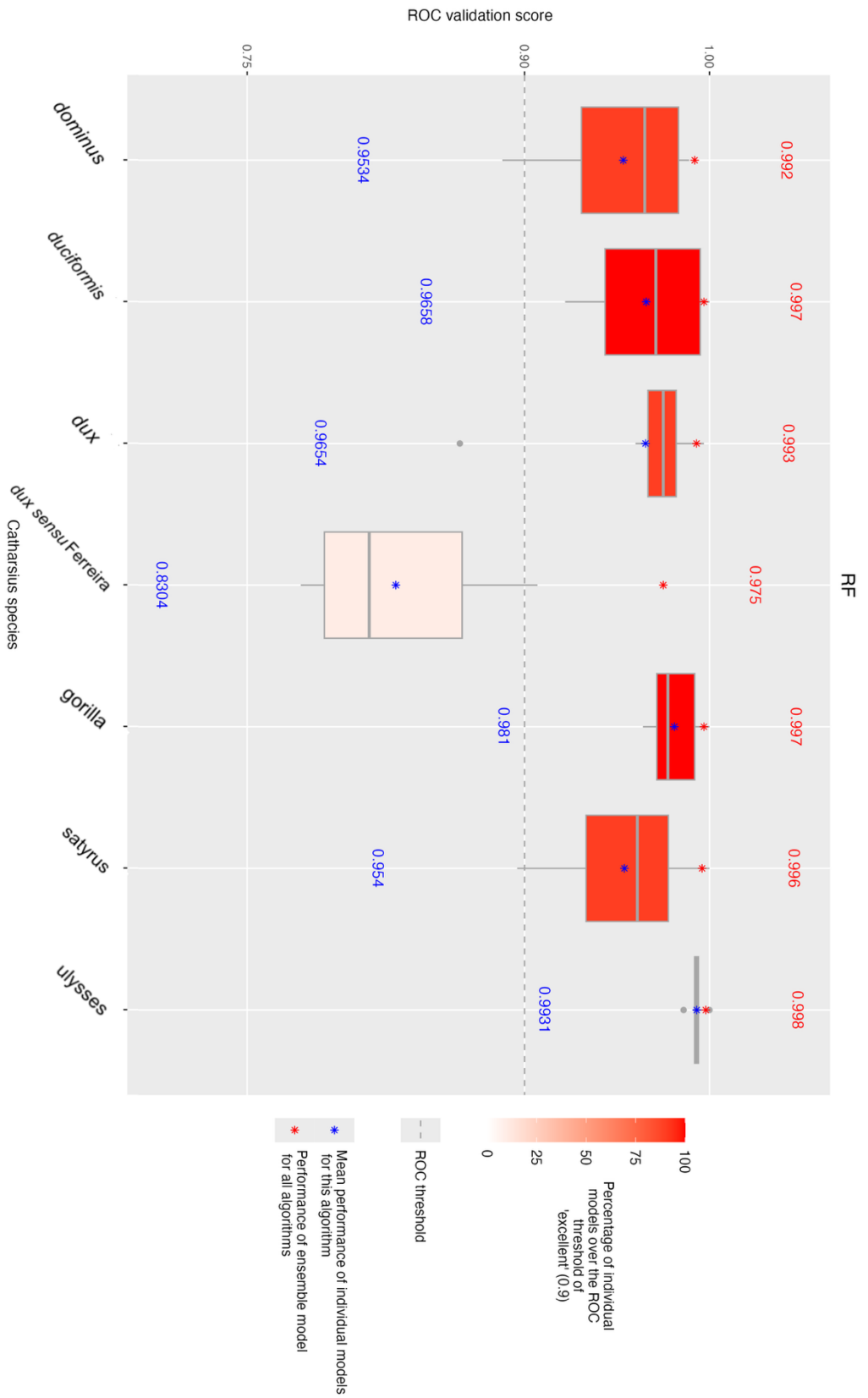

KAPPA validation score
0.25
0.50
0.75
0.80
1.00
MAXNET
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.7388
0.898
0.7781
0.933
0.7595
0.922
0.5163
0.874
0.7632
0.954
0.7444
0.964
0.8319
0.939
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
KAPPA threshold
0
25
50
75
100
Percentage of individual models over the KAPPA threshold of 'almost perfect' (0.8)

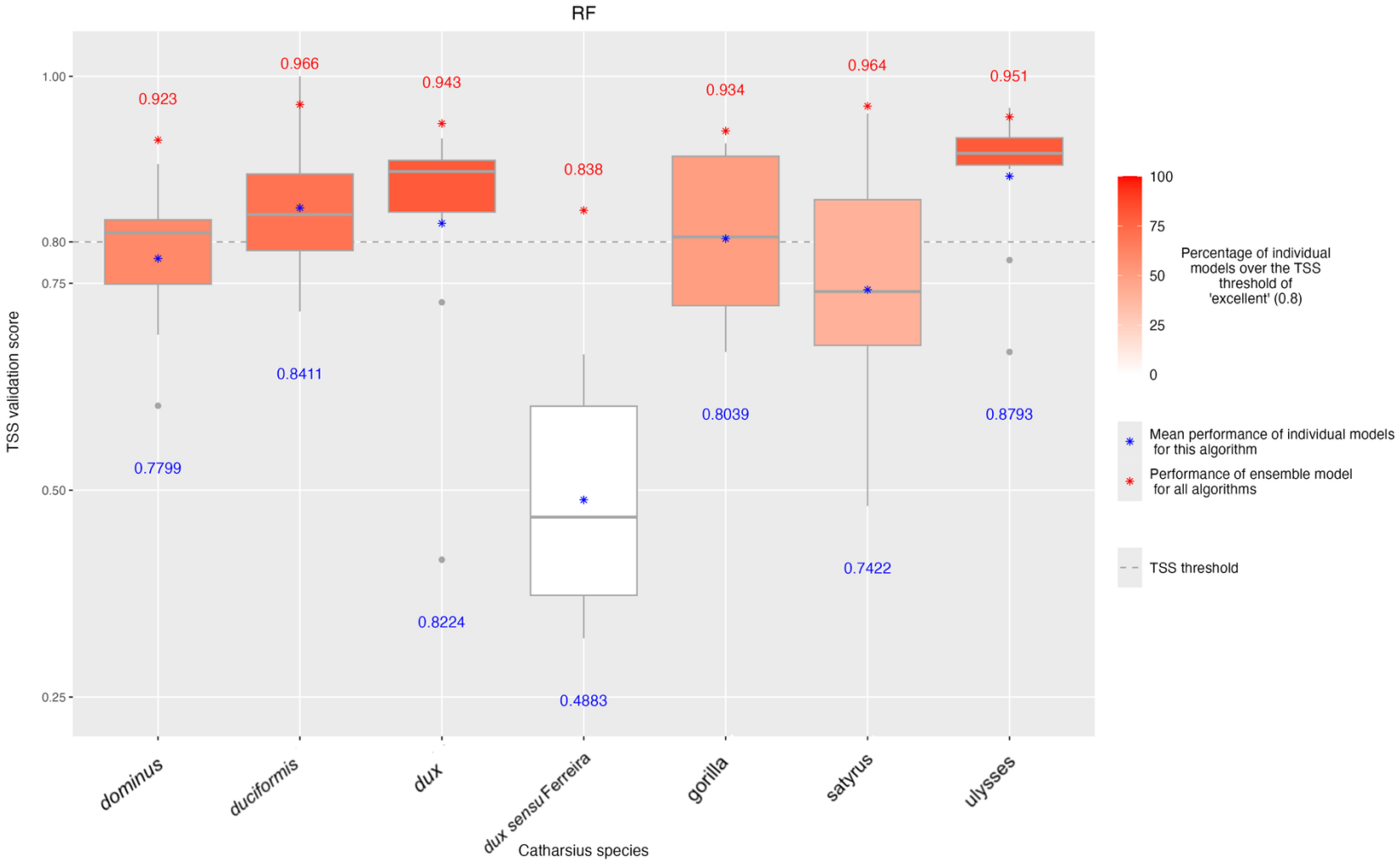

MAXNET
ROC validation score
1.00
0.90
0.75
0.992
0.997
0.993
0.975
0.997
0.996
0.998
0.9538
0.9598
0.9717
0.8008
0.975
0.9726
0.9884
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
100
75
50
25
0
Percentage of individual models over the ROC threshold of 'excellent' (0.9)
ROC threshold
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms


MAXNET
TSS validation score
1.00
0.80
0.75
0.50
0.25
0.923
0.966
0.943
0.838
0.934
0.964
0.951
0.7983
0.7795
0.8375
0.4748
0.8025
0.7461
0.8636
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
100
75
50
25
0
Percentage of individual models over the TSS threshold of 'excellent' (0.8)
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
TSS threshold

