Article / Discoveries

Evolutionary history of the Nesophontidae, the last unplaced Recent mammal family.

Selina Brace[1†], Jessica A. Thomas[2†], Love Dalén[3], Joachim Burger[4], Ross D. E. MacPhee[5], Ian Barnes[1*] & Samuel T. Turvey[6].

[1]Department of Earth Sciences, Natural History Museum, London SW7 5BD, UK.

[2]University of York, York, YO10 5DD, UK.

[3]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, SE-104 0, Sweden.

[4]Institute of Anthropology, Johannes Gutenberg-University, Anselm-Franz-von-Bentzel-Weg 7, D-55128 Mainz, Germany.

[5]Department of Mammalogy, American Museum of Natural History, New York, NY, 10024, USA.

[6]Institute of Zoology, Zoological Society of London, London NW1 4RY, UK.

[†]These authors contributed equally to this work.

[*]**Corresponding author:** Email i.barnes@nhm.ac.uk

Abstract:

The mammalian evolutionary tree has lost several major clades through recent human-caused extinctions. This process of historical biodiversity loss has particularly affected tropical island regions such as the Caribbean, an area of great evolutionary diversification but poor molecular preservation. The most enigmatic of the recently extinct endemic Caribbean mammals are the Nesophontidae, a family of morphologically plesiomorphic lipotyphlan insectivores with no consensus on their evolutionary affinities, and which constitute the only major Recent mammal clade to lack any molecular information on their phylogenetic placement. Here we use a palaeogenomic approach to place Nesophontidae within the phylogeny of recent Lipotyphla. We recovered the near-complete mitochondrial genome and sequences for 17 nuclear genes from a ~750-year-old Hispaniolan *Nesophontes* specimen, and identify a divergence from their closest living relatives, the Solenodontidae, more than 40 million years ago. Nesophontidae is thus an older distinct lineage than many extant mammalian orders, highlighting not only the role of island systems as "museums" of diversity that preserve ancient lineages, but also the major human-caused loss of evolutionary history.

Key words: Ancient DNA, Cenozoic, insectivore, systematics, tropics.

Introduction

Reconstructions of the evolutionary history of the radiation of modern mammals have been constrained by recent phylogenetically non-random extinctions, which have driven the loss of several major clades. Insular terrestrial vertebrate faunas in particular have been disproportionately affected in recent millennia by extensive human-caused extinctions (Turvey and Fritz 2011). This is problematic for understanding global patterns of mammal evolution, as although island faunas are typically narrow in taxonomic diversity, they may preserve "relict" clades that have mostly or completely disappeared elsewhere (e.g. New Zealand's tuataras, the only surviving rhynchocephalians), as originally emphasized by Wallace (1880). Substantial recent advances in understanding historical relationships and ages of island clades have been enabled by the increasing availability of molecular datasets, which have proved highly effective in identifying geographic regions of elevated evolutionary diversification and ancient endemism (Rosauer and Jetz 2014). However, past human-caused losses have severely restricted our capacity to collect high-quality genomic information for many island faunas, and limited our understanding of the evolutionary history of these taxa.

Unusually for an oceanic-type island system, the insular Caribbean witnessed multiple major endemic radiations of non-volant land mammals (Woods and Sergile 2001). These include two groups of lipotyphlan "insectivores", traditionally assigned to separate placental families due

to their morphological and assumed phylogenetic distinctiveness (Woods and Sergile 2001). Of these, the only survivors are the two relatively large-bodied (~1kg), venomous, and highly threatened solenodons (Solenodontidae: *Atopogale* and *Solenodon*). *Nesophontes* (Nesophontidae), a much smaller (~10-125g) and morphologically more plesiomorphic genus, was represented by eight nominal species formerly occurring in Cuba, Hispaniola, Puerto Rico and the Cayman Islands (Turvey and Fritz 2011). Radiometric studies indicate that *Nesophontes* became extinct shortly after European arrival ~500 years ago, probably due to introduction and subsequent ecological release of black rats (*Rattus rattus*) (MacPhee et al. 1999).

The pre-Pleistocene Caribbean lipotyphlan fossil record is extremely limited. However, molecular analyses have revealed that solenodons are not only the sister taxon to all other living lipotyphlans but also among the oldest surviving lineages within Placentalia (Roca et al. 2004). They may have reached landmasses in the central Caribbean during the late Mesozoic (Roca et al. 2004; Iturralde-Vinent 2006), which would make Solenodontidae by far the oldest family-level taxon of extant insular placentals. In sharp contrast, the phylogenetic position and paleobiogeographic history of *Nesophontes* has remained persistently obscure. Anthony (1916) originally considered *Nesophontes* to have "no evident close relationships with any known genus. Its characters are so strikingly different and of a nature so important that full justice may be done them only by erection of a new family". Its plesiomorphic craniodental features are consistent with it also representing a phylogenetically ancient lineage, but cladistic morphological analyses have provided little insight into its evolutionary relationships (see Supplementary Information). The dilambdodont dentition of *Nesophontes* superficially resembles that of soricids and talpids, while the zalambdodont dentition of solenodons was formerly thought to indicate close relationship with non-lipotyphlan tenrecs and chrysochlorids (see Supplementary Information). Whereas molecular analyses have now refuted the possibility of a close relationship between solenodons and insectivorous afrothere groups (Roca et al. 2004), no recent morphological phylogenetic analyses have recovered *Nesophontes* and solenodons as a monophyletic group (Asher et al. 2002).

For most of the twentieth century, relationships between Caribbean lipotyphlans were regarded as unresolved and perhaps unsolvable. The advent of palaeogenomic approaches has enabled reconstruction of phylogenetic relationships for several now-extinct Late Quaternary temperate and high-latitude mammals (Rohland et al. 2010; Heintzman et al. 2015). However, ancient DNA analysis of extinct tropical taxa remains highly challenging due to generally unfavourable environmental conditions for preservation of ancient biomolecules. Higher temperatures, such as those found in the Caribbean, lead to accelerated rates of

deamination and strand scission of DNA (Smith et al. 2003). As DNA fragment length shortens it becomes increasingly challenging to retrieve sufficient data, and to bioinformatically align and analyse those data. The difficulties of recovering DNA from *Nesophontes* are further compounded by small specimen size (i.e. small amounts of bone material), and because it is the only Recent mammalian family for which there is no closely-related reference genome.

