

From satellite pixels to fossil apes: targeting gaps in the hominid fossil record



João Pedro Valente de Oliveira Coelho
St. Catherine's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2023

Acknowledgements

I express my heartfelt gratitude to my parents, my brother, and my family for their firm support throughout this journey. To my beloved Joana, I am profoundly grateful for your love and for you choosing to embark on, not only the years we've shared, but also the forthcoming ones in Gorongosa. You've helped in more ways than I can count. To all the friends in my hometown, Ansião, and those I've encountered in Coimbra, your enduring closeness throughout this period is truly appreciated. Thank you for all the laughs and hugs when I felt lost.

I extend my gratitude to all my colleagues and friends I met at Banbury Rd 64 in Oxford, with special thanks to those who welcomed me—Kat, Dan, Adam, Arran, Jacob, and Hristo. The memories shared with the cohort of students who embarked on this journey alongside me, Fig, Lynn, and Lucy, will be cherished always. My warmest appreciation to everyone who has been or is a member of the PrimoBevo Lab. Thank you for your guidance, support, and the enjoyable moments we've shared. As I write these words in a particularly cold winter, I find myself missing the mulled wine from The Rose & Crown, even though my initial encounter with its flavour was not a symphony for my taste buds. Here's a toast to these unforgettable moments at Oxford—may our memories also age well—Cheers!

In the lab at Oxford or amidst the landscapes of Kenya and Mozambique, I've had the pleasure of crossing paths with numerous researchers, students, field assistants, and *fiscals* who not only became cherished friends but also played a part in one way or another in bringing my research to fruition. I owe a special debt of gratitude to Dr. David R. Braun, co-director of the Koobi Fora field school, for his pivotal role in granting access to the field and offering an abundance of counsel and insights about East Turkana. His extensive involvement, from in-depth discussions on my research proposal to continuous support before, during, and after fieldwork is much appreciated. Special thanks also go to Frances Forrest, Silindokuhle Mavuso, and Sharon Kuo for making me feel like part of the Koobi Fora family.

Every person I encountered in Chitengo, Gorongosa, has been incredibly gracious and welcoming. This magical place *from the Montane to the Mangrove* truly works as an open lab and a field university. Marc Stalmans, Jason Denlinger, Tongai Castigo, Miguel Lajas, and João Nhampoca provided crucial help with the logistics of field research and fossil surveying in the Miombo woodlands Gorongosa National Park. While there are many colleagues who have contributed to my delightful experience at Gorongosa, I want to highlight two particularly inspirational friends, Jacinto Mathe and Rassina Farassi. I wish to you both the very best as you continue to advance the fields of palaeoanthropology and primatology in Mozambique.

I'm deeply grateful for Dr. René Bobe's incredible generosity and guidance, which significantly shaped my dissertation by enriching both its depth and breadth. Learning anatomy and taxonomy under René's mentorship at the NMK in Nairobi was a priceless highlight of my life as a scientist. I also acknowledge Dr. Thomas Püschel key role as a senior research collaborator and field colleague, along with his contagious laughing and humour. Thank you, Thomas and Ara, for the friendship and sharing with me countless moments I'll forever treasure. Before arriving to Oxford, I would never expect that the best thing about the town would be all the “crazy” Chileans I've met there! I've learned a lot with all of you.

I am grateful to St Catherine's College and the School of Anthropology & Museum Ethnography for hosting me during my time at Oxford. Special thanks to FCT: *Fundação para a Ciência e Tecnologia* for its funding that supported my research and covered expenses throughout my DPhil. I also appreciate the generous support provided by The Boise Trust Fund, the Africa Oxford Initiative – AfOx, and the Centre for Functional Ecology. Their contributions were key in covering the expenses related to my fieldwork in the Rift, data collection at palaeontological museums, and participation in international conferences. This support proved indispensable for the success of my research endeavours over the past years.

Finally, I want to express my heartfelt appreciation for the invaluable feedback and mentorship provided by my supervisors. Prof Robert L. Anemone for all his steady support, sharp insights, and constructive feedback which have been instrumental for my growth as a researcher. The

initial moment I discovered Anemone's work online was bittersweet—first, I got a bit upset to learn I was not going to be the first person ever to uncover fossil sites from space using AI—but a split second after this sentiment was followed by immense joy as I realized I had found the best possible mentor for this project.

