Supplementary Information Appendix

Materials and Methods

1. GENOME SEQUENCING AND ASSEMBLY

1.1 Sample collection for genome sequence

A male *Psammomys obesus* was obtained from Hadassah Medical School, Israel. All animals were handled in accordance to the regulations specified under the Protection of Animals Act by the authority in Denmark, European Union and Novo Nordisk A/S. The animal was sedated, euthanized and multiple tissues were dissected and snap frozen in liquid nitrogen for genomic DNA preparation. Genomic DNA was extracted from liver using the QIAGEN Genomic-tip 500/G kit following manufacturer's instructions.

1.2 Initial Genome Sequencing

Multiple genomic libraries were prepared by standard procedures. Small insert-size libraries of 250bp, 500bp and 800bp were constructed and sequenced using 150bp paired end reads on the Illumina HiSeq2000 platform. After a basic genome survey the sequencing of the large insert-size libraries (2Kbp, 5Kbp, 10Kbp and 20Kbp) was performed (Table S1).

Table S1. Raw genomic DNA sequencing data metrics

Insert size	Read pairs	Read length	Raw bp
Paired end libraries			
250bp	197,653,655	2x150	59,296,096,500
250bp	208,984,386	2x150	62,695,315,800
500bp	171,439,814	2x150	51,431,944,200
800bp	137,629,829	2x150	41,288,948,700
TOTAL	**715,707,684**		**214,712,305,200**
GC-enriched libraries			
550bp	4,605,282	2x300	2,763,169,200
550bp	4,064,520	2x300	2,438,712,000
TOTAL	**8,669,802**		**5,201,881,200**
Mate pair libraries			
2Kb	206,747,237	2x49	20,261,229,226
2Kb	171,782,906	2x49	16,834,724,788
2Kb	152,607,390	2x49	14,955,524,220
5Kb	173,390,804	2x49	16,992,298,792
5Kb	167,200,992	2x49	16,385,697,216
5Kb	157,372,673	2x49	15,422,521,954
10Kb	196,302,639	2x49	19,237,658,622
10Kb	200,677,190	2x49	19,666,364,620
20Kb	190,067,357	2x49	18,626,600,986
20Kb	164,286,952	2x49	16,100,121,296
TOTAL	**1,780,436,140**		**174,482,741,720**

1.3 Predicted genome size and sequencing coverage

Our initial genome assembly was done using SOAPdenovo2 (1) with a k-mer size of 41 (see section 1.4). For genome size estimation, k-mer counting of 41-mers was done using Jellyfish (2) and the resulting k-mer spectrum was plotted using ggplot2 (3). A single peak can be seen at depth 23 (Figure S1). Using this 41-mer distribution we calculated the genome size using the equation G=N*(L-k+1)/D, where N is the number of reads, L the read length, k the k-mer size and D the depth of the peak.

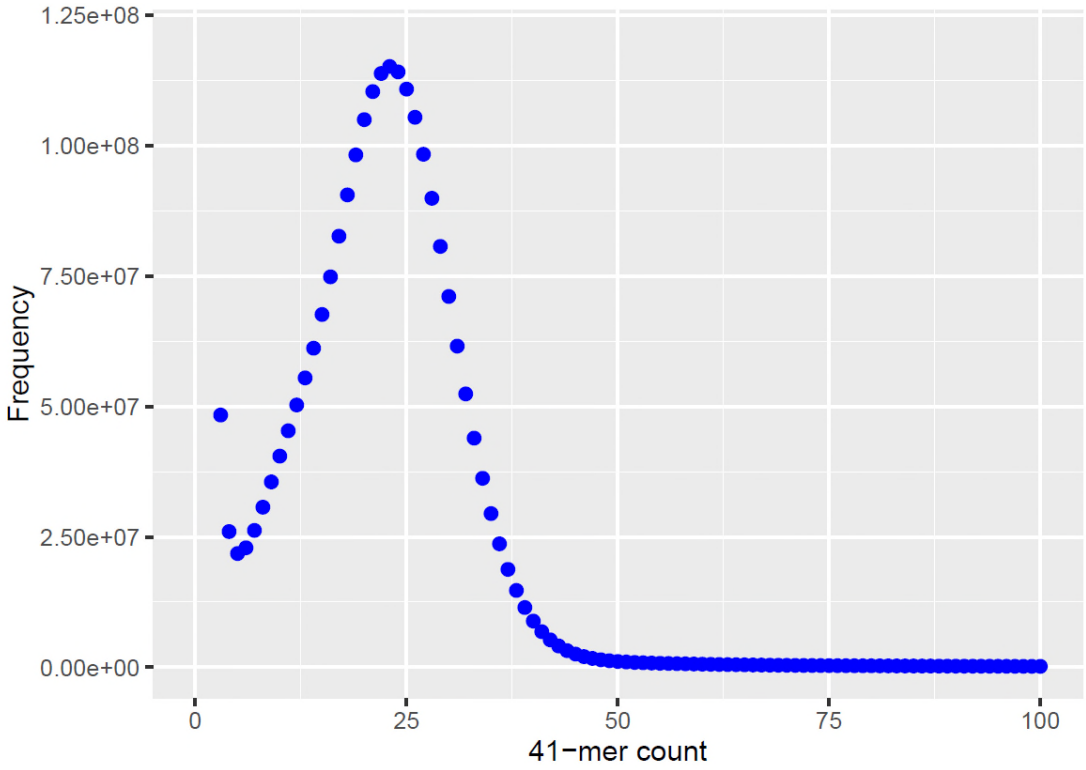


Figure S1. 41-mer distribution spectrum

The predicted genome size for *Psammomys obesus* using this method is 2.51 Gb, lower than the predicted genome sizes of other members of the Gerbillinae in the animal genome size database (4) but similar to the predicted genome sizes of other murid rodents, including the closely related *Acomys cahirinus*, and the total size of our final genome assembly (Table S2).

Table S2. Muridae predicted haploid genome sizes and karyotypes

Species	Predicted haploid genome size (Gb)	Chromosome number
Psammomys obesus	2.51	48
Meriones unguiculatus	3.64	44
Gerbillus pyramidum	4.00	40
Gerbillus campestris	3.64	56
Acomys cahirinus	2.64	38
Mus musculus	2.64	40
Rattus norvegicus	2.75	42

We generated approximately 394 Gb of raw genomic sequence data (Table S1). Based on the predicted genome size of 2.51 Gb we achieved a total sequencing depth of 87.6X according to the Lander/Waterman equation (5) ($C = LN/G$ where C is the estimated sequencing coverage, L is read length, N is the total number of reads sequenced and G represents the known (or estimated) haploid genome size).

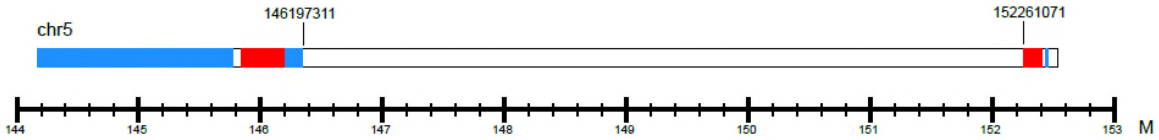
1.4 Initial Genome assembly

Raw genomic reads were filtered using the following steps: 1) Reads containing Ns or polyA tracts over 10% or more of their total length were discarded; 2) Small-insert library reads with 50 or more bases with a Q20 value of 7 or less and large-insert library reads with 15 or more bases below this threshold were filtered; 3) To remove reads with adapter contamination, reads with more than 10bp aligned to the adapter sequence (allowing up to 3bp mismatch) were discarded; 4) Small insert size reads in which read1 and read2 overlapped more than or equal to 10bp allowing 10% mismatch were discarded; 5) PCR duplicate reads, identified when read1 and read2 of two paired end reads are totally identical, were discarded; 6) Low quality bases at read ends were trimmed directly.

An initial genome assembly was generated using SOAPdenovo2 (1) with a k-mer size of 41 (see Table S3 for assembly metrics). Pairwise whole-genome alignments (WGA) of gerbil and mouse were carried out using LASTz (6) (version 1.02.00) with the parameters "--step=19 --hspthresh=2200 --inner=2000 --ydrop=3400 --gappedthresh=10000 C=2, T=2, H=2000, Y=3400, L=6000, and K=2200". We used the Chain/Net package for post treatment. The gerbil genome was masked with RepeatMasker (7) (version 3.2.9) repeats at "-s" setting and TRF tandem repeats of period ≤ 12. The mouse (mm9) repeat-masked genome was downloaded from UCSC (http://genome.ucsc.edu). Following whole genome alignment we identified a large contiguous region missing from our sand rat genome assembly, which contains 88 genes (orthologous between mouse and rat) and includes the ParaHox cluster (Figure S2).

