



Photo of a *Solenodon paradoxus* from the island of Hispaniola, courtesy of Eladio Fernandez, Caribbean Nature Photography

Grigorev et al.

## RESEARCH

# Innovative assembly strategy contributes to understanding the evolution and conservation genetics of the endangered *Solenodon paradoxus* from the island of Hispaniola

Kirill Grigorev <sup>1,†</sup>, Sergey Kliver <sup>2,†</sup>, Pavel Dobrynin<sup>2</sup>, Aleksey Komissarov<sup>2</sup>, Walter Wolfsberger<sup>1,3</sup>, Ksenia Krasheninnikova<sup>2</sup>, Yashira M. Afanador-Hernández<sup>1</sup>, Adam L. Brandt<sup>4,5</sup>, Liz A. Paulino<sup>6</sup>, Rosanna Carreras<sup>6</sup>, Luis E. Rodríguez<sup>6</sup>, Adrell Núñez<sup>7</sup>, Jessica R. Brandt<sup>4,8</sup>, Filipe Silva<sup>9,10</sup>, J. David Hernández-Martich<sup>11</sup>, Audrey J. Majeske<sup>1</sup>, Agostinho Antunes<sup>9,10</sup>, Alfred L. Roca<sup>4,12</sup>, Stephen J. O'Brien<sup>2,13</sup>, Juan Carlos Martínez-Cruzado<sup>1</sup> and Taras K. Oleksyk <sup>1,3,\*</sup>

<sup>1</sup>Department of Biology, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico, <sup>2</sup>Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia, <sup>3</sup>Biology Department, Uzhhorod National University, Uzhhorod, Ukraine, <sup>4</sup>Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, <sup>5</sup>Division of Natural Sciences, St. Norbert College, De Pere, Wisconsin, USA, <sup>6</sup>Instituto Tecnológico de Santo Domingo (INTEC), Santo Domingo, Dominican Republic, <sup>7</sup>Department of Conservation and Science, Parque Zoológico Nacional (ZOODOM), Santo Domingo, Dominican Republic, <sup>8</sup>Department of Biology, Marian University, Fond du Lac, Wisconsin, USA, <sup>9</sup>CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos, s/n, 4450–208 Porto, Portugal, <sup>10</sup>Department of Biology, Faculty of Sciences, University of Porto. Rua do Campo Alegre, 4169-007 Porto, Portugal, <sup>11</sup>Instituto de Investigaciones Botánicas y Zoológicas, Universidad Autónoma de Santo Domingo, Santo Domingo, Dominican Republic, <sup>12</sup>Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA and <sup>13</sup>Oceanographic Center, Nova Southeastern University, Fort Lauderdale, Florida, USA

\*Correspondence address. Departamento de Biología Carr. 108 Barrio Miradero Km 1.3. Entrada al Zoológico Mayagüez, PR 00680

E-mail: [taras.oleksyk@upr.edu](mailto:taras.oleksyk@upr.edu)

<sup>†</sup>These authors contributed equally.

Received: 20 July 2017; Revised: 26 January 2018; Accepted: 7 March 2018

© The Author(s) 2018. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

Solenodons are insectivores that live in Hispaniola and Cuba. They form an isolated branch in the tree of placental mammals that are highly divergent from other eulipotyphlan insectivores. The history, unique biology, and adaptations of these enigmatic venomous species could be illuminated by the availability of genome data. However, a whole genome assembly for solenodons has not been previously performed, partially due to the difficulty in obtaining samples from the field. Island isolation and reduced numbers have likely resulted in high homozygosity within the Hispaniolan solenodon (*Solenodon paradoxus*). Thus, we tested the performance of several assembly strategies on the genome of this genetically impoverished species. The string graph-based assembly strategy seemed a better choice compared to the conventional de Bruijn graph approach due to the high levels of homozygosity, which is often a hallmark of endemic or endangered species. A consensus reference genome was assembled from sequences of 5 individuals from the southern subspecies (*S. p. woodi*). In addition, we obtained an additional sequence from 1 sample of the northern subspecies (*S. p. paradoxus*). The resulting genome assemblies were compared to each other and annotated for genes, with an emphasis on venom genes, repeats, variable microsatellite loci, and other genomic variants. Phylogenetic positioning and selection signatures were inferred based on 4,416 single-copy orthologs from 10 other mammals. We estimated that solenodons diverged from other extant mammals 73.6 million years ago. Patterns of single-nucleotide polymorphism variation allowed us to infer population demography, which supported a subspecies split within the Hispaniolan solenodon at least 300 thousand years ago.

**Keywords:** genome assembly; de Bruijn; string graph; *Solenodon paradoxus*; Hispaniola; Caribbean; island evolution; natural selection; isolation; heterozygosity

## Background

The only 2 surviving species of solenodons, found on the 2 largest Caribbean islands, Hispaniola (*Solenodon paradoxus*) and Cuba (*Solenodon cubanus*), are among the few endemic terrestrial mammals that survived human settlement of these islands. Phenotypically, solenodons somewhat resemble shrews (Fig. 1), but molecular evidence indicates that they are actually the sister group to all other extant eulipotyphlan insectivores (hedgehogs, moles, shrews) from which they split in the Cretaceous Period [1–3]. These enigmatic species have various local names in Cuba and Hispaniola, including *oso* (bear), *hormiguero* (ant-eater), *jorón* (ferret), *milquí* (or *almiquí*), and *agouta* [4, 5], all pointing to the first impression made on the Spanish colonists by its unusual appearance. Today, the Hispaniolan solenodon (*S. paradoxus*) is difficult to find in the wild because of its nocturnal activity pattern and its low population numbers. Here, we report the assembly and annotation of the nuclear genome sequences and genomic variation of 2 subspecies of *S. paradoxus*. We used analytical strategies that will allow researchers to formulate hypotheses and develop genetic tools to assist future studies of evolutionary inference and conservation applications.

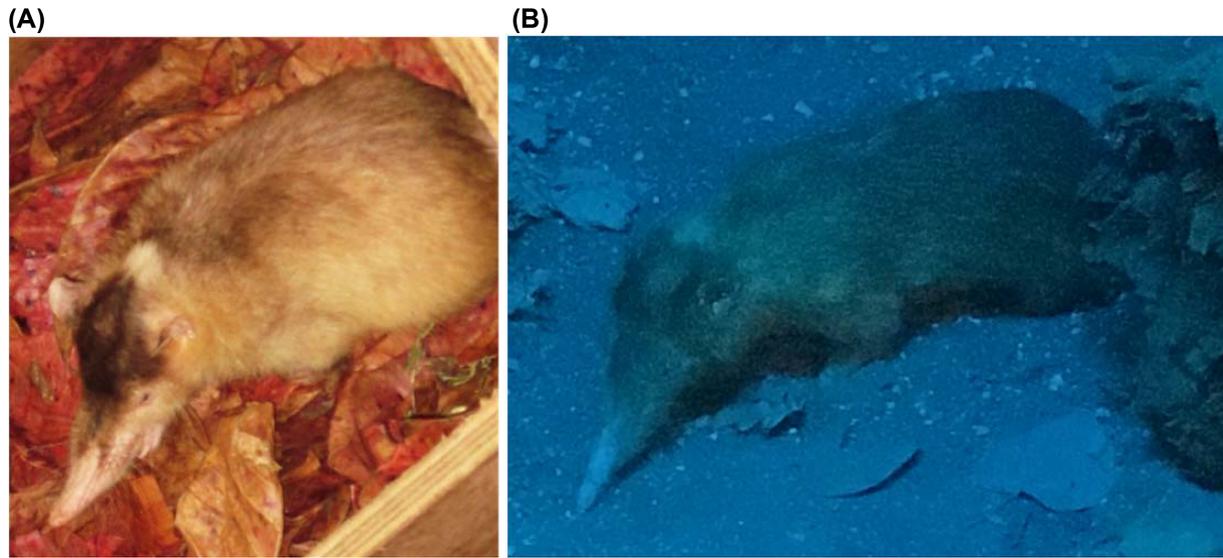
*Solenodon paradoxus* was originally described from a skin and partial skull at the St. Petersburg Academy of Sciences in Russia [6]. It has a large head with a long rostrum and tiny eyes and ears partially hidden by the dusky brown body fur that turns reddish on the sides of the head, throat, and upper chest. The tail, legs, snout, and eyelids of the *S. paradoxus* are hairless. The front legs are noticeably more developed, but all 4 have strong claws useful for burrowing (Fig. 1). Adult animals measure 49–72 cm in total length and weigh more than 1 kg [7]. Solenodons are social animals; they spend their days in extensive underground tunnel networks shared by family groups and come to the surface at night to hunt small vertebrates and large invertebrates [8]. A unique feature is the *os proboscidis*, a bone that extends forward from the nasal opening to support the snout cartilage [9]. Solenodons are venomous mammals that display a fascinating strategy for venom delivery. The second lower incisor of solenodons has a narrow, almost fully enfolded tubular channel, through

which saliva secreted by the submaxillary gland flows into the victim [10]. The genus name *Solenodon* means “grooved tooth” in Greek and refers to the shape of this incisor. Although solenodons rarely bite humans, the bites can be very painful (Nicolás Corona, personal communication), and even a small injection of venom has been shown to be fatal to mice in a matter of minutes [7]. The chemical composition of solenodon venom has not yet been resolved [11].

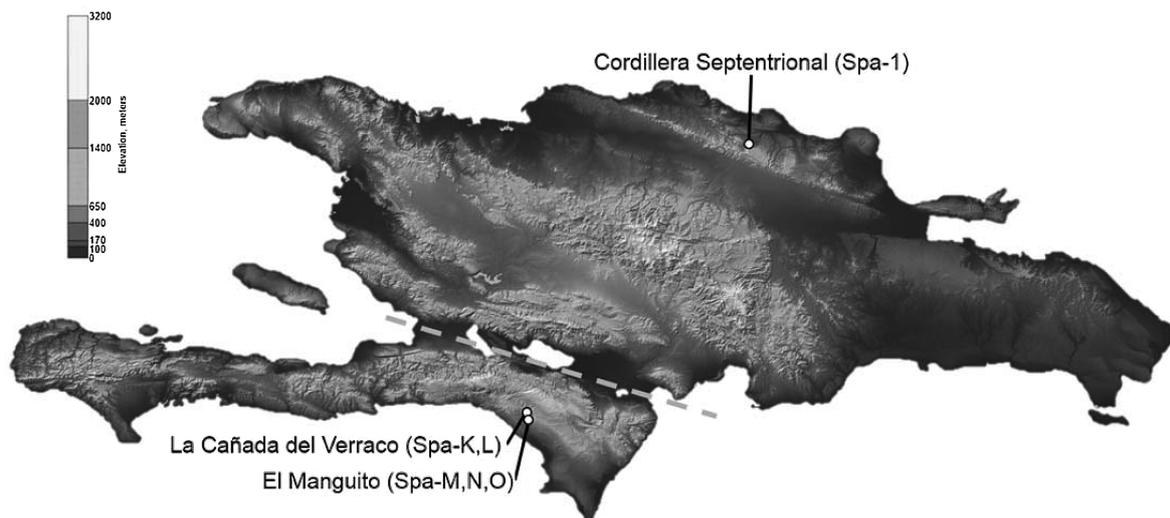
Roca et al. [3] sequenced 13.9 kb of nuclear and mitochondrial sequences of *S. paradoxus*, inferring that solenodon divergence from other eulipotyphlan mammals such as shrews and moles dates back to the Cretaceous Period, ~76 million years ago (Mya), before the mass extinction of the dinosaurs ~66 Mya. Brandt et al. [12] sequenced complete mitogenome sequences of 6 Hispaniolan solenodon specimens from the southern part of Hispaniola (Fig. 2), corroborating this conclusion, and estimated that *S. paradoxus* diverged from all other mammals ~78 Mya [12]. Other studies have reported similarly deep divergence dates (reviewed by Sato et al. [13]). Whole genome analysis of *S. paradoxus* could provide support and validation to the earlier evolutionary studies.

Morphometric studies suggest that southern and northern Hispaniolan solenodons may be distinctive enough to be considered separate subspecies [2, 14, 15], a notion supported by recent mitochondrial DNA studies [12, 16]. The southern Hispaniolan solenodons had less genetic diversity than those in the north, so that the control region sequences of all 5 southern specimens (the same individuals used in this study) were identical or nearly identical [12], indicating that Hispaniolan solenodons have a very low level of mitochondrial diversity.

It may now be imperative to study conservation genomics of solenodons because their extinction would mean the loss of an entire evolutionary lineage whose antiquity goes back to the age of dinosaurs. *Solenodon paradoxus* survived in spectacular island isolation despite the devastating human impact to biodiversity in recent centuries [3, 12]. Nevertheless, survival of this species is now threatened by deforestation, increasing human activity, and predation by introduced dogs, cats, and mongooses. It is declining in population, its habitat is severely fragmented, and it



**Figure 1:** The 2 subspecies of *Solenodon paradoxus*. (A) A captive Hispaniolan solenodon from the northern subspecies (*S. p. paradoxus*) photographed at the Santo Domingo Zoo (photo taken by Juan C. Martínez-Cruzado in 2014). (B) A mounted specimen of the southern subspecies (*S. p. woodi*) photographed at the Museo Nacional de Historia Natural prof. Eugenio de Jesús Marciano in Santo Domingo, Dominican Republic (photo taken by Taras K. Oleksyk in 2017).