RF


KAPPA validation score
0.25
0.50
0.75
0.80
1.00
RF
Catharsius species
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
0.7821
0.898
0.8181
0.933
0.8026
0.922
0.4973
0.874
0.8161
0.954
0.724
0.964
0.8755
0.939
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
KAPPA threshold
Percentage of individual models over the KAPPA threshold of 'almost perfect' (0.8)
0
25
50
75
100


ROC validation score
0.75
0.90
1.00
RF
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
0.9534
0.9658
0.9654
0.8304
0.981
0.954
0.9931
0.992
0.997
0.993
0.975
0.997
0.996
0.998
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
ROC threshold
Percentage of individual models over the ROC threshold of 'excellent' (0.9)
0
25
50
75
100


RF
TSS validation score
1.00
0.80
0.75
0.50
0.25
0.923
0.966
0.943
0.838
0.934
0.964
0.951
0.7799
0.8411
0.8224
0.4883
0.8039
0.7422
0.8793
dominus
duciformis
dux
dux sensu Ferreira
gorilla
satyrus
ulysses
Catharsius species
100
75
50
25
0
Percentage of individual models over the TSS threshold of 'excellent' (0.8)
Mean performance of individual models for this algorithm
Performance of ensemble model for all algorithms
TSS threshold

B

Supplementary Information Chapter 3

Appendix S1 – Study region

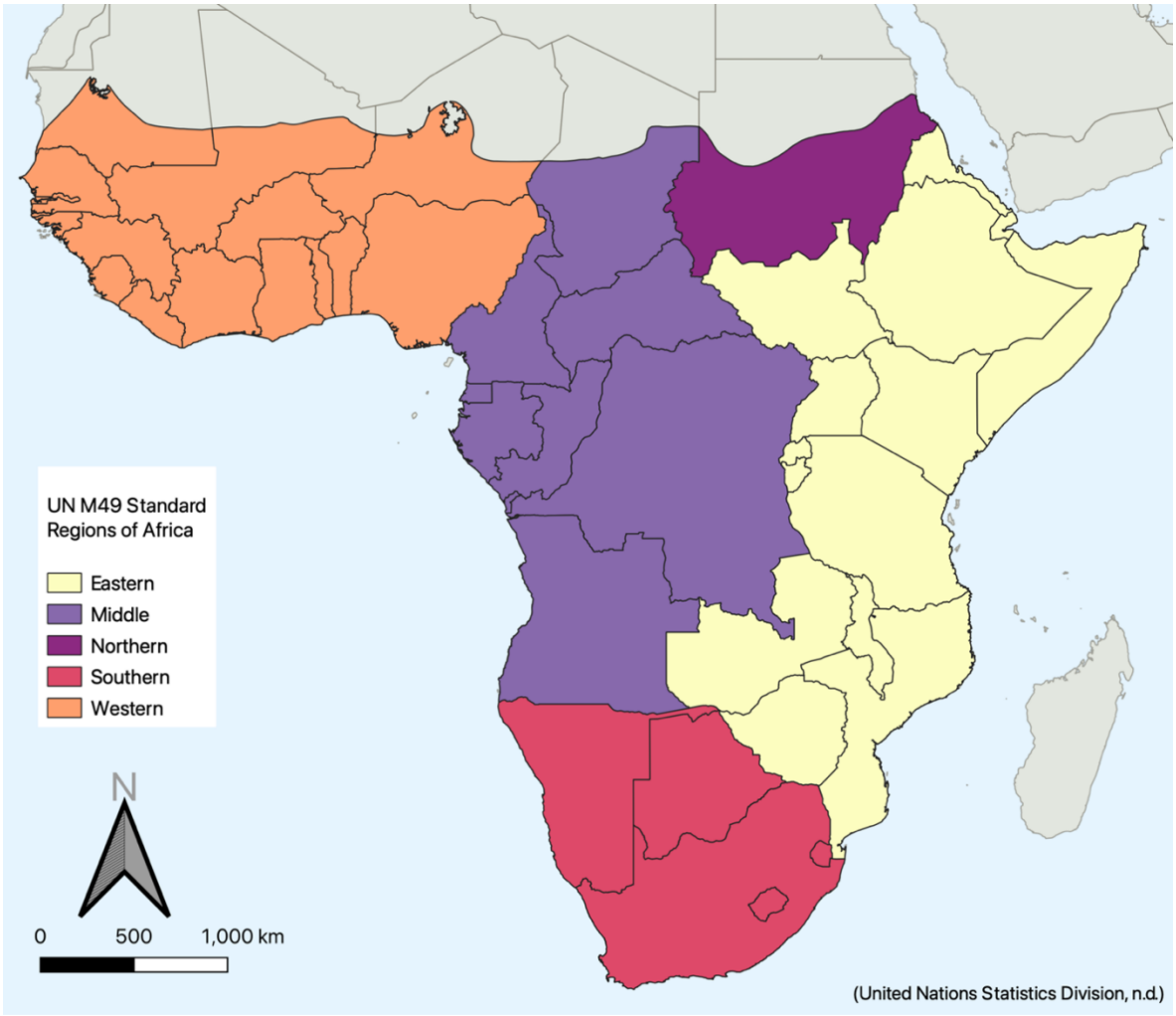


Figure S1.1: The mainland Afrotropical realm, created from World Wildlife Fund (2012), showing regions of Africa as designated by the UN M49 Standard (United Nations Statistics Division, n.d.). For the purposes of the study, the manuscript uses the term “Central Africa” in place of “Middle Africa”.

Appendix S2 – Occurrence data

Table S2.1: Pre-processing procedure **of GBIF data**

Action	Number of entries changed	Entries remaining
1. Original GBIF dataset (GBIF, 2023)		4270
2. Filtered to include only entries pertaining to species-level identifications or lower	566 removed	3704
3. Discard entries with no coordinate information, those with identical longitude and latitude, and those with both longitude and latitude equal to zero	535 removed	3169
4. Filtered to include only entries whose coordinates have one or more decimal places	39 removed	3130
5. Discard entries that fall exactly on a country centroid, in GBIF headquarters, and biodiversity institutions	Country centroid: 49 removed (GBIF HQ: 0) (Biodiversity institutions: 0)	3081
6. Temporarily filter entries which fall in countries other than the one listed in the record or in the sea (for subsequent steps)	123 temporarily removed	2958
7. From the 123 removed entries, filter those that fall in the sea. If these are within a buffer of 0.1 degree (approximately 10km) from land belonging to the country in the record, assign entry to the closest land and replace in dataset	(In the sea: 12) (Within buffer: 11) In correct country when assigned: 10 replaced	2968
8. Manual check of remaining records which fell in countries other than that in the record. Those with correctable errors rectified and replaced in the set (excluding those in Asia)	64 replaced (Asian species not included in this step)	3032
9. Crop to project extent	1333	**1699**
Note: Nineteen entries were removed from the revision dataset when cropped to the project extent, resulting in a final total of 4979		

Table S2.2: Taxonomic standardisation of GBIF data

Action	Number of entries changed	Reason
C. birmanensis removed	11	Not a recognised African species
C. fastidiosus removed	2	Placed into *incertae sedis*
C. ninus to *C. approximans*	1	*C.ninus* newly synonymised with *C. approximans*
C. oedipus to *C. polynices*	12	*C. oedipus* newly synonymised with *C. polynices*
C. platycerus to *C. obtusicornis*	23	*C. platycerus* designated as a junior primary homonym
C. simillimus to *C. polynices*	5	*C. simillimus* newly synonymised with *C. polynices*
C. vansoni to *C. longiceps*	1	*C. vansoni* newly synonymised with *C. longiceps*
C. philus to *C. vitulus*	125	*C. vitulus* newly synonymised *C. philus*
C. pseudooedipus to *C. princeps*	1	*C. pseudooedipus* newly synonymised to *C. princeps*
After these changes, 1686 entries remained in the GBIF set		
Note: no records were re-identified as part of this process, and only the species names of these records were changed. Only records whose original names have been subject to revision were altered in this way, and these changes do not encompass all alterations made in the revision. Taxonomic changes as a result of this revision are pre-publication.		

