Title

Bayesian analysis of presence-only data predicts 0.43% of Malta's Pleistocene fossil discoveries will be new species

Authors

Adrian Timpson[1,2], Alexandra A. E. van der Geer[3], Emily Y. Hallett[1], Victoria L. Herridge[4], Mark G. Thomas[2], Nicholas Vella[5], Ritienne Gauci[5,], Eleanor Scerri[1,5,6]

1. Pan-African Evolution Research Group, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, 07745, Jena, Germany.
2. UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College London, Gower St, London WC1E 6BT, United Kingdom
3. Naturalis Biodiversity Center, Vertebrate Evolution, Development and Ecology Research Group, Darwinweg 2, 2333 CR Leiden, Netherlands
4. Natural History Museum, London, U.K.
5. Department of Classics and Archaeology, University of Malta, Msida, Malta
6. Institute of Prehistoric Archaeology, University of Cologne, Cologne, 50931, Germany.

Abstract

1. Presence-only datasets are common in a wide range of ecological studies, including the identification of faunal or botanical species in fossil records. Inherently lacking sample sizes, these data are the information-impoverished analogue of count data. As a consequence, presence-only data are and important resource that are poorly exploited and typically only used qualitatively.
2. We propose a novel Bayesian method which accounts for all compositional count possibilities in presence-only data to probabilistically predict the future discovery of a new species, and test for differences between datasets. We apply this to the fossil record of Malta to test the widely assumed hypothesis that its Pleistocene fauna was similar to that of Sicily, given the frequent periods of low sea levels when a land bridge emerged between them.
3. We establish that the faunal composition of Malta and Sicily did differ significantly in all Pleistocene sub-periods. We predict 0.427% (95% CI = 0% to 2.1%) of future Maltese fossil recoveries will yield new undiscovered species, which is 13 times more probable than on Sicily.
4. Our results indicate substantial long-term ecological and evolutionary differences existed between Malta and Sicily despite frequent Pleistocene land bridges. Our prediction of new species discoveries provides a robust quantitative justification for further fieldwork on Malta.

Keywords

Bayesian; MCMC; mean likelihood estimator; missingness; presence-only data;

Introduction

Ecological context

The current state of Malta and Sicily as distinct islands is atypical. For 2.5-million-years Europe has been dominated by recurrent glacial periods when vast amounts of water locked in polar ice drastically depleted Mediterranean Sea levels, joining these islands together. By the Late Pleistocene, the Maltese Islands featured one of the highest rates of faunal endemism in the Mediterranean (Van Der Geer 2021), yet the remarkable natural history of this Mediterranean 'Galapagos' remains poorly understood. Attempts to use the palaeontological record to connect vertebrates with their environmental settings have been limited by the destruction of most of the original early fossil sites, rendering a modern, high-resolution evaluation of chronology and context impossible. Faunal turnovers are clearly attested, yet there is little understanding of the contemporaneity of different species, or the timing and causes of extinctions, since chronologies are largely only relative approximations. As a result, Malta has played only a minor role in the reconstruction of broader, regional biogeographic and evolutionary patterns, and has typically been viewed as an extension of a larger Sicilian story. Fossils of many faunal species have been found on both Malta and Sicily, suggesting migration during the frequent land-bridged glacial periods, although the direction, amount and timing of such events remains to be established. Observed differences between their fossil records have generally been attributed to limited sampling, particularly on Malta where discovered sites are sparce. However, substantial separation has also been hypothesised, with fossil similarities instead attributed to convergent evolution (Herridge 2010). Untangling these processes is particularly challenging given the problematic nomenclatures used (e.g. see Van Der Geer (2021) and Herridge (2010) for discussion), and ultimately will require further palaeontological investigation from the discovery and careful excavation of new sites. It remains unclear why some species have so far been discovered only on one island or the other. Therefore, it is both a starting point and a soluble statistical problem to use the current best available data to determine whether the difference between observational records is merely the consequence of a sampling bias or shows genuine faunal differences between Malta and Sicily during the Pleistocene.

Statistical context

Testing for a statistically significant difference in fauna between Malta and Sicily first assumes a null hypothesis that both had the same species, then evaluates if the observed number of differences could happen by chance given the number of samples. The fewer the samples, the greater the chance of observing more differences under this null. However, reliable sample sizes such as NISP (Number of Identified/Individual Specimens) are not available for the faunal fossil record of Malta and Sicily. This is not uncommon for prehistoric fauna assemblages, not only because many are hard to identify, but because research agendas are typically more interested in establishing where and when species existed, and less interested in quantifying the fossils once their presence has been established beyond doubt. Unfortunately, presence-only data pose a major hurdle for any statistical analysis. For example, compressing the observed sample sizes of species A (n = 50) and species B (n = 1) into merely the presence of A and B causes a massive and irreversible loss of information content, informing us merely that *at least one* fossil of each was identified. The true counts can be handled as free parameters in a model, but there remains an infinite number of count combinations of A and B that satisfy their presence.

As such, the key statistical contribution of this paper is in the design of an appropriate model that properly accounts for all possibilities that give rise to the observed presence/absence, allowing us to test these differences and provide quantitative predictions of future discoveries. This can be achieved effectively if we incorporate two further pieces of information. Firstly, we can include prior estimates of the *relative frequencies* of species that have been identified as present. Surprisingly this provides information about both the total number of samples, and about the number of fossils from new species that have not yet been discovered – the missingness. This is because the presence of a rare species in our data is better explained by a larger number of fossil samples, and a larger number of samples reduces the chance of any further undiscovered species. Secondly, we can include a prior estimate of the total number of fossil samples that have been recovered from each Island. Even the most conservatively broad range provides a hugely informative limit on what is otherwise only constrained to be equal or greater than the number of identified species.

The necessarily subjective nature of these two priors may be of concern to the frequentist who values only the data, but within a Bayesian inferential framework these prior beliefs work beautifully in conjunction with the data to reveal a tightly constrained parameter space. The precision of these estimates can be hugely improved if we also have *absence data* i.e., species that can be reasonably assumed to exist, but have not yet been observed. For example, under the null hypothesis that the true faunal composition of Sicily and Malta were the same, any species observed in Sicily but not on Malta can be considered absent from the Malta dataset. This generates a statistical tension whereby the absence of common species is better explained by fewer samples, whilst the presence of rarer species is better explained by more samples.

Geological context

Today, Malta is a small archipelago with an area of just 316 km^2. However, during periods of low sea levels it was either a much larger island or a south-eastern peninsula of a connected Sicily-Italy. The size of Malta and its land bridge or 'steppingstone' island connections to the European mainland are estimated using the bathymetry of the Malta Plateau, the now underwater region between Malta and Sicily. As a structurally elevated submarine shelf covering an area of 10,700 km^2, the Malta Plateau is a seaward extension of the Hyblean plateau from Sicily and is located at bathymetric depths ranging from 100m to 150m (Bishop & Debono 1996; Micallef, Berndt & Debono 2011; Reyes Suarez *et al.* 2019). The north-western side of the Plateau is bounded by submarine ridges with depressions reaching depths of 1700m. The Maltese archipelago represents the part of the Plateau that emerged in the early Messinian c.6Ma (Pedley & Clarke 2002). The Plateau's submarine morphology is characterised by its position on the African plate, which is compressively wedged in a north-western direction along the European continental plate at a relatively slow rate of convergence, averaging less than 1 cm per annum (Catalano *et al.* 2008; Micallef, Berndt & Debono 2011). However, the current Afro-Eurasian margin also represents the last stages of a large-scale and rapid process of subduction rollback, which affected the western and central Mediterranean during the past 30 million years (Galea 2019). The result of this structural setting is a complex tectonic-controlled mosaic of thrust faults along the northern and western margins of the Malta-Sicily ridge, a sequence of shallow shelves and elongated, fault-derived rift basins of Mio-Pliocene origin in its centre and the NNW-SSE trending Malta Escarpment on its eastern margin (Reuther & Eisbacher 1985).

Projecting the current bathymetry into the past is problematized by tectonic changes. Patterns of tectonic instability since the Late Pleistocene have been inferred in the region of western and south-eastern Sicily using seismic and global positioning system (GPS) data (Oldow *et al.* 2002; Anzidei *et al.* 2014), however over the last 125ka the Maltese archipelago has been generally observed as tectonically stable (Pedley & Clarke 2002; Galea 2007; Serpelloni *et al.* 2007; Furlani *et al.* 2013). Therefore, post 125ka estimates of the changing coastal morphology of Malta and its connections based on past periods of Mediterranean Sea-level changes are fairly reliable (Siddall *et al.* 2003; Lambeck & Purcell 2005; Lambeck *et al.* 2011; Zecchin *et al.* 2015; Benjamin *et al.* 2017; Antonioli *et al.* 2018; Antonioli *et al.* 2021).