Subsequent analysis of tissue transcriptomes (see Section 2) revealed that several missing genes were in fact expressed as gene transcripts, implying that our initial genome assembly was incomplete. As these transcripts were found to be high in Guanine and Cytosine base composition, we sought to re-sequence genomic DNA enriched for elevated GC content.

A



B

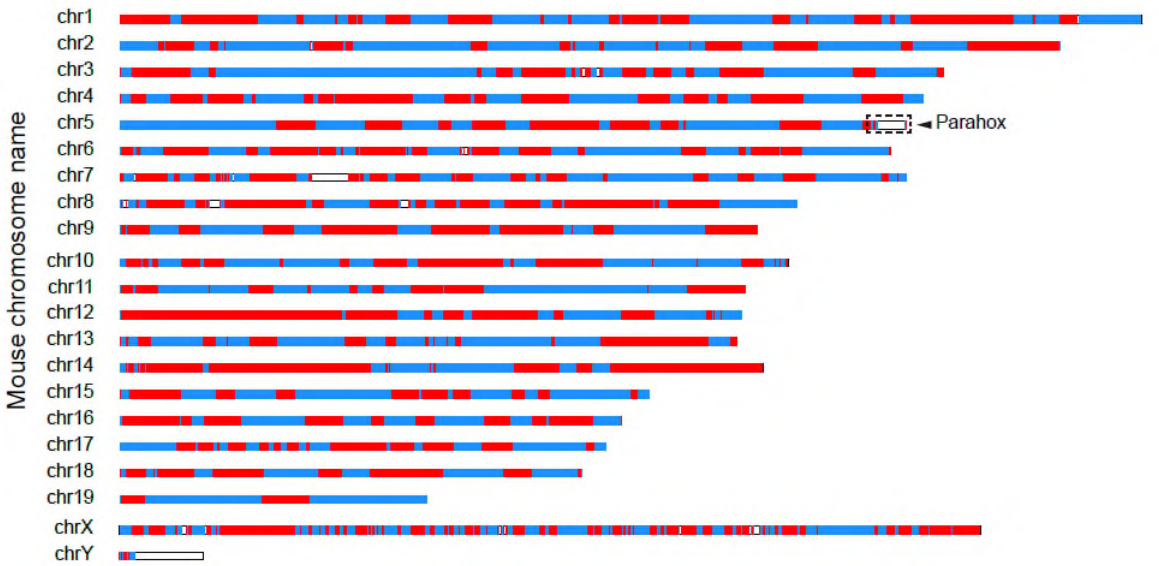


Figure S2. (A) Local liner map of the missing sand rat genomic region corresponding to the end of chromosome 5 and the beginning of chromosome 8 of mouse. (B) Local lastz alignment of the sand rat initial genome assembly to the mouse genome. The missing ParaHox region corresponding to the syntenic region on mouse chromosome 5 is indicated. Blue and red blocks indicate different sand rat genomic scaffolds aligning to the mouse genome.

1.5 GC-enriched DNA sequencing

GC-enriched DNA samples were obtained using Caesium Chloride (CsCl) gradient centrifugation. 5ug of Sand rat genomic DNA in 1M TE was sheared to a size of 10 Kb using Covaris g-tubes (#520079). In order to estimate the GC content of resulting sand rat DNA fractions, a mixture of *Escherichia coli* genomic DNA (Obtained in-house using the Roche DNA isolation kit for cells and tissues; Roche #11814770001) and *Micrococcus luteus* genomic DNA (Sigma #D8259-5MG) was also prepared in parallel to act as a "GC content reference" (*E. coli* 50% GC; *M. luteus* 70% GC). DNA samples were added to 13ml quick-seal tubes and filled with 1g/ml CsCl dissolved in 1M TE, without ethidium bromide. Tubes were subsequently sealed and spun using a Beckman Coulter Optima XPN-80

ultracentrifuge with a Beckman Type 70.1 Ti rotor at 40,000 RPM for 67 hours at 20˚C with no brake for deceleration so as not to disrupt the CsCl isopycnic density gradient. Each tube was then decanted into a series of Eppendorf tubes (roughly 5 drops per tube) after puncturing the tube top and bottom with a hypodermic needle. After verification on a 0.7% agarose gel and comparison to the *E. coli* + *M. luteus* DNA distribution, sand rat DNA samples were pooled to result in GC rich (60-65% GC) and very GC rich (+65%) DNA fractions. These samples were then purified by dialysis using Novagen D-tube Dialyzer Midi MWCO 3.5 kDa dialysis columns (Novagen #71506-3) in 2 litres of 1M TE for 24 hours (with intermittent TE changes), concentrated using the Qiagen QIAquick PCR purification kit and eluted into 20µl of Qiagen elution buffer EB. The two DNA samples were used for preparation of 550 bp insert TruSeq DNA sequencing libraries at the Oxford Genomics centre (Wellcome Trust Centre for Human Genetics, Oxford, UK) and sequenced on the Illumina MiSeq platform using 2x300bp paired-end reads (Table S1).

1.6 Sand rat Pdx1 initial discovery

The GC-enriched Miseq reads were merged using FLASH (8) (v1.2.11). Subsequently, all merged MiSeq reads and all Illumina paired-end reads were re-mapped to the initial genome assembly using Bowtie2 (9) (v2.1.0). Only unmapped reads were extracted from the resulting .bam file and then used to generate a local assembly using ABySS (10) with a k-mer size of 61. A local BLAST+ (11) survey detected a contig containing the *P. obesus* Pdx1 homeodomain.

1.7 Genome re-assembly and GC-enriched sequence incorporation

The GC-enriched Miseq reads were filtered and trimmed using the same approach as described in section 1.4. The genome was then re-assembled using SOAPdenovo2 with k-mer size 41 incorporating the initial Illumina reads and the GC-enriched Miseq reads. Gaps were subsequently filled using GapCloser (1) (version 1.12). A local BLAST search revealed that this final genome assembly contained a full *Pdx1* gene sequence.

To ascertain the extent of GC bias in the original genomic sequencing reads, the sequence coverage per scaffold was calculated by mapping the original HiSeq reads only to the *P.obesus* final genome assembly using bowtie2 (9) followed by using the 'genomecov' utility of bedtools (12). This was then subsequently summarised using in-house scripts and compared to the %GC per scaffold (Figure S3). This revealed a sharp drop in coverage when GC content per scaffold deviates away from 30-50%. However, there were HiSeq reads present for GC-rich scaffolds, and the subsequent MiSeq read number was relatively low, indicating that it was not solely increasing the read coverage that allowed assembly; it is probable that the longer MiSeq read length was also important.

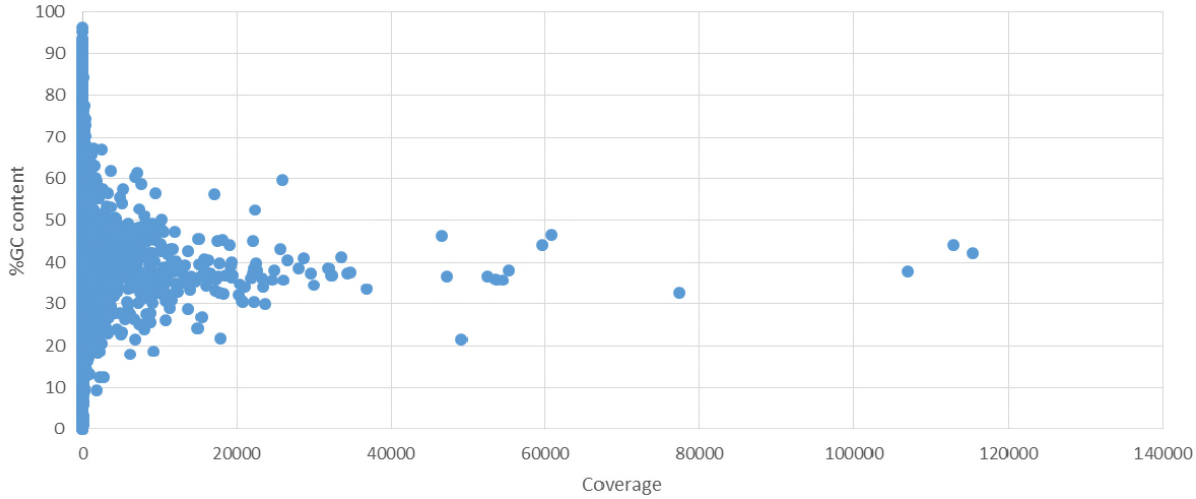


Figure S3. Scatter plot of coverage per scaffold against %GC content per scaffold for the final *Psammomys obesus* genome assembly.

Genome assembly assessment

Genome assembly metrics (Table S3) were obtained using a perl script from the Assemblathon 2 (13) (Currently available at https://github.com/ucdavis-bioinformatics/assemblathon2-analysis). BUSCO (Benchmarking Universal Single-Copy Orthologs) analysis (14) indicates that the final genome assembly is 89% complete.