Here we report success in extracting and characterizing endogenous DNA from a skull morphologically identified as *Nesophontes paramicrus*, collected in 2007 from a probable owl-accumulated deposit in the Dominican Republic that has been radiocarbon dated to ~750 years before present (Table 1).

Results

DNA quality and assembly

A total of 337,313,945 processed reads were obtained from two lanes of Illumina HiSeq 2000 sequencing of the *Nesophontes* specimen. In the absence of a reference genome, it is difficult to assess the proportion of endogenous DNA in the sample. We found 18,426,595 reads that had a top BLAST hit to Mammalia, giving a maximum endogenous content of approximately 5.6%; the remaining reads showed highest sequence similarity to bacteria and fungi. We additionally examined endogenous content by mapping back to the *Sorex* and *Erinaceus* genomes; a total of 13,167,917 and 11,884,165 reads mapped back respectively, corresponding to an endogenous content (underestimated, due to sequence similarity) of 3.9% and 3.5%.

Using *clc_denovo_assembler* (available in CLC Assembly Cell v.4.2), we recovered 5,832,418 contigs of between 70 and 37,489 base pairs (bp) in length (N50=154 bp). Unfortunately simply mapping reads to a reference genome would be inappropriate given the level of taxonomic difference shown by *Nesophontes* to other mammals, and could lead to potential artifacts. Post *de novo* contig assembly, we assembled *Nesophontes* mitochondrial and nuclear genes by alignment of the contigs to existing lipotyphlan datasets (see Supplementary Table 1). For the mitochondrial genome, we obtained 34 *de novo* contigs, which after alignment overlapped to leave 15 gaps ranging from 10-182 bp. Sanger sequencing resolved seven of the 15 gaps; unresolved gaps (all <50 bp) were coded as missing data in our alignment. Nuclear sequence coverage was comparatively low, and in many places only a single read in depth; to avoid incorrectly calling bases, only sequence regions covered by contigs (read depth >3) were used in phylogenetic analysis. Several features of the sequence data support its antiquity: median (62 bp) and mean (68.3 bp) lengths of merged read pairs were very short (Supplementary Fig. 1), and fragment misincorporation plots from MapDamage indicate

standard ancient DNA damage of C->T mutations towards the ends of sequences (Supplementary Fig. 2).

Phylogenetic Analyses

To place these results in a phylogenetic context, we assembled and analyzed relevant sequence information for 25 extant lipotyphlan species, representing all extant families and including the complete mitochondrial genome and additional nuclear genes for the Hispaniolan solenodon (*Solenodon paradoxus*) specifically generated for this study (see Methods and Supplementary Information). This is the most extensive dataset for studying lipotyphlan genomics assembled to date. Our maximum-likelihood phylogenetic analysis conclusively identifies *Nesophontes* as the sister taxon of Solenodontidae*,* which together comprise a monophyletic Caribbean clade (Fig. 1 and Supplementary Fig. 3). No previous molecular analysis has covered as many extant lipotyphlan subfamilies simultaneously. We find strong support for monophyly of the three accepted lipotyphlan families (Talpidae, Erinaceidae and Soricidae), with Talpidae recovered as sister to Erinaceidae+Soricidae. Alternative arrangements, either Talpidae+Soricidae or Talpidae+Erinaceidae, are strongly rejected (p<0.01) by the approximately unbiased (AU) test (Shimodaira and Hasegawa 2001). These phylogenetic relationships are similar to those previously found in most other more restricted analyses (Roca et al. 2004; Cabria et al. 2006; Dubey et al. 2007; Meredith et al. 2011), although Meredith et al. (2011) found conflicting placements when incorporating single representatives for each family, with DNA-based analyses recovering an Erinaceidae+Soricidae clade whilst an amino acid-based approach recognized a Talpidae+Soricidae clade. With our larger dataset we were able to fully resolve almost all within-family relationships, with the exception of the nodes grouping *Sorex* and *Blarinella* and, as in other studies (Cabria et al. 2006), star-nosed moles (*Condylura*), desmans (represented by *Galemys*), and the shrew-mole *Urotrichus*. Our dated phylogeny (see Methods and Fig. 2) exhibits a topology that is almost identical to that of the maximum-likelihood tree, with all nodes fully resolved at 100% posterior probability, except for the positions of *Condylura* (99) and *Blarinella* (55). The split between *Nesophontes* and *Solenodon* is estimated to have occurred 57.3 Ma (95% highest probability density=43.6–68.6 Ma), whereas the *Nesophontes-Solenodon* clade is estimated to have diverged from other lipotyphlans 72.8 Ma (95% highest probability density=67.5–78.9 Ma).

Discussion

Although the overall morphological (and now genomic) distinctiveness of solenodons and *Nesophontes* justifies their continued placement in different families, the deep division of this

clade from all other extant lipotyphlans has not been accurately identified before and warrants proper taxonomic recognition. Van Valen (1967) created the suborder Erinaceota for non-zalambdodont placental "insectivores" (including Erinaceidae, Soricidae, Talpidae, their extinct relatives, and Nesophontidae), but placement of Solenodontidae was left unresolved. We take this step now, by placing the monophyletic Nesophontidae+Solenodontidae in the new taxon **Solenodonota**, coordinate in rank with suborder Erinaceota, its sister taxon.