Studies focused on the Malta-Sicily channel have tended to reconstruct the most recent period of land connection and subsequent inundation, which can be used as a model for previous such events. The most recent cycle occurred during the Last Glacial Maximum (LGM, 20ka), when an exposed Malta Plateau acted as a connecting land bridge (90 km long and 40 km wide) between Sicily and Malta at 130m isobath (Furlani *et al.* 2013; Micallef *et al.* 2013; Foglini *et al.* 2016). Central to these studies, was the prior work of Lambeck *et al.* (2011) which reported estimates of sea-level changes for the past 20ka for south Sicily. This provided a reliable calibration for Malta due to both its close proximity and similar stable tectonic setting, and paleogeographic reconstructions in the Sicily-Malta channel at 130m isobath, supported by underwater archaeological surveys (Furlani *et al.* 2013; Micallef *et al.*

2013; Foglini *et al.* 2016; Furlani *et al.* 2018), are consistent with estimates for other Mediterranean stable coastlines (Lambeck *et al.* 2011).

Post-LGM glacial melting increased sea-level rise at an average rate of 5mm per annum (Lambeck *et al.* 2011) and by 14.4ka, the sea level in the Sicily-Malta channel rose to -100m, reducing the land bridge to less than 10km wide (Foglini *et al.* 2016). Disconnection between Malta and Sicily started from 12.9ka, when the Malta Plateau was completely submerged by successive events of glacio-eustatic sea-level rise. Coastal morphological features are thought to have been inundated by brief but rapid episodes of sea-level rise (i.e., melt-water pulses 1A and 1B), which occurred between 15 and 10 ka. Submerged features and deposits were observed to be located systematically at two specific water depth intervals, i.e., between −100 and −70m, and between −65 and −40m (Zecchin *et al.* 2015). Micallef *et al.* (2013) identified two prominent palaeo-shorelines at ca. 130m and 90m water depth off Malta, with the former linked to the low-stand LGM palaeo-shoreline, and the latter representing a drowned feature with the onset of the meltwater pulse (MWP) 1A. Similar evidence was found worldwide suggesting a consistent eustatic mechanism acted across the semi-enclosed Mediterranean Sea.

During the greatest glacial extremes of the Pleistocene, sea levels periodically dropped to at least 120 m below current sea level, coinciding with MIS 16 (c. 630ka; early Middle Pleistocene) (Bintanja, Van De Wal & Oerlemans 2005), MIS 12 (c. 460ka; middle Pleistocene), and a drop in sea-level to 130 isobath (MIS 6) c. 140ka which likely represents the most recent substantial land connection prior to the LGM (Benjamin *et al.* 2017). Other more recent dips such as c.100m around 80ka may not have been sufficient to maintain a complete land bridge, but rather may have exposed a series of islands linking a larger Malta with Sicily. Thus, Malta is likely to have had a semi-permeable relationship with Sicily between 140 and 20 ka, with steppingstone islands likely acting as a filter between faunal exchanges.

An important anomaly between the faunal records of Malta and Sicily is the absence of any fossils on Malta from the Early Paleolithic. This is not simply a matter of poor sampling, but rather a complete lack of any geological deposits, suggesting Malta was mainly submerged during this early period.

Evolutionary context

The faunal units or biozones of Malta and Sicily have been put on a par since the late 19th century, when the Maltese dwarf elephant fossils were compared with those from Sicily, and assigned to the same species (Busk 1868; Adams 1870). The subsequent biozones represent new colonisations from the Apennine mainland via what is today the Strait of Messina during the low sea level stands of glacial periods, therefore permitting faunal exchanges between Sicily and Malta during these periods. Significant sea level drops expose the shallow flat between Malta with Sicily's southern tip as dry land or shallow water forming a land bridge between the two islands (Vogiatzakis, Pungetti & Mannion 2008), increasing the geodispersal between these biozones in both directions.

However, the changing geology throughout the Pleistocene also provides a plausible mechanism for significant faunal differences between Malta and Sicily. Sea-level changes likely created a frequently changing ecological landscape, which in turn would change selective pressures. The separation of islands and loss of habitat diversity from higher sea levels during warmer interglacial periods would have permitted divergent evolutionary trajectories and independent extinctions, and this isolation coupled with a smaller terrestrial surface area would have reduced carrying capacity and population sizes, exacerbating evolutionary change further through increased genetic drift.

Materials and Methods

Pleistocene palaeontological faunal fossil data of Malta and Sicily

We collate from the literature faunal fossil records for broad sub-periods of the Pleistocene into an updated taxonomic framework. These are summarised for Malta and Sicily as Early Middle Pleistocene (EMP), Late Middle Pleistocene (LMP) and Late Pleistocene (LP). Additionally, data fossil records from the Early Pleistocene (EP) are available for Sicily, but despite extensive surveying, no geological deposits from this period have been found on Malta. Therefore, *a priori* we assume Malta was submerged during the EP and homed no terrestrial species.

The biostratigraphy on Malta is based on the cave of Għar Dalam which has a relatively well-preserved stratigraphy and has been extensively excavated (Van Der Geer 2021). The youngest layer in this cave is a cultural layer, which dates to the Neolithic and yielded remains of domestic animals, which we exclude. Għar Dalam yielded mainly megafauna (*Hippopotamus*, *Palaeoloxodon*, *Cervus*, *Canis*) and no dormice. As a consequence, their biostratigraphic position is inferred from co-occurring megafauna at their localities, in particular *Leithia melitensis* from Mnajdra, Tal-Ġnien fissure, Wied Inċita and Bengħajsa Gap, *L. cartei* from Mnajdra, *Maltamys gollcheri* from Mnajdra, and *M. wiedincitensis* from Wied Inċita. *Maltamys gollcheri* is sometimes considered a junior synonym of *Leithia cartei* (Zammit-Maempel & De Bruijn 1982), whilst we follow the view that *Maltamys* sp.–*wiedincitensis*–*gollcheri* are chronospecies and *Leithia cartei* is a distinct species (Masini *et al.* 2008). The elephant genus *Palaeoloxodon* is referred to as *Elephas* in older literature. We also drop subgenus level e.g., *Microtus* instead of *Pitimys* for the Maltese vole.

The current biozones of Sicily— also known as Faunal Complexes— are mainly based upon new arrivals and extinction events of megafauna in combination with stratigraphical data (Bonfiglio, Marra & Masini 2000; Bonfiglio *et al.* 2002; Masini *et al.* 2008). The earliest insular biozone, the Early Pleistocene Monte Pellegrino fauna, named after its single site (Kotsakis 1978) has so far no equivalent on Malta. Yet, it already contains early forms of the two giant dormouse lineages that are shared between Sicily and Malta starting in the early Middle Pleistocene. An even older biozone, the Messinian, late Turolian (MN 13) Gravitelli fauna (Seguenza 1902) is excluded here as it represents a Eurasian mainland fauna. The Gravitelli area (Messina) is located in the south-western end of the island, just a few kilometres separated from the Apennine mainland, of which it formed an integral part during the Messinian Salinity Crisis. Also excluded here is the Castello Faunal Complex of the terminal Pleistocene (likely coinciding with the Last Glacial Maximum) as this fauna is also a Eurasian mainland fauna, typical of the pan-Eurasian mammoth steppe fauna including aurochs (*Bos primigenius*) and horse (*Equus caballus*).

Initially, three dwarf Maltese elephant species were recognised: *Palaeoloxodon melitensis*, *P. falconeri* and *P. mnaidriensis*. The first is considered a junior synonym of the second (Ambrosetti 1968) as the material cannot be properly separated based on size. We follow this consensus view. Confusingly, the three elephant species names that were defined based on Maltese fossils, have been applied since their discovery to elephant remains from Sicily. In addition, a fourth, large-sized subspecies (*P. antiquus leonardi*) has been named for a Sicilian late Middle Pleistocene deposit (Aguirre 1969). Its fossils occur together with remains of middle-sized species at Contrada Fusco near Siracuse, indicating they might represent male individuals, as elephants are highly sexually dimorphic in size. We here exclude *leonardi* from our overviews.