Table S3. *Psammomys obesus* genome assembly metrics

	Initial assembly	Final assembly
Number of scaffolds	135,630	150,763
Number of scaffolds >2 kb	2,020	1,737
Total length of assembly (bp)	2,373,092,066	2,381,209,849
Longest scaffold (bp)	38,234,054	54,616,910
Mean scaffold length (bp)	17,497	15,794
Scaffold N50 (bp)	8,843,897	10,461,538
Scaffold L50	83	63
Contig N50 (bp)	36,780	83,904
Percentage of assembly in scaffolds (%)	98.7	98.6
G + C content (%)	41.87	41.04
Complete BUSCOs	1,508	2,233
Fragmented BUSCOs	791	437
Missing BUSCOs	710	353

2. TRANSCRIPTOME SEQUENCING AND ASSEMBLY

2.1 Sample preparation

Pancreatic islets

Pancreatic islets were isolated from nine *P. obesus* individuals belonging to three groups (three individuals per group). The three groups were: diabetic animals fed on a high energy Diet (HEDD), non-diabetic animals fed on a high energy Diet (HEDND) and animals maintained on a low energy Diet (LED). Islet cells were isolated by standard collagenase methods (15) and cultured overnight to remove all exocrine and ductal cell contamination. Isolated islet cells were subsequently collected, briefly centrifuged and placed directly in the lysis buffer provided by the Qiagen RNA Easy kit. Total RNA was extracted using the Qiagen RNeasy Mini kit with on-column DNase treatment. RNA quality was accessed using the Agilent chip methods and only samples with a RIN value greater than 8 were used for RNA sequencing. Total RNA samples were enriched for messenger RNA using PolyA priming, cDNA was synthesized and different size libraries between 300 and 500bp were constructed for RNAseq. Libraries were both run individually and as pools on 16 lanes of the Illumina GAII sequencing platform at the Institute for Systems Biology (Seattle, USA).

Liver

Liver samples from six *P.obesus* individuals (two from each group of HEDD, HEDND and LED; see above) were dissected and snap frozen in liquid nitrogen. Total RNA was extracted using the Qiagen RNeasy Mini kit with on-column DNase treatment. RNA-seq libraries were prepared using the Illumina TruSeq RNA library preparation kit (v2 chemistry) following PolyA enrichment. Libraries were pooled and run on 2 lanes of the Illumina HiSeq2000 at the Beijing Genomics Institute using 2x90bp paired-end reads.

Duodenum

A 7-week old diabetes prone *P. obesus* individual and an adult *Meriones unguiculatus* individual were obtained from colonies maintained at Bangor University, UK. The animals were euthanized according to Schedule 1 of the Animals (Scientific procedures) Act 1986. Anterior duodenum samples were dissected and immediately snap frozen in liquid nitrogen and stored at -80˚C until required. Total RNA was extracted using TRI Reagent (Sigma #T9424) with an additional on-column DNase step (Qiagen #79254) as described previously (16) and eluted into RNase-free water. Sequencing libraries were prepared using the Illumina TruSeq RNA library preparation kit with PolyA enrichment at the Wellcome Trust Centre for Human Genetics (Oxford, UK) and sequenced on 1/6th of a lane of the Illumina HiSeq4000 using 2x75bp paired-end reads (Table S4).

Table S4. RNA sequencing raw data metrics.

Tissue	Paired-end reads	Read length (bp)	Total raw sequence (bp)
Pancreatic islets	171,668,505	2x100	34,333,701,000
Liver	98,742,080	2x100	19,748,416,000
Duodenum (P. obesus)	62,563,996	2x75	9,384,599,400
Duodenum (M. unguiculatus)	43,543,534	2x75	6,531,530,100

2.2 Transcriptome assembly

Adapter contamination was removed using Trimmomatic (17) (version 0.33) and raw reads were trimmed using Sickle (18) (version 1.33). The transcriptomic data for pancreatic islets was assembled using Trans-ABySS (19) (version 1.2.5) with multiple k-mer sizes (k=41 up to k=79 in increments of 2). Read data for liver and duodenum was assembled using Trinity (20). Open Reading Frame prediction was carried out using TransDecoder (21) (version 2.0.1) which included first identifying ORFs with homology to known proteins in the UniProt Swiss-Prot database (22) (last downloaded July 6th 2016) using blastp (11) (Table S5).

Table S5. Tissue transcriptome assembly metrics

	Pancreatic islets	Liver	Duodenum (P. obesus)	Duodenum (M. unguiculatus)
Number of contigs	175,097	256,993	121,343	84,334
Total length	134,549,058	334,473,169	111,092,270	73,420,368
Longest contig	28,084	18,181	17,288	14,622
N50	1,286	3,208	2,036	1,758
Number of predicted ORFs	84,633	139,947	41,963	34,094
Number of predicted full coding sequences	17,989	98,462	21,365	13,009

2.3 Detecting presence of GC rich genes

Local blast surveys for the missing genes were carried out using BLAST+ (11) (version 2.2.31) with query protein sequences downloaded from Ensembl and Genbank databases. We were able to recover 52 out of the 88 originally missing GC-rich genes, 29 of which are present in the predicted proteins derived from our final genome assembly (Table S6).

Table S6. Presence or absence of GC-rich genes/transcripts in the *Psammomys obesus* genome and transcriptomes

	Tissue			
Gene	Pancreatic islets	Liver	Duodenum	Present in predicted proteins
Lmtk2	✓	✓	✓	✓
Bhlha15	✓	×	✓	✓
Tecpr1	✓	✓	✓	✓
Bri3	✓	✓	✓	
Baiap2l1	✓	✓	✓	✓

Nptx2	✓	×	×	
Tmem130	✓	×	×	✓
Zfp498	×	×	×	
Trrap	✓	✓	✓	✓
Smurf1	✓	✓	✓	✓
Kpna7	×	×	×	
Arpc1a	✓	✓	✓	✓
Arpc1b	×	✓	✓	✓
Pdap1	✓	×	✓	✓
Bud31	✓	✓	✓	✓
Ptcd1	✓	✓	✓	✓
Cpsf4	✓	✓	✓	
Atp5j2	×	✓	✓	
Rnf6	×	×	×	
Cdk8	✓	✓	✓	
Wasf3	×	×	×	
Gpr12	×	×	×	
Usp12	×	×	×	
Rpl21	✓	✓	✓	✓
Rasl11a	×	×	×	
Gtf3a	✓	✓	✓	
Mtif3	×	✓	×	
Lnx2	×	×	×	
Polr1d	✓	✓	✓	✓
Gsx1	×	×	×	
Pdx1	✓	×	✓	✓
Cdx2	×	×	✓	✓
Urad (Prhxnb)	×	×	×	
Flt3	×	×	×	
Pan3	×	✓	✓	✓
Flt1	×	×	×	
Pomp	✓	✓	✓	
Slc46a3	×	×	×	
Mtus2	×	×	×	
Slc7a1	×	×	×	
Ubl3	✓	✓	✓	
Katnal1	✓	✓	✓	✓
Hmgb1	✓	✓	✓	
Uspl1	✓	✓	×	
Alox5ap	×	✓	✓	✓
Medag (MEDA-4)	✓	✓	✓	✓
Tex26	×	×	×	
B3glct	✓	✓	✓	✓

Hsph1	×	×	×	
Wdr95	×	×	×	
Rxfp2	×	×	×	
Fry	×	×	×	
Zfp958	×	×	×	
Cers4	✓	✓	✓	
Prr36	×	×	×	
Evi5l	✓	✓	✓	
Lrrc8e	×	×	×	
Map2k7	✓	✓	✓	
Tgfbr3l	×	×	×	
Snapc2	✓	✓	×	
Ctxn1	×	×	×	
Timm44	✓	✓	✓	✓
Elavl1	✓	✓	✓	✓
Ccl25	×	×	✓	
Clec4g	×	×	×	
Fcer2	×	×	×	
TrappC5	✓	✓	✓	✓
Mcemp1	×	×	×	
Retn	×	×	×	
Stxbp2	✓	✓	×	
Pcp2	×	×	×	
Pet100	×	×	×	
Xab2	✓	✓	✓	✓
Camsap3	✓	✓	✓	
Pnpla6	✓	✓	✓	
Zfp358	×	×	×	
Mcoln1	✓	✓	✓	
Pex11g	✓	✓	×	✓
Arhgef18	×	×	×	
Insr	✓	✓	✓	✓
Rfc3	✓	✓	✓	✓
Stard13	×	×	×	
Kl	×	×	×	
Pds5b	✓	✓	✓	
N4bp2l2	✓	✓	✓	
N4bp2l1	✓	×	✓	
Brca2	×	✓	×	✓
Zar1l	×	×	×	

3. REPEAT ANNOTATION

Repetitive elements were annotated using RepeatMasker (23) (version open-4.0.5). A *de novo* repeat library was first generated using RepeatModeler (24) and was used for subsequent repeat annotation (Table S7). A total of 37.7% of bases in the sand rat genome were masked as repeats.