Table S2.3: Manual identification and removal of duplicates in creation of the combined set

When the combined set was created, duplicate records were manually identified and removed. Criteria used to identify duplicates are as follows:

Criterion	**Explanation**
Species name and year identical	If year = NA, this was automatically deemed not a duplicate. Many collecting locations were returned to year after year, and so, without this information, comparison between records was impossible
Lat and long <1 degree apart	Could not ensure that duplicates had completely identical coordinates given differing levels of precision used by collectors and institutions when records were digitised and / or uploaded to GBIF
If date listed with more precision than just year, these must be identical	
Collector and location the same, even if written in a different way	When digitised and / or uploaded to GBIF, label information is formatted in a way that may be different from the original specimen label. As the revision dataset uses original label information, the format of the contents may be different, but the content must be the same to be considered a duplicate
No more than two columns can differ in any way	Although more columns were included in each separate dataset, those that were comparable contained information on: Species, year, recorder / collector, longitude, latitude, collecting location (description), date and total number of specimens. Some records fulfilled all criteria for duplicates with the exception of total number of specimens. In these cases, if the revision dataset total was higher, the original revision dataset entry was removed and the difference in total re-entered as a new record. E.g. Revision entry with a total of five specimens fulfils all the criteria for being a duplicate of a GBIF entry with total of three specimens. The original revision entry is removed, but an identical record created with a total of two specimens, to ensure all extra value from the taxonomic revision is included. This was thought to be likely a consequence of the digitisation and mobilisation process, in which specimens were missed or even added to collections after this had taken place.

Appendix S3 – Environmental data

Table S3.1: WorldClim Variables used in the principal components analysis (Fick and Hijmans, 2017; WorldClim, 2020)

Table S3.1: WorldClim Variables used in the principal components analysis (Fick and Hijmans, 2017; WorldClim, 2020)	
Code	**Description**
BIO1	Annual Mean Temperature
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3	Isothermality (BIO2/BIO7) (×100)
BIO4	Temperature Seasonality (standard deviation ×100)
BIO5	Max Temperature of Warmest Month
BIO6	Min Temperature of Coldest Month
BIO7	Temperature Annual Range (BIO5-BIO6)
BIO8	Mean Temperature of Wettest Quarter
BIO9	Mean Temperature of Driest Quarter
BIO10	Mean Temperature of Warmest Quarter
BIO11	Mean Temperature of Coldest Quarter
BIO12	Annual Precipitation
BIO13	Precipitation of Wettest Month
BIO14	Precipitation of Driest Month
BIO15	Precipitation Seasonality (Coefficient of Variation)
BIO16	Precipitation of Wettest Quarter
BIO17	Precipitation of Driest Quarter
BIO18	Precipitation of Warmest Quarter
BIO19	Precipitation of Coldest Quarter

C

Supplementary Information Chapter 4

Tables

Supplementary Table 4.1: Full genetic distance matrix (Euclidean distance) showing Catharsius *species similarity based on whole genome SNPs. Species with a non-NA comparison to themselves are those which were represented by more than one individual in the dataset, and can be interpreted as intra-species variation. The reduced genetic distance matrix is as below, with* C. merrettorum *and* C. dux / duciformis *removed.*

***Catharsius* species**	***dux / duciformis***	***dux***	***merrettorum***	***satyrus***	***duciformis***	***peregrinus***	***orami***	***luluensis***	***machadoi***	***smithi***	***biconifer***	***pallas***
dux / duciformis	640.46	641.20	683.62	650.83	643.08	693.27	673.92	677.09	703.73	697.27	701.49	682.37
dux	641.20	639.61	687.21	652.73	643.00	695.90	677.20	680.67	706.76	700.73	704.56	685.26
merrettorum	683.62	687.21	NA	655.38	687.05	632.52	617.17	620.22	632.14	630.23	644.97	608.63
satyrus	650.83	652.73	655.38	576.44	654.77	665.76	642.99	648.56	675.55	669.84	674.29	655.43
duciformis	643.08	643.00	687.05	654.77	641.59	695.96	676.36	679.91	706.06	699.93	704.12	684.86
peregrinus	693.27	695.90	632.52	665.76	695.96	537.67	587.81	593.08	651.76	645.54	651.06	601.53
orami	673.92	677.20	617.17	642.99	676.36	587.81	NA	564.15	630.78	626.01	636.19	583.39
luluensis	677.09	680.67	620.22	648.56	679.91	593.08	564.15	535.22	635.83	630.63	643.63	588.53
machadoi	703.73	706.76	632.14	675.55	706.06	651.76	630.78	635.83	534.92	570.28	631.84	626.19
smithi	697.27	700.73	630.23	669.84	699.93	645.54	626.01	630.63	570.28	NA	627.14	622.84
biconifer	701.49	704.56	644.97	674.29	704.12	651.06	636.19	643.63	631.84	627.14	528.45	640.40
pallas	682.37	685.26	608.63	655.43	684.86	601.53	583.39	588.53	626.19	622.84	640.40	NA

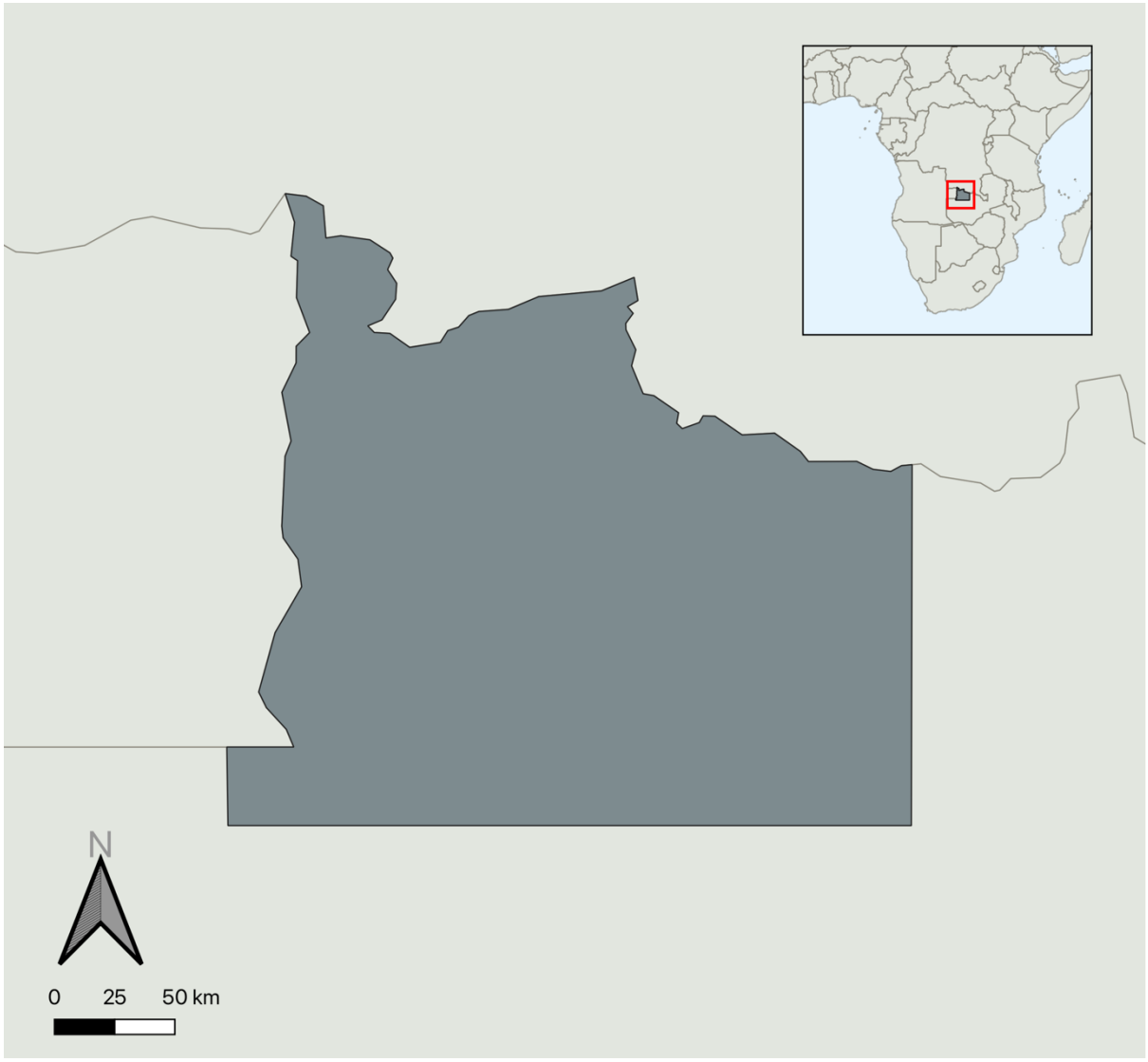
Supplementary Table 4.2: Distance matrix (Gower's distance) derived from Catharsius *morphological traits.*