The middle-sized dwarf elephant from Sicily likely represents a separate species from the similarly sized *P. mnaidriensis* from Malta (Ferretti 2008; Herridge 2010), thus in the literature is placed in quotation marks or prefixed with *aff*. to indicate taxonomic uncertainty. We here keep the name, pending a revision of the Siculo-Maltese elephant phylogeny. An additional dwarf species may be present on Sicily(Herridge 2010), given the size and geological age differences between the samples attributed to *P. falconeri* from Spinagallo and Luparello caves.

Prior beliefs of relative frequencies

Our estimates of the relative frequencies of each species are presented as integer ratios with respect to the rarest species (relative frequency = 1, see table 1). These prior beliefs are provided by author AAEvdG who formed these estimates using expert knowledge of trophic pyramids, predator-prey dynamics, modern species analogues, NISP counts of comparative fossil assemblages, and differential taphonomic loss. This prior knowledge is by definition subjective, so as a sensitivity test, we also use a second set of prior estimates from author EYH, ensuring complete independence between authors. They are similar in their ranking of the relative frequencies of species but differ quantitively by an order of magnitude (biggest ratio of 100:1 rather than 1000:1).

species	**name**	**Prior beliefs of relative species frequencies**		**Early Palaeolithic (EP)**		**Early Middle Palaeolithic (EMP)**		**Late Middle Palaeolithic (LMP)**		**Late Palaeolithic (LP)**	
		AAEvdG	**EYH**	**Sicily**		**Sicily**	**Malta**	**Sicily**	**Malta**	**Sicily**	**Malta**
Pellegrinia panormensis	Gundi	700	100	x							
Mustelercta arzilla	Marten	5	50	x							
Leithia sp.	Dormouse	800	100	x							
Hypolagus peregrinus	Hare	600	80	x							
Apodemus maximus	Giant fieldmouse	1000	100	x							
Asoriculus burgioi	Shrew	500	100	x							
Maltamys cf. gollcheri	Dormouse	800	100	x							
Leithia melitensis	Giant Dormouse	800	80			x	x				
Palaeoloxodon falconeri	Dwarf Elephant	300	10			x	x				
Leithia cartei	Giant Dormouse	800	80			x	x	x	x		
Maltamys gollcheri	Dormouse	800	80			x	x				
Crocidura esuae	White-toothed Shrew	500	100			x	x	x	x		
Lutraeximia trinacriae	Sicilian Otter	5	1			x		x			
Tyto mourerchauvireae	Large Barn Owl	5	1			x					
Cygnus equitum	Small swan	10	20			x					
Grus melitensis	Giant Crane	10	20			x	x	x	x		
Athene trinacriae	Owl	5	1			x					
Testudo hermanni	Hermann's Tortoise	7	10	x		x		x		x	
Testudininei sp.	Giant Tortoise	50	60			x					
Lutra euxena	Maltese Otter	5	1				x		x		
Maltamys wiedincitensis	Dormouse	800	100				x	x	x		x
Hippopotamus pentlandi	Dwarf Hippo	300	10					x			
Palaeoloxodon mnaidriensis	Dwarf Elephant	300	10					x	x	x	
Hippopotamus melitensis	Dwarf Hippo	300	10						x		
Microtus melitensis	Vole	800	80						x		
Grus grus	Common Crane	10	20						x		
Centrochelys robusta	Giant Tortoise	50	60						x		
Microtus pauli	Vole	800	80						x		
Dama carburangelensis	Fallow Deer	100	40					x		x	
Bos primigenius	Aurochs	100	20					x		x	
Cervus elaphus	Red Deer	100	40					x	x	x	
Sus scrofa	Wild Boar	100	40					x		x	
Ursus arctos	Brown Bear	1	1					x		x	x
Bison priscus	European Bison	100	20					x		x	
Canis lupus	Wolf	1	1					x	x	x	
Panthera leo	Lion	1	1					x		x	
Crocuta crocuta	Spotted Hyena	1	1					x		x	
Lepus europaeus	Brown Hare	600	80							x	
Erinaceus europaeus	European hedgehog	75	80					x			
Emys orbicularis	European pond turtle	7	60					x			
Cygnus falconeri	Giant Swan	10	20					x	x		
Equus hydruntinus	Wild Ass	100	10							x	
Apodemus sylvaticus	Common Fieldmouse	1000	100							x	
Microtus savii	Pine Vole	800	80							x	
Crocidura sicula	Sicilian Shrew	500	100							x	x
Equus caballus	Horse	100	10							x	
Crocidura sicula calypso	Gozitan Shrew	500	100								x
Cervus sp	Dwarf Deer (medium)	300	40								x
Cervus sp	Dwarf Deer (small)	300	40								x
Cervus sp	Dwarf Deer (tiny)	300	40								x
Microtus melitensis	Burrowing vole	800	80								x
Rhinolophus hipposideros	Lesser horseshoe bat	25	80								x

Vulpes vulpes	Red Fox	1	10							x	x

Table 1: Species identified as present (x) on the islands of Malta and / or Sicily during the four broad Pleistocene periods. Prior beliefs of the relative frequencies of each species were obtained independently from authors AAEvdG and EYG. These are qualitatively similar in terms of their ranking but differ quantitively by and order of magnitude.

Estimating missingness

Consider the simpler problem of how to use *counts* of fossils to quantify species that existed but have not yet been discovered. With the accumulation of many specimens we can become increasingly confident that no further undiscovered species exist, since the absence of evidence increasingly becomes evidence of absence as sample sizes increase. Equivalently, the chance of discovering a new species is greater if we have observed very few specimens so far.

We can model this as a vast urn containing balls (specimens) of many different colours (species). We then propose the existence of one further colour, which aggregates all remaining species that have not been observed, which we call the missingness. This missingness proportion of balls in the urn remains a free parameter in the model to be indirectly estimated (a percentage between 0% and 100%). Once a missingness percentage is proposed, it is trivial to convert the relative frequencies of the observed species to proportions, to ensure 100% of the balls have been appropriately assigned to either an observed species or missingness. For example, if two species are observed which we believe existed with a relative frequency of 1 to 10, and 3% missingness is proposed, the urn comprises 8.82%, 88.18% and 3% of each type respectively.

The likelihood of any proposed urn configuration (the probability of observed *counts* given some proposed urn proportions and missingness) can be calculated exactly using the multinomial distribution. However, there are many possible urn configurations that could lead to the same observed data, therefore there is always a range of possible true values of missingness. Fig 1 illustrates this for the simple case of just two species (A and B) believed to have been present with equal frequency, under all combinations of data between 0 and 10 samples each, and five discrete proposed values of missingness.

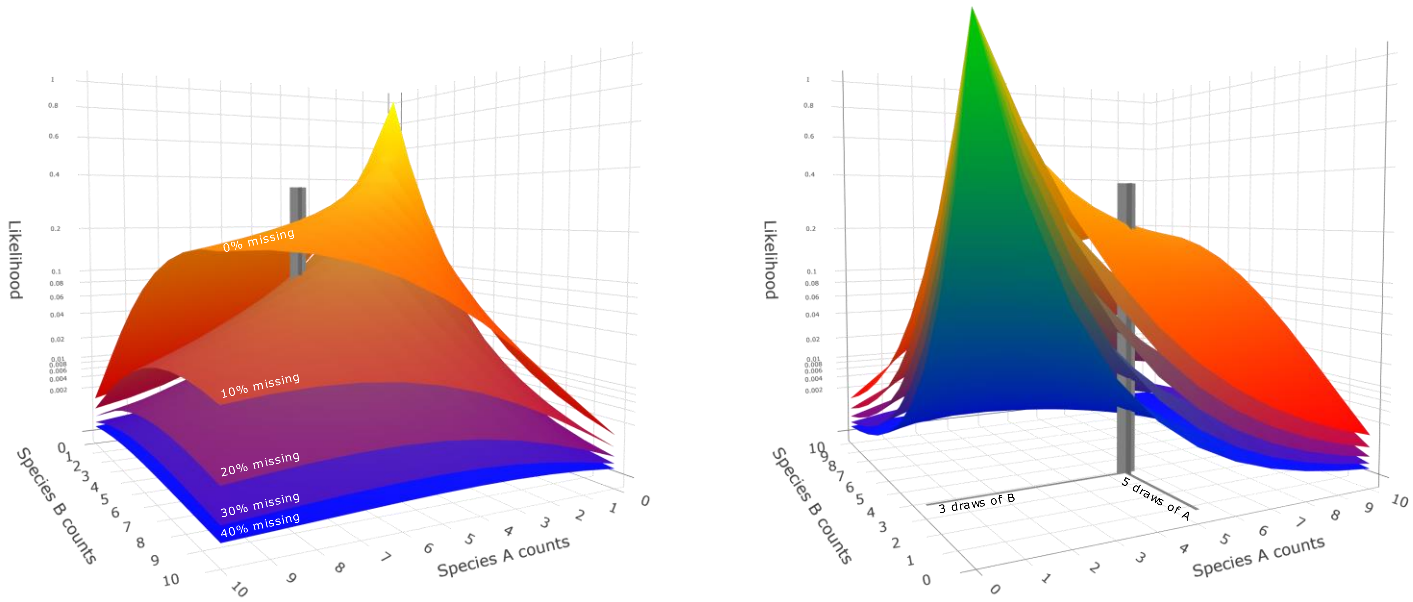

Fig 1: Two views of the same five likelihood surfaces, illustrating how confident we can be in a model of both species A and B being present with equal frequencies, under various combinations of observed counts and proposed missingness.