Table S7. Repeatmasker output for the sand rat final genome assembly.

Repeat/TE class	Number of elements	Length occupied (bp)	Percentage of sequence (%)
SINEs	1,420,596	187,571,594	7.88
Alu/B1	494,184	56,418,111	2.37
B2-B4	676,137	107,700,139	4.52
IDs	81,666	5,900,797	0.25
MIRs	103,584	12,497,978	0.52
LINEs	748,856	425,985,711	17.89
LINE1	678,340	414,208,719	17.39
LINE2	56,297	9,804,836	0.41
L3/CR1	10,394	1,460,127	0.06
LTR elements	577,946	183,123,703	7.69
ERVL	85,702	24,071,333	1.01
ERVL-MaLRs	283,757	90,679,360	3.81
ERV_classI	47,067	15,599,597	0.66
ERV_classII	156,826	51,810,173	2.18
DNA elements	143,219	28,587,362	1.20
hAT-Charlie	93,634	18,177,962	0.76
TcMar-Tigger	23,128	5,321,930	0.22
Unclassified	22,598	9,321,173	0.39
Small RNA	20,138	1,740,861	0.07
Satellites	8,572	1,418,652	0.06
Simple repeats	1,104,046	50,694,653	2.13
Low complexity	173,356	9,463,235	0.40

As RepeatMasker annotates and quantifies repetitive sequences post-assembly, and therefore could underestimate repeat content in instances of collapsed repeats, we also used dnaPipeTE (25) for repeat analysis. This pipeline assembles sequences from a down-sampled selection of raw genomic reads (using this low-coverage data, repeated sequences will successfully assemble whilst single-copy sequence will not) followed by annotation and abundance estimation. We find no evidence of any repeat expansions in the sand rat which would explain the emergence of a large genomic region of elevated GC content (Figure S4).

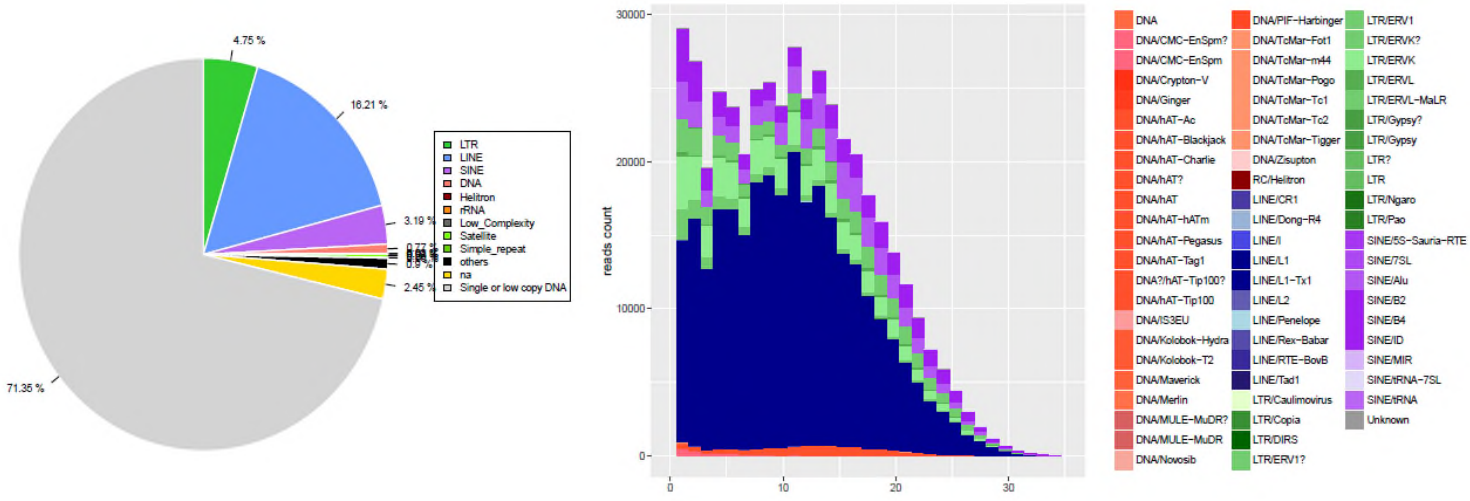


Figure S4. DnaPipeTE analysis of the sand rat raw genomic data. (A) Proportions of repeat element classes predicted in the sand rat genome and (B) repeat landscape of *Psammomys obesus*.

4. PROTEIN-CODING GENE ANNOTATION

In order to predict the protein coding genes in the sand rat genome, we performed a multi-faceted analysis comprising multiple annotation methods. Using this approach we predict a total of 21,807 protein-coding genes in the sand rat genome, with an average gene size of 28,336 bp, an average exon number per gene of 8.4, and an average exon and intron length per gene of 1,604 bp and 26,733 bp respectively.

Ab initio prediction

Repetitive elements in the final genome assembly were first masked using RepeatMasker (23) followed by *ab initio* gene prediction with AUGUSTUS (26) (version 2.5.5) using the *Homo sapiens* parameters.

Homolog-based prediction

Homologous proteins from well annotated mammalian species genomes (mouse and human from Ensemble release-68) were mapped to the sand rat genome using TBLASTN (11) (BLASTall version 2.2.23) with an E-value cutoff $1e^{-5}$. The aligned sequence as well as its query protein were then filtered and passed to GeneWise (27) (version 2.2.0) for searching for accurate spliced alignments.

GLEAN integration

Output generated from the approaches mentioned above was integrated using GLEAN (28) to produce a consensus gene set.

Refinement using RNA-Seq data

We next sought to combine our consensus gene set produced using GLEAN with our RNA-seq data to improve the accuracy and confidence of our gene predictions. First, we aligned reads to the genome using TopHat (29) to identify candidate exon regions. Second, we identified donors and acceptors according to the predicted splicing sites. Then, we assembled the transcripts using Cufflinks (30) (version 1.3.0). Finally, based on these assembled candidate transcript sequences, the Open Reading

Frame (ORF) was predicted to acquire reliable transcripts using the HMM-based training parameter. These predictions were then combined with our GLEAN consensus gene set.

Manual annotation refinement

Several genes of interest (most notably *Pdx1* and *Cdx2*) were initially incompletely annotated after these previous methods. We therefore manually annotated these genes based upon sequence information found in the tissue transcriptomes.

5. 'PROTEIN DEVIATION INDEX' DIVERGENCE CALCULATIONS

To calculate a Protein Deviation Index (PDI) of sequence divergence across the genome, we compared 1:1 orthologous proteins between sand rat, mouse and human (similar to a method used by Clark et al (31)). We first downloaded protein sets for mouse and human from Ensembl biomart (32). Reciprocal blasts were conducted using blastp (11) between the mouse and human orthologue sets and the predicted protein sequences from the *P. obesus* genome assembly with an e-value of 1e-6 and one target sequence specified. We then used a custom python script "reciblast.py" to extract lists of protein IDs for each reciprocal blast. We then used "build_triplets.py" to generate amino acid sequence from all three species for each orthologous gene. Protein sequences were then aligned using MAFFT (33) (v7.123b). The script "percent_id.py" was used to generate a table of the percentage identity between each aligned protein sequence. In order to increase stringency, any results from alignments which had an ungapped alignment length of less than 50% of the total alignment length were discarded. The ratios of mouse:human and mouse:sand rat were then divided to give the final protein deviation index value (Table S8). A histogram (Figure 2c) was plotted in R (34) using the package ggplot2 (3).

6. HOX COMPLEMENT AND DISTRIBUTION ON SCAFFOLDS

We searched for the Hox complement of *Psammomys obesus* by first conducting local BLAST surveys using homeodomain protein sequences from mouse downloaded from HomeoDB (35) as query sequences. Matches were confirmed by reciprocal BLAST against the nr database and by visual inspection of the annotated scaffold. We find a complement of 39 Hox genes in our assembly, arranged in 4 clusters (Figure S5). This is an identical complement to those found in other rodents such as mouse and rat.

Table S8. Top 20 most divergent genes in the *Psammomys obesus* genome.