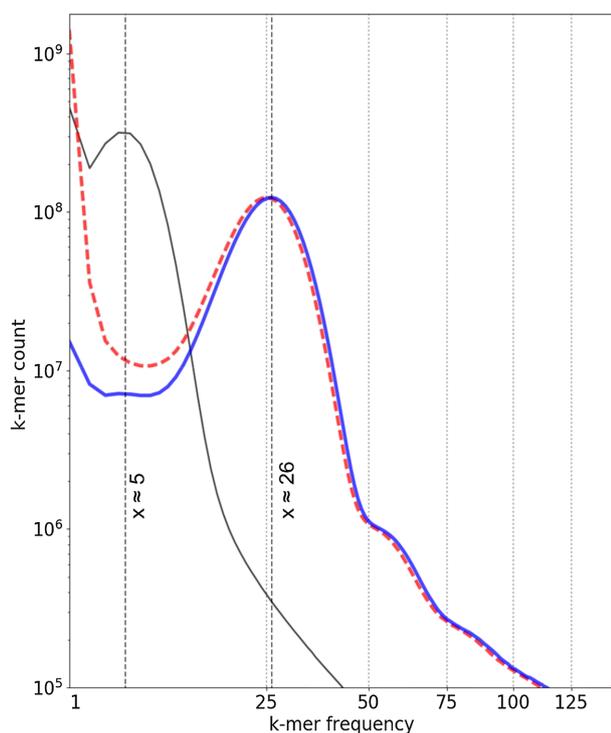


**Figure 2:** Origins of the genomic DNA samples of *Solenodon paradoxus* from the island of Hispaniola. Approximate locations of capture for 5 wild individuals of *S. p. woodi*: Spa-K and Spa-L from La Cañada del Verraco, as well as Spa-M, Spa-N, and Spa-O from the El Manguito location in the Pedernales Province in the southwest corner of the Dominican Republic bordering Haiti. In addition, 1 *S. p. paradoxus* sample (Spa-1) was taken from Cordillera Septentrional in the northern part of the island. Exact coordinates of each sample location are listed in [12]. The dashed line indicates the position of the Cul de Sac Plain and Neiba Valley; this region was periodically inundated by a marine canal that separated Hispaniola into north and south paleo-islands during the Pliocene and Pleistocene [15]. The original map is in the public domain (courtesy of NASA).

is listed as endangered by the International Union for Conservation of Nature (IUCN) Red List of Threatened Species (Red List category B2ab, assessed in 2008 [17]).

In this study, we assembled the genome of *S. paradoxus* using low coverage genome data ( $\sim 5\times$  each) from 5 individuals of *S. paradoxus woodi*. We took advantage of the low individual and population genetic diversity to pool individual data and apply a string graph assembly approach that resulted in a working genome assembly of the *S. paradoxus* genome from the combined paired-end dataset ( $\sim 26\times$ ; Fig. 3). Our methodology introduces a useful pipeline for genome assembly to compensate for the limited amount of sequencing that, in this instance,

performs better than the assembly by a traditional de Bruijn algorithm (SOAPdenovo2) [18]. We used the string-graph assembler Fermi [19] as a principal tool for contig assembly in conjunction with SSPACE [20] and GapCloser [18] for scaffolding. The resulting genome sequence data was sufficient for high-quality annotation of genes and functional elements, as well as for comparative genomics and population genetic analyses. Prior to this study, the string-graph assembler Fermi [19] had been used only in studies for annotation or as a complementary tool for *de novo* assemblies made with de Bruijn algorithms [21]. We present and compare genome assemblies for the southern subspecies (*S. p. woodi*) based on several combinations of assembly tools, provide



**Figure 3:** Heterozygosity and k-mer distribution. k-mer distributions for the *S. p. woodi* reads. Only 1 original sample (SPA-K) distribution is shown as a solid gray line, as the distributions were identical for each of the individual samples. The predicted mean genome coverage was approximately 5x for each sample ( $x = 5$ ). One example is plotted by a black solid line on the left. The combined uncorrected dataset is plotted in a dashed red line indicates a maximum at  $x = 26$ . The combined dataset corrected with QuorUM [25] is plotted in a solid blue line, also with a maximum at  $x = 26$ . A smaller local maximum on the left side for both combined distributions, corrected and uncorrected (representing k-mers found once or very few times), is expected from differences between overlapping reads, most likely the sequencing errors. Other local maxima (seen as a small bulge at the  $x = 5$ ) are interpreted as heterozygous sites. These proved to have almost no impact on the combined sample even after read correction, indicating a lack of heterozygous sites for this solenodon subspecies. The largest local maxima (to the right) are interpreted as projected coverage. For the combined samples, this value is  $x = 26$ .

a high-quality annotation of genome features and describe genetic variation in 2 subspecies (*S. p. woodi* and *S. p. paradoxus*), make inferences about recent evolution and selection signatures in genes, trace demographic histories, and develop molecular tools for future conservation studies.

## Data Description

### Sample collection and sequencing

Five adult individuals of *S. paradoxus woodi* (National Center for Biotechnology Information [NCBI] Taxon ID:1906352) from the southern Dominican Republic were collected in the wild following a general field protocol described earlier [12]; this included 2 specimens caught from La Cañada del Verraco and 3 from the El Manguito location in the Pedernales Province. The captured individuals were visually assessed for obvious signs of disease, weighed, measured, sexed, and released at the capture site, all within 10 minutes of capture. Geographic coordinates were recorded for every location. In addition, 1 *S. p. paradoxus* (Spa-1) sample was acquired through collaboration with ZooDom at Santo Domingo that originated in the Cordillera Septentrional

in the northern part of the island. Figure 2 highlights geographical locations of sample collection points for the samples used in this study.

The 5 *S. p. woodi* samples were sequenced using HiSeq2000 technology (Illumina Inc.), resulting in an average of 151,783,327 paired-end reads 101 bp long, or 15.33 Gb of sequence data, per individual. In addition, DNA extracted from the northern solenodon (*S. p. paradoxus*; Spa-1) was sequenced using MiSeq V3 technology (Illumina Inc.) and produced 52,358,830 paired-end reads, equating to approximately 13.09 Gb of sequence data. Only the samples of *S. paradoxus woodi* were used for assembly since the northern subspecies (*S. paradoxus paradoxus*) did not have sufficient coverage for the *de novo* assembly.

Further details about sample collection, DNA extraction, library construction, and sequencing can be found in the Methods section. The whole genome shotgun data from this project have been deposited at DDBJ/ENA/GenBank under the accession NKTL00000000. The version described here is version NKTL01000000. The genome data has also been deposited into the NCBI under BioProject PRJNA368679 and to the GigaScience GigaDB repository [22].

### Read correction

After the reduction of adapter contamination with Cookiecutter [23], the k-mer distribution in the reads for the 5 individuals of *S. paradoxus woodi* was assessed with Jellyfish (Jellyfish, RRID:SCR.005491) [24]. The predicted mean genome coverage was approximately 5x for each sample (Fig. 3), which is too low for individual *de novo* genome assembly. However, because of the extremely low levels of genetic diversity suggested by the earlier study of mitochondrial DNA in the southern subspecies [12] and in order to increase the average depth of coverage, the reads from the 5 samples were combined into a single dataset. As a result, the projected mean genome coverage for the combined genome assembly was 26x. Error correction was applied with QuorUM (QuorUM, RRID:SCR.011840) [25] using the value  $k = 31$ . The k-mer distribution analysis by Jellyfish in the combined and error-corrected dataset indicated very low levels of heterozygosity in accordance with the hypothesis (see Fig. 3 legend), allowing use of the combined dataset for further genome assembly. Using KmerGenie [26], the genome size has been estimated to be 2.06 Gbp.

## Analyses

### Assembly tool combinations

We used several alternative combinations of tools to determine the best approach to an assembly of the combined genome data, outlined in Table 1. First, the combined libraries of paired-end reads were assembled into contigs with Fermi, a string graph-based tool [19]. Second, the same libraries were also assembled with SOAPdenovo2, a de Bruijn graph-based tool (SOAPdenovo2, RRID:SCR.014986) [18]. The optimal k-mer length parameter for SOAPdenovo2 was determined to be  $k = 35$  with the use of KmerGenie [26]. For the scaffolding step, we used either SSPACE (SSPACE, RRID:SCR.005056) [20] or the scaffolding module of SOAPdenovo2 [18]. For all instances, the GapCloser module of SOAPdenovo2 was used to fill in gaps in the scaffolds (GapCloser, RRID:SCR.015026) [18]. After assembly, datasets were trimmed; scaffolds shorter than 1 Kbp were removed from the output. In Table 1, the 4 possible combinations of tools used for the assembly are referred to with capital letters A, B, C, and D for brevity.

**Table 1:** Description of the assembly strategies and comparison of metrics for the resulting assemblies.

Assembly names	A	B	C	D
<b>Assembly tools</b>				
Contig assembly tool	Fermi	Fermi	SOAPdenovo2	SOAPdenovo2
Scaffolding tool	SOAPdenovo2	SSPACE	SOAPdenovo2	SSPACE
Gap closing tool	GapCloser	GapCloser	GapCloser	GapCloser
<b>Assembly metrics</b>				
<b>Total contigs (&gt;1,000 bp)</b>	<b>71,429</b>	<b>71,429</b>	<b>189,566</b>	<b>189,566</b>
Contig N50	54,944	54,944	4,048	4,048
Contig CEGMA (%)*	96.37(77.42)	96.37(77.42)	68.15(33.06)	68.15(33.06)
Contig BUSCO (%)	86(65)	86(65)	42(21)	42(21)
<b>Total scaffolds (&gt;1,000 bp)</b>	<b>14,417</b>	<b>40,372</b>	<b>20,466</b>	-
Final N50	555,585	110,915	331,639	-
Final CEGMA (%)	95.56(81.85)	95.97(88.71)	95.97(90.73)	-
Final BUSCO (%)	91(74)	86(64)	94(80)	-
<b>Quality</b>				
Percentage of Ns (%)	0.06322	0.0135	0.02622	-
REAPR error-free bases (%)	96.46	95.35	94.98	-
REAPR low-scoring regions	18	16	71	-
REAPR incorrectly oriented reads	11,543	5,329	28,964	-

\*BUSCO [29] and CEGMA [30] percentages are reported for all genes (complete and partial), while the percentage of complete genes are shown in parentheses.

However, SOAPdenovo2 introduces artifacts at the contig construction stage, which it is specifically designed to mitigate at later stages, and SSPACE is not aware of such artifacts [27]. For this reason, the assembly produced by combination D (contig assembly with SOAPdenovo2 and scaffolding with SSPACE) was not reported.

### Quality control (QC) and structural comparisons between the assemblies

We used QCAST (QCAST, [RRID:SCR.001228](#)) [28] to estimate the common metrics of assembly quality for all combinations of assembly tools: N50 and gappedness (the percentage of Ns (Table 1)). Fermi-assembled contigs (A and B) were overall longer and fewer in number than the SOAPdenovo2 (C and D). The assembly completeness was also evaluated with both benchmarking universal single-copy orthologs (BUSCO, [RRID:SCR.015008](#)) [29] and core eukaryotic genes mapping approach (CEGMA, [RRID:SCR.015055](#)) [30] for completeness of conservative genes. Fermi assemblies (A and B) showed high levels of completeness compared to SOAPdenovo2 (86% vs. 42%) at the contig level. However, this difference was partially mitigated at the scaffolding step where SOAPdenovo2 increases completeness for Fermi assembly (A) and more than doubles it for the SOAPdenovo2 assembly (C). To directly evaluate the quality of all assemblies, we applied REAPR [31]. From the REAPR metrics presented at the bottom of Table 1, it appears that even though the scaffolding step increased the final N50 for the C assembly, it contains significantly more regions with high probability of mis-assemblies (low-scoring regions), less error-free bases, and 3 to 6 times higher number of incorrectly oriented reads compared to the Fermi-based assemblies (A and B) (Table 1).

We hypothesized that aligning the 3 genome assemblies to each other would allow us to detect some of these mis-assemblies. A comparison to the best, most closely related genome assembly (i.e., *Sorex araneus*) will reveal several rearrangements that, in many cases, reflect real evolutionary events. It is reasonable to assume that if all the rearrangements that are detected are real and not due to the assembly artifacts, the num-

ber of detected rearrangements vs. *Sorex* assembly would be the same for all 3 *Solenodon* assemblies (A, B, and C). Following the parsimony principle, an assembly that shows rearrangements is also likely to contain the most assembly artifacts. Conversely, we expected that the best of the 3 assemblies of the *Solenodon* genome would contain the least number of reversals and transpositions when compared to the best available closely related genome (*Sorex araneus*).

To test this hypothesis, the 3 completed assemblies of *Solenodon* (A, B, and C) were aligned to each other and to the out-group, which was the *Sorex* genome (SorAra 2.0, NCBI accession GCA.000181275.2), using Progressive Cactus [32]. Custom scripts were used to interpret binary output of the pairwise genome by genome comparisons; the resulting coverage metrics are presented in Table 2. In this comparison, all 3 *Solenodon* genome assemblies had a substantial overlap and resulted in similar levels of synteny when compared against the *Sorex* reference assembly. However, assemblies A and B had the fewest differences with *Sorex*, while assembly C had more differences vs. A, B, and *Sorex*. Next, syntenic blocks between each of the 3 *Solenodon* assemblies (A, B, and C) were compared to the *Sorex* assembly. The 50-Kbp syntenic blocks were identified using the ragout-maf2 synteny module of the software package Ragout [33], and the number of scaffolds that contained syntenic block rearrangements was determined. As a result, assembly B had the lowest number of reversals and transpositions compared to the *S. araneus* reference genome (Table 2). Based on the combined results of the evaluations by REAPR [31], Progressive Cactus [32], and Ragout [33], assembly C (generated by the complete SOAPdenovo2 run) was not included in further analysis.

### Genome annotation and evaluation of assembly completeness

Repeats in assemblies A and B were identified and soft masked using RepeatMasker (RepeatMasker, [RRID:SCR.012954](#)) [34] with the RepBase library [35]. The total percentage of all interspersed repeats masked in the genome was lower than in *S. araneus* (22.53% vs. 30.48%). One possible reason could be that a low

**Table 2:** Pairwise genomic coverage for the 3 assemblies and the *Sorex araneus* genome (SorAra 2.0, NCBI accession number GCA.000181275.2) obtained from the Progressive Cactus [32] alignments.

Assembly	vs <i>S. paradoxus woodi</i>			vs <i>S. araneus</i>		
	Pairwise genome coverage (%)*			# Inversions	# Translocations	
	A	B	C			
<i>S. araneus</i>		42.1	42.2	42.3	-	-
<i>S. p. woodi</i>	A	-	99.4	98.5	35.5	87
	B	99.3	-	99.3	35.5	34
	C	98.4	98.5	-	35.5	81

While all 3 assemblies have similar amounts of syntenic coverage to the *Sorex* genome, assembly B contains the least number of structural rearrangements (inversions and translocations) compared to the other 2 assemblies (A and C).

\*Values in cells at the intersection of rows and columns represent the percentage (%) of coverage between the 2 compared genome assemblies. Syntenic blocks between each of the 3 solenodon assemblies (A, B, and C) were compared to the *S. araneus* assembly, and 50 Kbp syntenic blocks were identified using the *ragout-maf2synteny* module of the software package Ragout [33].