My most profound appreciation goes to Prof Susana Carvalho, who not only afforded me the opportunity to undertake this research under her guidance but also infused the journey with her boundless energy, unwavering dedication, and compassionate kindness. During these years many of my dreams have turned into reality due to her inspiring vision and superhuman work ethic. But the immense surprises during this wild learning experience were even greater than the dreams I had! I'm sincerely grateful for her consistent presence and support, especially during the moments when I needed it the most. I also want to acknowledge Prof Eugénia Cunha, who supervised my master's with great care, and played a pivotal role in connecting me with Prof Susana Carvalho for my DPhil in Anthropology, ultimately enabling me to pursue a career in Human Evolution. Their guidance has been instrumental in shaping my research journey, and I am profoundly grateful for their wisdom, support, and encouragement throughout this chapter of my life.

Abstract

Palaeoanthropological research encounters several limitations that impact our understanding of hominin origins. Fossil records are often incomplete and geographically biased, concentrating findings in specific regions. Temporal gaps and taphonomic biases lead to further uncertainties in evolutionary timelines. Nevertheless, ongoing interdisciplinary advancements shape and refine our understanding of human evolution. In Chapter 1, I contextualize these spatiotemporal gaps and propose potential solutions, including the implementation of remote fossil site detection methods. Chapter 2 offers a comprehensive review of the *Geospatial Palaeontology* literature, providing a thorough examination of existing research in this field. The discoveries in each chapter carry significant implications for future research and practical applications. In Chapter 3, I introduce a new meta-analytical approach for dating time trees and test it on the *Pan-Homo* divergence event. This analysis indicates a high likelihood of the split occurring before 8 Ma. This method aids in fine-tuning our comprehension of *Pan* and hominin origins and directs efforts towards pinpointing fossil deposits with greater chronological precision. The breakthroughs presented in Chapter 4 regarding remote detection methods offer promising prospects for the efficient identification of fossil sites. This is especially relevant for regions characterized by limited sampling and challenging surveying environments like Gorongosa. We found new highly fossiliferous deposits that harbour Miocene apes and many other mammalian species using this approach. The incorporation of innovative tools for large-scale automated habitat predictions, as pioneered in Chapter 5, significantly enhances our ability to reconstruct ancient landscapes. This advancement provides valuable insights into the ecological context of hominins and, for the first time, enables the creation of a comprehensive palaeoenvironmental indicator for most of the Koobi Fora formation in East Turkana, spanning both spatial and temporal dimensions. In the concluding Chapter 6, I delve into the primary contributions, acknowledge limitations, and chart potential future directions for research. At its core, this thesis seamlessly integrates a diverse array of interdisciplinary methods and concepts from fields such as data science, geospatial learning, palaeobiogeography, and computational palaeontology. This fosters a holistic approach to fossil site exploration, opening new avenues for future research in this dynamic and ever-evolving field.