Notice that likelihoods always increase as the number of observations reduces. This is because it is more probable to observe one of each from two draws, than two of each from four draws, and certain to observe nothing from zero draws. Notice also that the surface with zero missingness always has a higher likelihood. This might seem strange – surely if I draw only two samples and each is a different species, I should expect to find a new species soon if I continue drawing from the urn?

The solution to this counterintuition is two-fold. Firstly, we are interested in the probability of drawing these two *specific* species, not *any* two (we will later specify prior beliefs on the proportions of *specific* species), so the chances of drawing two specific species from a possible two species is higher than drawing the same two from a possible three species. Secondly, and most importantly, this type of problem is served poorly by a point estimate, and even if a point estimate was desired, the maximum likelihood estimator (MLE) is not the best estimator. This is because the MLE always sits at the limit of the distribution, and therefore does not represent the expected value which is somewhere within the distribution. Instead, Fig 2 shows the Mean likelihood estimator (MELE) provides the expected value, which is equivalent to the Bayes estimator with a uniform prior (McLeod & Quenneville 2001).

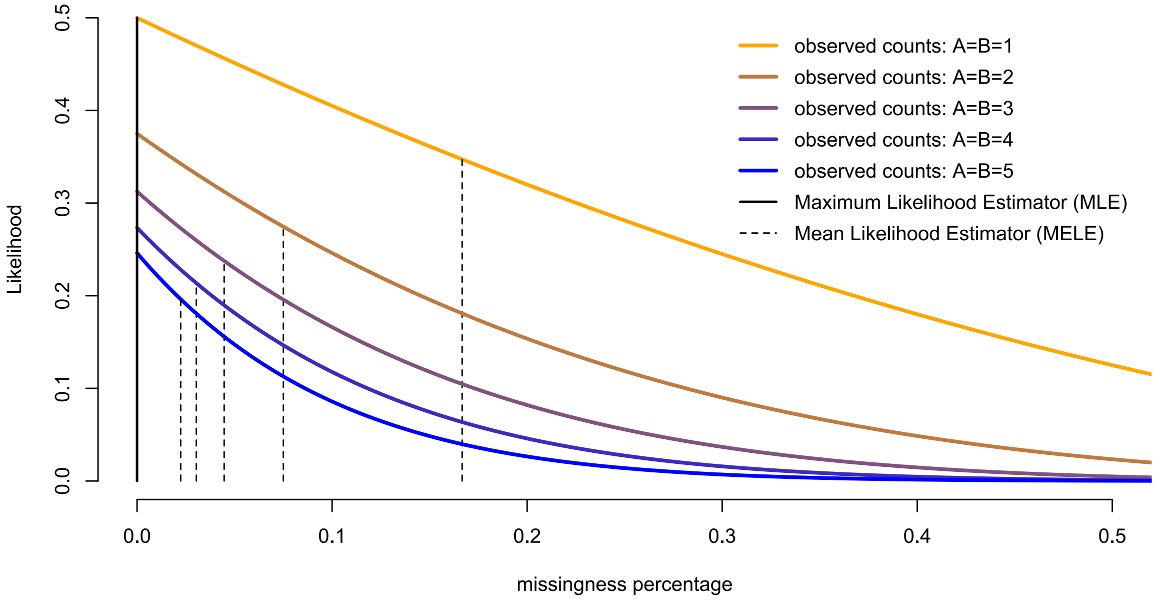

Fig 2: Intuitively there is a high chance of a new species C existing if we have only observed one count each of A and B, and a tiny chance of a new species C existing if we have observed thousands of A's and B's. Yet the MLE is always zero (solid vertical line). In contrast, the MELE provides an estimate of the expected missingness value (dashed vertical line), which is much greater with fewer observations.

Calculating likelihoods from Presence-only data

By definition, presence-only data does not include the number of samples for each species, and there are an infinite number of possible draws that satisfy the observation of being present. We resolve this by first introducing one further parameter – the total number of draws D, which must be an integer greater or equal to the number of species observed. For any proposed value for D we can then generate all possible combinations of species counts that sum to D and calculate a likelihood for each combination. For example, if we propose four draws (D=4) from two species (A and B), the three (unordered) sets AAAB, AABB, ABBB are plausible explanations for the observation that species A and

B were both present. Since all these combinations are possible, the overall likelihood becomes the sum of the likelihoods of each combination.

In practice this approach intractable for even moderate values of D due to the huge number of compositional sets determined by $C(n,r)$ where $n = D - 1$ and $r = number\ of\ species$. Instead we devise an efficient algorithm written in R (see SI for functions written in R)(Team 2013) that utilises the inclusion – exclusion principle so that only 2^{r-1} combinations need be generated, no matter how large D. For example, there are $C(499,10)$ = 2.4 x 10^{20} compositional sets that satisfy 500 draws from 10 species, but our algorithm only requires 2^9 = 512 combinations to be calculated.

The need for this additional parameter (D) couples with the poorer information content of presence data to generate huge equifinality, with a vast array of parameter combinations being almost identically probable. For example, Fig 3 illustrates that if five species have been observed, likelihoods are almost certain under zero missingness with any number of draws above c.50. Clearly point estimates cannot represent this uncertainty, nevertheless the maximum likelihood explanation is that we have exhaustively sampled (a huge number of balls were drawn) and there are no further species to be found. However, a *better* explanation using the MELE is that 13 fossils were drawn and 6.6% of fossils in the urn belong to as yet undiscovered species (Fig 3 left), or in the case of one species *a priori* believed to be 10 times more common than the others, the better explanation is that 34 fossils have been drawn with 2.7% belonging to new species (Fig 3 right).