**Table 3:** Repeat content of the *Solenodon paradoxus* genome (assembly B), annotated by RepeatMasker [34] with the RepBase library [35].

Class	Number	Length (bp)	Percentage (%)
<b>Total interspersed repeats</b>		<b>461,754,432</b>	<b>22.53</b>
SINEs	271,839	36,271,455	1.77
Alu/B1	6	341	<0.0001
MIRs	264,319	35,557,190	1.73
LINEs	610,079	304,823,409	14.87
LINE1	425,750	260,176,709	12.7
LINE2	157,422	39,432,276	1.92
L3/CR1	22,172	4,293,335	0.21
RTE	4,122	839,744	0.04
LTR elements	246,305	78,108,726	3.81
ERVL	61,150	24,158,692	1.18
ERVL-MaLRs	94,934	30,075,905	1.47
ERV_classI	57,674	19,259,649	0.94
ERV_classII	24,454	2,840,874	0.14
DNA elements	204,413	42,015,054	2.05
hAT-Charlie	112,664	21,168,194	1.03
TcMar-Tigger	43,950	11,141,107	0.54
Small RNAs	4,772	456,810	0.02
Satellites	46,734	20,910,815	1.02
Simple repeats	644,811	28,549,871	1.39
Low complexity regions	114,188	5,933,786	0.29
Unclassified	3,051	535,788	0.03

coverage assembly may perform better in nonrepetitive regions. Alternatively, if the repeat content in *S. paradoxus* is indeed lower, this would have to be evaluated using a higher-quality assembly with the use of long read data. The total masked repeat content of the *S. paradoxus* genome including simple/tandem repeats, satellite DNA, low complexity regions, and other elements is presented in Table 3. The repeat content can be retrieved from Database S1.

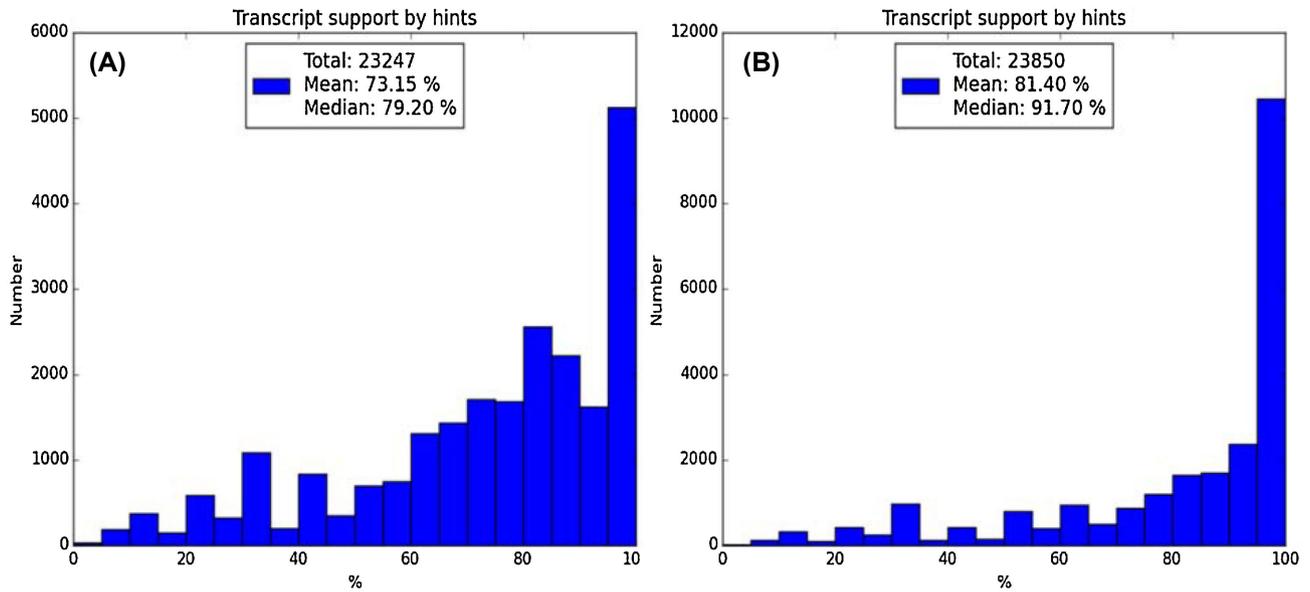
The annotation of protein-coding genes was performed using a combined approach that synthesized both homology-based and *de novo* predictions, where *de novo* predictions were used to fill gaps and extend homology-based predictions. Gene annotation was performed for both assemblies (A and B) independently. Proteins of 4 reference species *S. araneus* (SorAra 2.0, GCA.000181275.2), *Erinaceus europaeus* (EriEur2.0, GCA.000296755.1), *Homo sapiens* (GRCh38.p7), and *Mus musculus*

(GRCm38.p4) were aligned to a *S. paradoxus* assembly with Exonerate [36] with a maximum of 3 “hits” (matches) per protein. The obtained alignments were classified into the top (primary) hit and 2 secondary hits; the coding sequence (CDS) fragments were cut from each side by 3 bp for the top hits and by 9 bp for secondary hits. These truncated fragments were clustered and supplied as hints (local pieces of information about the gene in the input sequence, such as a likely stretch of coding sequence) of the potential protein-coding regions to the AUGUSTUS software package (Augustus: Gene Prediction, RRID:SCR.008417) [37], which predicted genes in the soft-masked *Solenodon* assembly. Proteins were extracted from the predicted genes and aligned by HMMER (Hmmer, RRID:SCR.005305) [38] and Basic Local Alignment Search Tool for Proteins (BLAST; NCBI, RRID:SCR.004870) [39] to Pfam (Pfam, RRID:SCR.004726) [40] and Swiss-Prot [41] databases, respectively. Genes supported by hits to protein databases and hints were retained; the unsupported sequences were discarded. The annotated genes can be retrieved from Database S2.

Assembly B showed a higher support compared to assembly A (91.7% vs. 79.2%) for the protein-coding gene predictions by extrinsic evidence, even though assembly A had a larger N50 value (Table 1). These values were calculated as a median fraction of exons supported by alignments of proteins from reference species to genome (Fig. 4). In other words, assembly B is more useful for gene predictions and is likely to contain better gene models that can be used in the downstream analysis. Therefore, based on 2 lines of evidence, low rearrangement counts (Table 2) and high support to gene prediction for the assembly B, it was chosen for the subsequent analyses as the most useful current representation of the *Solenodon* genome.

### Noncoding RNA genes

For all noncoding RNA genes except for tRNA and rRNA genes, the search was performed with INFERNAL (Infernal, RRID:SCR.011809) [42] using the Rfam (Rfam, RRID:SCR.007891) [43] BLASTN hits as seeds. The tRNA genes were predicted using tRNAscan-SE (tRNAscan-SE, RRID:SCR.010835) [44], and rRNA genes were predicted with Barrnap ((Basic Rapid Ribosomal RNA Predictor), version 0.6 (Barrnap, RRID:SCR.015995) [45]). Additionally, RNA genes discovered by RepeatMasker at the earlier stages of the analysis were used to cross-reference the findings of rRNA- and tRNA-finding software. The list of the noncoding RNA genes can be accessed in Database S3.



**Figure 4:** Distribution of the gene prediction support by extrinsic evidence for *Solenodon* assemblies A (on the left) and B (on the right). Proteins of 4 reference species *S. araneus* (SorAra 2.0, GCA.000181275.2), *Erinaceus europaeus* (EriEur2.0, GCA.000296755.1), *Homo sapiens* (GRCh38.p7), and *Mus musculus* (GRCm38.p4) were aligned to a *S. paradoxus* assembly with Exonerate [36] with a maximum of 3 best matches per protein. Coding sequences were cut from each, clustered, and uploaded into the AUGUSTUS software package [37] to predict genes in the soft-masked *Solenodon* assembly. Proteins from the predicted genes were aligned by HMMER [38] and BLAST [39] to Pfam [40] and Swiss-Prot [41] databases. Genes supported by matches to protein databases and “hints” (see definition in main text) were retained; the rest were discarded. Substantially more transcripts have higher hint support in assembly B. The annotated genes can be retrieved from Database S2. Assembly C has not been evaluated.

**Table 4:** The weighted coverages of the genomes in the Progressive Cactus alignment [32], as calculated against the *C. familiaris* genome.

Query genome	Weighted coverage
Dog ( <i>Canis familiaris</i> )	(1.14)*
Cow ( <i>Bos taurus</i> )	1.06
Common shrew ( <i>Sorex araneus</i> )	1.05
Star-nosed mole ( <i>Condylura cristata</i> )	1.04
Hispaniolan solenodon ( <i>Solenodon paradoxus</i> )	1.05

\*The weighted coverage of a genome to itself is parenthesized as it is not a comparative value.

The weighted coverage of the *S. paradoxus* genome assembly from our study is comparable to other high-coverage mammalian genome assemblies. The cladogram used for multiple genome alignment with Progressive Cactus is shown in Fig. S1.

### Multiple genome alignment, synteny, and duplication structure

To compare the *Solenodon* genome assembly with other mammalian genomes, a multiple alignment with genomes of related species was performed using Progressive Cactus [32]. Currently available genomic assemblies of cow (*Bos taurus*, BosTau 3.1.1, NCBI accession number DAAA00000000.2), dog (*Canis familiaris*, CanFam 3.1, GCA.000002285.2), star-nosed mole (*Condylura cristata*, ConCri 1.0, GCF.000260355.1), common shrew (*S. araneus*, SorAra 2.0, GCA.000181275.2), and *S. paradoxus woodi* (assembly B from this study) were aligned together, guided by a cladogram representing branching order in a subset of a larger phylogeny (Fig. S1). We evaluated the *S. paradoxus* coverage by comparing it to the weighted coverages of other genomes in the alignment to the *C. familiaris* genome (Table 4). Custom scripts were used to interpret the binary output of Progressive Cactus

(“Cactus”) [32]. Cactus genome alignments were used to build a “sparse map” of the homologies between a set of input sequences. Once this sparse map is constructed, in the form of a Cactus graph, the sequences that were initially unaligned in the sparse map are also aligned [32]. Weighted coverage of a genome by genome comparison was calculated by binning an alignment into regions of different coverage and averaging these coverages, with lengths of bins as weights. The weighted coverage of *S. paradoxus* to *C. familiaris* was 1.05, which indicated that the present genome assembly is comparable in quality and duplication structure to other available mammalian assemblies, which are close to each other and are close to 1.0 (Table 4).

### Detection of single-copy orthologs

Single-copy orthologs (single gene copies) are essential for the evolutionary analysis since they represent a useful conservative homologous set, unlike genes with paralogs, which are difficult to compare across species. The longest polypeptide coded by each gene of *S. paradoxus* and of 3 other *Eulipotyphla*—*Erinaceus europaeus*, *S. araneus*, and *C. cristata*—were aligned to profile hidden Markov models of the TreeFam database (Tree families database, RRID:SCR.013401) [46, 47] using HMMER [38]. Top hits from these alignments were extracted and used for assignment of corresponding proteins to families. The same procedure was performed in order to assign proteins to orthologous groups using profile hidden Markov models of orthologous groups of the maNOG subset from the eggNOG database (eggNOG, RRID:SCR.002456) [48] as referenced. Orthologous groups and families for which high levels of error rates were observed while testing assignment of proteins to them were discarded; the rest of the orthologous groups and families were retained for further analysis. Proteins and the

**Table 5:** Fossil-based priors associated with mammalian evolution used for calibration of divergence times [54–57].

Node	Calibration prior on clade	Node min. age (Mya)	Node max. age (Mya)	Evidence
Opossum—placental mammals split	<i>Eutheria—Metatheria</i>	157.3	169.6	Fossil [55]
Human—mouse	<i>Archonta—Glires</i>	61.5	100.5	Biostratigraphy [55]
Primates, mouse—dog, horse, cow	<i>Euarchontaglires—Laurasiatheria</i>	61.6	100.5	Fossil [55]
Dog—ferret	<i>Canidae—Arctoidea</i>	35	45	Fossil [56, 57]
Solenodon—hedgehog, shrew, mole	<i>Lipotyphla</i>	61.6	100.5	Fossil [55]
Cow—horse	<i>Artiodactyla</i> as soft minimum	52.4	100.5	Fossil [55]

The 4,416 single-copy orthologs identified in our assembly were used for phylogeny inference via 4-fold degenerate sites with programs RAXML [49] and PAML [50]. The resulting phylogenetic tree was plotted with FigTree [58] and is presented in Fig. 5.

corresponding assignments were obtained from the maNOG database for 7 other species: *H. sapiens*, *M. musculus*, *B. taurus*, *C. familiaris*, *Equus caballus*, *Mustela putorius furo*, and *Monodelphis domestica*. Inspection of assignments across all the species yielded 4,416 orthologous groups containing single-copy orthologous genes (Database S5).

### Species tree reconstruction and divergence time estimation

We used our genome assembly to infer phylogenetic relationships between *S. paradoxus* and other eutherian species with known genome sequences and estimated their divergence time using the new data. Based on the alignments of the single-copy orthologous proteins for the species included in the analysis, a maximum likelihood tree was built using RAXML [49] with the PROTGAMMAAUTO option and the JTT fitting model tested with 1,000 bootstrap replications. From the codon alignments of single-copy orthologs of the 11 species, 461,539 4-fold degenerate sites were extracted. The divergence time estimation was made by the MCMCtree tool from the software package PAML (PAML, RRID:SCR.014932) [50] with the HKY+G model of nucleotide substitutions and 2,200,000 generations of MCMC (of which the first 200,000 generations were discarded as burn-in). A test for substitution saturation [51, 52] was performed using DAMBE6 [53] for both all third codon positions and only 4-fold degenerated sites. In both cases the index of substitution saturation was significantly lower than the threshold value for both symmetrical and asymmetrical trees indicating low saturation level. Therefore, saturation was not detected for any of the third positions nor for the 4-fold degenerated sites.