Table of Contents

| | |
|---|----|
| 1. Introduction | 1 |
| 1.1 Background | 1 |
| 1.2 Biogeography of Gorongosa | 5 |
| 1.3 Palaeobiogeography of Gorongosa | 7 |
| 1.4 Gorongosa for testing coastal models of hominin evolution | 10 |
| 1.4.1 “East African Ground Ape” hypothesis – Kingdon, 2003 | 10 |
| 1.4.2 Coastal Refuge hypothesis – Joordens et al, 2019 | 13 |
| 1.4.3 Adaptations and resource exploration in the littoral | 15 |
| 1.5 Targeting the gaps – time beats timing | 17 |
| 1.6 Targeting the gaps – remote site detection | 20 |
| 1.7 Large-scale and automated palaeoecological reconstruction: a case-study in East Turkana..... | 22 |
| 2. Literature Review | 25 |
| 2.1 Traditional discovery of fossil sites | 25 |
| 2.2 Searching localities from space | 28 |
| 2.2.1 The early years of archaeological predictive modelling | 28 |
| 2.2.2 Predictive modelling by means of environmental data in Palaeontology | 31 |
| 2.2.3 Other GIS applications in Palaeontology | 33 |
| 2.2.4 Geospatial fossil hunting for hominins | 34 |
| 2.2.5 Predictive modelling through remotely sensed imagery in Palaeontology | 36 |
| 2.2.6 New approaches: hybrid-datasets, continental-wide taxa-specific discovery, and individual fossil targeting | 41 |
| 3. Parting ways: <i>Pan-Homo</i> Divergence Revisited | 47 |
| 3.1 Introduction | 48 |
| 3.2 Materials and Methods | 52 |
| 3.2.1 Pre-processing | 52 |
| 3.2.2 Fossil thresholds | 53 |
| 3.2.3 Regression analyses | 55 |
| 3.2.4 Meta-analyses | 56 |
| 3.2.5 Data and materials availability | 57 |
| 3.3. Results | 57 |
| 3.3.1 Dataset structure | 57 |
| 3.3.2 The 4.4 Ma threshold: The last <i>Ardipithecus</i> and the first <i>Australopithecus</i> | 58 |
| 3.3.3 The 6.2 Ma threshold: <i>Orrorin tugenensis</i> and <i>Ardipithecus kadabba</i> | 59 |
| 3.3.4 The 7.3 Ma threshold: <i>Sahelanthropus tchadensis</i> | 60 |
| 3.3.5 Regression analyses of molecular estimates from 1967–2021 | 61 |
| 3.3.6 Bayesian meta-analyses | 63 |
| 3.4 Discussion | 65 |
| 3.4.1 Conclusions | 69 |
| 3.5 Supplementary Text | 71 |
| 3.5.1 S1: Dataset distribution | 71 |
| 3.5.2 S2: Bayesian meta-analysis: further details | 72 |
| 3.5.3 Data S1 | 77 |

| | |
|---|-----|
| 4. Unsupervised learning of satellite images enhances discovery of late Miocene fossil sites in the Urema Rift, Gorongosa, Mozambique | 79 |
| 4.1 Introduction | 81 |
| 4.2 Materials & Methods..... | 87 |
| 4.2.1 Applying the <i>k</i> -means algorithm to satellite images | 89 |
| 4.2.2 Clusters as survey guides for fossil site discovery..... | 90 |
| 4.3 Results | 91 |
| 4.3.1 Digital validation | 91 |
| 4.3.2 Ground-truthing | 93 |
| 4.4 Discussion | 97 |
| 4.4.1 Conclusions..... | 99 |
| 5. Geospatial palaeoecology: estimating aquatic and terrestrial fossil abundance in East Lake Turkana | 103 |
| 5.1 Introduction | 105 |
| 5.1.1 Geospatial Learning: Supervised vs Unsupervised..... | 105 |
| 5.1.2 Learning in the absence of counter-examples..... | 108 |
| 5.1.3 Automated palaeoecological reconstructions: taking geospatial palaeontology a step further | 109 |
| 5.2 Materials and Methods | 111 |
| 5.2.1 Study area..... | 111 |
| 5.2.2 Layered dataset | 112 |
| 5.2.3 Data collection in the field: bone walks..... | 116 |
| 5.2.4 BetaReg: Aquatic-to-terrestrial ratio..... | 117 |
| 5.2.5 MaxEnt: Fossil distribution..... | 118 |
| 5.2.6 PalaeoEnv: Filtered ensemble..... | 119 |
| 5.3 Results | 120 |
| 5.3.1 BetaReg results | 120 |
| 5.3.2 MaxEnt results | 122 |
| 5.3.3 BetaReg vs MaxEnt | 124 |
| 5.3.4 Ileret | 125 |
| 5.3.5 Karari | 127 |
| 5.3.6 Koobi Fora | 130 |
| 5.4 Discussion | 132 |
| 5.4.1 Remote palaeoenvironmental reconstruction: spatiotemporal patterning..... | 132 |
| 5.4.2 Short-wave infrared: soil content and moisture | 137 |
| 5.4.3 Near-infrared: a “fossil band”? | 139 |
| 5.4.4 The visible bands: RGB and ultrablue | 140 |
| 5.4.5 Modern environmental and topographic contexts..... | 141 |
| 5.4.6 Conclusion | 143 |
| 5.5 Supporting Info | 145 |
| 5.5.1 SF1. A leave-one-out cross-validation (LOOCV) procedure for the beta regression model..... | 145 |
| 5.5.2 SF2. Calculating variable importance in Beta regression..... | 147 |
| 5.5.3 SF3. Response curves for single variable MaxEnt independent models | 149 |
| 5.5.4 SF4. Response curves of the final MaxEnt model for remote detection of fossiliferous deposits | 150 |