Recognition of Solenodonota as a single deep-branching lipotyphlan clade contrasts with the evolutionary history of some other Quaternary Caribbean vertebrate groups (e.g. leptodactylid frogs, mabuyid skinks, megalonychid sloths, oryzomyine rice rats), which have been shown to comprise multiple distantly related lineages representing separate colonization events (Woods and Sergile 2001; Hedges et al. 2012; Brace et al. 2015), and raises several paleobiogeographical questions. Where and when did Solenodonota diverge from its closest relative? Divergence within Gondwanan landmasses can be excluded: South America lacked lipotyphlans until the Great American Biotic Interchange, when soricids (*Cryptotis*) invaded from the north (Woodburne 2010), while neither solenodons nor *Nesophontes* display molecular affinities with afrotheres despite early morphology-based views to the contrary (see Supplementary Information). More plausibly, solenodonotans may be closely related to a set of early-middle Cenozoic North American zalambdodont lipotyphlans (Asher et al. 2002).

Although it is parsimonious to conclude that solenodonotans evolved in North America, there was probably only one solenodonotan colonization of the Caribbean, with later divergences occurring on landmasses comprising the Cretaceous proto-Antillean island arc. Our estimated age for Solenodonota is consistent with previous continent-island vicariance hypotheses for late Mesozoic divergence of both Caribbean lipotyphlans and other ancient Caribbean endemics, which assume colonization from southernmost North America as the proto-Antilles were tectonically displaced from the Pacific into the Caribbean basin (Iturralde-Vinent 2006; Woodburne 2010). Our estimate for *Nesophontes*-solenodon divergence (95% highest probability density=43.6–68.6 Ma) is a period of resurgent regional volcanic activity associated with the temporary formation of subaerial landmasses (Iturralde-Vinent and MacPhee 1999), and thus compatible with allopatric fragmentation and isolation of terrestrial vertebrate populations. Conversely, as recent representatives of Nesophontidae and Solenodontidae differ markedly in body size, divergence between these lineages may have resulted from sympatric adaptive radiation associated with niche separation (e.g. adaptation towards different prey bases).

High levels of recent extinction have occurred in almost all island systems. The insular Caribbean has lost ~100 endemic mammal species or island populations through anthropogenic activities, representing the world's highest levels of Holocene and historical-era mammal extinction (MacPhee 2009; Turvey and Fritz 2011). The demonstrated antiquity of nesophontids, a family-level lineage as old or older than estimates for many extant mammal orders (dos Reis et al. 2012; Beck and Lee 2014) rather than a recent divergence from solenodons, highlights the role of island systems as "museums" of diversity that preserve ancient lineages as well as promoting adaptive radiation in younger clades. The complete loss of Nesophontidae in recent times, together with the disappearance of most or all representatives of other ancient Caribbean mammal groups such as sloths, primates and capromyid rodents (Woods and Sergile 2001; Fabre et al. 2014), unfortunately also highlights the magnitude of ongoing human-induced loss of mammalian evolutionary history.

Methods

Collection locality and age of *Nesophontes* and *Solenodon* samples

The sample of *Nesophontes paramicrus* used in this study was collected by S.T.T. on 5 May 2007 from Cueva de Bosque Humido, Los Haitises National Park, Hato Mayor Province, Dominican Republic (19.077389°N, 69.477389°W, 44 m above sea level). This specimen formed part of a rich surficial accumulation of small vertebrate skeletal elements which probably represent prey remains from an ancient barn owl roost, and which also contained abundant other subfossil material of *Nesophontes paramicrus, N. hypomicrus* and *N. zamicrus*, as well as two other now-extinct native non-volant small mammals (*Brotomys voratus* and *Isolobodon portoricensis*), bats, birds and reptiles, and non-native *Rattus rattus*. AMS dates for specimens of native Hispaniolan taxa from this vertebrate assemblage (including a *Nesophontes* specimen not used for ancient DNA analysis) indicate a late Holocene, immediately pre-European age, whereas *Rattus rattus* material is modern and intrusive (Table 1).

The *Solenodon paradoxus* sample, comprising a small tissue/hair sample from a live-caught individual that was subsequently re-released, was collected in the field by S.T.T. in April 2009 from Mencia, Sierra de Bahoruco, Dominican Republic (Secretaría de Estado de Medio Ambiente y Recursos Naturales, Subsecretaría de Estado de Áreas Protegidas y Biodiversidad, project authorisation no. 00826, export authorisation no. 0120; United Kingdom Department for Environment, Food and Rural Affairs, import licence no. PATH/110/2009/2). Due to prolonged storage under non-refrigerated conditions, the sample

was treated as a degraded tissue sample and processed using ancient DNA protocols (see Methods).

Selection of taxa and molecular markers

To examine the position of *Nesophontes* within the Lipotyphla, we constructed an alignment of 25 lipotyphlan and 10 outgroup taxa for which mitochondrial and nuclear sequences were available on Genbank (http://ncbi.nlm.nih.gov), including additional nuclear genes and the complete mitochondrial genome for the Hispaniolan solenodon generated for this study. In addition to Hispaniolan solenodon, our total lipotyphlan dataset comprised members of the Erinaceidae (three hedgehogs (*Erinaceus, Hemiechinus* and *Atelerix*), moonrat (*Echinosorex*), and three gymnures (*Neotetracus, Hylomys* and *Podogymnura*)), Soricidae (two white-toothed shrews (*Suncus* and *Crocidura*) and six red-toothed shrews (*Sorex, Nectogale, Episoriculus, Blarinella, Anourosorex* and *Neomys*)), and Talpidae (six moles (*Talpa, Euroscaptor, Mogera, Condylura, Scalopus* and *Scapanulus*), a desman (*Galemys*) and two shrew-moles (*Urotrichus* and *Uropsilus*)). Several laurasiatherian and an afrotherian outgroup taxa were selected: *Canis, Felis, Manis, Tursiops, Bos, Sus, Equus, Myotis, Pteropus* and *Loxodonta.* The whole mitochondrial genome, except for the hypervariable control region (consisting of two ribosomal genes: 12S and 16S, 13 protein coding sequences: COI, COII, COIII, CytB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6, ATP6 and ATP8, and 22 tRNA sequences, in total approximately 14,450 bp), and 21 orthologous nuclear genes: ADORA3, ADRA2B, ADRB2, APOB, APP, ATP7A, BCHE, BDNF, BMI1, BMP4, BRCA1, CREM, EDG1, GHR, PLCB4, RAG1, RAG2, RHO, TTN, TYR and VWF, in total approximately 20,170 bp) were targeted for phylogenetic analysis (Supplementary Table 1). Gene selection was based on availability of genetic data from extant taxa and represents the standard set of orthologous nuclear genes for phylogenetic analyses. Genes were aligned manually. We used a supermatrix approach to sequence alignment, meaning that some sequences were missing for some taxa. In order to maximize taxonomic coverage, for a few taxa, chimeric taxon units were made up of two or three species from the same genus (Supplementary Table 1). Furthermore, we obtained the whole mitochondrial genome for all except six taxa (*Atelerix, Condylura, Euroscaptor*, *Podogymnura, Scalopus* and *Solenodon*), three of which we recovered through analysis of NGS data (see below).