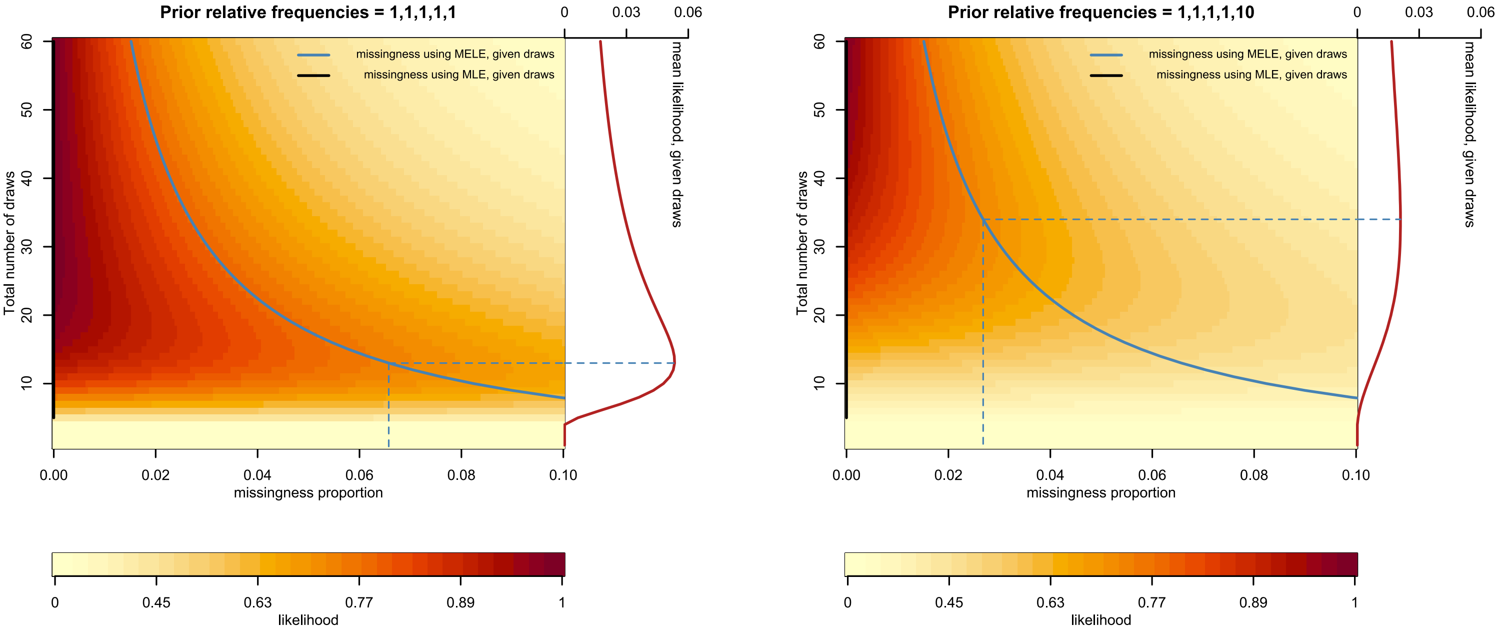

Fig 3: Likelihood surfaces illustrate the substantial equifinality in presence data, since there are many possible combinations of missingness and the number of draws that can result in the same observation of five species being present. Nevertheless, if point estimates are required, mean likelihood estimates (MELE) are superior to maximum likelihood estimates (MLE) for this type of problem. Left: five species are observed where all are *a priori* believed to be present with equal frequencies. Right: the same five species where one is *a priori* believed to be ten times more common than the others. In both cases the best MLE point estimate is an infinite number of draws with zero missingness (an exhaustive search where all species have been found). In contrast, the best MELE point estimate is 13 draws with a further 6.6% of samples in the urn belonging to as yet unobserved species (left), and 34 draws with a further 2.7% in the urn belonging to as yet unobserved species.

Known absence

If more than one sample set is available *for the same ecological unit*, we gain the advantage that a species that is present in one set but absent in another provides additional constraints to the parameters. For example, under the null hypothesis that land-joining between Sicily and Malta was substantial and persistent enough to ensure the same fauna on both, species present on Sicily that have not been observed on Malta are now known to be absent from the Maltese data, despite being hypothesised to have existed. Therefore, the failure to have observed them is best explained by fewer draws, which generates more statistical tension. Fig 4 illustrates this effect for just two toy presence sample sets, such that set 1 comprises species A, B, D, F, G whilst set 2 comprises species A, B, C, E. We apply our prior beliefs in the relative frequencies of A, B, C, D, E, F, G as 10, 8, 4, 3, 2, 1, 1 respectively, and an uninformative prior on the total number of draws. This results in a 3D likelihood manifold with regard to the total draws from set 1, total draws from set 2, and the common missingness (species in neither set), which we illustrate using 2D and 1D marginal distributions.

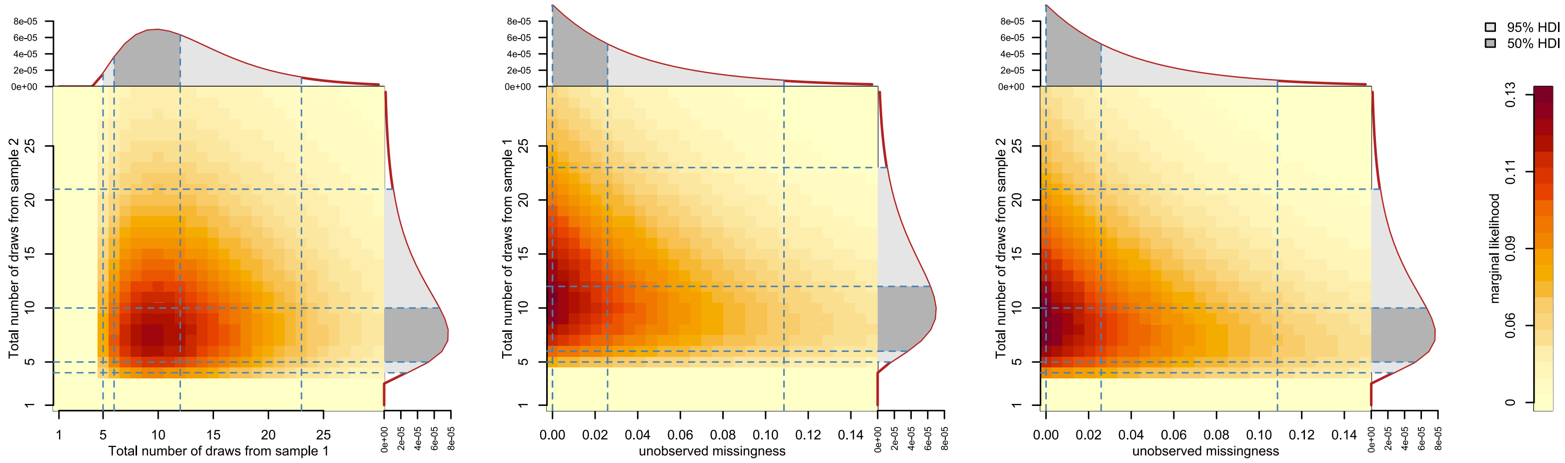

Fig 4: All three marginal 2D and 1D distributions of the 3D likelihood manifold, illustrating the joint likelihoods of i) the total number of draws from sample set 1; ii) the total number of draws from sample set 2; and iii) the proportion of missingness, given two pieces of information: a) the observation species A, B, D, F and G were present in sample set 1, and species A, B, C and E were present in sample set 2; and b) prior beliefs that the true relative proportions of species A, B, C, D, E, F, G are 10, 8, 4, 3, 2, 1, 1 respectively. This heavily constrains 3D parameter space, providing estimates that the total number of draws in sample 1 is 5 to 23 (95% HDI), total number of draws from sample set 2 is (4 to 21 95% HDI), and the unobserved missingness is 0 to 11% (percentage of the urn that belong to new species). 95% and 50% highest density estimates are illustrated for 1D marginals only.

Prior beliefs of total number of fossils

Parameter estimates can be greatly improved with the inclusion of a prior estimate of the total number draws D. These can be easily incorporated within the Bayesian framework by multiplying the likelihood distribution by a probability distribution that describes our *a priori* belief of D, which we use within an MCMC framework to sample from the posterior distribution. Our prior estimates (independent for each period for each island) were provided by author AAEvdG, who has expert knowledge of the archiving of fossil assemblages (see table 2). The ranges are deliberately wide to ensure conservative results and represent AAEvdG's 95% confidence intervals of the true number of fossils that have been discovered. In their raw form these are simply upper and lower boundaries, but we also incorporate a central tendency to represent AAEvdG's belief that values near the middle of the range are more probable than values near the boundaries. This is achieved by converting the raw ranges to the 95% quantiles of a unimodal distribution. A Gaussian distribution is inappropriate for this since the true number of samples cannot be negative, instead we use a Gamma probability distribution.

Period	***Island***	***Prior number of fossil samples***		***Gamma probability distribution parameters***	
		lower	***upper***	***shape***	***rate***
EP	Sicily	500	2000	14.44314	0.01305
EMP	Malta	20	150	7.15700	0.10596
	Sicily	2000	4000	55.88176	0.01925
LMP	Malta	500	2000	14.44314	0.01305
	Sicily	1000	5000	10.87494	0.00424
LP	Malta	500	2000	14.44314	0.01305
	Sicily	5000	10000	55.88176	0.00770

Table 2: Prior beliefs of the total number of fossil samples recovered in each period for each island. Lower and upper values represent the raw ranges that encompass the true number of fossils, with 95% confidence. These are then converted to Gamma distribution parameters (shape and rate) to also incorporate the central tendency of this prior probability distribution, such that values near the boundary are less probable.

Parameter estimates using MCMC

All parameters were estimated using Markov Chain Monte Carlo (MCMC) using the Metropolis-Hastings algorithm (Hastings 1970). The three independent search types are summarised in Table 3, repeated for each period independently (EMP, LMP and LP) plus EP for the independent hypothesis on Malta. Each was performed using AAEvdG's prior frequencies and repeated as a sensitivity test using AYH's prior frequencies. We ran 100 chains to avoid the need for thinning (Link & Eaton 2012), and each chain was run for 100,000 steps, with the first 1000 discarded for burn-in. Both the starting parameters and the jump sizes of each parameter proposal were tuned by literately repeating this entire process until convergence and reasonable acceptance ratios (c.20-55%).