***Catharsius* species**	***dux***	***duciformis***	***machadoi***	***smithi***	***biconifer***	***pallas***	***peregrinus***	***orami***	***luluensis***	***satyrus***
dux	0.00	0.01	0.34	0.35	0.33	0.36	0.42	0.34	0.38	0.04
duciformis	0.01	0.00	0.34	0.34	0.34	0.36	0.42	0.33	0.38	0.05
machadoi	0.34	0.34	0.00	0.12	0.19	0.38	0.35	0.39	0.38	0.35
smithi	0.35	0.34	0.12	0.00	0.13	0.34	0.37	0.34	0.35	0.35
biconifer	0.33	0.34	0.19	0.13	0.00	0.36	0.38	0.34	0.37	0.34
pallas	0.36	0.36	0.38	0.34	0.36	0.00	0.15	0.05	0.05	0.38
peregrinus	0.42	0.42	0.35	0.37	0.38	0.15	0.00	0.17	0.18	0.44
orami	0.34	0.33	0.39	0.34	0.34	0.05	0.17	0.00	0.06	0.36
luluensis	0.38	0.38	0.38	0.35	0.37	0.05	0.18	0.06	0.00	0.41
satyrus	0.04	0.05	0.35	0.35	0.34	0.38	0.44	0.36	0.41	0.00