Next Generation sequencing

DNA extractions of *Nesophontes* (bone sample from Dominican Republic) and *Solenodon* (degraded tissue sample from Dominican Republic) were carried out in a dedicated ancient DNA laboratory at Royal Holloway, University of London, and followed previously described

protocols (Brace et al. 2012). *Nesophontes* and *Solenodon* NGS libraries were constructed in a separate dedicated ancient DNA laboratory at Johannes Gutenberg University, Mainz, using a modified protocol (Meyer and Kircher 2010). Library modifications: the initial DNA fragmentation step was not required; all clean-up steps used MinElute PCR purification kits. Blunt-end repair step: Buffer Tango and ATP were replaced with 0.1mg/mL BSA and 1 x T4DNA ligase buffer. The proceeding clean-up step was replaced by an inactivation step, heating to 75˚C for 10 minutes. Adapter ligation step: 0.5mM ATP replaced the T4 DNA Ligase buffer. The index PCR step used AmpliTaq Gold DNA polymerase and the addition of 0.4mg/mL BSA. The index PCR was set for 20 cycles with three PCR reactions conducted per library. The *Nesophontes* and *Solenodon* libraries were sequenced on an Illumina HiSeq platform (Mainz, Germany); *Nesophontes* on a single lane, single-end read sequencing run and *Solenodon* on a single lane, paired-end read sequencing run. The *Nesophontes* library was additionally sequenced on a single lane, paired-end read sequencing run on an Illumina HiSeq platform (Stockholm, Sweden).

Sanger sequencing

Post NGS sequencing, assembly of the *Nesophontes* mitochondrial genome identified 15 gaps ranging from 10-182 bp long. To resolve gaps that were over 20 bp long, 9 primer pairs (Supplementary Table 2) were designed to amplify the missing fragments. PCR reactions using the remaining *Nesophontes* DNA extract were set up in a dedicated ancient DNA laboratory at the Natural History Museum, London. PCR reactions and amplicon purification were performed as previously described (Brace et al. 2012). Sanger sequencing reactions were performed at the Natural History Museum.

Bioinformatic analysis

NGS data generated for this analysis consisted of three Illumina runs: one 50 bp single-end and one paired-end (2x100 bp) run for *Nesophontes*, and one 100 bp paired-end (2x100 bp) run for *Solenodon*. In addition, we downloaded several further Illumina datasets from the short-read archive (SRA) for two previously sequenced lipotyphlan taxa, in order to obtain mitochondrial and nuclear sequences that were not available on Genbank. These were for *Condylura cristata* (comprising six single-end datasets: SRR869610_TG1.fastq.gz, SRR869613_DRG1.fastq.gz (76 bp), SRR869614_DRG2.fastq.gz, SRR869611_TG2.fastq.gz, SRR869612_TG3.fastq.gz (36 bp), SRR869615_DRG3.fastq.gz (100 bp)), and *Atelerix albiventris* (consisting of four paired-end 100 bp datasets: SRR765874_1.fastq.gz,

SRR765874_2.fastq.gz, SRR765875_1.fastq.gz, SRR765875_2.fastq.gz, SRR765876_1.fastq.gz, SRR765876_2.fastq.gz, SRR765877_1.fastq.gz, SRR765877_2.fastq.gz).

For tagged sequences (the *Nesophontes* single-end 50 bp run), *fastx_barcode_splitter* (from the Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) was used to pull out sequences with the correct barcode tag. Single-end reads (*Nesophontes* and *Condylura*) were first quality trimmed (q=10), and then Illumina adapter sequences were trimmed with *cut-adapt* (Martin 2011). Paired-end reads (*Nesophontes, Solenodon* and *Atelerix* data) were also quality trimmed (q=10) with *cut-adapt*, but sequences were then simultaneously adapter trimmed and the paired reads (with an overlap of 10 bp) joined together with *Seq-Prep* (https://github.com/jstjohn/SeqPrep). All reads shorter than 30 bp as well as unmerged reads were discarded. Due to the absence of any close phylogenetic relative (required for the accurate genomic mapping of reads), processed reads were instead *de-novo* assembled into contigs using *clc_denovo_assembler* (available in CLC Assembly Cell v.4.2), with contigs shorter than 70 bp discarded.

In order to determine whether there were any endogenous mammalian DNA sequences suitable for phylogenetic analysis in our dataset, we used two approaches. Firstly, for *Nesophontes*, we assessed the processed reads in the dataset using BLAST, with the *max_target_seqs* set to 1. This enabled us to estimate the endogenous content of the sample. Secondly, for each taxon, a BLAST database of the contigs was generated using the *makeblastdb* command in BLAST. These databases could then be searched using lipotyphlan mitochondrial and nuclear phylogenetic query sequences of interest using *blastn*. For each sequenced taxa, the mitochondrial genome and each nuclear gene sequence were assembled from individual contigs, by alignment to this dataset (Table 2).

To assess the sequencing depth and coverage of reads (all taxa) making up each contig, reads were mapped back to sequences using Bowtie2 (Langmead and Salzberg 2012), with the mapping parameters set to *--very-sensitive-local* (corresponding to -D 20 -R 3 -N 0 -L 20 -i S,1,0.50). Duplicate reads were removed with the *rmdup* command in samtools (http://samtools.sourceforge.net). Alignments were visualised in Tablet (Milne et al. 2013). For *Nesophontes*, reads were additionally mapped back to *Condylura*, *Erinaceus, Solenodon* and *Sorex* sequences, to pull out as many reads as possible (Table 2).