Hypothesis	**Parameters**	**Parameter description**
Null	missingness D_{malta} D_{sicily}	Proportion of undiscovered fossils belonging to species not yet found on Malta or Sicily Total number of fossils so far drawn from Malta Total number of fossils so far drawn from Sicily
Malta Independent	missingness D	Proportion of undiscovered fossils belonging to species not yet found on Malta Total number of fossils so far drawn from Malta
Sicily Independent	missingness D	Proportion of undiscovered fossils belonging to species not yet found on Sicily Total number of fossils so far drawn from Sicily

Table 3: Summary of the independent parameter searches performed for each of the Palaeolithic time periods.

Null hypothesis test

Our objective is to test if the number of species differences between Sicily and Malta (present on one but not the other) could have occurred under the null hypothesis that Sicily and Malta had the same fauna. We test this for each period (EMP, LMP, LP) independently, by first performing a parameter search under the null (see Table 3), then randomly sample parameter sets the joint posterior distribution. Each parameter set is used to generate an integer number of observations for each species (for Malta and Sicily separately) by randomly sampling from the multinomial distribution such that size = D and the multinomial probabilities use the species frequency priors and the missingness. Integers are then compressed to presence – absence, and our summary statistic is the total number of species differences between Sicily and Malta. When repeated for 100,000 sampled parameter sets, this provides a summary statistic distribution under the null hypothesis, which can be compared with the same statistic from the observed data. Finally, we calculate the p-value as the proportion of null samples that generate more (or the same) differences as the observed data.

Alternative hypothesis

If the null hypothesis can be rejected, we instead accept the alternative hypothesis that the faunal compositions of Malta and Sicily differed, and we can instead directly use our missingness parameter to infer the proportion of new fossils belonging to undiscovered species. From this we derive the number of new fossils to be found before we should expect a better than even chance of discovering a new species.

Results

Null hypothesis test

The null hypothesis can be unequivocally be rejected for all three time periods tested (EMP, LMP, LP). Fig 5 illustrates the observed number of species differences between Malta and Sicily compared to the null distribution of differences, generated from the joint posterior parameter distributions. In the cases of LMP and LP, not a single iteration of the MCMC chain under the null model generated more differences than observed (18 differences from 26 species, and 22 differences from 25 respectively) and in the case of EMP around 0.04% of simulations were as extreme (8 differences). Since we reject the null hypothesis, no direct inferences can be drawn from the parameter estimates under the null, nevertheless the influence of the additional absence data clearly has a huge constraining influence on parameter space, and shows that precise estimates can be achieved in other applications where there are known absences (Fig SI1). As a sensitivity test, we repeated using our independent set of priors for the species' frequencies from author AMH (Fig SI2), which also rejected the null hypothesis for all three time periods.

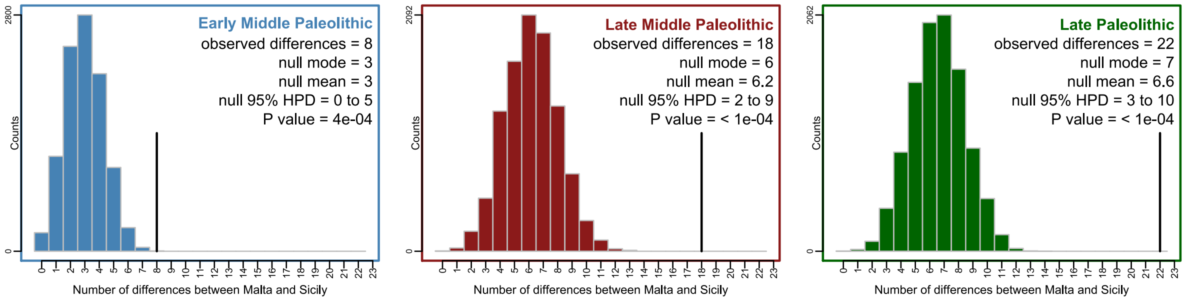

Fig 5: Null distributions of the number of species differences between Sicily and Malta, for each time period tested independently. In each case the observed number of species differences (vertical black line) is significantly greater than can be expected under the null.

Alternative hypothesis parameter estimates

Having rejected the null hypothesis, the alternative hypothesis that species composition to differed between islands requires independent parameters for Sicily and Malta. Fig 6 illustrates and summarises these posterior estimates. Missingness represents the probability of the next fossil being a new species, which is substantially higher for Malta in all periods. Most notably, this probability is slightly over 1% for the Early Palaeolithic. A quantitative comparison with Sicily's missingness indicates this potential for the discovery of a new species is several times greater on Malta than Sicily, with mean missingness of Malta divided my mean missingness of Sicily = 30.6, 2.4 and 6.4 for EMP, LMP and LP respectively. Overall (equally weighting all three periods) this is 13.16 times greater probability of discovering a new species on Malta.

These missingness probabilities can be used in the Geometric distribution to predict how many fossil draws should be expected before there is a better than even chance of discovering a new species. In the case of Malta EMP, the mode estimate is just 16 new draws (50% CI = 13 to 88), suggesting further excavation work on Malta is highly likely to yield important new discoveries.

Our sensitivity test repeats all these independent hypothesis tests using our alternative prior beliefs provided by author EYH (Fig SI3), and result in remarkably similar estimates despite the magnitude of these priors differing by an order of magnitude. This suggests that although this Bayesian framework has incorporated several relevant prior beliefs, ultimately these results are dominated by the data. Indeed, the alternative priors provide slightly higher estimates of missingness on Malta (1.127%, 0.098% and 0.098% for EMP, LMP and LP respectively), suggesting our results may even be slightly too conservative.

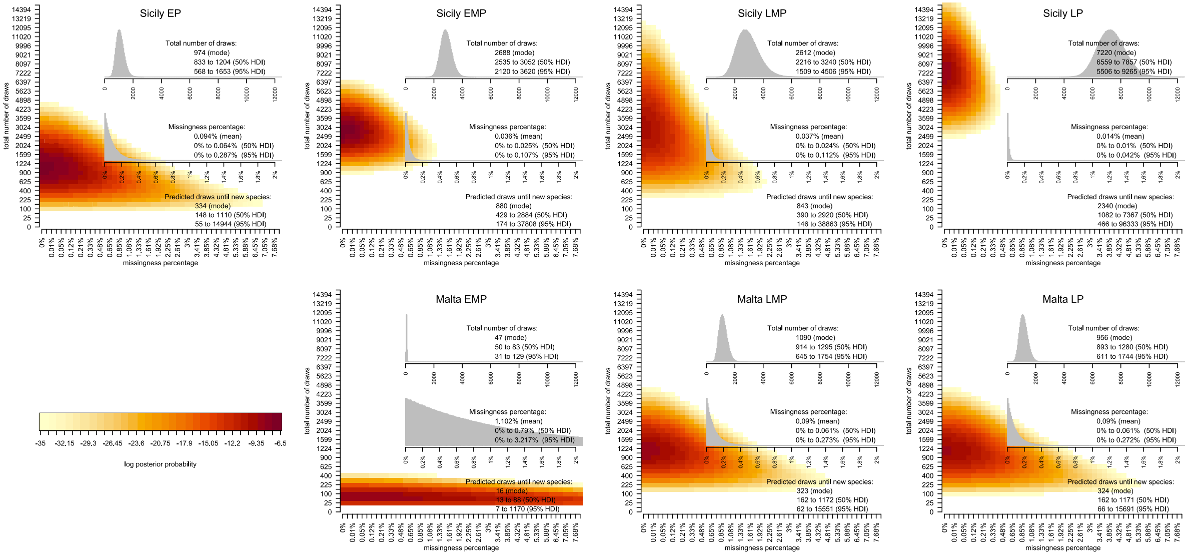

Fig 6: Posterior parameter estimates under the alternative hypothesis that the faunal composition of Sicily and Malta differed.

Discussion

Acknowledgements

The authors acknowledge the use of the UCL Myriad High-Performance Computing Facility (Myriad@UCL), and associated support services, in the completion of this work.

Conflict of interest statement

The authors have no conflicts of interest to declare.