| | |
|--|-----|
| 6. Discussion..... | 151 |
| 6.1 Summary | 151 |
| 6.2 Key contributions | 153 |
| 6.2.1 Through the hourglass: the double-helix and clocks | 153 |
| 6.2.2 Filling gaps: the sands of time | 155 |
| 6.2.3 The satellite's guide to Gorongosa's fossils..... | 161 |
| 6.2.4 Searching for Kingdon's coastal ape | 163 |
| 6.2.5 Do orbiting machines dream of ancient environments? | 164 |
| 6.3 Future directions..... | 167 |
| 6.3.1 Dating the tree of life | 167 |
| 6.3.2 Gorongosa's unfinished symphony: a work of decades | 168 |
| 6.3.3 Remote cave detection | 171 |
| 6.3.4 Geospatial palaeoecology: an emerging new discipline | 173 |
| 6.4 Conclusion..... | 174 |
| 7. Works Cited..... | 177 |
| Appendix I | 224 |
| Appendix II | 263 |
| Appendix III..... | 294 |

List of Figures

| | |
|---|----|
| Figure 1.1 – Chronological timeline of hominin species and associated stratigraphic ranges, highlighting the emergence of the earliest known techno-complexes (adapted from Bobe & Reynolds, 2022). | 3 |
| Figure 1.2 – Geological formations of the African Late Miocene , including preliminary dates for Gorongosa sites (Bobe <i>et al.</i> , 2023). Notice the sparsity of deposits spanning the 9–7 Ma period within Africa (adapted from Bobe & Reynolds, 2022). | 4 |
| Figure 1.3 – Gorongosa environments and the location of the southernmost fossil sites in the EARS . Generated in R using the Leaflet package (Cheng, Karambelkar & Xie, 2021). ... | 6 |
| Figure 1.4 – Palaeontological study area currently being surveyed by the PPPG within the geological context of Gorongosa National Park. | 7 |
| Figure 1.5 – Palaeogeographic reconstruction and depositional environments indicate a Miocene Gorongosa by the sea (after Habermann <i>et al.</i> , 2019). | 9 |
| Figure 1.6 – Sketch drawings of eastern Africa coastal apes from Jonathan Kingdon’s Lowly Origin (2003) overlapped to a photo of modern day estuarine coastal forest in Macaneta, Mozambique. | 11 |
| Figure 1.7 – The role of orbitally-induced wet-dry cycles in the Pliocene CFEA : A) Present day Indian Ocean coastal forests and Montane forests distribution. B) The humid-dry cycle oscillating between arid barriers and riverine corridors 5–2.5 Ma (after Joordens, 2011). | 14 |
| Figure 1.8 – Primates foraging aquatic resources : a) chacma baboon fishing in a desert waterhole (Hamilton & Tilson, 1985); b) chacma baboon eating shellfish in the littoral (Lewis & O’Riain, 2017); c) Japanese macaques are well known to explore coastal settings, but recently have also been found fishing salmon upstream (Takenaka <i>et al.</i> , 2022); d) bonobos bipedally wade when fishing for iodine-rich herbs (Hohmann <i>et al.</i> , 2019); e) chimpanzees forage for crabs in forest streams (Koops <i>et al.</i> , 2019); f) crab-eating macaques use stone tools to access marine invertebrates (Luncz <i>et al.</i> , 2017). | 16 |