DNA fragmentation patterns (*Nesophontes* merged read lengths) are shown in Supplementary Fig. 1, and DNA misincorporation patterns (aligned *Nesophontes* reads) were examined using mapDamage v2.0 (Jónsson et al. 2013; Supplementary Fig. 2).

Mitochondrial genome assembly (*Solenodon*, *Atelerix* and *Condylura*)

Solenodon, *Atelerix* and *Condylura* mitochondrial genomes were assembled from six, four and two overlapping contigs, respectively. *Atelerix* and *Condylura* data were retrieved from the SRA, and *Solenodon* NGS data were sequenced for this study. The whole mitochondrial genome was recovered for each, because the CR1 region (a section which can be hard to assemble, due to repetitive regions and a lack of sequence similarity between taxa) was recovered in the middle of a contig. Sequence depth was high for each sample, ranging between 75X for *Solenodon* to >20,000X for *Atelerix.*

Nuclear sequence assembly (*Solenodon* and *Atelerix*)

Supplementary Table 1 shows the contig information for each gene. For *Solenodon* and the previously sequenced *Atelerix* dataset, individual nuclear genes were comprised of predominantly single contigs, except where the coding sequence consisted of more than one exon. Average coverage varied dramatically between sequences for *Atelerix*, which ranged from 731X for GHR to 3.2X for ADORA3 (variation in coverage between different genes is likely to arise from the nature of the *Atelerix* SRA datasets, which comprise transcriptomic and targeted genomic sequence data). Average sequence coverage for *Solenodon* nuclear sequences varied between 10X and 16X. As for the mitochondrial genome, we assessed whether our data were likely to be 'authentic' or the result of contaminants following the four steps outlined below.

Assessment of contamination

In addition to assessing the sequence similarity of our contigs to lipotyphlan sequences during the BLAST search, we further examined whether the contig sequences could have arisen from contaminant rather than lipotyphlan DNA, in several ways: 1) comparing contigs to previously (Sanger) sequenced genes from the same species available on Genbank; 2) examining for evidence of pseudogenes in the form of frame-shift mutations or unexpected stop codons for protein-coding sequences; 3) assessing the GC content of contigs to exclude potential bacterial sequences; and 4) assessing for human contamination by mapping all reads in each dataset to the human mitochondrial genome and human nuclear genes in this study. In all cases we found no evidence that contigs were composed of contaminant sequence.

Phylogenetic analysis

Partition Finder (Lanfear et al. 2012) was used to select the best-fit partitioning scheme for the mitochondrial and nuclear sequences. This corresponded to 13 separate partitions for the

mitochondrial data and 15 for the nuclear sequences (Supplementary Table 3). Mitochondrial and nuclear sequences were concatenated, and a maximum likelihood (ML) phylogeny was estimated in RAxML v.7.0.4 (Stamatakis 2014). The general time-reversible (GTR) model of sequence evolution (Rodríguez et al. 1990) with across-site rate variation modelled to a gamma distribution (corresponding to the GTR+ Γ model in RAxML) was applied to each partition. Twenty separate ML analyses were performed (using the "–f d" command in RAxML), and the tree with the highest likelihood was chosen from this set. A separate bootstrap analysis with 100 replicates was also executed to examine the relative support for each clade. In order to address the possibility that post-mortem damage (i.e. C-T transitions) lead to an overestimate of substitution rates in our phylogeny, we used RY-recoding to estimate a ML tree using the transversional mutations only (all A and G bases were recoded as R, and C and T bases as Y). The resulting topology (not shown) was almost identical (differing in only one place within the Talpidae), with equivalent statistical support to the original tree, indicating that our topology is not derived from the effects of post-mortem damage. Tests of alternative arrangements were made by submitting the sitewise log-likelihood values obtained from RAxML to CONSEL, to calculate the p-values for each topology by using the AU test. We also estimated a phylogeny using only mitochondrial sequences, and recovered an identical topology to that obtained with the combined dataset (Supplementary Fig. 4).

Dated phylogenies were constructed in BEAST v.1.7 (Drummond et al. 2012). As a result of convergence problems in BEAST for large genomic datasets, the Hasegawa-Kishino-Yano (HKY) model of sequence evolution (Hasegawa et al. 1985) with the distribution of rates across sites (approximating a gamma distribution) was applied to each partition (rather than those estimated in PartitionFinder). Four chains were run under a relaxed clock, with the Yule model of speciation (the birth-death model was also run for comparison, and topologies were identical), for 100 million generations (sampled every 10,000 generations), with separate relaxed clock rates estimated for the mitochondrial and nuclear partitions. Tree parameters were linked across partitions. Prior distributions on the root and seven other nodes were applied based on previous interpretation of the mammalian fossil record (Benton et al. 2009; dos Reis et al. 2012). The clock rate priors were set as uninformative uniform distributions (upper=E100, lower=E-12). All other priors were left as the default values in BEAUti v.1.8.0. Details of all prior distributions for divergence times are found in Supplementary Table 4. Convergence and effective sampling were assessed using Tracer 1.7 (http://beast.bio.ed.ac.uk/Tracer), with all ESS values >100. A Maximum Clade Credibility (MCC) tree was constructed using TreeAnnotator (available with BEAST) from the trees sampled in the posterior distribution. In order to explore the possibility that placental ordinal divergences occurred deeper into the Cretaceous than supported by the fossil record, we

conducted additional analyses calibrated with previously published, molecular-clock derived node time estimates (Meredith et al. 2011). This analysis provides an identical topology (Supplementary Fig. 5), with a deeper point estimate of the divergence of *Nesophontes* and solenodons 68.2 Ma (95% highest probability density=47.55-81.14 Ma). This date range is compatible with that derived from fossil-only calibration; however, this tree has both lower posterior probability support for nodes and broader date ranges.