Divergence times were calibrated using fossil-based priors associated with mammalian evolution, listed in Table 5 and based on [54–57]. FigTree [58] was used to plot the resulting tree, shown in Fig. 5. According to this analysis, *S. paradoxus* diverged from other mammals 73.6 Mya (95% confidence interval of 61.4–88.2 Mya). This is in accordance with earlier estimates based on nuclear and mitochondrial sequences (e.g., [3, 12]) as reviewed by Springer et al. [59]. This date is also much older than the timeframe of molecular estimates of divergence times between most island taxa and their closest mainland relatives [60]. Our data support solenodons forming a sister group to other eulipotyphlans, i.e., hedgehogs, shrews, and moles [61–64], with a divergence date as old as splits between some pairs of mammalian orders, such as between rodents and primates or between carnivores and artiodactyls (Fig. 5).

### Positively selected genes

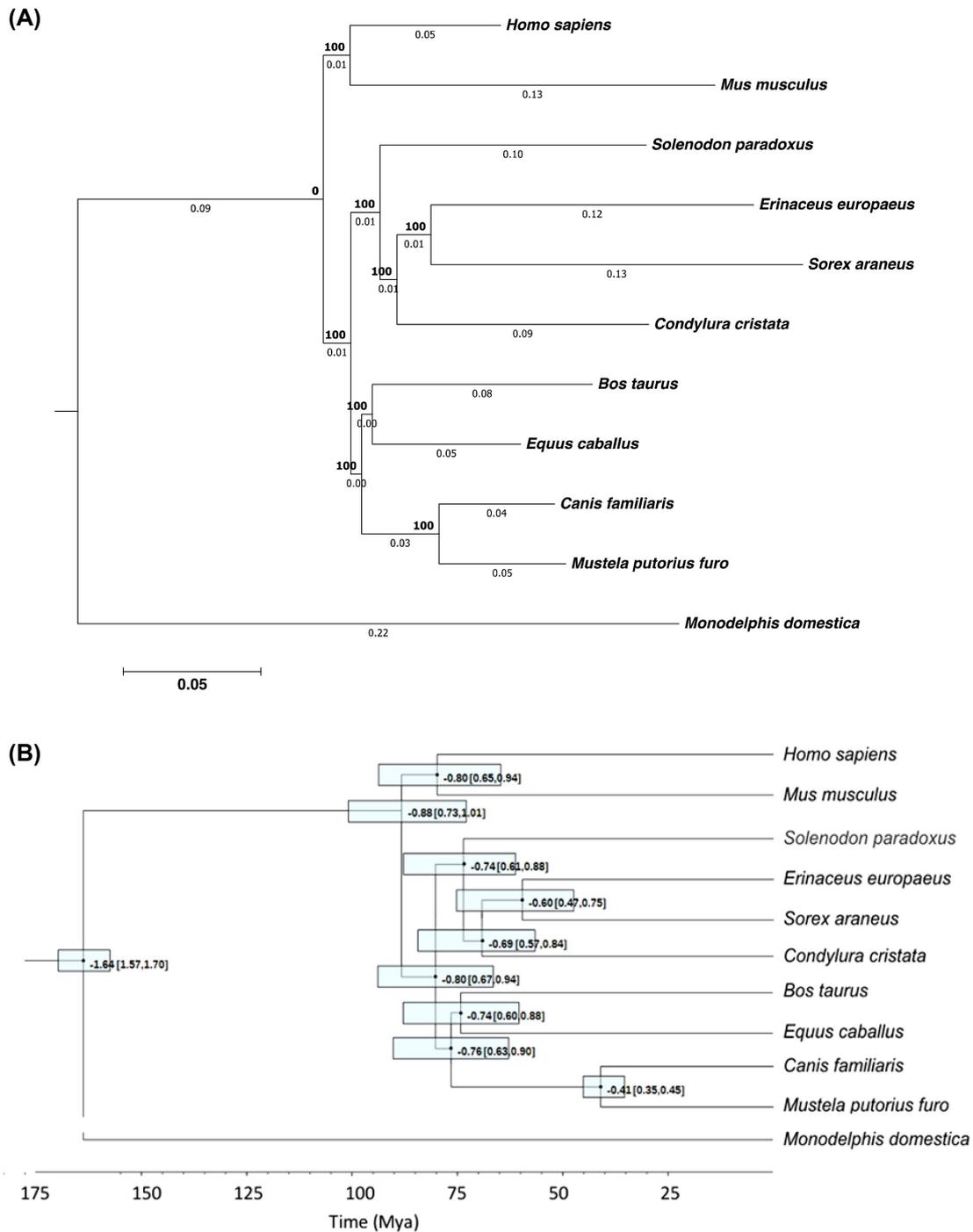
To evaluate signatures of selection in the assembled genomes, we used a dataset of 4,416 orthologous groups containing single-copy orthologous genes of the mammalian species described earlier. Single-copy orthologs were used as a conservative set necessary for comparing coding sequences that only arose 1 time in order to avoid the uncertainties associated with paralogs and lineage-specific gene duplications. First, we translated DNA sequences into amino acids, aligned them in MUSCLE (MUSCLE, RRID:SCR.011812) [65], and then translated them back into DNA code using the original nucleotide sequences by PAL2NAL [66]. Genic dN/dS ratios were estimated among the 11 mammalian species (including *Solenodon*) used in constructing the phylogeny represented in Fig. 5.

To estimate dN/dS ratios, we used the *codeml* module from the PAML package [50]. The dN/dS ratios were calculated over the entire length of a protein-coding gene. The branch-site model was not included in the current analysis because of the risk of reporting false positives due to sequencing and alignment errors [67], especially on smaller datasets, and additional uncertainties could be introduced from the lack of power under synonymous substitution saturation and high variation in the GC content [68].

All the single-copy orthologs were plotted in the dN to dS coordinates and color-coded according to the 96 gene ontology generic categories (Fig. 6). We retrieved values of dN, dS, and  $w$  ( $w = dN/dS$ ) for all single-copy orthologs and used human annotation categories to assign all the genes with their gene ontologies (GO) using the Python package *goatools* [69] and the GO Slim generic database [69] to assign the genes to the major GO categories.

The dN/dS values for the 12 genes exhibiting positive selection (Table 6) are visible above the line showing  $dN = dS$ . Three of these genes belong to the plasma membrane GO category (GO:0005886), while cytosol (GO:0005829), mitochondrial electron transport chain (GO:0005739), cytoplasm (GO:0005737), and generation of precursor metabolites (GO:0006091) were represented by 1 gene each. Five of the genes exhibiting positive selection signatures could not be assigned to GO categories. Some of these are also associated with the plasma membranes (*TMEM56*, *SMIM3*), and 1 gene (*CCRNL4*) encodes a protein highly similar to the nocturnin, a gene identified as a circadian clock regulator in *Xenopus laevis* [70]. The full list of genes, GO annotations, and associated dN/dS values are included in Database S6.

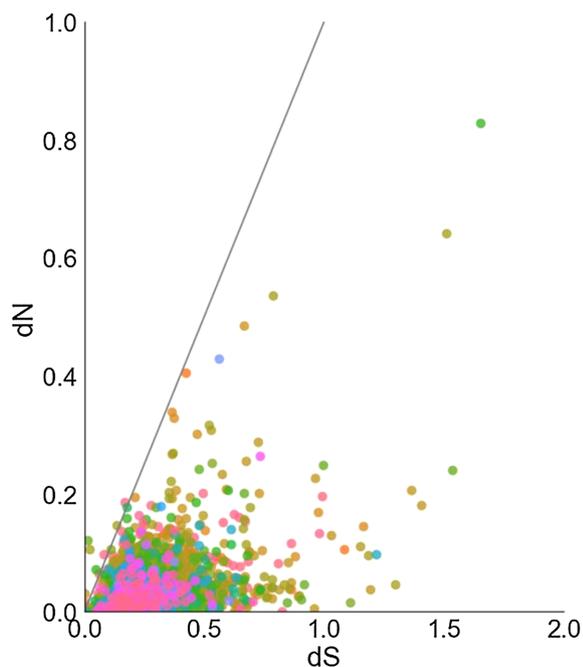
Traditionally, one of the most commonly used signatures of selection is the ratio of nonsynonymous (dN) to synonymous



**Figure 5:** Phylogenetic relationships of *Solenodon paradoxus* and other mammals from whole-genome data. (A) Maximum likelihood phylogeny showing branch lengths. The tree was built using RAxML [49] with the PROTGAMMAAUTO option and the JTT fitting model tested with 1,000 bootstrap replicates. (B) Divergence time estimates based on 461,539 4-fold degenerate sites from the codon alignments of single-copy orthologs and using fossil-based priors (Table 5). The divergence time estimation was made by the MCMCtree tool from the software package PAML [50] with the HKY+G model of nucleotide substitutions and 2,200,000 generations of MCMC (of which the first 200,000 generations were discarded as burn-in). The 95% confidence intervals are given in square brackets and depicted as semitransparent boxes around the nodes. The inferred divergence time of *S. paradoxus* from other mammals is 73.6 Mya (95% confidence interval of 61.4–88.2 Mya).

(dS) substitutions, dN/dS [71]. The synonymous rate (dS) expresses the rate of unconstrained, neutral evolution, so that when dN/dS < 1, the usual interpretation is that negative selection has taken place on nonsynonymous substitutions. Otherwise, when dN/dS > 1, the interpretation is that the positive

selection is likely to have accelerated the rate of fixation of nonsynonymous substitutions. It is possible to quantify the proportion of nonsynonymous substitutions that are slightly deleterious from the differences in dN/dS between rare and common alleles [72,73]. In our comparison, a subset of single-copy or-



**Figure 6:** The dN/dS ratios for 4,416 single-copy orthologous genes. The dN and dS ratios were calculated with the *codeml* module from the PAML package [50] and calculated over the entire length of a protein coding gene. Values are color-coded by GO term aggregated by the GO Slim generic database [69,106]; the color code legend is presented in Fig. S2. The solid black line represents  $dN = dS$ ; dots above it represent genes showing signatures of positive selection. The figure is truncated at  $dN = 1$  and  $dS = 2$ , so larger values are not shown on the graph, but all  $\omega$ , dN, and dS values are available in Database S6.

thologs dN/dS compared to the 10 mammalian species (Fig. 5) is estimated to be  $\sim 0.18$  or 18%, on average, compared to  $\sim 0.25$  reported for the human-chimp and  $\sim 0.13$  reported for the mouse-rat comparisons [74]. In other words, it suggests that up to 82% of all amino acid replacements in *S. paradoxus* are removed by purifying selection [74].

Note that purifying selection is the conservative force in molecular evolution, whereas positive selection is the diversifying force that drives molecular adaptation. Overall, the list of

positively selected genes is relatively short compared to numbers of positively selected genes reported in other studies (e.g., human-to-chimpanzee comparison yields several hundreds of human-specific genes under selection [75–77]). This observation could be a consequence of the averaging effect of a large comparison group that included mammals very distantly related to solenodons.

The dN/dS ratios can also be used as a proxy to illustrate the rate of evolution for proteins. By looking at the trends in fast evolved genes ( $dN/dS > 0.25$ ), we can make inferences about the factors that shaped the genome of this species during the millions of years of island isolation. To summarize the functional contributions, we used the PANTHER Overrepresentation Test and GO Ontology database based on the *H. sapiens* (Supplementary Table S1) and *M. musculus* (Supplementary Table S2) genes [78]. Interestingly, genes involved in the inflammatory response and located on cell surfaces were among those overrepresented among the rapidly evolving genes in the Solenodon genome compared to either the human or mouse databases (Supplementary Tables S1 and S2).

### Venom gene identification

Since solenodon is one of very few venomous eutherian mammals, of special interest in the solenodon genome were the putative venom genes. While there was no saliva sample in our possession that could be analyzed for the expressed toxin genes, a comparative genome approach could be applied as an indirect way to find venom genes orthologous to genes expressed in venom for other species. First, we identified 6,534 toxin and venom protein representatives (Tox-Prot) from Uniprot (UniProt, RRID:SCR.002380) [79] and queried them with BLAST against the current *S. paradoxus* genome assembly. The hit scaffolds were then extracted from the AUGUSTUS CDS prediction file. The same Tox-Prot sequences were used for Exonerate with the protein-to-genome model. The hits were used as queries against the NCBI database to ensure gene identity, further examined through phylogenetic analyses with select model mammalian and venom reptile genes (also adding randomly selected sequences for each gene to reduce clade bias). The retrieved sequences were aligned with MUSCLE [65], followed

**Table 6:** The putative targets of positive selection in the solenodon genome.

Solenodon gene	dS	dN	dN/dS	GO category description	Human ortholog
ENOG410UG5H	0.000003	0.002563	$\geq 999$	Plasma membrane	KLF9
ENOG410USMX	0.000011	0.010830	$\geq 999$	Plasma membrane	TNFSF13B
ENOG410UWRE	0.000015	0.014790	$\geq 999$	-	SMIM3
ENOG410UNED	0.000174	0.030411	174.84	-	CCRN4L
ENOG410UJP8	0.013214	0.120449	9.12	Cytosol	PLK4
ENOG410UWA9	0.020955	0.104972	5.01	Mitochondrion	NDUFC1
ENOG410V3Q6	0.047538	0.071112	1.50	Plasma membrane	SYT16
ENOG410UQAM	0.078543	0.096445	1.23	-	WBP2NL
ENOG410UKXY	0.168982	0.185535	1.10	-	TIGIT
ENOG410UKXJ	0.134581	0.146926	1.09	Cytoplasm	LRRC66
ENOG410UIAB	0.060622	0.065402	1.08	-	TMEM56
ENOG410UG23	0.176174	0.177344	1.01	Generation of precursor metabolites and energy	THTPA

The dN/dS values and the GO categories for the 12 genes that showed signatures of positive selection in the *Solenodon paradoxus woodi* genome ( $dN > dS$ ). All other genes are reported in Database S6.

**Table 7:** Homologous matches for the most relevant protein venom classes in the *Solenodon paradoxus* genome.

Protein groups found in animal venoms	Number of matches in the <i>S. paradoxus</i> genome
Metalloproteinase; serine protease	8 each
Hyaluronidase	6
(Acetyl) cholinesterase	2
Calglandulin; nerve growth factors	4 each
Lipase	3
Hydrolase; Kunitz serine protease inhibitor; nucleotidase; O-methyltransferase; oxidase; peptidase; phosphodiesterase; phospholipase; vascular endothelial growth factor	1 each

Genes were identified by querying 6,534 toxin and venom protein representatives found in animal venoms in Tox-Prot from Uniprot [79]. All of the protein groups are present in snake venoms. The sequences of the putative venom genes from *S. paradoxus* are available in the Database S7.

by a maximum likelihood (WAG+I+G) phylogenetic reconstruction. Hits were matched against their respective references in an alignment and visually inspected.