Rank on PDI index	Gene name	Present in GC rich region
1	Pdx1	✓
2	Pex11C	✓
3	Medag	✓
4	Imp4	×
5	Polr1d	✓
6	TrappC5	✓
7	Insr	✓
8	Gsx1	✓
9	Apbb3	×
10	Ldoc1	×
11	Sft2d3	×
12	Arglu1	×
13	Cmtm4	×
14	Ift46	×
15	Golga7b	×
16	March5	×
17	Trim56	×
18	Pdap1	✓
19	Pianp	×
20	Znf3	×

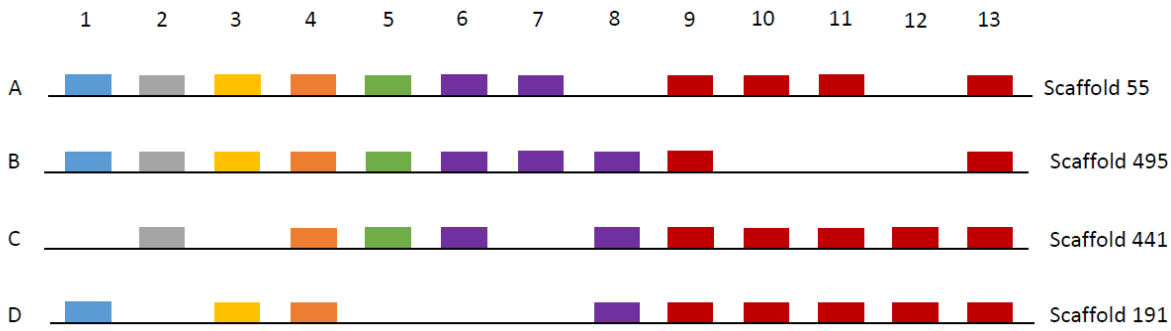


Figure S5. The Hox gene complement of *Psammomys obesus*.

7. IDENTIFYING SAND RAT PDX1

Following initial discovery of the putative highly divergent sand rat *Pdx1* homeodomain in the local ABySS genome assembly (section 1.6), we sought to confirm that this was indeed a real Pdx1 sequence. Firstly, the contig containing the putative homeodomain was used as a query for a local blast search of the pancreatic islets and duodenum transcriptomes. Multiple contigs were found encoding a full coding sequence in both transcriptome assemblies. Reciprocal BLAST identifies this sequence as *Pdx1*, with a top hit to Yak (*Bos mutus*). We identified 47 point mutations in the homeodomain (compared to the closely related *Acomys cahirinus Pdx1*; Genbank Accession GQ179992), 27 of which were A/T to G/C mutations. Out of the 15 amino acid residue changes in the *P. obesus Pdx1* homeodomain (Figure 2a), 14 are caused by these G/C mutations. We also note extensive deletions throughout the *Pdx1* coding sequence in *P. obesus* (621 nucleotides) when compared to mouse (855 nucleotides), rat (852 nucleotides) and the spiny mouse *Acomys cahirinus* (861 nucleotides). Along with the presence of a homeodomain we also find a conserved hexapeptide domain present in the sand rat coding sequence, further suggesting that this transcript encodes Pdx1. Secondly, phylogenetic analysis was conducted using predicted amino acid sequences for sand rat and several other vertebrate and cephalochordate species aligned using MUSCLE (36) with maximum likelihood trees generated using MEGA5 (37) using the WAG model of protein evolution and 1,000 bootstraps. The resulting tree (Figure S6) puts sand rat Pdx1 in a clade with other Pdx1 sequences on a long branch. Finally, we found the putative *Pdx1* in our final genome assembly on a scaffold along with the genes *Rpl21*, *Cdx2*, *Cdk8*, *Atp5j2*, *Pdap1*, *Bud31*, *Ptcd1* and *Cpsf4*, confirming that it shares a syntenic genomic location when compared to other rodents (Table S9). We are therefore confident that we have identified a bone fide *Pdx1* gene in the sand rat.

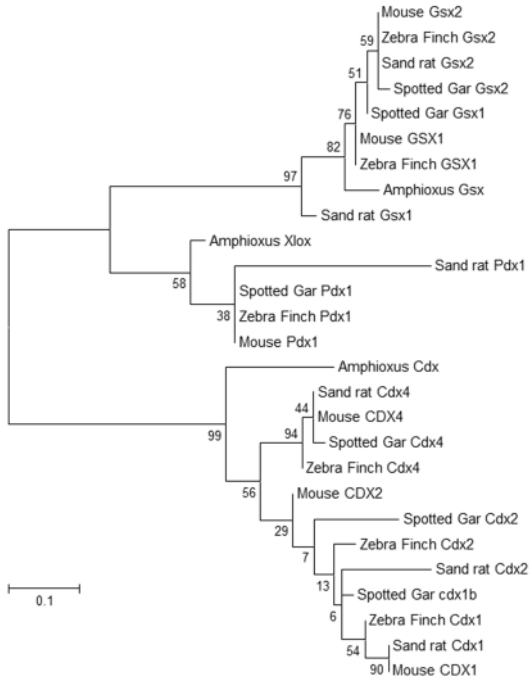


Figure S6. Maximum likelihood tree of ParaHox genes in sand rat (*Psammomys obesus*), mouse (*Mus musculus),* zebra finch (*Taeniopygia guttata*), spotted gar (*Lepisosteus oculatus*) and amphioxus (*Branchiostoma floridae*).

Table S9. Genes located on scaffold 966 in the sand rat genome assembly and their corresponding chromosomal locations in human, mouse and rat.

Sand rat Scaffold 966 genes	Mouse	Rat	Human
Pdap1	Chr5, 85.04 cM	Chr12 p11	7q22.1
Bud31	Chr5, 85.05 cM	Chr12 p11	7q22.1
Ptcd1	Chr5, 85.06 cM	Chr12 p11	7q22.1
Cpsf4	Chr5, 85.08 cM	Chr12 p11	7q22.1
Atp5j2	Chr5, 85.09 cM	Chr12 p11	7q22.1
Cdk8	Chr5, 85.19 cM	Chr12 p11	13q12.13
Rpl21	Chr5, 86.50 cM	Chr12 p11	13q12.2
Pdx1	**Chr5, 86.84 cM**	**Chr12 p11**	**13q12.2**
Cdx2	Chr5, 86.86 cM	Chr12 p11	13q12.2

8. TYPE 2 DIABETES-RELATED GENES

To investigate if sand rats have accumulated deleterious mutations in additional genes affecting glucose metabolism or pancreatic function, we compiled a list of 45 candidate genes from human studies, including genes implicated in monogenic diabetes and genes for which coding sequence variants have been strongly associated with T2D in association studies (38-41). For each of the 45 genes, we identified the sand rat orthologue and calculated its divergence index as described above in Section 5. Nine of these genes are not in the 1:1:1 orthologue set generated in Section 5 and so it is not possible to assign them a PDI value. Moreover 3 genes were not found in our predicted protein set. Out of the total of 45 genes, Pdx1 is ranked as the number 1 most divergent gene, and is considerably more divergent when compared to other type 2 diabetes-related genes (Table S10).

9. GC CONTENT ACROSS THE MUTATIONAL HOTSPOT

Sand rat transcripts encoding the GC rich genes were identified by performing reciprocal blast surveys of the tissue transcriptomes using protein sequences downloaded from Ensembl and Genbank from mouse (*Mus musculus*), rat (*Rattus norvegicus*) and chinchilla (*Chinchilla lanigera*) as query sequences. Detected sand rat sequences were extracted and annotated manually to give full coding sequence. In any instance where using a full coding sequence was not possible, we aligned the sand rat sequence with the other remaining rodent sequences using ClustalW (42) and then only used contiguous aligned sequence between all species to generate GC content values. To get GC values we used the online GC calculator (http://www.endmemo.com/bio/gc.php). The gene order shown is that inferred for the ancestor of rodents, obtained through comparison of human, mouse and rat genomes; the sand rat gene order may not be identical and a lack of a single, contiguous scaffold containing this large genomic region precludes any definitive synteny analyses.

Table S10. Genes for which coding sequence variants are implicated in human type 2 diabetes, shown in order of "Protein Deviation Index" value.