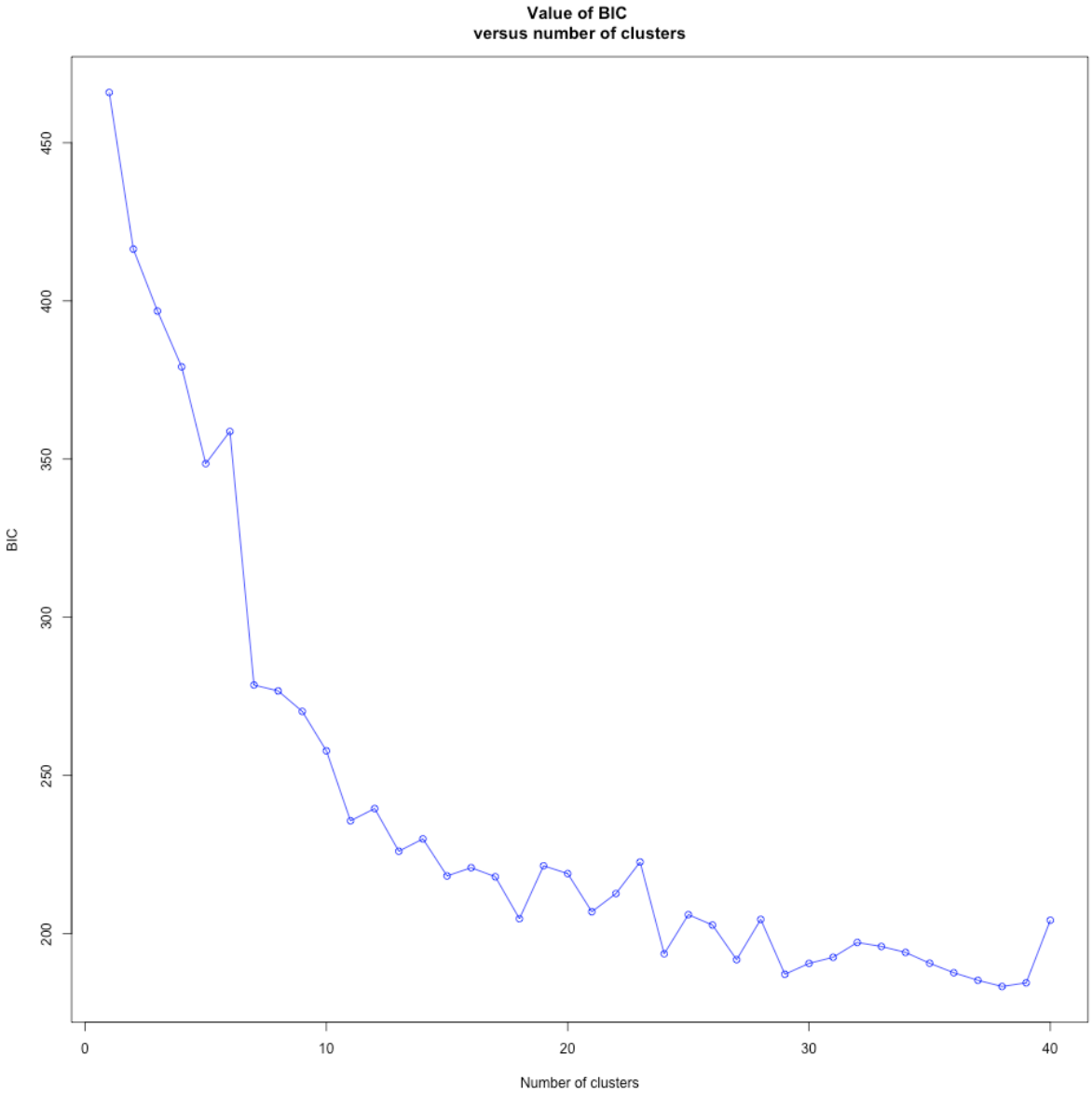
D

Supplementary Information Chapter 5

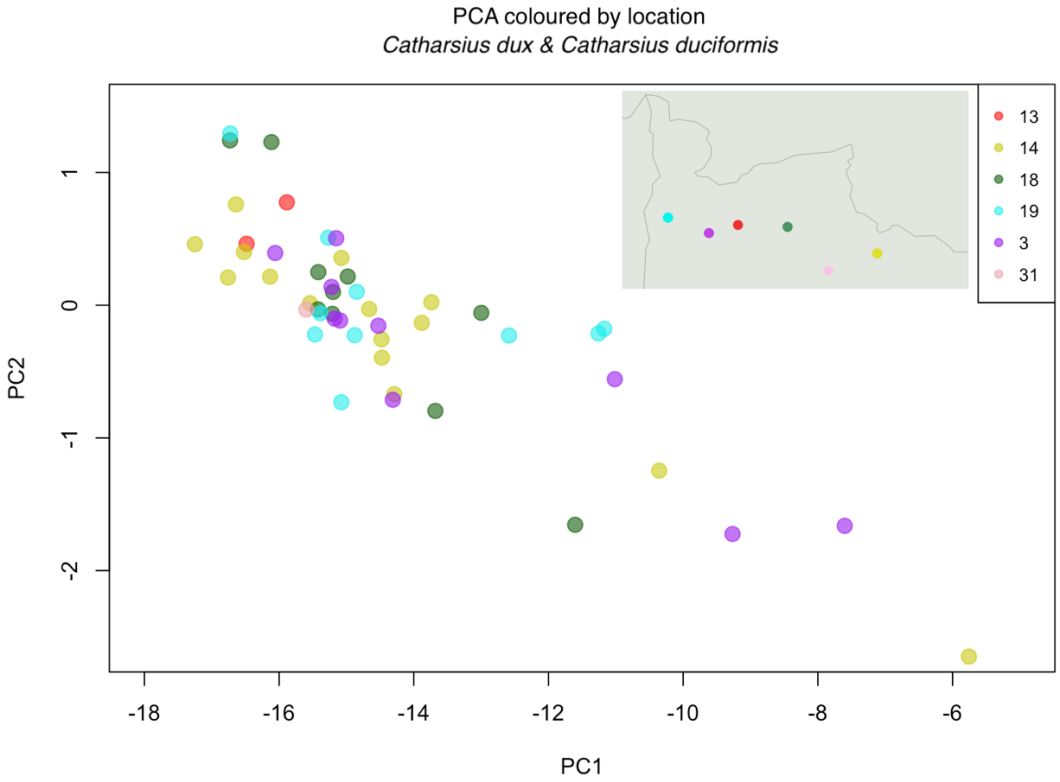
Figures



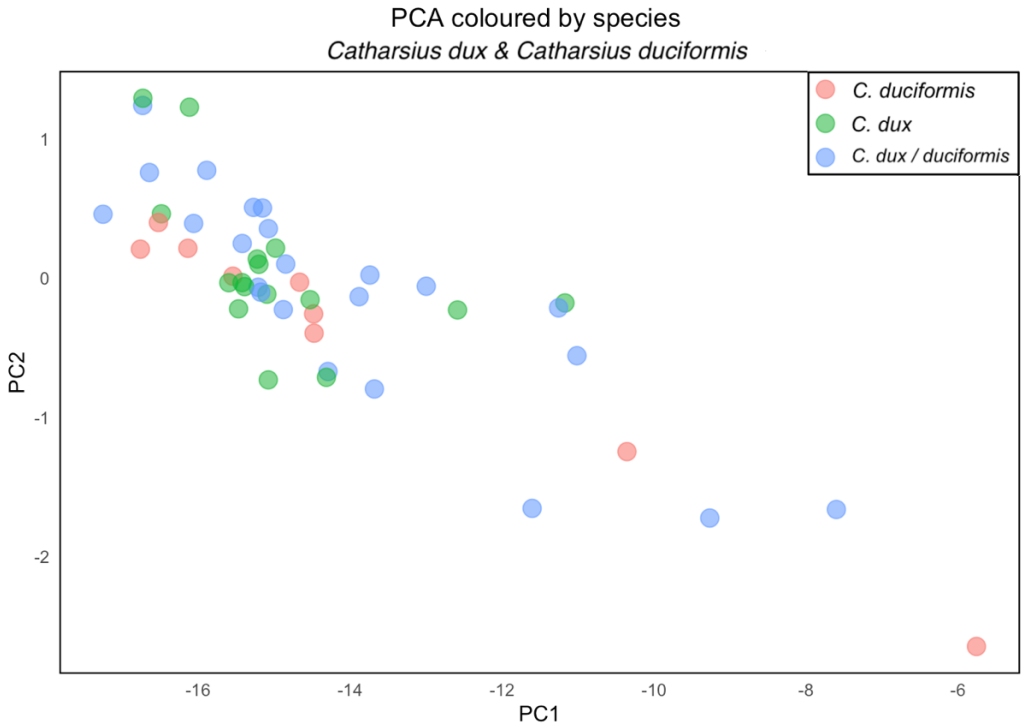
Supplementary Figure 5.1: The study area shapefile that was created to crop environmental variables to the extent of the fieldwork area. The inset map shows the shapefile's situation in the wider continental geographic context.



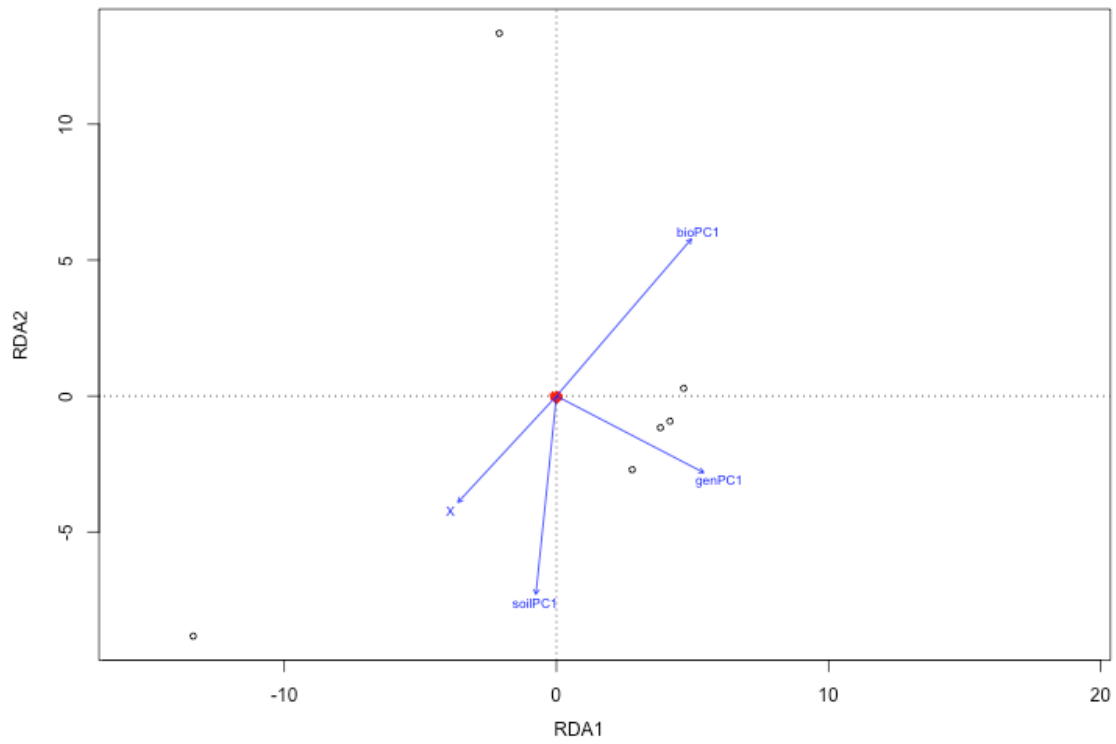
Supplementary Figure 5.2: Goodness-of-fit, as measured by Bayesian Information Criteria (BIC), for the number of clusters (k) Catharsius *individuals are split into during k-means clustering. From this, k=5 and k=11 were chosen for further analyses. The cluster membership for each specimen can be found in Table 5.2.*



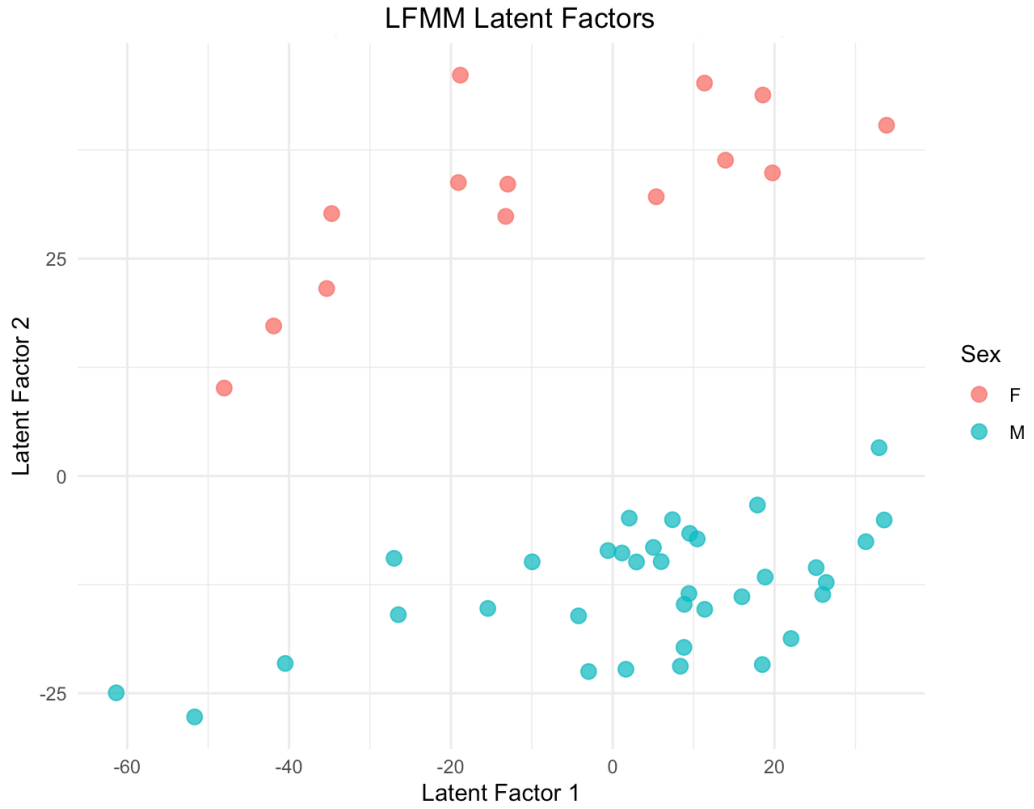
Supplementary Figure 5.3: The first and second principal components of the PCA of Catharsius *SNPs showing the main axes of genetic variation, zoomed into the* C. dux *and* C. duciformis *cluster (not including specimen Z3_18). Individuals are coloured by collecting location, showing no clear pattern.*