Acknowledgements

Funding was provided by the Royal Society (UF080320/130573, RG100902), the Natural Environment Research Council (NE/D009456/1, NE/J010480/1), SYNTHESYS2 (SYNthesis of SYStematic resources), made available by the European Community Research Infrastructure under FP7 (Synthesis of Systematic Resources, 226506-CP-CSA-Infra) and the British Ecological Society (771/899). Fieldwork in the Dominican Republic was assisted by R.P. Scofield, J. Brocca and H.M.R. Meredith. The authors would also like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure (NGI), Sweden, the Knut and Alice Wallenberg Foundation, and UPPMAX for providing assistance in massively parallel DNA sequencing and computational infrastructure.

References

Anthony HE. 1916. Preliminary diagnosis of an apparently new family of insectivores. *Bull Am Mus Nat Hist* 35:725–728.

Asher RJ, McKenna MC, Emry RJ, Tabrum AR, Kron DG. 2002. Morphology and relationships of *Apternodus* and other extinct, zalambdodont, placental mammals. *Bull Am Mus Nat Hist* 273:1–117.

Beck RMD, Lee MSY. 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. *Proc Roy Soc B* 281:20141278.

Benton MJ, Donoghue PCJ, Asher RJ. 2009. Calibrating and constraining molecular clocks. In: Hedges SB, Kumar S, editors. The Timetree of Life. Oxford: Oxford University Press. p. 35–86.

Brace S, Barnes I, Powell A, Pearson R, Woolaver LG, Thomas MG, Turvey ST. 2012. Population history of the Hispaniolan hutia *Plagiodontia aedium* (Rodentia: Capromyidae): testing the model of ancient differentiation on a geotectonically complex Caribbean island. *Mol Ecol* 21:2239–2253.

Brace S, Turvey ST, Weksler M, Hoogland MLP, Barnes I. 2015. Unexpected evolutionary diversity in a recently extinct Caribbean mammal radiation. *Proc Roy Soc B* 282:20142371.

Cabria MT, Rubines J, Gómez-Moliner B, Zardoya R. 2006. On the phylogenetic position of a rare Iberian endemic mammal, the Pyrenean desman (*Galemys pyrenaicus*). *Gene* 375:1–13.

dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc Roy Soc B* 279:3491–3500.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29:1969–1973.

Dubey S, Salamin N, Ohdachi SD, Barrière P, Vogel P. 2007. Molecular phylogenetics of shrews (Mammalia: Soricidae) reveal timing of transcontinental colonizations. *Mol Phylogenet Evol* 44:126–137.

Fabre PH, Vilstrup JT, Raghavan M, Der Sarkissian C, Willerslev E, Douzery EJP, Orlando L. 2014. Rodents of the Caribbean: origin and diversification of hutias unravelled by next-generation museomics. *Biol Lett* 10:20140266.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174.

Hedges SB, Conn CE. 2012. A new skink fauna from Caribbean islands (Squamata, Mabuyidae, Mabuyinae). *Zootaxa* 3288:1–244.

Heintzman PD, Zazula GD, Cahill JA, Reyes AV, MacPhee RD, Shapiro B. 2015. Genomic data from extinct North American *Camelops* revise camel evolutionary history. *Mol Biol Evol* 32:2433–2440.

Iturralde-Vinent M. 2006. Meso-Cenozoic Caribbean paleogeography: implications for the historical biogeography of the region. *Int Geol Rev* 48:791–827.

Iturralde-Vinent MA, MacPhee RDE. 1999. Paleogeography of the Caribbean region: implications for Cenozoic biogeography. *Bull Am Mus Nat Hist* 238:1–95.

Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. 2013. MapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinform* 29:1682–1684.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol* 29:1695–1701.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:357–359.

MacPhee RDE. 2009. Insulae infortunatae: establishing a chronology for Late Quaternary mammal extinctions in the West Indies. In: Haynes G, editor. American megafaunal extinctions at the end of the Pleistocene. Dordrecht: Springer. p. 169–193.

MacPhee RDE, Flemming C, Lunde DP. 1999. “Last occurrence” of the Antillean insectivoran *Nesophontes*: new radiometric dates and their interpretation. *Am Mus Novit* 3261:1–21.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 17:10–12.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TLL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* 334:521–524.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* 10.1101/pdb.prot5448.

Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 14:193–202.

Roca AL, Bar-Gal GK, Eizirik E, Helgen KM, Maria R, Springer MS, O'Brien SJ, Murphy WJ. 2004. Mesozoic origin for West Indian insectivores. *Nature* 429:649–651.

Rodríguez FJ, Oliver JL, Marín A, Medina JR. 1990. The general stochastic model of nucleotide substitution. *J Theor Biol* 142:485–501.

Rohland N, Reich D, Mallick S, Meyer M, Green RE, Georgiadis NJ, Roca AL, Hofreiter M. 2010. Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biol* 8:1000564.

Rosauer DF, Jetz W. 2014. Phylogenetic endemism in terrestrial mammals. *Global Ecol Biogeogr* 24:168–179.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.

Smith CI, Chamberlain AT, Riley MS, Stringer C, Collins MJ. 2003. The thermal history of human fossils and the likelihood of successful DNA amplification. *J Hum Evol* 45:203–217.

Stamatakis A. 2014. RAxML Version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinform* 30:1312–1313.

Turvey ST, Fritz SA. 2011. The ghosts of mammals past: biological and geographical patterns of global mammalian extinction across the Holocene. *Phil Trans Roy Soc B* 366:2564–2576.

Van Valen L. 1967. New Paleocene insectivores and insectivore classification. *Bull Am Mus Nat Hist* 135:217–284.

Wallace AR. 1880. Island life. London: Macmillan & Co.

Woodburne MO. 2010. The Great American Biotic Interchange: dispersals, tectonics, climate, sea level and holding pens. *J Mammal Evol* 17:245–264.

Woods CA, Sergile FE. Editors. 2001. Biogeography of the West Indies: patterns and perspectives. Boca Raton: CRC Press.