Author contributions

AAEvdG and EYH collated the data, generated the taxonomic framework and provided prior beliefs; ES, AAEvdG and AT led the writing of the manuscript; AT conceived and implemented the methodology. All authors contributed critically to the drafts and gave final approval for publication.

References

Adams, A.L. (1870) *Notes of a Naturalist in the Nile Valley and Malta: A Narrative of Exploration and Research in Connection with the Natural History, Geology, and Archaeology of the Lower Nile and Maltese Islands*. Edmonston and Douglas.

Aguirre, E. (1969) Revisión de Elephantidae por su morfología y morfometría dentaria (tercera parte). *Estudios Geológicos***,** 3-4.

Ambrosetti, P. (1968) *The Pleistocene dwarf elephants of Spinagallo (Siracusa, south-eastern Sicily)*. Geologica romana.

Antonioli, F., Ferranti, L., Stocchi, P., Deiana, G., Presti, V.L., Furlani, S., Marino, C., Orru, P., Scicchitano, G. & Trainito, E. (2018) Morphometry and elevation of the last interglacial tidal notches in tectonically stable coasts of the Mediterranean Sea. *Earth-Science Reviews,* **185,** 600-623.

Antonioli, F., Furlani, S., Montagna, P. & Stocchi, P. (2021) The Use of Submerged Speleothems for Sea Level Studies in the Mediterranean Sea: A New Perspective Using Glacial Isostatic Adjustment (GIA). *Geosciences,* **11,** 77.

Anzidei, M., Lambeck, K., Antonioli, F., Furlani, S., Mastronuzzi, G., Serpelloni, E. & Vannucci, G. (2014) Coastal structure, sea-level changes and vertical motion of the land in the Mediterranean. *Geological Society, London, Special Publications,* **388,** 453-479.

Benjamin, J., Rovere, A., Fontana, A., Furlani, S., Vacchi, M., Inglis, R.H., Galili, E., Antonioli, F., Sivan, D. & Miko, S. (2017) Late Quaternary sea-level changes and early human societies in the central and eastern Mediterranean Basin: An interdisciplinary review. *Quaternary International,* **449,** 29-57.

Bintanja, R., Van De Wal, R.S. & Oerlemans, J. (2005) Modelled atmospheric temperatures and global sea levels over the past million years. *Nature,* **437,** 125-128.

Bishop, W. & Debono, G. (1996) The hydrocarbon geology of southern offshore Malta and surrounding regions. *Journal of Petroleum Geology,* **19,** 129-160.

Bonfiglio, L., Mangano, G., Marra, A.C., Masini, F., Pavia, M. & Petruso, D. (2002) Pleistocene calabrian and sicilian bioprovinces. *Geobios,* **35,** 29-39.

Bonfiglio, L., Marra, A.C. & Masini, F. (2000) The contribution of Quaternary vertebrates to palaeoenvironmental and palaeoclimatological reconstructions in Sicily. *Geological Society, London, Special Publications,* **181,** 171-184.

Busk, G. (1868) Description of the Remains of three extinct Species of Elephant, collected by Capt. Spratt, CB, RN, in the Ossiferous Cavern of Zebbug, in the Island of Malta. *The Transactions of the Zoological Society of London,* **6,** 227-306.

Catalano, S., De Guidi, G., Romagnoli, G., Torrisi, S., Tortorici, G. & Tortorici, L. (2008) The migration of plate boundaries in SE Sicily: Influence on the large-scale kinematic model of the African promontory in southern Italy. *Tectonophysics,* **449,** 41-62.

Ferretti, M. (2008) The dwarf elephant Palaeoloxodon mnaidriensis from Puntali Cave, Carini (Sicily; late Middle Pleistocene): Anatomy, systematics and phylogenetic relationships. *Quaternary International,* **182,** 90-108.

Foglini, F., Prampolini, M., Micallef, A., Angeletti, L., Vandelli, V., Deidun, A., Soldati, M. & Taviani, M. (2016) Late Quaternary coastal landscape morphology and evolution of the Maltese Islands (Mediterranean Sea) reconstructed from high-resolution seafloor data. *Geological Society, London, Special Publications,* **411,** 77-95.

Furlani, S., Antonioli, F., Biolchi, S., Gambin, T., Gauci, R., Presti, V.L., Anzidei, M., Devoto, S., Palombo, M. & Sulli, A. (2013) Holocene sea level change in Malta. *Quaternary International,* **288,** 146-157.

Furlani, S., Piacentini, D., Troiani, F., Biolchi, S., Roccheggiani, M., Tamburini, A., Tirincanti, E., Vaccher, V., Antonioli, F. & Devoto, S. (2018) Tidal notches (Tn) along the western Adriatic coast as markers of coastal stability during late Holocene. *Geogr. Fis. Din. Quat,* **41,** 33-46.

Galea, P. (2007) Seismic history of the Maltese islands and considerations on seismic risk. *Annals of geophysics*.

Galea, P. (2019) Central Mediterranean tectonics—a key player in the geomorphology of the Maltese Islands. *Landscapes and Landforms of the Maltese Islands*, pp. 19-30. Springer.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* **57,** 97-109.

Herridge, V.L. (2010) Dwarf elephants on Mediterranean islands: a natural experiment in parallel evolution. UCL (University College London).

Kotsakis, T. (1978) Sulle mammalofaune quaternarie siciliane.

Lambeck, K., Antonioli, F., Anzidei, M., Ferranti, L., Leoni, G., Scicchitano, G. & Silenzi, S. (2011) Sea level change along the Italian coast during the Holocene and projections for the future. *Quaternary International,* **232,** 250-257.

Lambeck, K. & Purcell, A. (2005) Sea-level change in the Mediterranean Sea since the LGM: model predictions for tectonically stable areas. *Quaternary Science Reviews,* **24,** 1969-1988.

Link, W.A. & Eaton, M.J. (2012) On thinning of chains in MCMC. *Methods in ecology and evolution,* **3,** 112-115.

Masini, F., Petruso, D., Bonfiglio, L. & Mangano, G. (2008) Origination and extinction patterns of mammals in three central Western Mediterranean islands from the Late Miocene to Quaternary. *Quaternary International,* **182,** 63-79.

McLeod, A.I. & Quenneville, B. (2001) Mean likelihood estimators. *Statistics and Computing,* **11,** 57-65.

Micallef, A., Berndt, C. & Debono, G. (2011) Fluid flow systems of the Malta Plateau, central Mediterranean Sea. *Marine Geology,* **284,** 74-85.

Micallef, A., Foglini, F., Le Bas, T., Angeletti, L., Maselli, V., Pasuto, A. & Taviani, M. (2013) The submerged paleolandscape of the Maltese Islands: Morphology, evolution and relation to Quaternary environmental change. *Marine Geology,* **335,** 129-147.

Oldow, J., Ferranti, L., Lewis, D., Campbell, J., d'Argenio, B., Catalano, R., Pappone, G., Carmignani, L., Conti, P. & Aiken, C. (2002) Active fragmentation of Adria, the north African promontory, central Mediterranean orogen. *Geology,* **30,** 779-782.

Pedley, M. & Clarke, M.H. (2002) *Limestone isles in a crystal sea: the geology of the Maltese Islands*. Publishers Enterprises Group.

Reuther, C.-D. & Eisbacher, G. (1985) Pantelleria Rift—crustal extension in a convergent intraplate setting. *Geologische Rundschau,* **74,** 585-597.

Reyes Suarez, N.C., Cook, M.S., Gačić, M., Paduan, J.D., Drago, A. & Cardin, V. (2019) Sea Surface Circulation Structures in the Malta-Sicily Channel from Remote Sensing Data. *Water,* **11,** 1589.

Seguenza, L. (1902) I vertebrati fossili della provincia di Messina: Parte II. Mammiferi e geologia del piano Pontico. *Bollettino della Società Geologica Italiana,* **21,** 115-172.

Serpelloni, E., Vannucci, G., Pondrelli, S., Argnani, A., Casula, G., Anzidei, M., Baldi, P. & Gasperini, P. (2007) Kinematics of the Western Africa-Eurasia plate boundary from focal mechanisms and GPS data. *Geophysical Journal International,* **169,** 1180-1200.

Siddall, M., Rohling, E.J., Almogi-Labin, A., Hemleben, C., Meischner, D., Schmelzer, I. & Smeed, D. (2003) Sea-level fluctuations during the last glacial cycle. *Nature,* **423,** 853-858.

Team, R.C. (2013) R: A language and environment for statistical computing.