| | |
|--|----|
| Figure 1.9 – Polynomial regression showing the trend for hominin brain expansion (dataset provided in Du <i>et al.</i> , 2018)..... | 18 |
| Figure 2.1 – The two main approaches of data collection for geospatial predictive modelling. Either using spectral bands (each band as a variable) or environmental variables as inputs..... | 29 |
| Figure 2.2 – Potential localities (red) in the Great Divide Basin, Wyoming as estimated by an artificial neural network. Notice that prospecting the whole basin would take centuries even with large teams. This method severely reduces the area of prospection (after Emerson & Anemone, 2012)..... | 38 |
| Figure 2.3 – Supervised and Unsupervised Learning are the two main conceptual modes of thinking statistical problems within the machine learning paradigm. | 40 |
| Figure 2.4 – Ensemble Learning approach to identify potential fossil areas with combined models (after Block <i>et al.</i> , 2016). For a given taxon, the areas yield new fossils (red map) are those where the species used to live (brown map), where its fossils could be preserved (blue map), and where it is now possible to find its fossils (green map). | 44 |
| Figure 3.1 – Molecular estimates histogram. Dashed vertical lines represent fossil thresholds. Note that instead of following a normal distribution, studies seem to cluster in excess around important fossil discoveries, and there is an unexpected gap of mean-estimates at 9–9.5 Ma. | 58 |
| Figure 3.2 – Interquartile range boxplots for divergence estimates filtered by different fossil thresholds. The arithmetic means and the medians are represented by white diamonds and black bars, respectively. All boxplots fit within the late Miocene (11.6–5.3 Ma). | 60 |
| Figure 3.3 – Polynomial regression “full-dataset model” fitting the sample of Panini/Hominini split estimates by date of publication. Vertical dashed bars represent dates of publications of possible early hominins. A) <i>Au. anamensis</i> , <i>Ar. ramidus</i> ; B) <i>O. tugenensis</i> , <i>Ar. kadabba</i> ; C) <i>S. tchadensis</i> . The “ <i>Sahelanthropus</i> -restricted model” linear regression in purple fits all data above the <i>Sahelanthropus</i> filter (7.2 Ma) and since its publication (2002). The late Miocene is within the horizontal dotted black bars. Dotted error-bars are standard | |

errors, presented when available (studies with confidence or credible intervals were transformed to s.e.).61

Figure 3.4 – **Polynomial regression with molecular estimates, excluding the first two decades and the studies in the shaded areas in red, orange and yellow (filtered-by-thresholds model).** The linear regression in green (genomics-specific model) includes only the studies based on genomic data (without pruning). Dotted error-bars are standard errors, presented when available (studies with confidence or credible intervals were transformed to s.e.).63

Figure 3.5 – **Meta-analysis of the Pan/Homo divergence estimates.** Forest plot depicting the specific effect size and sampled posterior distribution of each study after applying the 7.3 Ma threshold.....64

Figure 4.1 – **The great gaps of the African late Miocene.** a) Time gap: during this key period the African fossil record of primates is very incomplete (evaluated through species richness); but notice the split estimates from genomics; b) Spatial gap: virtually no fossils of this age are known in southeastern Africa, notice the strategic location of Gorongosa. Data extracted from palaeobiodb.org, map adapted from Bobe et al. (2018); c) study area for *k*-means within the geological context of Gorongosa, adapted from Habermann et al. (2019).82

Figure 4.2 – **The miombo woodland and the challenges it presents to fossil prospecting: Gorongosa Palaeontological Location 1 (GPL-1).** a) GPL-1 outcrops, notice how the surrounding vegetation is far more dense and extensive than in typical fossil sites from the EARS; b) GPL-1 in high-resolution satellite image, extracted from bing.com, shows a reduction of vegetation, but outcrops are barely noticeable; c) GPL-1, in a black rectangle, appears brighter than surrounding areas, when being mapped by lower resolution Landsat 8 false colour (infrared) image, and the same happens with other fossil sites, suggesting the infrared bands might be a useful indicator of fossiliferous deposits.84

Figure 4.3 – **Surveying for fossils in a densely vegetated context.** a) Despite the ground foliage and dense vegetation, in situ and surface evidence of fossils abound in the gully valley connecting GPL-12 to GPL-12B; b) Systematic mapping and collecting of surface fossil finds

by students of the field school; c) Side gully (~3 m deep) exposure and shovel test pit at GPL-12. Photographs are from the Paleo-Primate Project Gorongosa archive.....85

Figure 4.4 – **Flowchart of the algorithmic pipeline used for remote fossil site detection:** 1) Example of one of the seven spectral bands satellite images used in this study; 2) false colour map based on the infrared bands, after cropping to study area; 3) results of clustering using all seven spectral bands; 4) Binarize clusters for classification by selecting the cluster that contains most fossil sites as the target class (“walking back the cat”) versus all other clusters combined into a single class.89