Gene	P. obesus predicted protein ID	Blast confirmation	% ID to top BLAST hit	Position in divergence table	PDI ratio
PDX1	PobPdx1Augustus	Y	42	1	1.66
PTF1A	Pob_R017023	Y	82	225	1.04
PAX6	Pob_R017876	Y	97	800	1
NeuroD1	Pob_R021476	Y	99	1,503	0.99
PAX4	Pob_R008854	Y	82	1,949	0.99
SLC19A2	Pob_R009579	Y	80	1,909	0.99
SIRT1	Pob_R018113	Y	93	1,576	0.99
KCNJ11	Pob_R006022	Y	98	3,104	0.98
CISD2	Pob_R020901	Y	100	3,275	0.98
PPIP5K2	Pob_R000491	y	97	2,340	0.98
HNF1A	Pob_R000788	Y	97	4,007	0.97
ABCC8	Pob_R006023	Y	97	3,767	0.97
PPARG	Pob_R011165	Y	99	4,225	0.97
GATA4	Pob_R000356	Y	94	3,653	0.97
PAM	Pob_R000489	Y	93	3,995	0.97
HNF4A	Pob_R005138	Y	99	4,284	0.96
ASCC2	Pob_R008713	Y	93	5,828	0.95
GCKR	Pob_R002481	Y	93	6,564	0.94
MTNR1B	Pob_R005880	Y	84	6,084	0.94
RFX6	Pob_R003880	Y	93	6,404	0.94
MNX1	Pob_R021615	Y	96	6,434	0.94
EIF2AK3	Pob_R011489	Y	93	6,582	0.94
FOXP3	Pob_R002782	Y	93	6,103	0.94
WFS1	Pob_R008797	Y	96	7,692	0.92
MTMR3	Pob_R008714	Y	96	7,534	0.92
BLK	Pob_R000357	Y	96	7,977	0.91
GLIS3	Pob_R005447	Y	95	8,408	0.9
THADA	Pob_R000142	Y	87	8,310	0.9
RREB1	Pob_R012867	Y	92	8,598	0.9
NGN3	Pob_R012754	Y	99	9,021	0.89
KLF11	Pob_R002344	Y	91	9,552	0.87
COBLL1	Pob_R006508	Y	81	10,114	0.84
AIRE	Pob_R017441	Y	88	10,195	0.83
SLC30A8	Pob_R019570	Y	91	Not in list of 1:1	
G6PC2	Pob_R006488	Y	90	Not in list of 1:1	
GCK	Pob_R008369	Y	95	Not in list of 1:1	
HNF1B	Pob_R010710	Y	91	Not in list of 1:1	
CEL	Pob_R018439	Y	68	Not in list of 1:1	
INS	Pob_R016749	Y	99	Not in list of 1:1	
GATA6	Pob_R015971	Y	87	Not in list of 1:1	
SLC19A3	Pob_R002765	Y	85	Not in list of 1:1	
GPSM1	Pob_R018481	Y	99	Not in list of 1:1	
IER3IP1		N			
TM6SF2		N			
SLC2A2		N			

10. MOLECULAR MODELLING

We downloaded the structure of hamster Pdx1 homeodomain bound to DNA (PDB ID 2h1k) and used chains b, e and f (43) as recently done by other groups (44-45). The amino acid sequence of the hamster Pdx1 homeodomain is 100% identical to that of mouse, human and most vertebrates, and the inferred ancestral vertebrate Pdx1 sequence, so can be considered the 'normal' Pdx1 structure. For hamster the sequence of the simulated target DNA was 5'-TCTCTAATGAGTTTC-3' in complex with 5'-AGAAACTCATTAGAG-3'. For sand rat (*Psammomys obesus*) the sequence of the simulated target DNA was 5'-TCCTTAATGGGCCAA-3' in complex with 5'-ATTGGCCCATTAAGG-3'; this sequence is derived from the orthologous region of the *insulin* gene in the sand rat genome determined in this study. For hamster (= normal vertebrate) the sequence of the simulated protein used was RTRTAYTRAQLLELEKEFLFNKYISRPRRVELAVMLNLTERHIKIWFQNRRMKWKK.
For sand rat the sequence of the simulated protein was
RTRTLYTRAQRLELEKEFLFSRYVARPRRVELARALNLTEKHVKVWFQNRRMRWKR.
The simulated protein sequences correspond to residues 3 to 58 of the commonly accepted nomenclature for the homeodomain proteins. We used the software package Rosetta (46) to make *in silico* point mutations on the protein and on the DNA to obtain the sand rat system starting from the hamster structure. We also prepared two "hybrid systems", HYB-1 and HYB-2, to understand the affinity of binding for all the combinations of protein-DNA complexes. The HYB-1 system consists of the protein with the hamster sequence bound to target the DNA with the sand rat sequence. The HYB-2 system of the protein with the sand rat sequence bound to the target DNA with the hamster sequence.

We performed molecular dynamics simulations with the Gromacs package (47) and the atomistic force-field Amber99SB-ildn-star for protein (48-51), with Parmbsc0 corrections for DNA and RNA ported from Amber to Gromacs format (52-53), with Chi-OL3 corrections for RNA (54) and ions from (55) as downloaded from the Bussi github page (56). The protein-DNA complexes were solvated and counter-ions added to neutralise the total charge.

The molecular dynamics simulations were performed following a standard procedure that consists of two phases: the first phase consists of equilibrating the system to reach thermodynamic stability at 300 K and 1 atm; the second phase is the 'production phase' that is a molecular dynamics simulation of 200 ns at 300 K and 1 atm from which 10,000 conformers of the protein-DNA complexes were extracted to run the analysis. The enthalpy of binding between Protein and DNA and the contribution per residue to the enthalpy of binding were calculated from the selected conformers with MMPBSA (Molecular Mechanics Poisson Boltzmann Surface Area) using the single trajectory approach (57-58). Calculations were performed using a reduced system with respect to the simulated one to have 14 complete Watson-Crick pairs per strand: for hamster the sequence of the DNA was 5'-CTCTAATGAGTTTC-3' in complex with 5'-GAAACTCATTAGAG-3', for sand rat the sequence of the DNA simulated was 5'-CCTTAATGGGCCAA-3' in complex with 5'-TTGGCCCATTAAGG -3'.

Figures of protein-DNA complexes have been generated using VMD (59). The statistical analysis was performed with the Python programming language (60).