As a result, we identified 44 gene hits of the 16 most relevant protein venom classes (all present in snakes) in the *S. paradoxus* genome (Table 7). Inspection of pairwise MUSCLE alignments of the putative *Solenodon* venom genes (Database S7) with their animal homologs revealed several interesting cues. The putative venom genes could not be confirmed through genomic information alone, yet they cannot be discarded given that they were matched to high homology regions of closely related genes, such as those originally recruited into venom. There were also unusual insertions not found in other species' venom genes. Specifically, an insertion in a serine protease, a gene with a role in coagulation (namely, coagulation factor X), is not present in known homologs. The insertion seems to be located at the start of the second exon. This particular gene was further analyzed to understand the insertion and its potential functional consequences (Fig. 7). Finally, none of the known venom genes from the closest related venomous insectivore (*Blarina brevicauda*) were found in our study. Our results indicate that a more detailed study of *Solenodon* venom genes using a transcriptome obtained from a fresh saliva sample is needed to address their molecular evolution and function.

### Genomic variation and demographic history inference

Once the reference alignment was assembled as a consensus between the sequences obtained from the 5 *S. p. woodi* individuals, polymorphisms were identified in the 6 individual genomes by aligning them to the combined reference. Single-nucleotide and short variants and indels were identified in 5 southern and 1 northern individual using Bowtie2 (Bowtie, RRID:SCR\_005476) [80], SAMtools and Bcftools (SAMTOOLS, RRID:SCR\_002105) [81], and VCFtools (VCFtools, RRID:SCR\_001235) [82]. The *S. p. woodi* individuals differed from the reference by an average of 1.25 million polymorphisms, and the *S. p. paradoxus* individual differed by 2.65 million from the reference assembly.

Whole *solenodon* genome single nucleotide variation (SNV) rates, defined as a ratio of all observed SNVs to all possible SNV sites in the genome, were calculated and found to be comparatively low relative to other mammals (Fig. 8) [83–86]. To enable this comparison, the same calculations were used, where SNVs

were not filtered by repetitive regions or mappability mask, and the number of possible SNV sites was defined as the genome assembly size minus the number of unknown base pairs ('N').

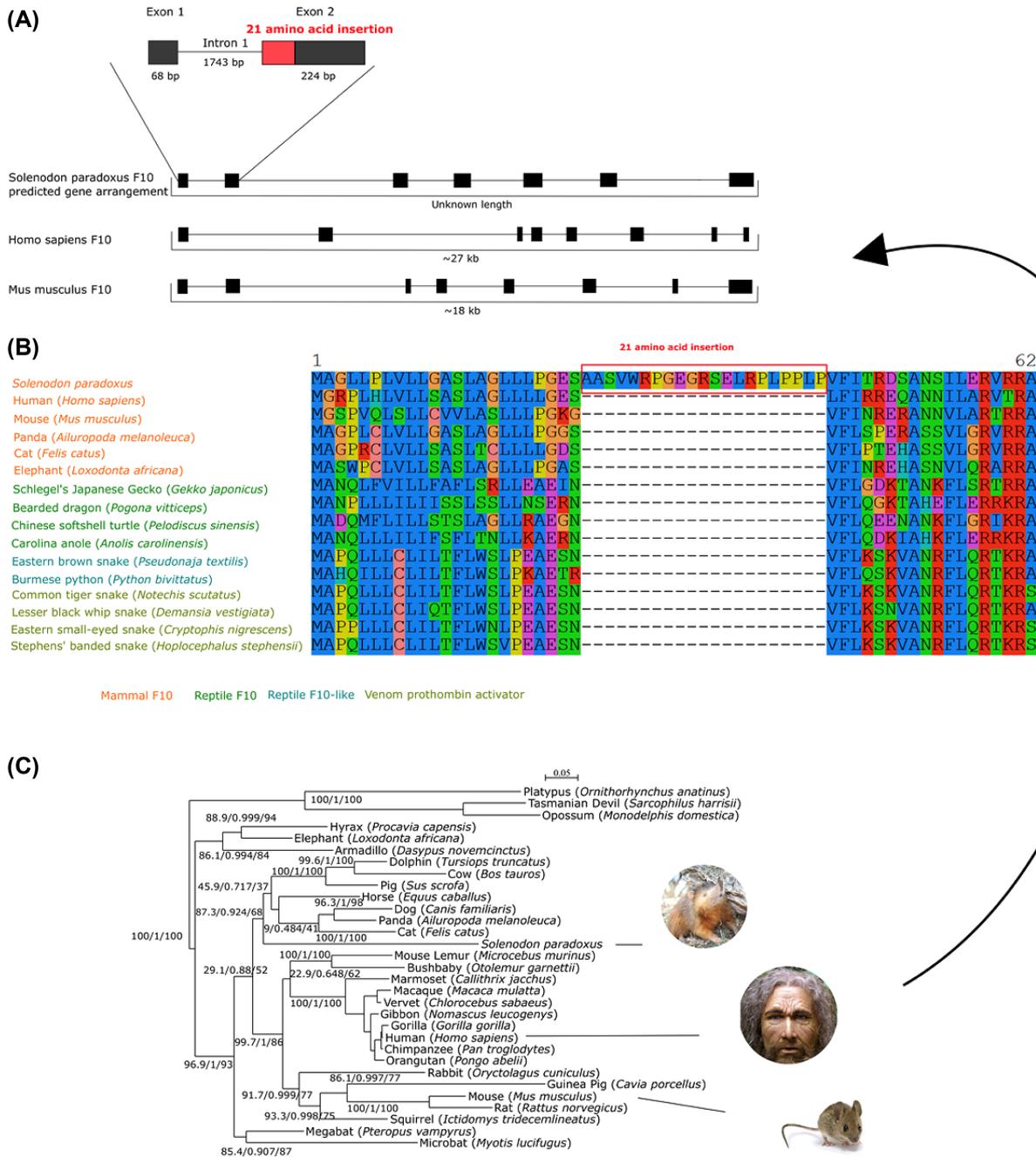
Based on the variation data from the genomes of 2 subspecies (*S. p. woodi* and *S. p. paradoxus*), we estimated population dynamics using pairwise sequentially Markovian coalescent (PSMC) model [87]. PSMC uses a coalescent approach to estimate changes in population size that allowed us to create a (The Most Recent Common Ancestor) TMRCA distribution across the genome and estimate the effective population size ( $N_e$ ) in recent evolutionary history (e.g., from 10,000 to 1 million years).

Demographic history was inferred separately for *S. p. woodi* and *S. p. paradoxus*, and the resulting plots revealed differences in demographic histories of the 2 subspecies (Fig. 9). Each southern individual was considered separately and their demographic histories overlapped. The difference in demographic history provides another argument in favor of a subspecies split, as evidenced by distinctly different effective population sizes at least since 300 Kya. According to this analysis, the northern *solenodon* subspecies currently has a much larger  $N_e$ , which has expanded relatively recently, between 10,000 and 11,000 years ago (Fig. 9). Prior to that, it was the southern subspecies (*S. p. woodi*) that had a larger  $N_e$ . At the same time, the demographic history inferred for both populations showed similar cyclical patterns of expansion and contraction around the mean of 6,000 "effective" individuals for the southern subspecies (*S. p. woodi*) and 3,000 for the northern subspecies (*S. p. paradoxus*). One unusual result of this analysis is that the northern subspecies shows a much lower  $N_e$  for all but the most recent time period.

### Development of tools to study population and conservation genetics of *S. paradoxus*

The presence of genome-wide sequences of multiple individuals from 2 subspecies created a possibility for the development of practical tools for conservation genetics of this endangered species. Generally, microsatellite loci are both abundant and widely distributed throughout the genome, while usable loci are characterized by a unique flanking DNA sequence so that a single locus can be independently amplified in many individuals [88–90]. The major advantages of microsatellite markers are well known: codominant transmission, high levels of polymorphisms leading to the high information content, high mutation rates that allow differentiation between individuals or populations within a species, and ease of genotyping. While a genome obtained from one individual can be searched for potentially variable microsatellite loci, this would (1) miss the majority of variable loci not represented in the individual's 2 chromosomes and (2) result in many positives that may be monomorphic following laboratory tests (usually by electrophoresis of the amplified fragments from population samples). The availability of several genomes can allow generation of a more comprehensive set of variable markers, while reducing false positives.

All 3 assemblies from this study (A, B, and C) were independently analyzed using a short tandem repeat (STR) detection pipeline. A, B, and C assemblies were analyzed separately with Tandem Repeats Finder to locate and display tandem repeats [91]. Each of the 6 individual samples from the 2 *solenodon* subspecies (5 from *S. p. woodi* and 1 from *S. p. paradoxus*) were aligned to the reference assemblies A, B, and C by Burrow-Wheelers Aligner [81]. Each set of individual alignments was analyzed with HipSTR [92]. Only loci that shared more than 20 reads in the sample alignments were considered for further steps in the search

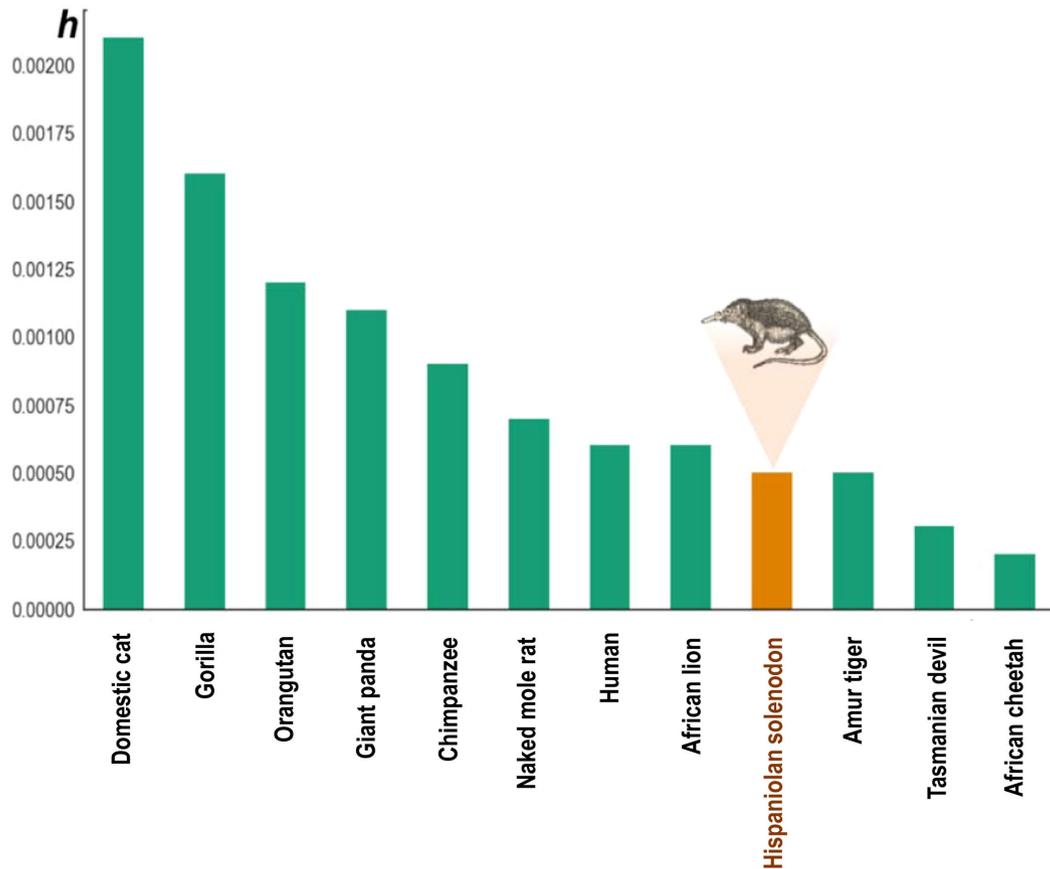


**Figure 7:** (A) Predicted coagulation factor X (F10) gene structure arrangement from the structure of known homologs (due to the scaffolding, the total gene length is unknown in solenodon). The 21 codon insertion is highlighted in red on exon 2 of the solenodon F10 gene. Exons are represented as black boxes and introns as lines connecting exons. (B) F10 protein sequence alignment showing an unusual insertion in the *Solenodon paradoxus* genome absent in all other mammalian and reptilian genes retrieved from the Tox-Prot from Uniprot [79]. The insertion of 21 amino acids is indicated with a red-boxed line in the alignment. (C) Reconstructed mammalian F10 phylogenetic maximum likelihood tree using the model GTR+I+ $\Gamma$ , 1,000 bootstrap replicates (1,590 bp long alignment). The numbers set indicate approximate likelihood-ratio branch test, Bayesian-like modification of the aLRT, and bootstrap percentage, respectively.

for variable microsatellite loci. The result of this search was saved in a variant call format file that includes annotations of all loci that had variation between samples and passed the minimum qualification of the reads parameter: to be successfully identified *in silico* in the data from at least 1 individual. The loci that did not pass these criteria were labeled as *unsuccessfully verified* and excluded from the list.

The remaining loci were subjected to additional filtering; all genotypes that had less than 90% posterior probability according

to HipSTR [92], genotypes with a flank indel in more than 15% of reads, and genotypes with more than 15% of reads with detected polymerase chain reaction (PCR) stutter artifacts were discarded. The final set containing loci that have at least 2 allele calls in 2 different individuals after filtering has been deposited in the polymorphic microsatellite database (Database S8). This database contains a list of variable microsatellites discovered, a total of 1,200 bp flanking sequences for primer construction, and the information on whether and where it was found to be



**Figure 8:** Low genome heterozygosity in *Solenodon paradoxus woodi* compared to other mammalian taxa. The SNV rate in the *S. p. woodi* genome is shown relative to other mammal genomes as an estimate of genome diversity ( $h$ ). The value for each sequenced individual was estimated using all variant positions, with repetitive regions not filtered. The SNVs are deposited in Database S9.

variable—between subspecies or within 1 of the subspecies. We also report the type (di-, tri-, etc.), number of repeats, number of variants, and % variable and provide up to 600 bp flanking sequences on each side that can be used to develop primer sequences (Database S8).