Figure 4.5 – **Output from the *k*-means algorithm for data mining.** All geolocations “Vertebrates + Invertebrates” and “Invertebrates” recorded by the PalaeoPrimate Project Gorongosa team during 2016 and 2017 (Habermann et al., 2019) are plotted over the clusters, as well as the “Single vertebrate find” and “Fossilized wood” localities reported by Pickford (2012, 2013). You can see the cluster 1 (white) tends to be represented in locations with fossil vertebrates, indicating that it has some potential as a new feature/variable for finding new fossil sites. The map displays a 6 by 6 km square; axes scales are in meters.92

Figure 4.6 – **Violin-boxplots comparing sample distribution of spectral bands between clusters.** Each spectral band is represented in its own cell, with nine bars comparing the range of spectral bands values at the geocoordinates of the fossil sites, plus the eight clusters generated by *k*-means. Notice how overall, cluster 1 tends to approximate better the true spectral range of known fossil sites in Gorongosa.....94

Figure 4.7 – **Binarized classification plotting cluster 1 vs all other clusters.** New fossil sites GPL-10, 11, 12 and 12B are documented here for the first time. Trackways of surveys during 2018 are drawn in black. Clusters 2-8 are merged into a single cluster and compared against cluster 1 (predictive cluster). Total area = 36 km². One grid square = 1 km². One pixel-cell = 900 m².95

Figure 4.8 – **Variable importance of spectral bands for clustering.** Bars represent relative importance of the spectral predictors for optimally classifying all clusters as calculated by a

| | |
|---|-----|
| supervised random forest algorithm (Breiman, 2001). Specific variable importance for detecting cluster 1 is shown with open circles..... | 96 |
| Figure 5.1 – Study area geographic context: a) East Turkana as depicted in a preview of scene ID = LC81690572018036LGN00, Landsat 8 image from 2018-Feb-5. The dashed rectangle was the cropped section used in all analyses. Notice the image has no clouds on the region of interest; b) the raw satellite image overlaid on its larger geographic context, covering part of northern Kenya and southern Ethiopia. | 113 |
| Figure 5.2 – All the layers of the stacked dataset used during training for predictive modelling | 115 |
| Figure 5.3 – Bonewalks raw data: A) Geocoordinates of the points of interest (POI) over a map of the Paleontological Collection Areas (Brown and Feibel, 1990, revised by Bobe et al., 2022), also including the training extent of the models in a pink dashed rectangle, and three smaller rectangles with the Ileret (purple), Karari (teal), and Koobi Fora (yellow) sub-regions; B) Ratio of aquatic-to-terrestrial fauna sampled by systematic bone walks in each POI..... | 118 |
| Figure 5.4 – Outputs of the main models for the East Turkana region: A) aquatic-to-terrestrial ratio predictions; B) estimated distribution of fossils; C) Final paleoenvironmental reconstruction based on BetaReg’s output filtered by MaxEnt. | 121 |
| Figure 5.5 – A test of the MaxEnt model using 1903 georeferenced fossil presences extracted from the PaleoTurkana database (Bobe et al., 2022); * - absences were generated by randomly selecting another set of 1903 xy points across the map..... | 124 |
| Figure 5.6 – Comparison of variable importance metrics for A) BetaReg and B) Maxent, as calculated by percent contribution, as defined in varImp() function from R (Liaw & Wiener, 2002). | 125 |
| Figure 5.7 – Predictive modelling of Ileret: A) BetaReg with raw output before masking; B) MaxEnt model predicting exposure of fossiliferous deposits; C) PaleoEnv model overlapped by the geological members of the Koobi Fora Fm (Gathogo, 2003; Gathogo & Brown, 2006). | 126 |

Figure 5.8 – **Boxplot of PaleoEnv model outputs** (BetaReg predictions after MaxEnt masking). Each box width is relative to the total number of cells after filtering (i.e., total fossiliferous pixels). Fill colour is the ratio of cells remaining by the total cells in each area (i.e., fossiliferous density). Diamonds are average means; and boxes are interquartile of aquatic ratio (25%; median; and 75%). Boxes are ordered from BetaReg’s median prediction of most terrestrial to most aquatic (after filtering by MaxEnt). 128