Table 1. AMS dates and associated radiocarbon ages for small mammal remains from Cueva de Bosque Humido, Dominican Republic.

Species	**Laboratory number**	**∂13C**	**Date (BP)**	**Radiocarbon age, 2σ (95.4%)**
Rattus rattus (femur)	OxA-19720	-21.74	—	post-AD 1950
Nesophontes hypomicrus (skull)	OxA-19721	-19.89	734 ± 24	AD 1247–1294
Artibeus jamaicensis (mandible)	OxA-19722	-18.68	846 ± 24	AD 1159–1254

Table 2. Mapping information for the NGS analyses. The genes used in the analyses, together with the number of mapped reads, the number of contigs assembled from the reads, and the mean coverage of reads to individual genes, are given for the four taxa for which NGS sequences were available: *Nesophontes*, *Solenodon*, *Atelerix* and *Condylura*.

Taxa	Gene	Sequence length	% Sequence covered	Number of contigs	Number of mapped reads	Mean coverage
Nesophontes	Mitogenome	16460	96.0	34	3288	12.672
	ADORA3	321	17.8	0	1	0.178
	ADRA2B	1064	57.0	3	53	1.827
	ADRB2	852	80.8	1	30	2.128
	APOB	2322	33.3	4	30	0.660
	APP	717	41.2	2	11	0.887
	ATP7A	678	37.3	1	3	0.481
	BCHE	987	31.2	1	6	0.507
	BDNF	555	96.8	2	31	2.927
	BMI1	324	43.7	0	3	0.570
	BMP4	1381	59.0	4	22	0.965
	BRCA1	3981	34.5	7	22	0.508
	CREM	472	73.4	1	9	1.404
	EDG1	978	86.0	4	51	2.734
	GHR	903	43.4	4	13	0.776
	PLCB4	342	38.0	1	4	0.529
	RAG1	1086	85.5	5	36	1.995
	RAG2	705	58.7	2	17	1.329
	RHO	261	88.5	1	37	5.027
	TTN	921	33.6	2	8	0.527
	TYR	426				
	VWF	1236				
Solenodon paradoxus	Mitogenome	17779	100	6	13,098	75.015
	BMP4	1381	100	1	199	13.735
	BRCA1	3981	100	1	374	11.963
	GHR	903	100	1	107	12.379
	RAG1	1086	100	1	165	16.619
	RAG2	705	100	1	91	12.847
	RHO	261	100	1	44	10.33
	TYR	426	100	1	44	11.575
	VWF	1236	100	5	199	13.646
Atelerix albiventris	Mitogenome	16460	100	4	3,513,587	21,028.05
	ADORA3	321	100	1	11	3.234
	ADRA2B	1064	96.5	2	41	4.094
	ADRB2	852	100	1	74	10.55
	APOB	2322	100	1	467	22.626
	ATP7A	678	100	1	560	88.826
	BCHE	987	100	1	123	13.633
	BDNF	555	100	1	108	19.038
	BMP4	1381	100	1	9,325	692.072
	BRCA1	3981	100	1	5,321	152.393
	EDG1	978	100	1	1,060	115.397
	GHR	903	100	1	6,699	731.011
	RAG1	1086	14.7	1	3	1.6
	RAG2	705	47.5	2	3	1.086
	VWF	1236	100	5	68	6.434
Condylura cristata	Mitogenome	16283	100	2	30,858	137.823

Fig. 1. Relationship of *Nesophontes* to other lipotyphlan mammals, showing maximum likelihood tree estimated for the concatenated mitochondrial and nuclear dataset of lipotyphlan and outgroup taxa using the GTR+ Γ model in RAxML. Nodal values represent bootstrap support (100 replicates). Only lipotyphlan taxa are depicted; see Methods for outgroup taxa.

Fig. 2. Molecular timescales for the Lipotyphla estimated in BEAST and based on seven fossil constraints (see Supplementary Table 4). Posterior probability values are shown to the left of nodes; estimates of median divergence dates are shown to the right of nodes, with 95% highest probability density values indicated in brackets. Only lipotyphlan taxa are depicted; see Methods for outgroup taxa. X-axis denotes time in millions of years. Pli = Pliocene.

Fig.1.