## Discussion

In this study, we sequenced and assembled the genome of an endangered Antillean mammal that survived tens of millions of years of island isolation but nevertheless is currently threatened with extinction due to anthropogenic activities. Our approach demonstrated sequencing, assembly, and annotation of a genome of a highly divergent lineage within the placental mammal tree, delivering an important phylogenetically diverse mammalian genome for analysis in a comparative context [93]. Although the full description of genome diversity of this rare enigmatic mammal needs to be further improved with more samples and analyses, our initial assembly of the solenodon genome contributes information and tools for future studies of evolution and conservation. Future studies can combine the current genome annotations with the inclusion of additional genetic and ecological data from further sampling.

With the new genome-wide assembly, we inferred a phylogeny that validates previous estimates of the time of divergence of *Solenodon* from other eulipotyphlan insectivores [3, 12],

also providing a window into genetic underpinnings of adaptive features, including genes responsible for inflammation and venom and how these may reflect its adaptation. In addition, we developed tools that will help guide future genome studies as well as conservation surveys of the remaining solenodon populations on the island of Hispaniola. In this study, we have made the first step into the whole-genome analysis of the *Solenodon*. A more complete genome sequence may provide a better picture of its evolutionary history, possible signatures of selection, and clues about the genetic basis of adaptive phenotypic features facilitating life on Caribbean islands and contribute to a better insight into island evolution and possible responses to current and future climate change.

## The string graph assembly approach for homozygous genomes

The advantages of the string graph assemblies in our particular case can be understood by looking at the nature of the underlying algorithms. The de Bruijn graph is a mathematical concept that simplifies genome assembly by reducing information from short next-generation sequencing reads, of which there can be billions, to an optimized computational problem that can be solved efficiently [94]. However, some information may

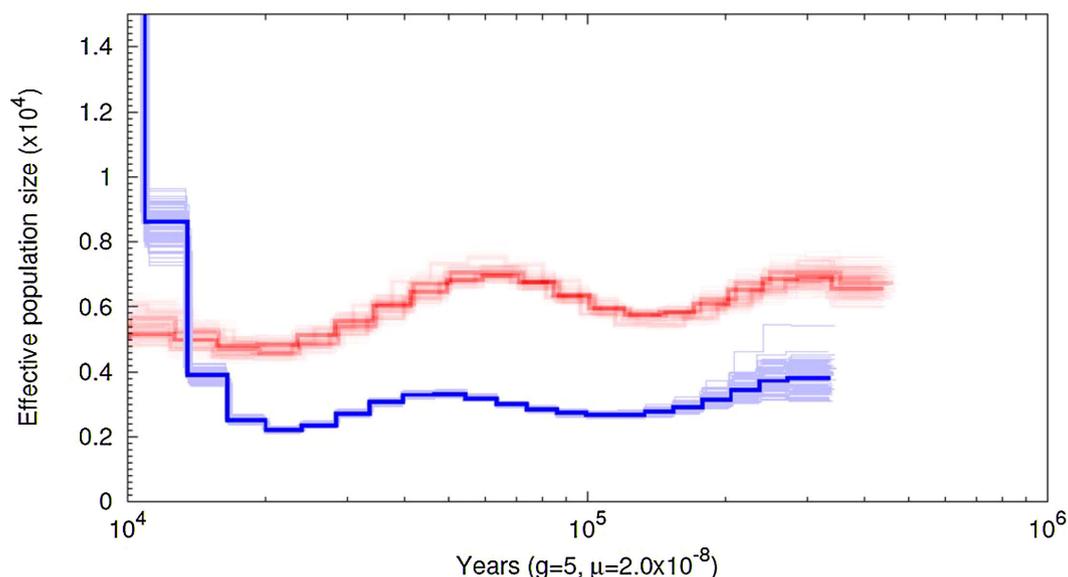


Figure 9: Demographic history inference for the southern *S. p. woodi* (red) and the northern *S. p. paradoxus* (blue) subspecies using the pairwise sequentially Markovian coalescent model [87].

indeed be lost, as the set of reads is effectively replaced with a set of much shorter  $k$ -mers to produce an optimal assembly path. Usually, this is compensated by overwhelming amounts of data in high coverage assemblies, and the difference in effectiveness between this and other types of algorithms, barring speed, becomes less evident. While sequencing becomes cheaper, genome projects continue to rely on the increased high-quality coverage, increasing the cost of the sequence data rather than trying to increase the efficacy of the assembly itself. In contrast, the string graph-based algorithms for genome assembly are intrinsically less erroneous than de Bruijn graph-based ones, since building and resolving a string graph does not require breaking reads into  $k$ -mers and therefore does not sacrifice long-range information [19]. This also helps reduce the probability of misassemblies; in theory, any path in a string graph represents a valid assembly [95, 96]. String graph-based approaches have already been applied successfully to assemblies from high coverage read sets; and one example is the Assemblathon 2 [97]. In projects with lower genome coverage like ours, adoption of a string graph-based approach might be of benefit to the genome assembly because it uses more information from the sequences. However, there are 2 major downsides for its widespread use: (1) it is more computationally intensive than methods utilizing de Bruijn graph algorithms and (2) the implementation of the string graph model is sensitive to sequence variation, and the effectiveness of this approach may depend on the level of heterozygosity in a DNA sample. It is worth noting that [19] was primarily intended for variant annotation via *de novo* local assembly, and not for whole genome assembly. Nevertheless, the new genome-wide data produced by our pipeline were sufficient for the comparative analysis and have been annotated for the genes and repetitive elements and interrogated for phylogeny, demographic history, and signatures of selection. In addition, using the current genome assembly, we were able to annotate large transpositions and translocations in the *Solenodon* in relation to the closest available high-quality genome assembly (*S. araneus*).

## Potential Implications

### Comparative genomics

We have taken advantage of the fact that the genome of this mammal shows reduced heterozygosity [12], which made it feasible to combine samples of multiple individuals in order to provide higher coverage and achieve a better assembly using Illumina reads. The current assembly was performed without the use of mate pair libraries and without high-quality DNA, nevertheless, it is comparable in quality to other available mammalian assemblies. In terms of contig N50 as a measure of contiguity, our assembly resulted in contig N50 of 54,944, while the most closely related available genome sequences of *Sorex araneus* (SorAra2.0) assembly features a contig N50 of 22,623, and the *Condylura cristata* (ConCri1.0) assembly has contig N50 of 46,163. It should be noted that scaffold N50 values are not to be compared as this study used only paired-end reads, as opposed to *S. araneus* and *C. cristata*. More importantly, the assembly provided complete or partial annotation for more than 95% of the genes based on the evolutionarily informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB v9 by BUSCO [29]. Among these, 4416 single-copy genes that have clear 1-to-1 orthologs across species (single-copy orthologs) [98, 99] were chosen for a subsequent comparative analysis involving genes in different mammalian species.

Specifically, the repetitive composition of the solenodon genome was evaluated. Compared to the estimates based on the reference human genome [100], very conspicuous is the lower numbers of SINES (no *Alu* elements) and a substantially lower number of LINES as well. Transpositions and translocations between the genomes of *S. paradoxus* and *S. araneus* were identified; very few rearrangements and translocations between the assembly and the *S. araneus* genome were found. At the same time a higher coverage would be needed to do more detailed analyses, for instance, to address the relative length and similarity of indels and copy number polymorphisms between solenodon populations [101].

## Evolutionary genomics

Using the nuclear genomes, we were able to confirm earlier divergence time estimates based on sets of genes [3], as well as full mitochondrial sequences [12]. The whole genome analysis points to a split between *Solenodon* and other eulypotiplans that occurred around 74 Mya (Fig. 5), which is very close to our earlier estimate of 78 Mya based on the full mitochondrial genome [12]. Our result does not support the 60 Mya point estimate made by a phylogenetic analysis based on sequences of 5 slowly evolving nuclear genes [13].

Our assembly provided enough gene sequences to gain insights into the evolution of functional elements in the solenodon genome. It is reasonable to suggest that this species historically had low effective population sizes, if they remained close to those estimated by this study, or about 4,000 on average (Fig. 9). Among the 4,416 single-copy orthologs analyzed for dN/dS ratios over the entire length of a protein-coding gene between *S. paradoxus* and 10 other mammals, 12 genes were identified as positively selected. Among these, the majority were membrane proteins, with 1 gene (*CCRNL4*) similar to a circadian clock regulator (Table 6). It is possible that the short list of the positively selected genes could be a consequence of the large comparison group that included mammals very distantly related to solenodon, and its genes need to be compared with more closely related species, e.g., once the genome of *S. cubanus* is reported and better gene annotations for *Sorex araneus* become available.

*Solenodon* is one of few mammals that use venomous saliva to disable prey. It delivers its venom similarly to snakes—using its teeth to inject venomous saliva into its target. Different approaches could be used to characterize venom genes, such as the use of noncurated databases to widen the search spectrum, which may include different molecules that could be found in *Solenodon*. For example, 6,534 toxin and venom protein representatives can be found in the UniProt database. It is also important to note that the database of venom gene sequences may not include those relevant to solenodons given their deep divergence from any other venomous mammalian species. The venom of *Solenodon* may contain novel protein modifications with unknown potential or application, making it valuable for future detailed characterization.

Genes associated with venom, such as serine proteases involved in coagulation (namely, the coagulation factor X), are of major interest since factor X in solenodon exhibited unusual insertions when compared to its homologs (Fig. 7). The detection of an unusual insertion in a serine protease has been previously found in another venomous mammalian species, the shrew *Blarina brevicauda*, but in a different gene than in solenodon. The coagulation factor X is involved in the circulatory system and is responsible for activating thrombin and inducing clotting. The insertion in the coagulation factor X gene seems to be a hydrophilic alpha helix with 3 potential protein–protein interaction sites. It occurs at the end of the region annotated as the signal peptide, while having a signal peptide cleavage site itself at the beginning of its sequence. The factor X protein structure was successfully modeled by Swiss-Model based on the venomous elapid snake *Pseudonaja textilis* (pdb: 4bxs) to have a heavy chain that contains the serine protease activity, which was modeled with a high degree of confidence (Fig. 10). The venom prothrombin activator has an advantage as a toxin in part due to modifications in inhibition sites, making it difficult to stop its activity. Another advantage is that the molecules are always found in an active form (Kinin). We hypothesize that the inser-

tion could allow a more successful interaction with molecules capable of activating the F10 protein. In mice, venom extracted from solenodons and venom prothrombin activator injections can both be lethal in minutes [7, 102]. The insertion was also searched against possible mobile DNA elements, but no matches were found. Our results should be followed in the future by detailed pharmacological studies.

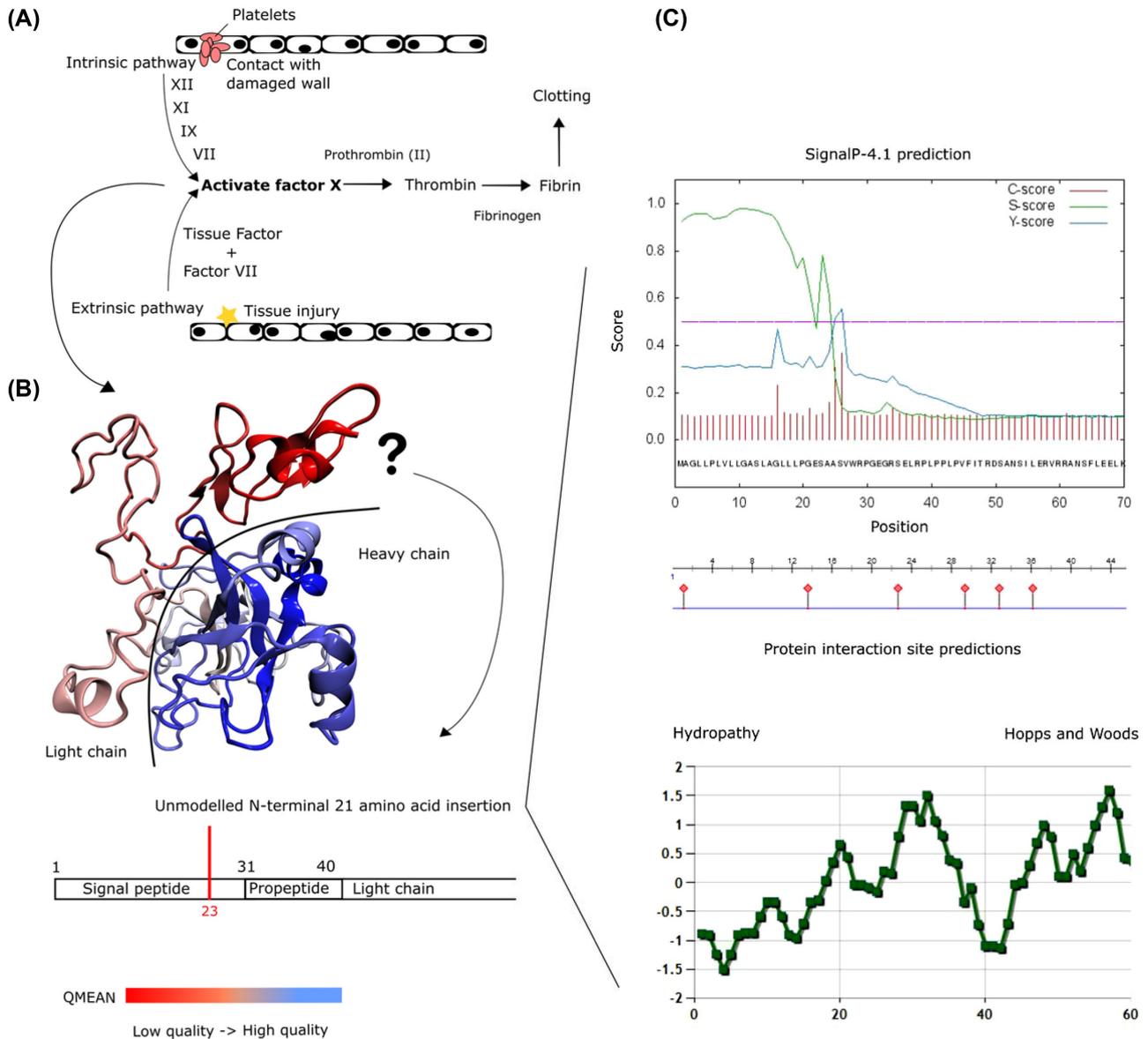
## Conservation genetics

The low variation that exists between the solenodon sequences is hardly surprising, because the theoretical consensus in conservation genetics predicts that small populations lose genetic diversity more rapidly than large populations [103] and measures of genetic diversity have been explicitly suggested to IUCN as a factor to consider in identifying species of conservation concern [104]. The historical  $N_e$  for each subspecies was examined by our analysis (Fig. 9) and showed lower levels recently in *S. p. woodi*. Due to the limitations of PSMC, the most recent  $N_e$  cannot be calculated from the genome sequences [87]. Therefore, this estimate of diversity does not reflect the recent impact on the solenodon population caused by anthropogenic factors in the last 10,000 years (Fig. 9).