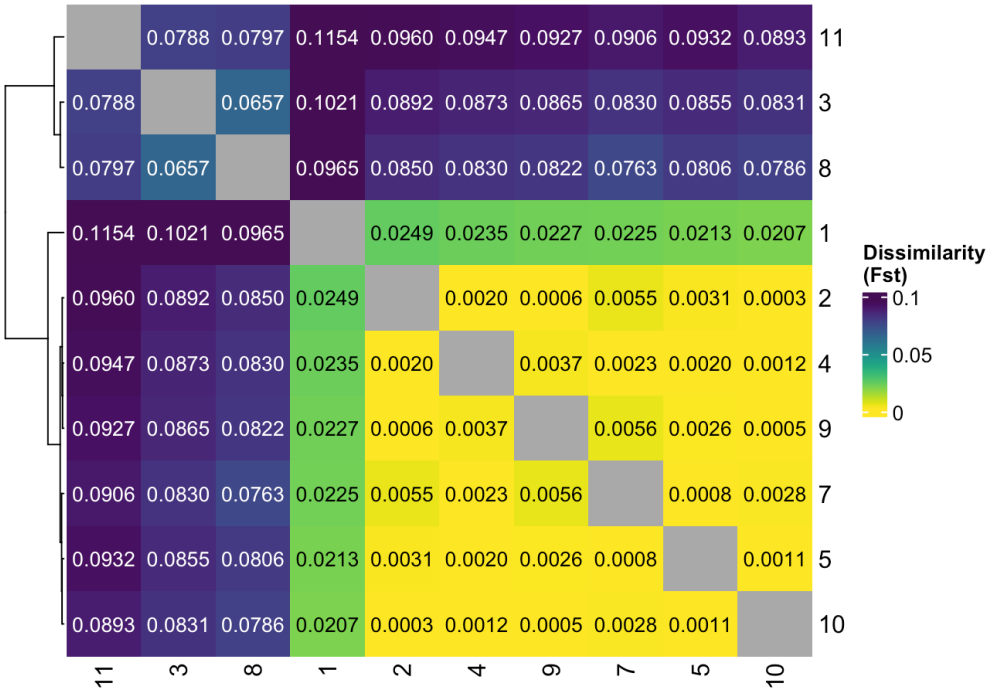
Supplementary Figure 5.4: The first and second principal components of the PCA of Catharsius *SNPs showing the main axes of genetic variation, zoomed into the* C. dux *and* C. duciformis *cluster (not including specimen Z3_18). Individuals are coloured by collecting species identification, showing no clear pattern.*



Supplementary Figure 5.5: Biplot of Partial Redundancy Analysis (pRDA) Model 1 illustrating the relationships between response variables (collecting locations displayed as points) and explanatory variables (arrows). A degree of correlation is seen between Soil (soilPC1) and Geography (x).

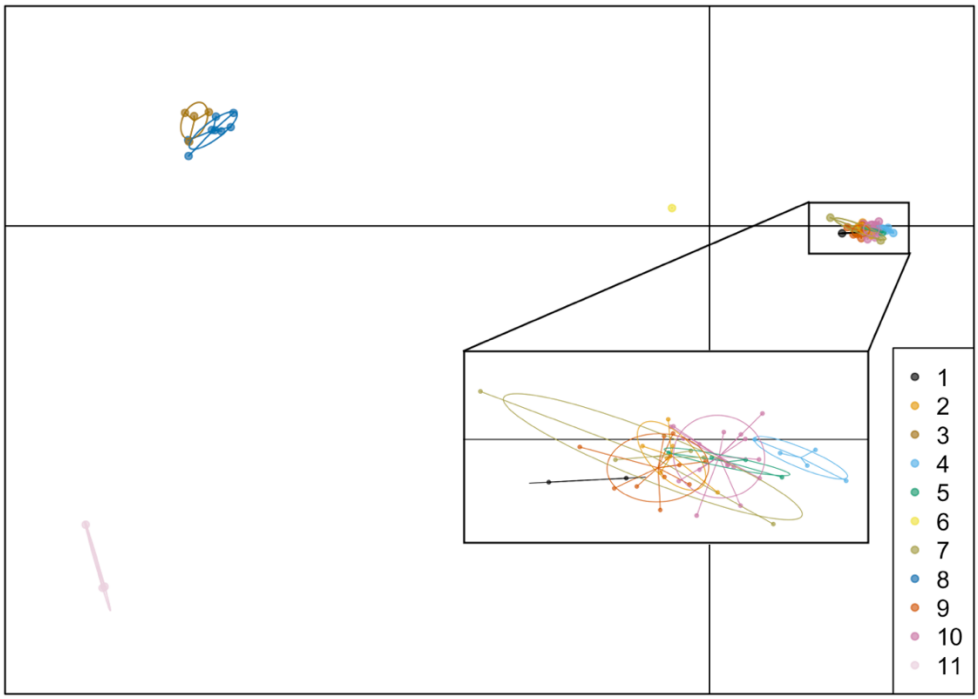


Supplementary Figure 5.6: Scatterplot of the first two latent factors from the Latent Factor Mixed Model (LFMM). Points represent individual Catharsius *specimens, coloured by sex. Latent factor 2 distinctly separates female and male specimens.*

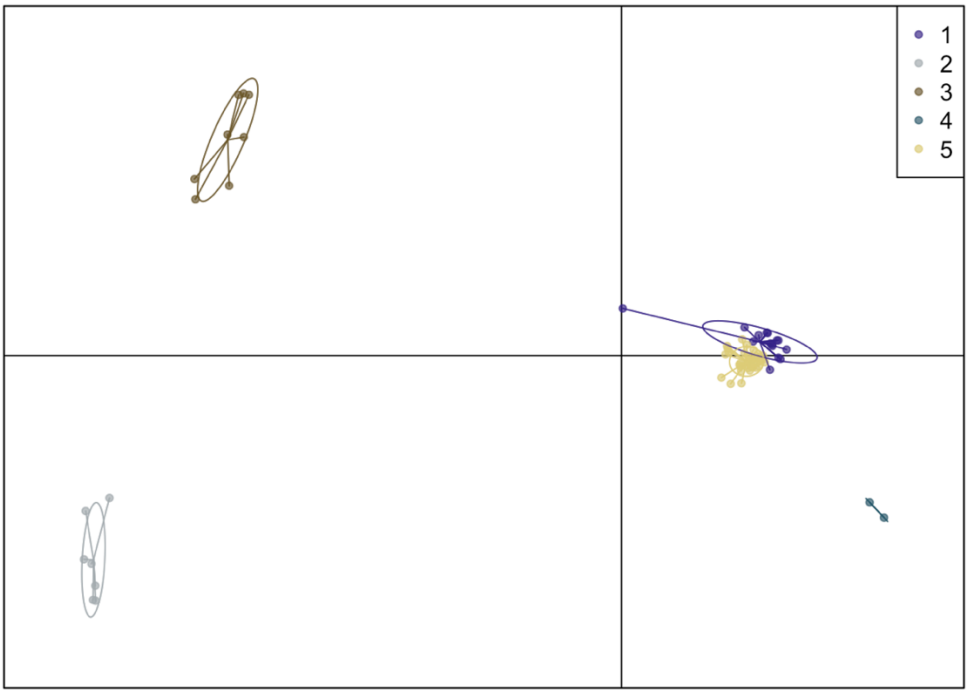


Supplementary Figure 5.7: Heatmap showing pairwise comparisons between clusters obtained from k-means clustering with k = 11. Clusters 2, 4, 5, 7, 9, and 10 contain C. dux *and / or* C. duciformis*, cluster 1 is composed of* C. satyrus*, while clusters 3, 8, and 11 include all other* Catharsius *species. Dark blue values indicate high levels of dissimilarity, and yellow indicate low levels. Cluster 6, containing just specimen Z3_18 is not included as fixation index (Fst) is a measure of population differentiation due to genetic structure and could not be computed with a single individual.*