Figure 5.9 – **Predictive models of Karari's Area 105**. A) BetaReg output, before masking; B) MaxEnt model predicting exposure of fossiliferous deposits; C) PaleoEnv model with geological layers. 129

Figure 5.10 – **Predictive models of Koobi Fora’s Area 103**. A) BetaReg with raw output, before masking; B) MaxEnt model predicting fossiliferous deposits; C) PaleoEnv with geological layers. 131

Figure 6.1 – **Preliminary temporal intervals for Gorongosa fossil sites** based on three-point estimates of $^{10}\text{Be}/^9\text{Be}$ dating using authigenic values based on different ecological assumptions (Appendix II), combined with the geological time scale from the Holocene to the Oligocene (GSA, 2022), and median divergence estimates of the literature, with respective confidence intervals, were obtained from timetree.org (Kumar *et al.*, 2022). 155

Figure 6.2 – **Challenges in deducing divergence times from the fossil record**. A) Primary obstacles in using palaeontological data to calibrate lineage splits: 1. Identifying and dating the clade's earliest known fossil (FAD); 2. Calculation of the temporal discrepancy ΔT_{Gap} between the FAD and the emergence of the lineage's initial fossilizable trait; 3. The interval $\Delta T_{\text{Div-1stApo}}$ between the clade’s first fossilizable characteristics and the actual genetic point of divergence. The latter, $\Delta T_{\text{Div-1stApo}}$, cannot be known, as it pertains to a period for which the clade's fossils are indistinguishable and thus unidentifiable within the fossil record. B) The estimation of ΔT_{Gap} is complicated by the diminishing likelihood of fossil discovery near the lineage’s origin, compounded by the inherently sporadic and fragmentary nature of fossils and stratigraphic records. Adapted from Marshall (2019) in order to illustrate the *Pan/Homo* split scenario. 160

| | |
|--|-----|
| Figure 6.3 – BetaReg predictions on grid cells known to contain A) <i>Homo</i> and <i>Paranthropus</i> ; B) Antilopini and Bovini. Geocoordinates sampled from the PaleoTurkana Database (Bobe et al., 2022). | 165 |
| Figure 6.4 – Morphospace for upper left first incisors , a preliminary analysis to assess the phenetic affinities of the Gorongosa cf. Hominoidea UI1 PPG2022-P-091 to other Primates via PCA..... | 170 |
| Figure 6.5 – Karstic map of the Afro-Arabian continental landmass . Vectorial dataset layer is the World Karst Aquifer Map (Goldscheider <i>et al.</i> , 2020), and the map was generated using the leaflet package for R (Cheng <i>et al.</i> , 2021). Zooming 10x: A) Total extension of cave systems in and around the Cradle of Humankind, South Africa; B) Gorongosa and associated limestone areas..... | 172 |

Authorship Statements


Collaborator Statement – Susana Carvalho

I hereby grant permission for the inclusion of the following chapters in the thesis submitted by João Pedro Valente de Oliveira Coelho to the University of Oxford for the degree of Doctor of Philosophy. I confirm that João Pedro Valente de Oliveira Coelho is the main author and contributor of these chapters including conceptualization, methodology, primary data collection and statistical analyses, as well as interpretation, data curation, visualization, and writing of the original manuscripts.

Chapter 3: Parting ways: *Pan-Homo* Divergence Revisited

Chapter 4: Unsupervised learning of satellite images enhances discovery of late Miocene fossil sites in the Urema Rift, Gorongosa, Mozambique

Chapter 5: Geospatial paleoecology: estimating aquatic and terrestrial fossil abundance in East Lake Turkana

Signature 

Name: Susana Cláudia Ribeiro Marques de Carvalho

Date: 15/12/2023

University: University of Oxford

Email: susana.carvalho@anthro.ox.ac.uk

Collaborator Statement – Robert L. Anemone

I hereby grant permission for the inclusion of the following chapters in the thesis submitted by João Pedro Valente de Oliveira Coelho to the University of Oxford for the degree of Doctor of Philosophy. I confirm that João Pedro Valente de Oliveira Coelho is the main author and contributor of these chapters including conceptualization, methodology, primary data collection and statistical analyses, as well as interpretation, data curation, visualization, and writing of the original manuscripts.