Many endangered species with small populations also have reduced heterozygosity across their genomes and would benefit from a computational approach that reduces the cost and optimizes the amount of data for the genome assembly. The real-life scenarios where no high-quality DNA can be produced because of the remoteness of sampling location, when transportation and storage are difficult, or when the high coverage cannot be produced due to the limited funds are well known to many, especially in the field of conservation genetics. The difficult field conditions and international regulations make it difficult to obtain samples with high-molecular-weight DNA. To aid future conservation studies, we have mined the current dataset for microsatellite markers that are useful within and between subspecies, to be used as tools for studies on population diversity, censoring, and monitoring.

The comparative analysis of the number and the length of microsatellite alleles pointed once more to the advantage of assembly B over A and C. The average length of microsatellite short tandem repeats in assembly B was the highest: 20.95 (assembly A) vs. 21.14 (assembly B) vs. 18.86 (assembly C). This may be a direct consequence of the high number of microsatellite alleles that were successfully genotyped in all of the southern samples for assembly B (2,660), as well as microsatellites that proved variable between the 2 subspecies but fixed within the southern samples (639). The low number of variable microsatellites between the 2 subspecies was likely due to the reduced amount of information obtainable from a single low coverage genome of the northern subspecies (*S. p. paradoxus*) used in this study. Venn diagrams showing overlap in microsatellite variation in 3 assemblies are presented in Fig. 11.

Recently, a genetic survey using mitochondrial cytochrome b and control region sequences from 34 solenodon samples identified distinct haplotypes in northern and southern Hispaniola [16], along with a distinctive third group, a small remnant population at the Massif de la Hotte in the extreme western tip of Haiti [16, 105] not sampled for this study. The north–south subspecies subdivision within *S. paradoxus* was further supported by mitogenomic sequences [12]. The island of Hispaniola has been divided into 3 main biogeographic regions that differ in climate and habitat. The north and center of the island provide the

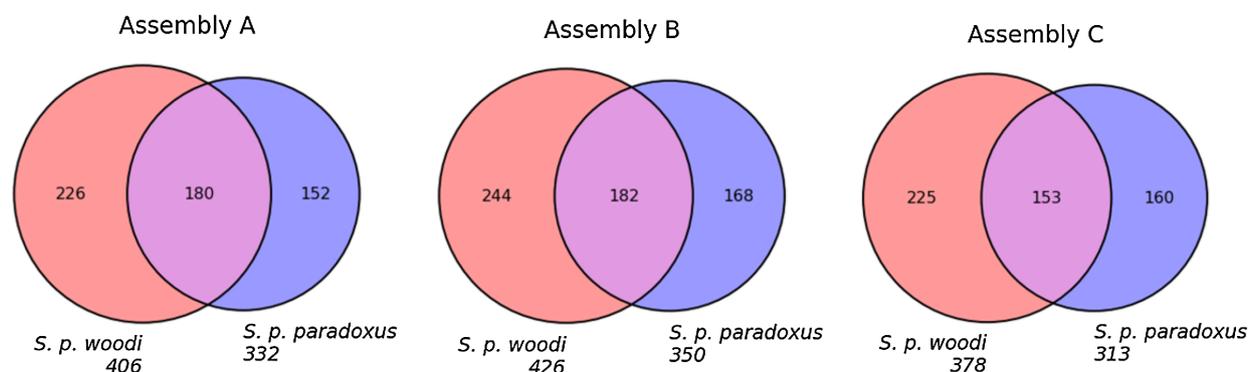


**Figure 10:** (A) Simplified version of the coagulation cascade, indicating key steps involving the coagulation factor X (F10). (B) Protein modeling of solenodon sequence data using SWISS-MODEL. The target model (4bxs) used was the F10-like protease of the venomous elapid snake *Pseudonaja textilis*. Due to its location, the insertion cannot be represented in the model (its location is indicated according to the PDB annotation). Colors indicate model quality, with red being low-quality and blue high-quality modeling. Colors also separate F10's light chain (EGF-like domain) in red from the heavy chain (serine protease domain) in blue (the half circle line in black separates both domains). (C) Amino acid sequence properties calculated for the solenodon F10 translated gene, with focus on the insertion region 23–43. One signal peptide cleavage site was detected between positions 25 and 26. Predicted protein interaction sites at position 26, 29–30 and 32–40. Hydrophathy analysis showed a relatively hydrophilic structure for the insertion.

largest area with known solenodon populations and shows no discontinuity with the southeast. However, the solenodon populations in the southwestern part of the island are currently geographically isolated by Cordillera Central and may have been isolated in the past by the ancient marine divide across the Neiba Valley (Fig. 2). This geographic isolation is likely the reason why the *S. p. paradoxus* in the larger northern area and *S. p. woodi* in the southwest show morphological differences suggestive of separate subspecies [15]. Future conservation strategies directed at protecting and restoring solenodon populations on Hispaniola should take into consideration this subdivision and treat the 2 subspecies as 2 separate conservation units.

## Methods

Provenance of the samples is shown on the map (Fig. 2), with coordinates listed in Supplementary Table S4. Solenodons were caught with help of local guides (Nicolás Corona and Yimell Corona). During the day, potential locations were inspected in daylight for animal tracks, burrows, droppings, and other signs of solenodon activity. At dawn, ambushes were set up in the forested areas along the potential animal trails. The approaching solenodons were identified by sound and chased with flashlights when approached. Since solenodons move slowly, animals were picked up by their tails, which is the only way to avoid potentially venomous bites. All wild-caught animals were released



**Figure 11:** Numbers of variable microsatellite alleles discovered in *S. paradoxus* assemblies. The diagrams were built independently for Fermi-based assemblies (A) and (B) and 1 SOAPdenovo2-based assembly (C). The red circle indicates microsatellites that were successfully genotyped in all samples with at least 1 alternative allele in the southern subspecies (*S. p. woodi*). The blue circle indicates microsatellites that were successfully genotyped in all samples with at least 1 alternative allele in the northern subspecies (*S. p. paradoxus*). The overlap indicates microsatellite loci with at least 1 alternative variant found in both subspecies. All alleles discovered, number of fixed alleles in each population, and number of unique alleles in each population are presented in Table S3. All the candidate microsatellite loci discovered in this study, along with their 5' and 3' flanking regions, are listed in the Database S8.

back into their habitats within 10 minutes after their capture. Before the release, the animals' tails were marked with a Sharpie pen to avoid recapturing.

Blood was drawn by a licensed ZooDom veterinarian (Adrell Núñez) from the *vena jugularis* using a 3-mL syringe with a 23G x 1-inch needle. The blood volume collected never exceeded 1% of body weight of animals. Before the draw, an aseptic technique was applied using a povidone-iodine solution, followed by isopropyl alcohol. Once collected, the samples were transferred to a collection tube with anticoagulant (BD Microtainer, 1.0 mg K2EDTA for 250–500 L volume). Collection tubes were refrigerated and transported to the lab at the Instituto Tecnológico de Santo Domingo where DNA was extracted from samples using the DNeasy Blood and Tissue kit (Qiagen, Hilden, Germany).

This study was reviewed and approved by the Institutional Animal Care and Use Committee of the University of Puerto Rico at Mayagüez. All the required collection and permits had been obtained before any field work was started. The samples were collected and exported in compliance with export permit VAPB-00909 (Dominican Republic Environment and Natural Resources Ministry Viceministry of Protected Areas and Biodiversity Department of Biodiversity) and imported in compliance with CITES/ESA import permit 14US84465A/9 (University of Illinois Board of Trustees).

## Sequencing

Sequences for *S. p. woodi* were generated by Illumina HiSeq 2000 (Illumina Inc) with 100 bp paired-end reads. The Illumina HiSeq generated raw images utilizing HiSeq Control Software (HCS) v2.2.38 for system control and base calling through an integrated primary analysis software called RTA (Real Time Analysis. v1.18.61.0). The base call binaries were converted into FASTQ utilizing the Illumina package bcl2fastq (v1.8.4). Sequences for *S. p. paradoxus* were generated by the Illumina MiSeq V3 (Illumina Inc.) at the Roy J. Carver Biotechnology Center, University of Illinois. The sequencing data for each sample used in this study are presented in Supplementary Table S5.

## Availability of supporting data

Database S1: Lists of repeats in the solenodon genome (assemblies A and B)

<http://public.dobzhanskycenter.ru/solenodon/repeats/solpar-a.txt>

<http://public.dobzhanskycenter.ru/solenodon/repeats/solpar-b.txt>

Database S2: List of protein coding genes in the solenodon genome (assembly B)

<http://public.dobzhanskycenter.ru/solenodon/genes/solpar-b.gff> also cds for each gene and translated sequences

Database S3: List of the annotated non-coding RNAs in the solenodon genome <http://public.dobzhanskycenter.ru/solenodon/rna>

Database S5: List of single-copy orthologs in the solenodon genome (columns include: ENOG id, gene name) <http://public.dobzhanskycenter.ru/solenodon/monoorthologs.txt>

Database S6: List of genes with dN/dS values and GO annotations

<http://public.dobzhanskycenter.ru/solenodon/selection.xls>

Database S7: List of venom genes [http://public.dobzhanskycenter.ru/solenodon/venom\\_genes.HitGeneDB.fasta](http://public.dobzhanskycenter.ru/solenodon/venom_genes.HitGeneDB.fasta)

Database S8: Microsatellite loci discovered in genomes of two solenodon subspecies *Solenodon paradoxus paradoxus* (northern) and *S. p. woodi* (southern), alleles, 600bp flanking regions (a total of 1,200 bp per locus), and frequency information for the two subspecies <http://public.dobzhanskycenter.ru/solenodon/STRs.xlsx>

Database S9: Lists of single nucleotide differences from the assembled individual genome of Spa-1 (from *Solenodon paradoxus paradoxus*) and Spa K,—L,—M,—N,—O (from the five *S. p. woodi*) used to show estimates of heterozygosity in Fig. 8 (see explanation in text)

<http://public.dobzhanskycenter.ru/solenodon/variants>

Supporting raw data is in the NCBI Sequence Read Archive [ENA: NKTLO1000000, PROJECT: PRJNA368679] and genome assemblies, custom codes and annotations are in the GigaScience GigaDB database [22].

## Additional file

Figure S1: The phylogenetic tree used for multiple genome alignment with Progressive Cactus (Paten et al. 2011). The taxa have been chosen based on their availability and the quality of genome assembly, not to make inferences about mammalian phylogeny. This cladogram only shows tree topology, and the

branches do not represent evolutionary time, and do not assume the basal position of *Solenodon paradoxus*.

Figure S2: The colors assigned to GO terms represented in Fig. 4.

Table S1: Classification of the fast-evolving genes ( $dN/dS > 0.25$ ) in solenodon genome using PANTHER Overrepresentation Test (release 20160715) and GO Ontology database (Released 2017-02-28) based on the *Homo sapiens* genes. Genes are only represented if  $P < 0.05$  after the Bonferroni correction for multiple testing (Treangen and Salzberg, 2012).

Table S2: Classification of the fast-evolving genes ( $w > 0.25$ ) in solenodon genome using PANTHER Overrepresentation Test (release 20160715) and GO Ontology database (Released 2017-02-28) based on the *Mus musculus* genes. Genes are only represented if  $P < 0.05$  after the Bonferroni correction for multiple testing (Treangen and Salzberg, 2012).

Table S3: Microsatellite alleles discovered in genomes of two solenodon subspecies *Solenodon paradoxus paradoxus* (northern) and *S. p. woodi* (southern).

Table S4: Locations for the samples used in this study as described in Fig. 2. Adopted from Brandt et al., (2016).

Table S5: Sequencing data for the samples used in this study.

## Abbreviations

BLAST: Basic Local Alignment Search Tool for Proteins; BUSCO: benchmarking universal single-copy orthologs; CDS: coding sequence; GO: gene ontology; IUCN: International Union for Conservation of Nature; Kya: thousand (kilo) years ago; Mya: million years ago; NCBI: National Center for Biotechnology Information; PSMC: pairwise sequentially Markovian coalescent; SNV: single nucleotide variation; STR: short tandem repeat; UCN: International Union for Conservation of Nature.

## Competing interests

The authors declare that they have no competing interests.

## Author contributions

KG, TKO, JCMC and ALR conceived and designed the study. KG, SK, PD, AK, WW, FS, and TKO did the analysis. YMAH, ALB, LAP, RC, AN, JRB, JCMC and TKO collected, processed, and contributed samples in the field and the laboratory, KG, AJM, KG, AA, ALR, SJO, JCMC and TKO wrote the draft. All authors contributed to finalizing of the manuscript.

## Funding

Authors at the University of Puerto Rico at Mayaguez were supported in part by a National Science Foundation award (1432092). Authors at the Theodosius Dobzhansky Center for Genome Bioinformatics were supported by a Russian Ministry of Science mega-grant (11.G34.31.0068) and a St. Petersburg State University grant (1.50.1623.2013).

## Acknowledgements

The authors thank Nicolás Corona and Yimell Corona for assistance in collecting samples. A special thanks to Kara Fore for helping to edit the manuscript. Sample collection in the Dominican Republic was performed under permit 00201171139 from the Ministry of Environment and Natural Resources.