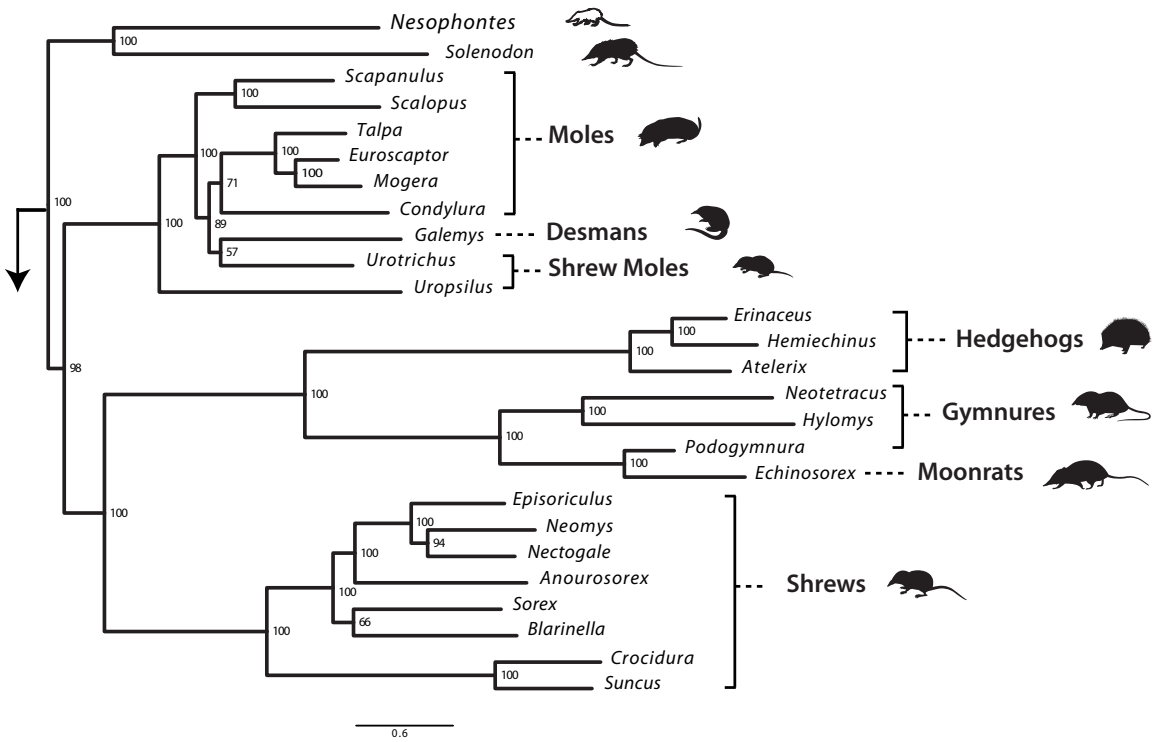
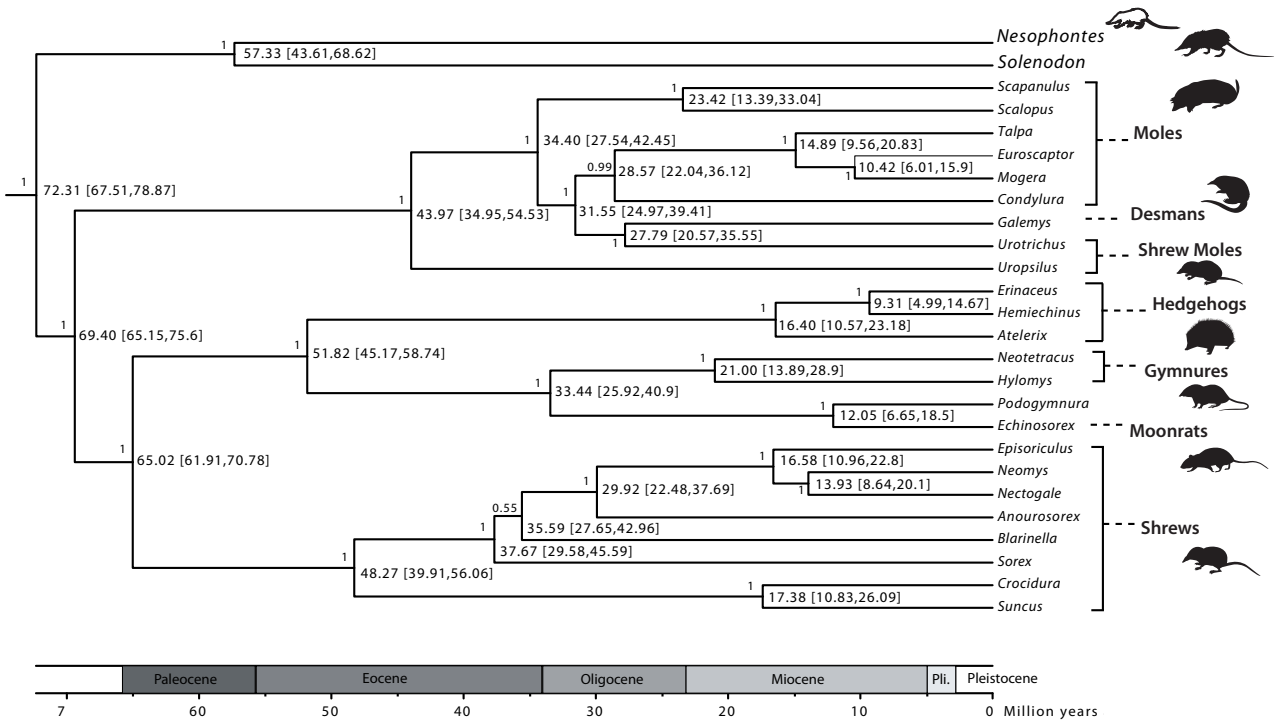

Nesophontes
Solenodon
Scapanulus
Scalopus
Talpa
Euroscaptor
Mogera
Condylura
Moles
Galemys
Desmans
Urotrichus
Uropsilus
Shrew Moles
Erinaceus
Hemiechinus
Atelerix
Hedgehogs
Neotetracus
Hylomys
Gymnures
Podogymnura
Echinosorex
Moonrats
Episoriculus
Neomys
Nectogale
Anourosorex
Sorex
Blarinella
Crocidura
Suncus
Shrews
0.6

Fig.2.


Nesophontes
Solenodon
57.33 [43.61,68.62]
Scapanulus
Scalopus
23.42 [13.39,33.04]
Talpa
Euroscaptor
Mogera
Condylura
34.40 [27.54,42.45]
14.89 [9.56,20.83]
10.42 [6.01,15.9]
28.57 [22.04,36.12]
31.55 [24.97,39.41]
27.79 [20.57,35.55]
Moles
Galemys
Desmans
Urotrichus
Uropsilus
Shrew Moles
72.31 [67.51,78.87]
43.97 [34.95,54.53]
Erinaceus
Hemiechinus
Atelerix
9.31 [4.99,14.67]
16.40 [10.57,23.18]
Hedgehogs
69.40 [65.15,75.6]
51.82 [45.17,58.74]
Neotetracus
Hylomys
21.00 [13.89,28.9]
33.44 [25.92,40.9]
Gymnures
Podogymnura
Echinosorex
12.05 [6.65,18.5]
Moonrats
65.02 [61.91,70.78]
Episoriculus
Neomys
Nectogale
Anourosorex
Blarinella
Sorex
Crocidura
Suncus
16.58 [10.96,22.8]
29.92 [22.48,37.69]
13.93 [8.64,20.1]
35.59 [27.65,42.96]
37.67 [29.58,45.59]
48.27 [39.91,56.06]
17.38 [10.83,26.09]
Shrews
Paleocene
Eocene
Oligocene
Miocene
Plio.
Pleistocene
7
60
50
40
30
20
10
0 Million years