Chapter 3: Parting ways: *Pan-Homo* Divergence Revisited

Chapter 4: Unsupervised learning of satellite images enhances discovery of late Miocene fossil sites in the Urema Rift, Gorongosa, Mozambique

Chapter 5: Geospatial paleoecology: estimating aquatic and terrestrial fossil abundance in East Lake Turkana

Signature: 

Name: Robert L. Anemone

Date: Dec 17th, 2023

University: University of North Carolina at Greensboro

Email: Robert.anemone@uncg.edu

Collaborator Statement – René Bobe

I hereby grant permission for the inclusion of the following chapters in the thesis submitted by João Pedro Valente de Oliveira Coelho to the University of Oxford for the degree of Doctor of Philosophy. I confirm that João Pedro Valente de Oliveira Coelho is the main author and contributor of these chapters including conceptualization, methodology, primary data collection and statistical analyses, as well as interpretation, data curation, visualization, and writing of the original manuscripts.

Chapter 3: Parting ways: *Pan-Homo* Divergence Revisited

Chapter 5: Geospatial paleoecology: estimating aquatic and terrestrial fossil abundance in East Lake Turkana

Signature: 

Name: René Bobe

Date: 18-12-2023

University: Gorongosa National Park

Email: renebobe@gmail.com

Collaborator Statement – David R. Braun

I hereby grant permission for the inclusion of the following chapters in the thesis submitted by João Pedro Valente de Oliveira Coelho to the University of Oxford for the degree of Doctor of Philosophy. I confirm that João Pedro Valente de Oliveira Coelho is the main author and contributor of these chapters including conceptualization, methodology, primary data collection and statistical analyses, as well as interpretation, data curation, visualization, and writing of the original manuscripts.

Chapter 5: Geospatial paleoecology: estimating aquatic and terrestrial fossil abundance in East Lake Turkana

Signature: 

Name: David R. Braun

Date: December 19th 2023

University: George Washington University

Email: David_braun@gwu.edu

Covid Statement

Balancing the academic pressures, teaching and tutoring, personal commitments, fieldwork, scientific conferences, and the rigorous research required for a doctoral thesis, has occasionally overwhelmed me. But never in my wildest nightmares I thought I would also have to deal with a worldwide pandemic or the second and third-order effects it caused. My work was difficulted, interrupted, or delayed multiple times due to the Covid-19 pandemic and the prolonged government lockdowns. I was unable to access the laboratories or even an office space for about two years, which severely impacted my productivity. My scholarship also ended abruptly during this challenging period, and the extensions for Covid-impacted students were limited to a few months. The lack of social contact with my colleagues in real life, and the increased load of virtual events also impacted my stress levels. The fact that I had to alter multiple times and miss most of my final stages of field work plans were also very difficult to handle and interfered with my motivation and well-being. Adaptability and strategic planning played pivotal roles in this endeavour, and I hope that, despite the challenges, I managed to successfully craft a cohesive manuscript.

1

Introduction

1.1 Background

When and where did our ancestors, the earliest hominins, originate? The current and possible consensus from molecular and fossil evidence points to the late Miocene forests and woodlands of Africa (White *et al.*, 2015; Püschel *et al.*, 2021; Alméjida *et al.*, 2021). Yet the late Miocene (11.6–5.3 Ma) is an extensive time period ranging over 6 million years, and Africa is a massive continent containing ~12.7% of the Earth’s land (Kingdon, 1990, 2003, 2023). Regrettably the geographic distribution of African fossil sites during this key period is highly biased and the associated hominin fossil record is likewise sparse (Cote, 2004, 2018; Senut, 2015; Wood & Smith, 2022). Given the significant gaps and uncertainties in the hominin fossil record, it is pertinent to question the feasibility of achieving a more precise resolution regarding the window of time when hominins originated (Bobe & Wood, 2022). This thesis merges methods and techniques from data science, biogeography, and computational palaeontology to offer a novel contribution to understanding the contexts of hominin evolution. A set of goals is proposed to address the following outstanding issues: to define a more precise chronological range for the divergence event between the Hominini and Panini clades, so that we can target sediments of correct age (Chapter 3); to increase the accuracy of remote detection of fossil