## References

- MacPhee RDE, Flemming C, Lunde DP. "Last occurrence" of the Antillean insectivoran *Nesophontes*: new radiometric dates and their interpretation. *American Museum novitates*; no. 3261. New York, NY: American Museum of Natural History; 1999.
- Ottenwalder JA. Systematics and biogeography of the West Indian genus *Solenodon*. *Biogeogr. West Indies Patterns Perspect.* Second Ed. CRC Press; 2001, 253–329.
- Roca AL, Bar-Gal GK, Eizirik E et al. Mesozoic origin for West Indian insectivores. *Nature*, Nature Publishing Group; 2004;**429**(6992):649–51.
- Verill AH. Notes on the habits and external characters of the solenodon of San Domingo (*Solenodon paradoxus*). *American Journal of Science* 1907;**XXIV**(139):55–57.
- Allen JA. Notes on *Solenodon paradoxus* Brandt. *Bull Am Museum Nat Hist* 1908;**XXIV**:505–5017.
- Brandt JF. De *Solenodonte*: novo mammalium insectivororum genere. *Mem. l'Académie Impériale des Sci. St. Pétersbg. l'Académie Impériale des Sciences de St. Pétersbourg*; 1833;**2**:459–78.
- Derbridge JJ, Posthumus EE, Chen HL et al. *Solenodon paradoxus* (Soricomorpha: Solenodontidae). *Mammalian Species* 2015;**47**(927):100–6.
- Feldhamer GA, Drickamer LC, Vessey SH et al. *Mammalogy: Adaptation, Diversity, and Ecology*. 4th ed. Johns Hopkins University Press, Baltimore, Maryland. 2015, p. 768
- Wible JR. On the cranial osteology of the Hispaniolan solenodon, *Solenodon paradoxus* Brandt, 1833 (Mammalia, Lipotyphla, Solenodontidae). *Annals of Carnegie Museum* 2008;**77**(3):321–402.
- Folinsbee KE, Müller J, Reisz RR. Canine grooves: morphology, function, and relevance to venom. *Journal of Vertebrate Paleontology* 2007;**27**(2):547–51.
- Dufton MJ. Venomous mammals. *Pharmacology & Therapeutics*. Elsevier; 1992;**53**(2):199–215.
- Brandt AL, Grigorev K, Afanador-Hernández YM et al. Mitogenomic sequences support a North-South subspecies subdivision within *Solenodon paradoxus*. *Mitochondrial DNA Part A* [Internet]. Taylor & Francis; 2017;**28**(5):662–70. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/27159724>.
- Sato JJ, Ohdachi SD, Echenique-Diaz LM et al. Molecular phylogenetic analysis of nuclear genes suggests a Cenozoic over-water dispersal origin for the Cuban solenodon. *Sci Rep.* 2016;**6**(1):31173.
- Ottenwalder JA. The Distribution and Habitat of *Solenodon* in the Dominican Republic. MS thesis Univ. Florida, Gainesville. 1985, p. 128.
- Ottenwalder JA. The Systematics, Biology and Conservation of *Solenodon*. Ph.D. dissertation, University of Florida, Gainesville 1991, p. 281.
- Turvey S, Incháustegui S. 2008. *Solenodon paradoxus*. The IUCN Red List of Threatened Species 2008: e.T20321A9186243. <http://dx.doi.org/10.2305/IUCN.UK.2008.RLTS.T20321A9186243.en>. Downloaded on 07 April 2018.
- <http://www.iucnredlist.org/details/20321/0>.
- Luo R, Liu B, Xie Y et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaSci* 2012;**1**(1):589–602.
- Li H. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*. Oxford University Press; 2012;**28**(14):1838–44.

20. Boetzer M, Henkel CV, Jansen HJ et al. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*. Oxford University Press; 2011;**27**(4):578–9.
21. Wang Y, Lu Y, Zhang Y et al. The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. *Nat Genet* 2015;**47**(6):625–31.
22. Grigorev K, Kliver S, Dobrynin P et al. Supporting data for “Innovative assembly strategy contributes to understanding the evolution and conservation genetics of the endangered *Solenodon paradoxus* from the island of Hispaniola.” *GigaScience Database*. 2018; <http://dx.doi.org/10.5524/100422>.
23. Starostina E, Tamazian G, Dobrynin P et al. Cookiecutter: a tool for kmer-based read filtering and extraction. *bioRxiv Cold Spring Harbor Labs Journals* 2015;24679. Available from: <https://doi.org/10.1101/024679>.
24. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;**27**(6):764–70.
25. Marçais G, Yorke JA, Zimin A. QuorUM: an error corrector for Illumina reads. *PLoS One* 2015;**10**(6):e0130821.
26. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* [Internet]. 2014;**30**:31–7. Available from: <http://dx.doi.org/10.1093/bioinformatics/btt310>.
27. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* 2010;**11**(5):473–83.
28. Gurevich A, Saveliev V, Vyahhi N et al. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;**29**(8):1072–5.
29. Simão FA, Waterhouse RM, Ioannidis P et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;**31**(19):3210–2.
30. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. 2007;**23**(9):1061–7.
31. Hunt M, Kikuchi T, Sanders M et al. REAPR: a universal tool for genome assembly evaluation. *Genome Biol*. 2013;**14**(5):R47.
32. Paten B, Earl D, Nguyen N et al. Cactus: algorithms for genome multiple sequence alignment. *Genome Research*. 2011;**21**(9):1512–28.
33. Kolmogorov M, Raney B, Paten B et al. Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*. 2014;**30**(12):i302–9.
34. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2010 Accessible at: <http://www.repeatmasker.org>.
35. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 2015;**6**(1):11.
36. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;**6**(1):31.
37. Stanke M, Keller O, Gunduz I et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. 2006;**34**(Web Server):W435–9.
38. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*. 2011;**39**(Web Server issue):W29–37. Available at: doi:10.1093/nar/gkr367.
39. Altschul SF, Gish W, Miller W et al. Basic local alignment search tool. *Journal of Molecular Biology Elsevier*; 1990;**215**(3):403–10.
40. Bateman A, Coin L, Durbin R et al. The Pfam protein families database. *Nucleic Acids Res*. 2004;**32**(90001):138D–141.
41. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*. 2015;**43**(Database issue):D204–12. Available at: doi:10.1093/nar/gku989.
42. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. Oxford University Press; 2013;**29**(22):2933–5.
43. Nawrocki EP, Burge SW, Bateman A et al. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*. 2015;**43**(Database issue):D130–7. doi:10.1093/nar/gku1063.
44. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;**25**(5):955–64.
45. Seemann T, Booth T. BARNAP: BASIC Rapid Ribosomal RNA Predictor [Internet]. Berlin: GitHub; 2013. p. <https://github.com/tseemann/barnap>. Accessed 1st March 2018.
46. Ruan J, Li H, Chen Z et al. TreeFam: 2008 update. *Nucleic Acids Res*. 2008;**36**(Database):D735–40.
47. Li H, Coghlan A, Ruan J et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 2006;**34**(90001):D572–80.
48. Huerta-Cepas J, Szklarczyk D, Forslund K et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*. 2016;**44**(Database issue):D286–93. Available at: doi:10.1093/nar/gkv1248.
49. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014;**30**(9):1312–3.
50. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 2007;**24**(8):1586–91.
51. Xia X, Xie Z, Salemi M et al. An index of substitution saturation and its application. *Mol Phylogenet Evol* 2003;**26**(1):1–7.
52. Xia X, Lemey P. Assessing substitution saturation with DAMBE. In: Lemey Philippe, Salemi, Marco, Vandamme, Anne-Mieke eds. *Phylogenetic Handb. A Pract. Approach to DNA Protein Phylogeny*. 2nd ed. Cambridge University Press; 2009, 615–30.
53. Xia X. DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. *J Hered* 2017;**108**(4):431–7.
54. Ksepka DT, Parham JF, Allman JF et al. The fossil calibration database—a new resource for divergence dating. *Syst Biol* 2015;syv025.
55. Benton MJ, Donoghue PCJ, Asher RJ et al. Constraints on the timescale of animal evolutionary history. *Palaeontol Electron, Paleontological Society*; 2015;**18**:1–106.
56. Munthe K. Canidae. *Evol. Tert. Mamm. North Am*. Cambridge University Press, Cambridge. 1998, p. 124–43.
57. Wang X, Whistler DP, Takeuchi GT. A new basal skunk *Martinogale* (Carnivora, Mephitinae) from late miocene dove spring formation, California, and origin of New World mephitines. *Journal of Vertebrate Paleontology* 2005;**25**(4):936–49.
58. Rambaut A. FigTree [Internet]. 2016. Available from: <http://tree.bio.ed.ac.uk/software/figtree/> Accessed 1 March 2018.

59. Springer MS, Murphy WJ, Roca AL. Appropriate fossil calibrations and tree constraints uphold the Mesozoic divergence of solenodons from other extant mammals. *Molecular Phylogenetics and Evolution* 2018;121:158–65.
60. Hedges SB. Vicariance and dispersal in Caribbean biogeography. *Herpetologica* 1996;52:158–65.
61. McDowell SB. The greater antillean insectivores. *Bull Am Museum Nat Hist [Internet]*. 1958;115:117. Available from: <http://hdl.handle.net/2246/1199>.
62. Butler PM. Phylogeny of the insectivores. In: Benton MJ, eds. *Phylogeny Classif. Tetrapods*. Oxford: Clarendon; 1988, 117–41.
63. MacPhee RD, Novacek M. Definition and relationships of the Lipotyphla. In: Soule F, Novacek M, McKenna M, eds. *Mamm. Phylogeny, Vol. 2, Placentals*. New York: Springer-Verlag; 1993, 13–31.
64. McKenna M, Bell S, Simpson S. *Classification of Mammals Above the Species Level*. New York: Columbia University Press; 1997.
65. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7.
66. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006;34(Web Server):W609–12.
67. Soto-Girón MJ, Ospina OE, Massey SE. Elevated levels of adaption in *Helicobacter pylori* genomes from Japan; a link to higher incidences of gastric cancer? *Evolution, Medicine, and Public Health*. 2015;2015(1):88–105.
68. Gharib WH, Robinson-Rechavi M. The Branch-Site Test of Positive Selection Is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC. *Molecular Biology and Evolution*. 2013;30(7):1675–86.
69. Tang H, Klopfenstein D, Pedersen B et al. GOATOOLS: Tools for gene ontology [Internet]. Zenodo 2015. Available from: <https://doi.org/10.5281/zenodo.31628>.
70. Baggs JE, Green CB. Nocturnin, a Deadenylase in *Xenopus laevis* Retina. *Current Biology* 2003;13(3):189–98.
71. Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2010;365(1537):185–205.
72. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. *Genetics* 2001;158:1227–34.
73. Fay JC, Wu C-I. The neutral theory in the genomic era. *Current Opinion in Genetics & Development* 2001;11(6):642–6.
74. Ellegren H. Evolution: natural selection in the evolution of humans and chimps. *Current Biology* 2005;5(22):R919–22.
75. Gayà-Vidal M, Albà M. Uncovering adaptive evolution in the human lineage. *BMC Genomics* 2014;15(1):R919–22.
76. Bakewell MA, Shi P, Zhang J. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proceedings of the National Academy of Sciences* 2007;104(18):7489–94.
77. Olson MV, Varki A. Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat Rev Genet* 2003;4(1):20–28.
78. Mi H, Huang X, Muruganujan A et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*. 2017;45(Database issue):D183–9. Available at doi:10.1093/nar/gkw1138.
79. Jungo F, Bougueleret L, Xenarios I et al. The UniProtKB/Swiss-Prot Tox-Prot program: a central hub of integrated venom protein data. *Toxicon* 2012;60(4):551–7.
80. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357–9.
81. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754–60.
82. Danecek P, Auton A, Abecasis G et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156–8.
83. Dobrynin P, Liu S, Tamazian G et al. Genomic legacy of the African cheetah, *Acinonyx jubatus*. *Genome Biol* 2015;16(1):277. doi: 10.1186/s13059-015-0837-4.
84. Gordon D, Huddleston J, Chaisson MJP et al. Long-read sequence assembly of the gorilla genome. *Science*. 2016;352(6281):aae0344-. doi:10.1126/science.aae0344.
85. Li R, Fan W, Tian G et al. Erratum: the sequence and de novo assembly of the giant panda genome. *Nature*. 2010;463(7284):1106-.
86. Cho YS, Hu L, Hou H et al. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature Communications*. 2013;4:2433.
87. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011;475(7357):493–6.
88. Weber JL. Human DNA polymorphisms and methods of analysis. *Current Opinion in Biotechnology* 1990;1(2):166–71.
89. Weber JL, Wong C. Mutation of human short tandem repeats. *Hum Mol Genet* 1993;2(8):1123–8.
90. Weber JL, May PE. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* 1989;44:388–96. PMC1715443
91. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573–80.
92. Willems T, Zielinski D, Yuan J et al. Genome-wide profiling of heritable and de novo STR variations. *Nat Meth* 2017;14(6):590–2.
93. Koepfli K-P, Paten B, O'Brien SJ. The Genome 10K Project: a way forward. *Annu Rev Anim Biosci*; 2015;3:57–111.
94. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 2011;29:987–91.
95. Myers EW. Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology* 1995;2:275–90.
96. Myers EW. The fragment assembly string graph. *Bioinformatics* 2005;21:ii79–85.
97. Bradnam KR, Fass JN, Alexandrov A et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaSci* 2013;2:10.
98. Gogarten JP, Olendzenski L. Orthologs, paralogs and genome comparisons. *Current Opinion in Genetics & Development* 1999:630–6.
99. Creevey CJ, Muller J, Doerks T et al. Identifying single copy orthologs in metazoa. *PLoS Comput Biol* 2011;7:e1002269.
100. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2012;13:36–46.
101. Volfovsky N, Oleksyk TK, Cruz KC et al. Genome and gene alterations by insertions and deletions in the evolution of

- human and chimpanzee chromosome 22. *BMC Genomics* 2009;10:51.
102. Rabb GB. Toxic salivary glands in the primitive insectivore Solenodon. *Nat Hist Misc* 1959;170:1–3.
103. Allendorf FW, Luikart G. Conservation and the Genetics of Populations. John Wiley & Sons; 2009.
104. Willoughby JR, Sundaram M, Wijayawardena BK et al. The reduction of genetic diversity in threatened vertebrates and new recommendations regarding IUCN conservation rankings. *Biological Conservation* 2015;191:495–503.
105. Turvey ST, Meredith HMR, Scofield RP. Continued survival of Hispaniolan solenodon *Solenodon paradoxus* in Haiti. *ORX*. Cambridge University Press; 2008;42:611–4.
106. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. Oxford Univ Press; 2004;32:D258–61.