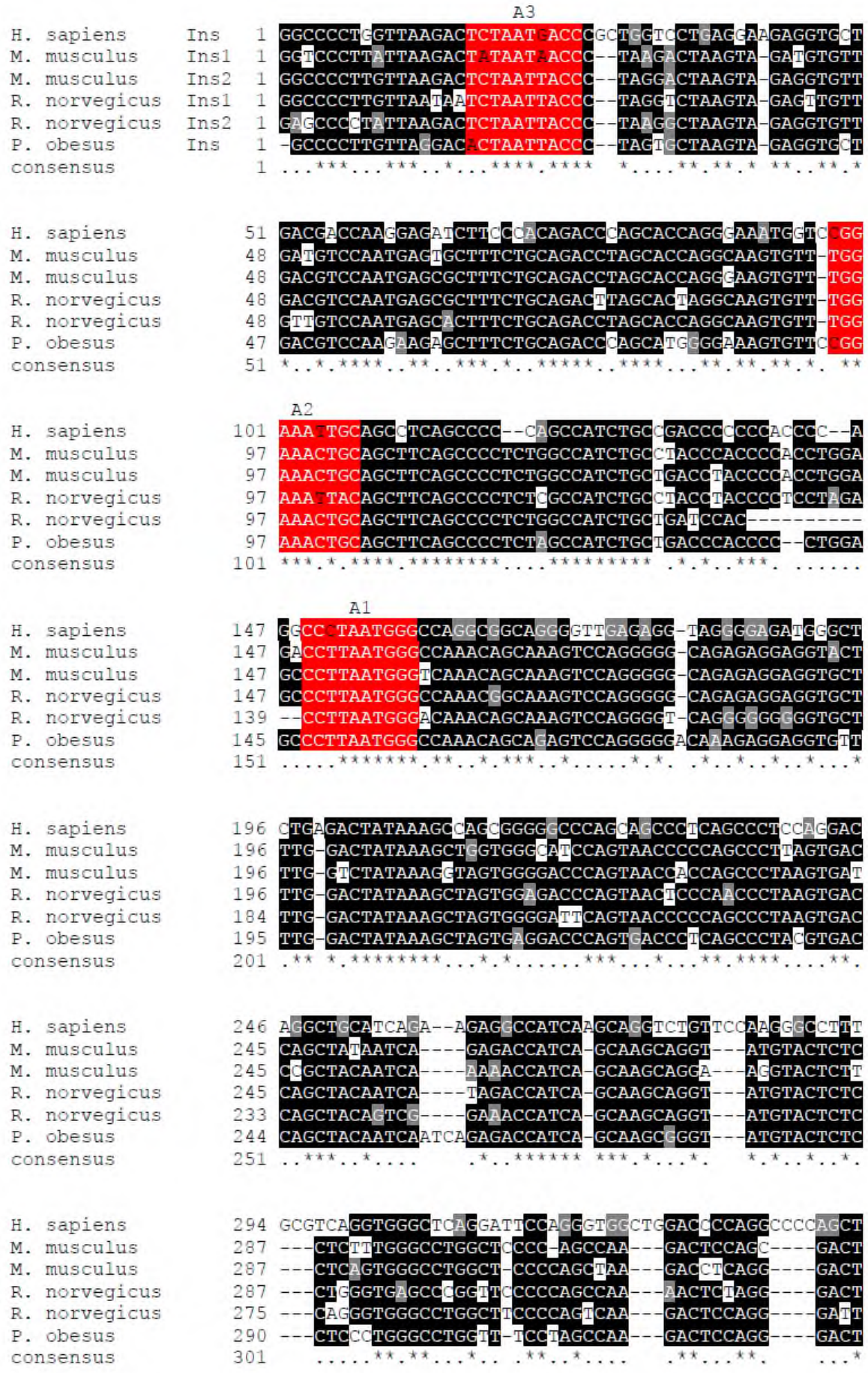
11. PROMOTER CONSERVATION OF PDX1 DOWNSTREAM TARGET GENES

We downloaded the upstream promoter regions from three Pdx1 target genes, namely *insulin* (*Ins*), *somatostatin* (*Sst*) and *glucokinase* (*Gck*). Sequence from mouse (*Mus musculus*), rat (*Rattus norvegicus*) and chinchilla (*Chinchilla lanigera*) were downloaded from Ensembl and the sequences for sand rat (*Psammomys obesus*) were extracted from our final genome assembly. All sequences terminate at the start codon. Multiple sequence alignment was carried out using ClustalW (41) and conserved Pdx1 binding sites were subsequently annotated manually. We detect negligible mutation in the conserved Pdx1 binding sites in the *insulin* (A1, A2 and A3 (61)), *somatostatin* (UE-B, TAAT1 and TAAT2 (62)) and *glucokinase* (UPE1 and UPE3 (63)) promoters (Figures S7-S9).

12. SELECTION ANALYSIS

GC-biased mutation in the sand rat

We counted the A/T and G/C mutations in a pairwise manner between sand rat, rat and mouse in genes located in the mutation hotspot region and randomly selected locations. Coding and intron sequences of sand rat, rat and mouse were extracted and aligned using MUSCLE (36) (version v3.8.31) and MAFFT (33) (version v7.222), respectively, with mutation types counted according to the generated alignments (Table S11). For genes located within the mutation hotspot, both protein coding and intronic regions presented a significantly elevated “W to S” (“weak” A/T to “strong” G/C) mutation rate in the sand rat compared to gene sequences from this region in Mouse and Rat. We also compared this with 100 randomly selected genes in the sand rat genome (Table S12), which showed that genes in the hotspot region are strongly biased towards GC when compared to genes located in other genomic regions.


A3
A2
A1
H. sapiens Ins 1
M. musculus Ins1 1
M. musculus Ins2 1
R. norvegicus Ins1 1
R. norvegicus Ins2 1
P. obesus Ins 1
consensus 1
H. sapiens 51
M. musculus 48
M. musculus 48
R. norvegicus 48
R. norvegicus 48
P. obesus 47
consensus 51
H. sapiens 101
M. musculus 97
M. musculus 97
R. norvegicus 97
R. norvegicus 97
P. obesus 97
consensus 101
H. sapiens 147
M. musculus 147
M. musculus 147
R. norvegicus 147
R. norvegicus 139
P. obesus 145
consensus 151
H. sapiens 196
M. musculus 196
M. musculus 196
R. norvegicus 196
R. norvegicus 184
P. obesus 195
consensus 201
H. sapiens 246
M. musculus 245
M. musculus 245
R. norvegicus 245
R. norvegicus 233
P. obesus 244
consensus 251
H. sapiens 294
M. musculus 287
M. musculus 287
R. norvegicus 287
R. norvegicus 275
P. obesus 290
consensus 301

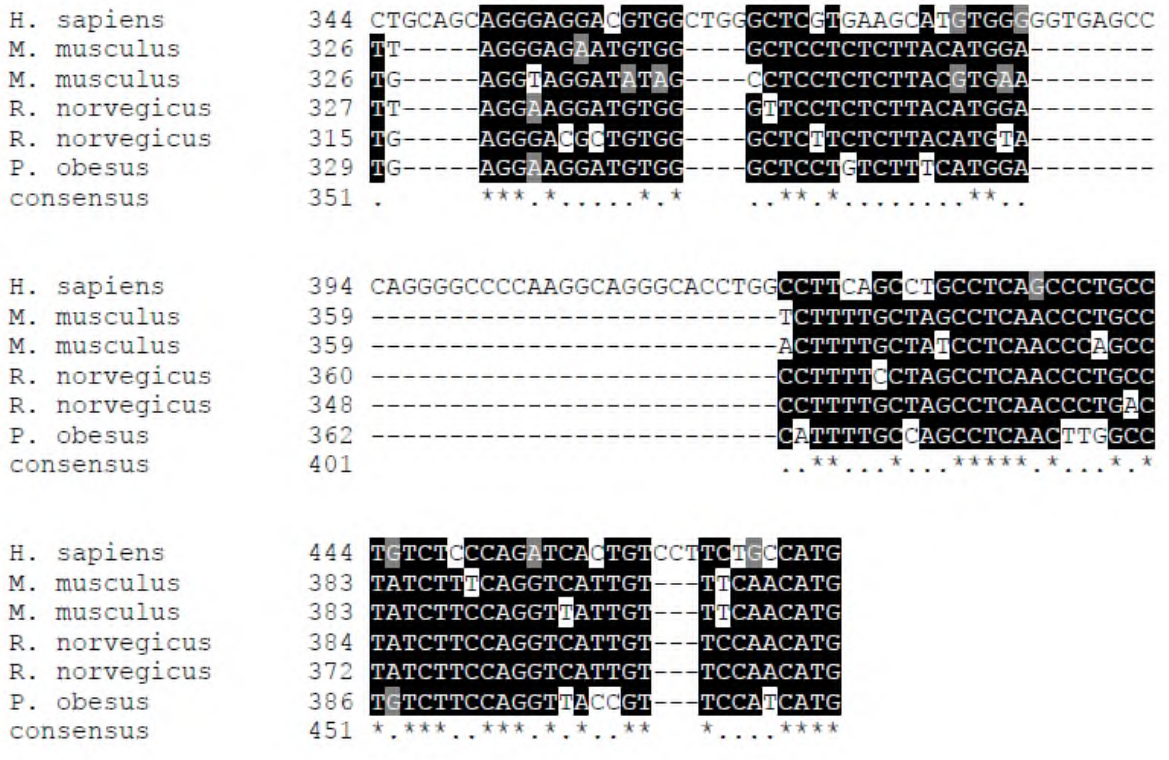
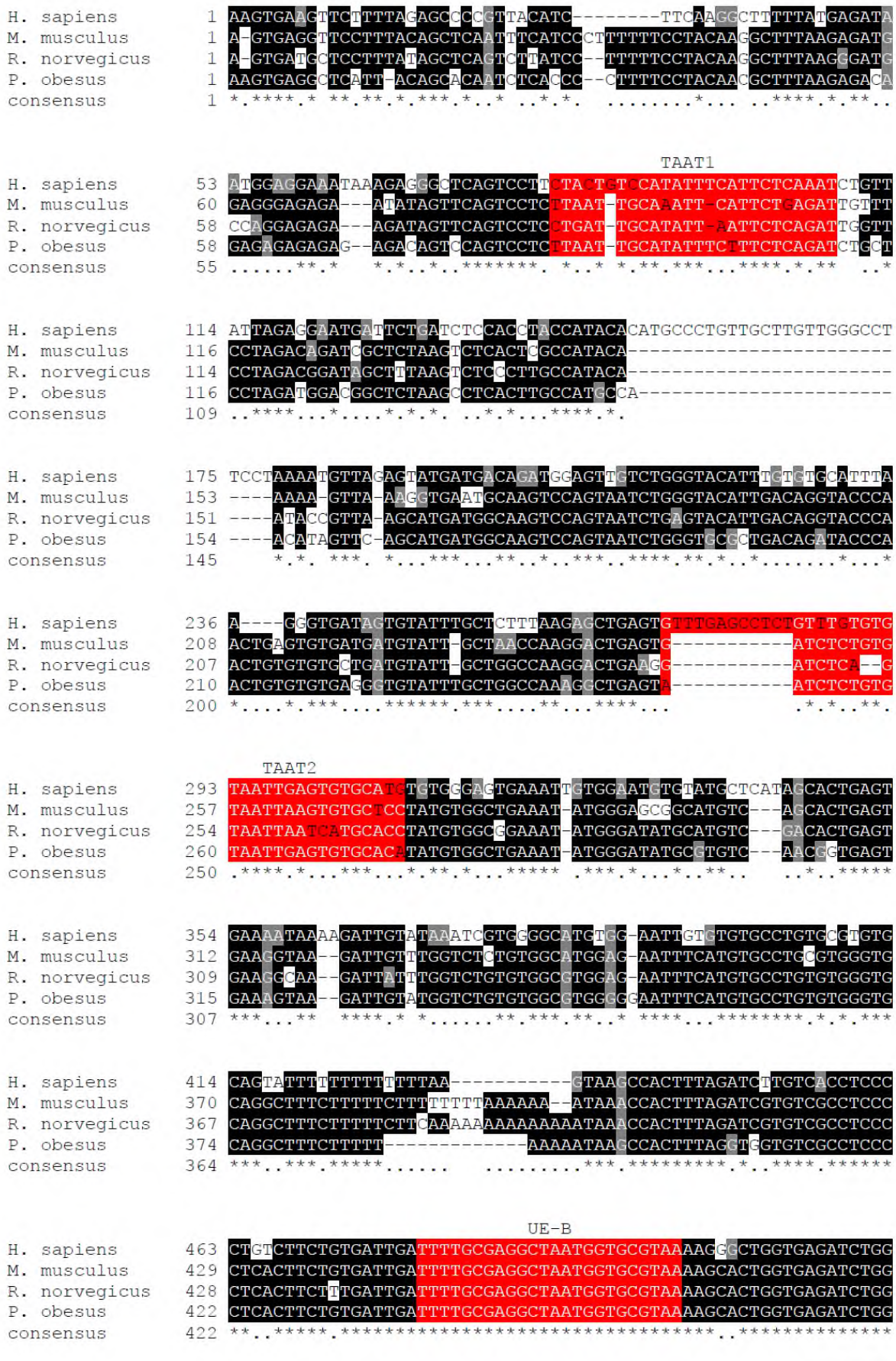