Van Der Geer, A. (2021) *EVOLUTION OF ISLAND MAMMALS*. WILEY-BLACKWELL, [S.l.].

Vogiatzakis, I.N., Pungetti, G. & Mannion, A.M. (2008) *Mediterranean island landscapes: natural and cultural approaches*. Springer Science & Business Media.

Zammit-Maempel, G. & De Bruijn, H. (1982) The Plio/Pleistocene Gliridae from the Mediterranean Islands reconsidered. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series B,* **85,** 113-128.

Zecchin, M., Praeg, D., Ceramicola, S. & Muto, F. (2015) Onshore to offshore correlation of regional unconformities in the Plio-Pleistocene sedimentary successions of the Calabrian Arc (central Mediterranean). *Earth-Science Reviews,* **142,** 60-78.

Supporting Information

Algorithm to calculate likelihood of presence-only data

The function *method.1()* provides an intuitive compositional approach to calculating the probability of presence-only data, where *draws* = the total number of samples, and *probs* is a vector of *S* + 1 probabilities defining the relative frequency of all *S* observed species and the last value of *probs* is the proposed missingness. In contrast, the function *method.2()* utilizes the inclusion – exclusion principle to provide the same result with massively less computational cost. However, *method.2()* can be vulnerable to catastrophic cancellation from floating point limits, therefore our complete algorithm also incorporates minor modifications to handle the final summation more precisely (see functions.R).

```
method.1  <- function(draws, probs){
        require(arrangements)
        x <- cbind(compositions( draws, length(probs) - 1), 0)
        res <- sum(apply(x, 1, dmultinom, prob=probs))
return(res)}

method.2 <- function(draws, probs){
        require(arrangements)
        N <- length(probs)
        expr <- list()
        coef <- rep(c(1,-1),length.out=N)
        if(N==1){
                k1 <- 1 - probs
                res <- sum(k1^draws)
                }
        if(N>1){
                for(n in 1:N){
                        k1 <- 1 - rowSums(cbind(v=combinations(probs[1:(N-1)], k=n-1), probs[N]))
                        expr[[n]] <- coef[n] * (k1^draws)
                        }
                res <- sum(unlist(expr))
                }
return(res)}
```

Null parameter estimates

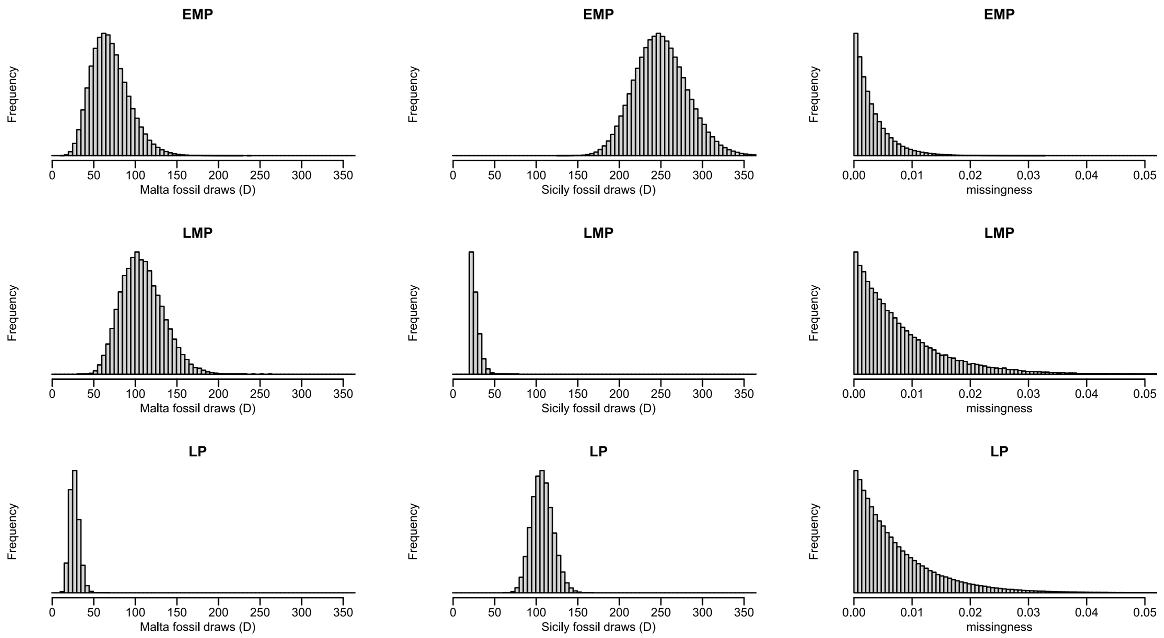

Fig SI1: Marginal posterior parameter estimates under the null hypothesis.

Null hypothesis sensitivity test

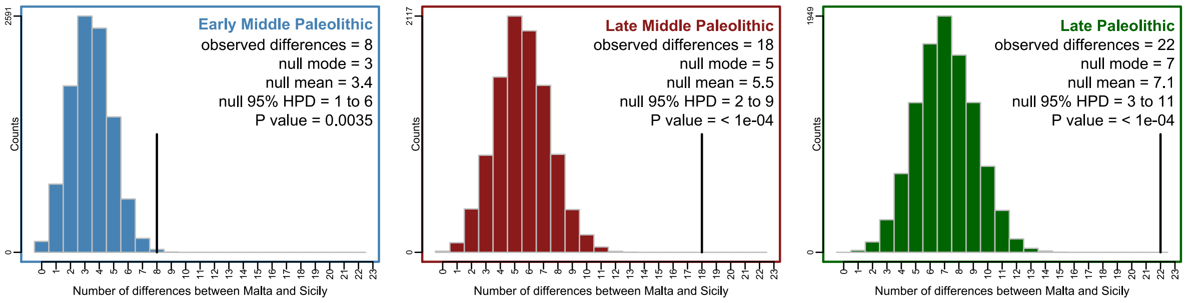

Fig SI2: Sensitivity test using our second independent set of prior beliefs of species frequencies from author EYH. Null distributions of the number of species differences between Sicily and Malta, for each time period tested independently. In each case the observed number of species differences (vertical black line) is significantly greater than can be expected under the null.

Independent hypothesis sensitivity test

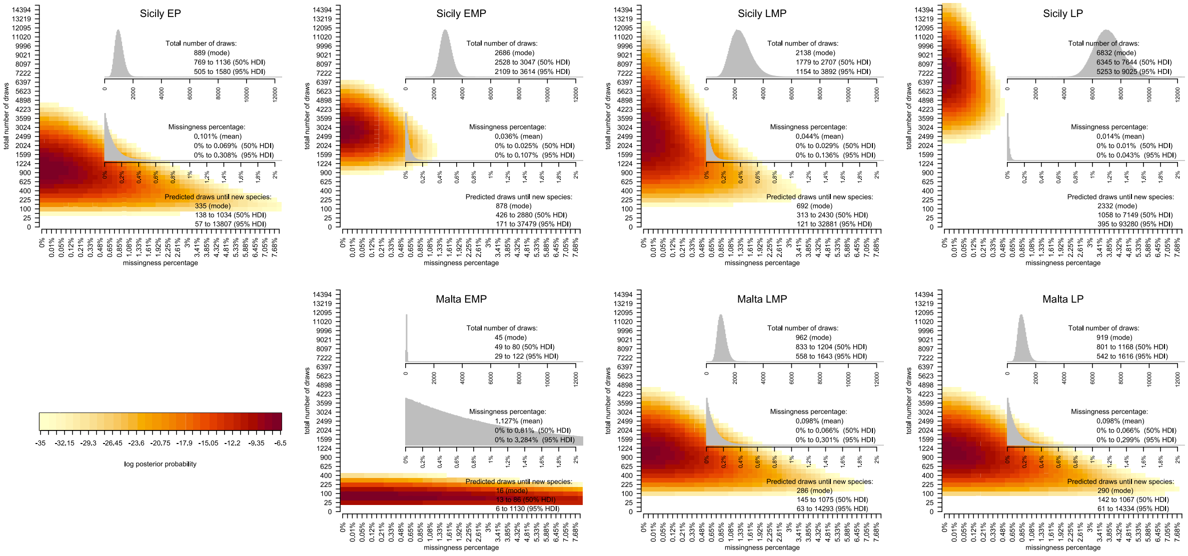

Fig SI3: Sensitivity test using our second independent set of prior beliefs of species frequencies from author EYH.