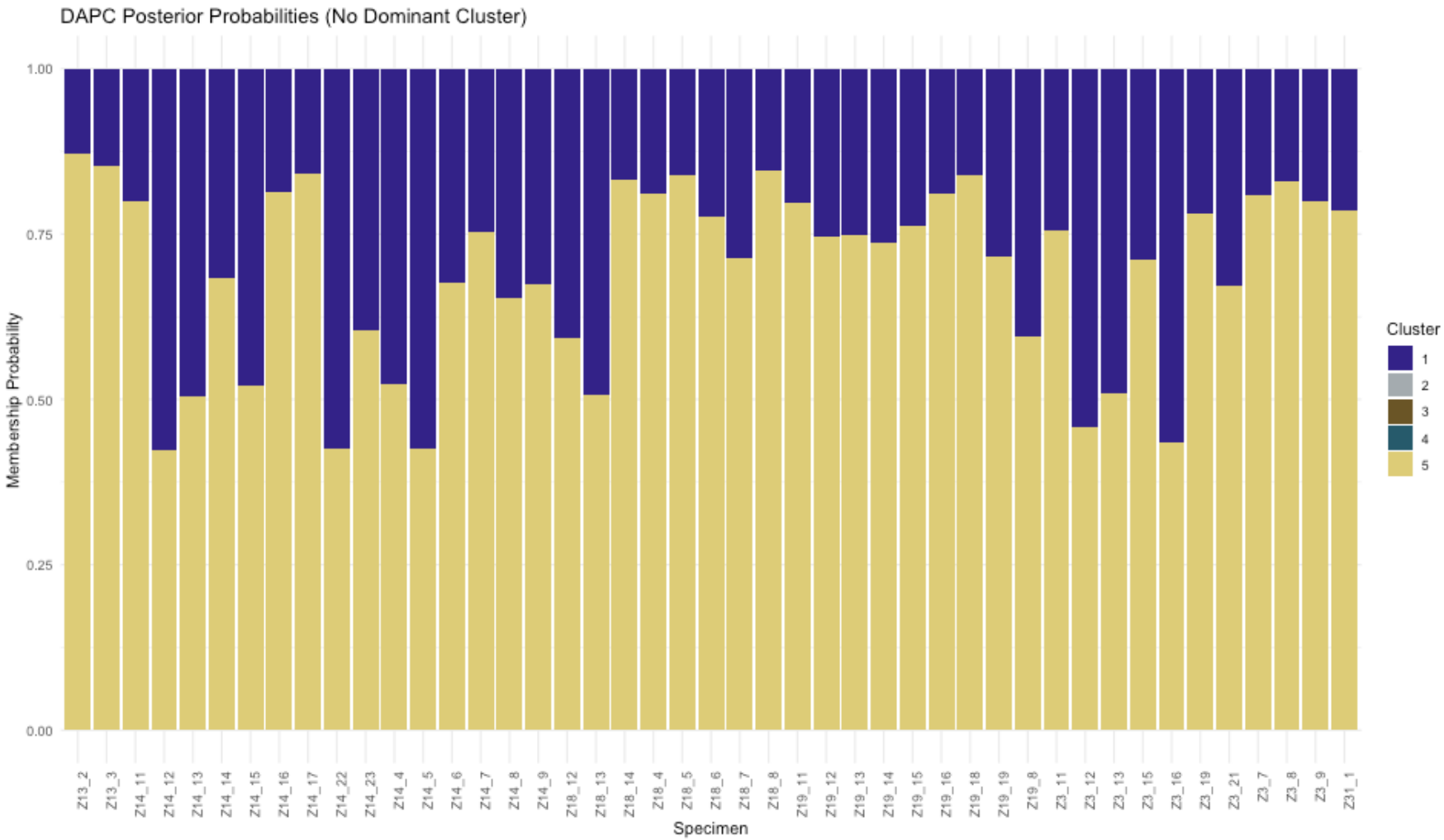
Supplementary Figure 5.8: Heatmap showing pairwise comparisons between clusters obtained from k-means clustering with k = 5. Clusters 1 and 5 contain C. dux *and / or* C. duciformis*, cluster 4 is composed of* C. satyrus*, while clusters 2 and 3 include all other* Catharsius *species. Dark blue values indicate high levels of dissimilarity, and yellow indicate low levels, as measured by fixation index (Fst).*



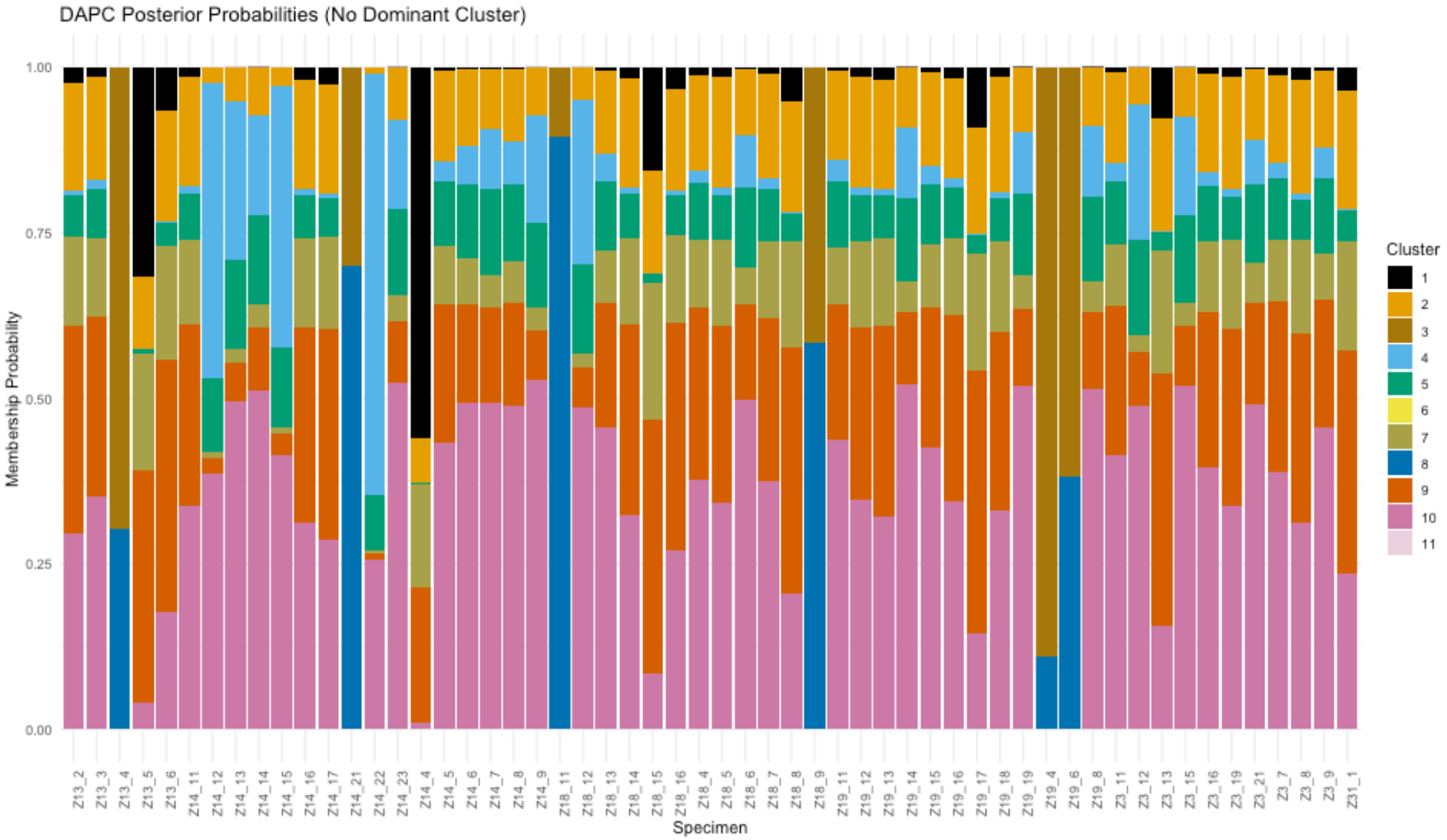
Supplementary Figure 5.9: Scatterplot of Discriminant Analysis of Principal Components (DAPC) showing Catharsius *individuals plotted along the first two discriminant functions. Points are coloured by prior clusters for k = 11, illustrating the genetic differentiation captured by the DAPC axes and the clustering structure in the dataset. The inset box is all specimens of C. dux and / or C. duciformis, with the exception of Z3_18, shown in yellow. Colours correspond with Table 5.2. Ovals and lines are to illustrate distance to cluster PCA average, and individual membership can be found in Table 5.2.*



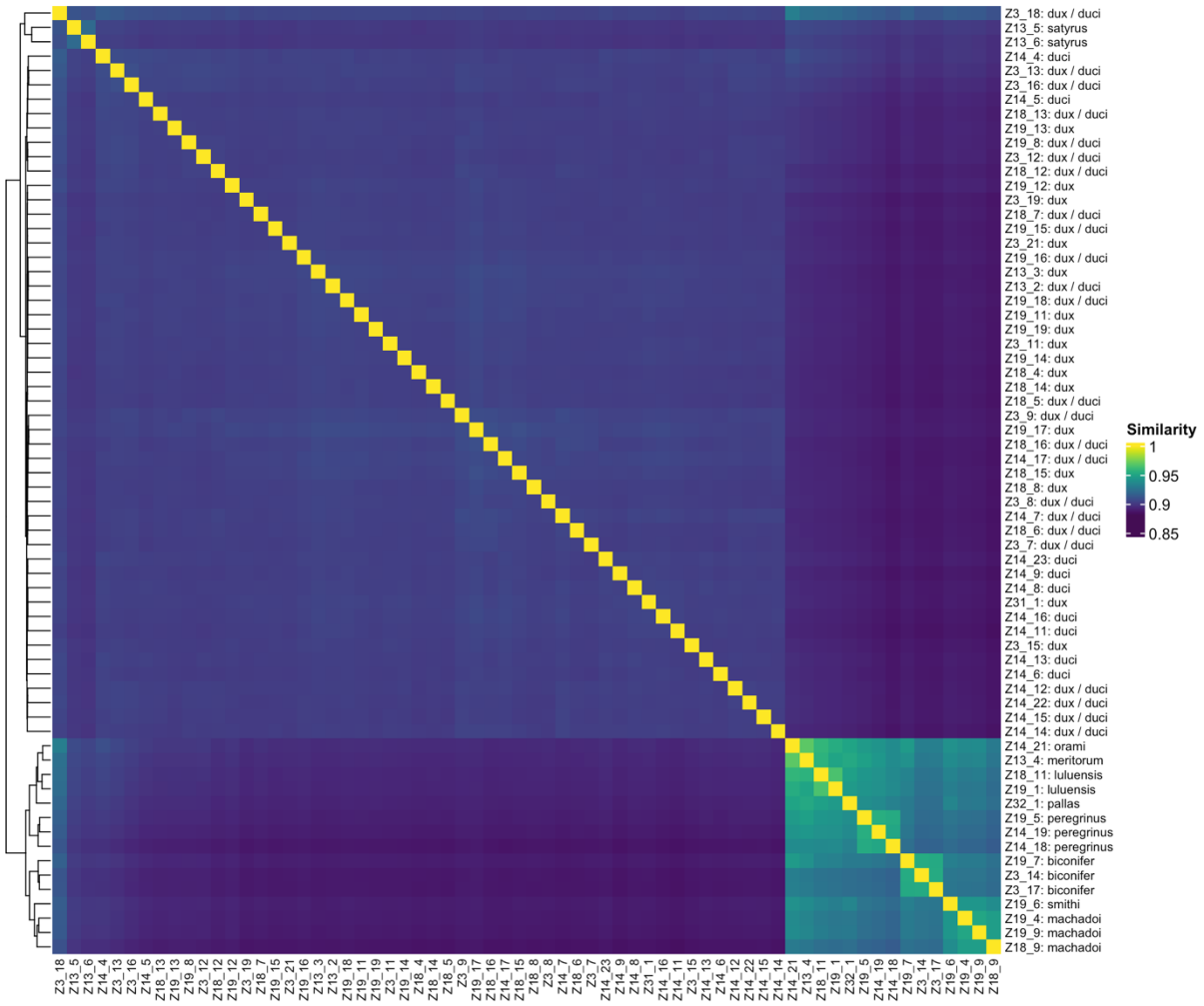
Supplementary Figure 5.10: Scatterplot of Discriminant Analysis of Principal Components (DAPC) showing Catharsius *individuals plotted along the first two discriminant functions. Points are coloured by prior clusters for k = 5, illustrating the genetic differentiation captured by the DAPC axes and the clustering structure in the dataset. Colours correspond with Table 5.2. Ovals and lines are to illustrate distance to cluster PCA average, and individual membership can be found in Table 5.2.*



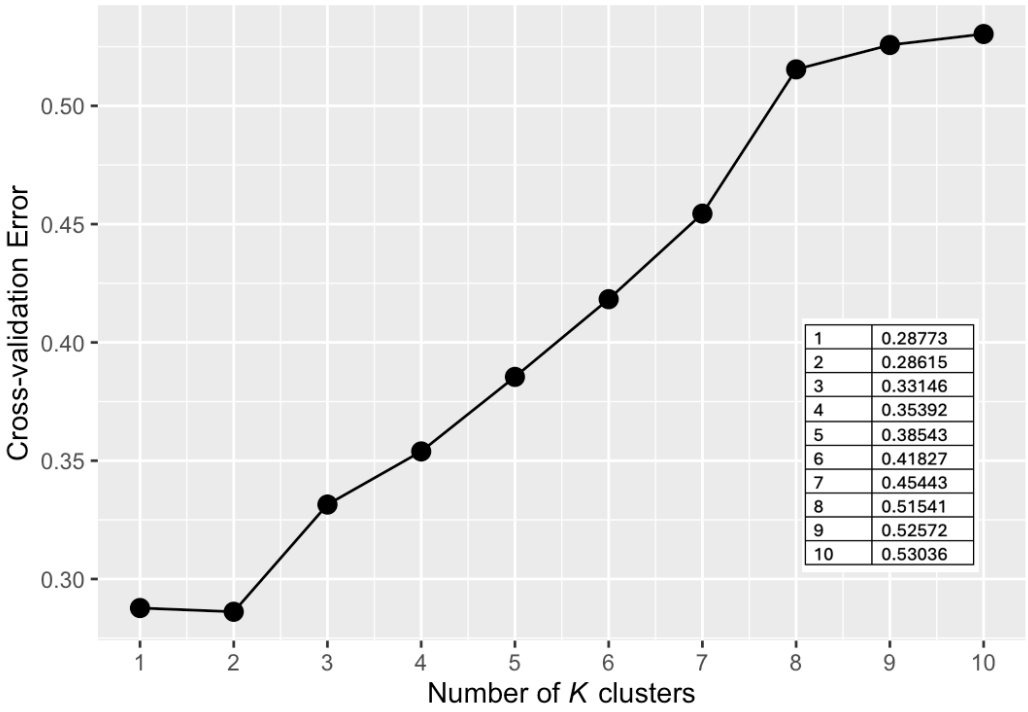
Supplementary Figure 5.11: Composition plot of DAPC posterior probabilities for "unsure" Catharsius individuals without a posterior probability greater than 0.9 for any cluster when k = 5, highlighting uncertainty in cluster assignment. Each bar represents an individual, and colours indicate the probability of membership of each cluster. Colours correspond with Table 5.2.



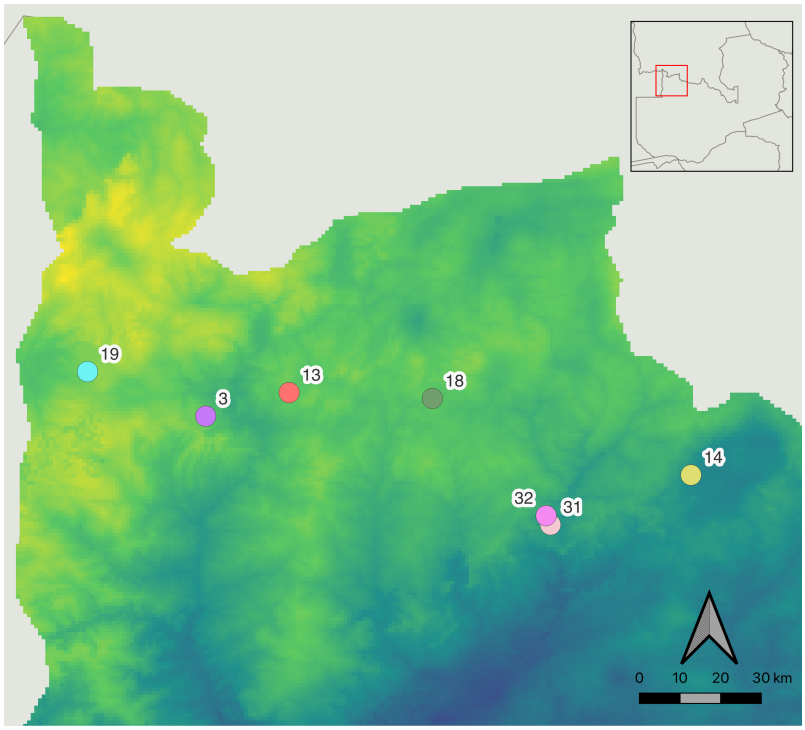
Supplementary Figure 5.12: Composition plot of DAPC posterior probabilities for " unsure" Catharsius *individuals without a posterior probability greater than 0.9 for any cluster when k = 11, highlighting uncertainty in cluster assignment. Each bar represents an individual, and colours indicate the probability of membership of each cluster. Colours correspond with Table 5.2.*



Supplementary Figure 5.13: Heatmap depicting genetic similarity based on shared alleles between Catharsius *individuals. Yellow indicates a high number of shared alleles, reflecting genetic relatedness, and dark blue indicates a low number. Individuals are labelled according to their specimen number, and on the righthand axis include their identification, where "duci" refers to "duciformis".*



Supplementary Figure 5.14: Cross-validation error rates for ADMIXTURE analyses across varying numbers of clusters (K). The inset box displays the specific cross-validation error values for each tested K.



Supplementary Figure 5.15: Geographic projection of the first principal component axis of the climatic PCA. The map reveals a pronounced environmental change oriented NE-SW at the eastern end of the sampling transect. Catharsius *sampling locations are indicated by points (colours corresponding to Table 5.2), illustrating the spatial distribution of collected specimens in relation to the climatic variation. This first climatic PCA axis predominantly captures the transition from wetter (yellow) to drier climate (blue) that occurs with increasing distance from the tropical environment found to the north in the DRC.*