Figure S7. Alignment of the upstream insulin promoter from human, mouse, rat and sand rat. The conserved Pdx1 binding regions A1, A2 and A3 are highlighted in red.


TAAT1
TAAT2
UE-B

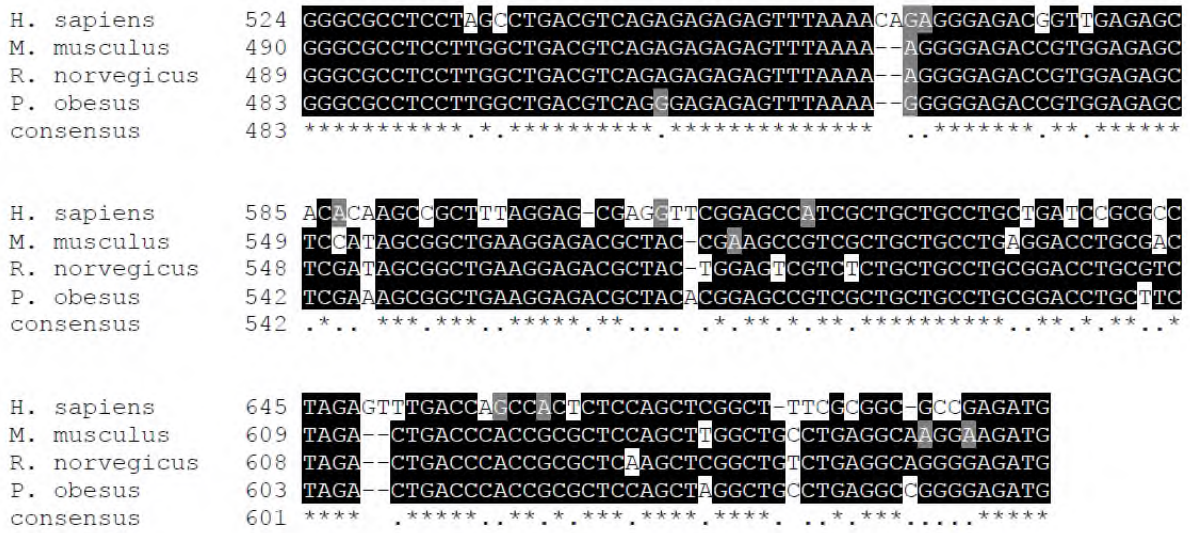


Figure S8. Alignment of the upstream somatostatin promoter from human, mouse, rat and sand rat. The conserved Pdx1 binding regions UE-B, TAAT1 and TAAT2 are highlighted in red.

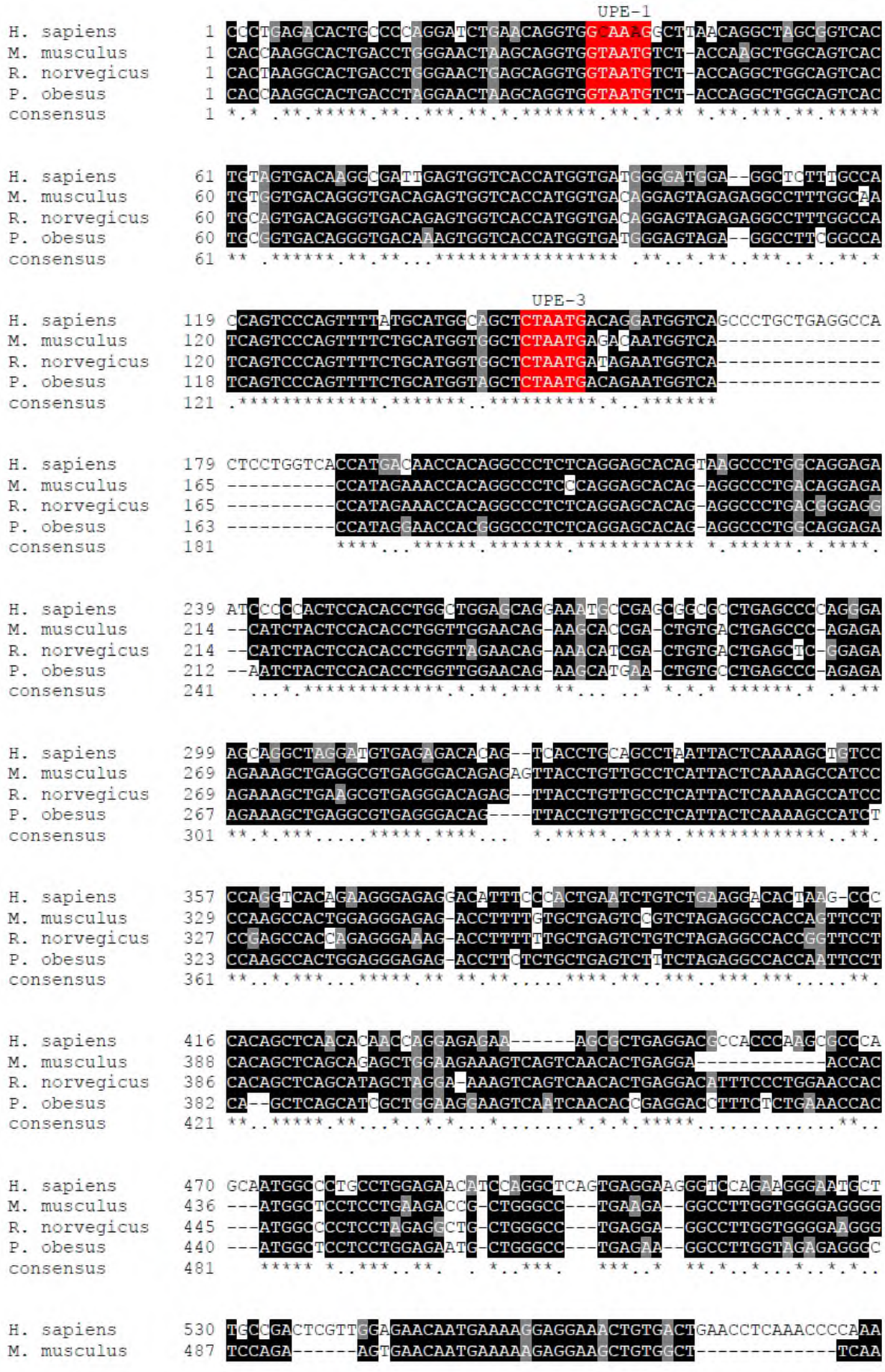

UPE-1
UPE-3
H. sapiens
M. musculus
R. norvegicus
P. obesus
consensus

Figure S9. Alignment of the upstream glucokinase promoter from human, mouse, rat and sand rat. The conserved Pdx1 binding regions UPE1 and UPE3 are highlighted in red.

Table S11. Comparison between the sand rat mutational hotspot and the corresponding syntenic Mouse/Rat genomic region. A significantly strong "W to S" mutation bias in the sand rat was observed in this region compared to Mouse and Rat.

		S to W	W to S	Chisq test p-value
Coding region	Mouse to sand rat	2,361	5,571	< 2.2e-16
	Mouse to Rat	1,695	2,028	
	Rat to sand rat	2,414	5,320	< 2.2e-16
	Rat to Mouse	2,028	1,695	
Intron region	Mouse to sand rat	43,817	64,272	< 2.2e-16
	Mouse to Rat	51,946	64,757	
	Rat to sand rat	45,370	60,488	< 2.2e-16
	Rat to Mouse	64,757	51,946	

Table S12. Comparison between the mutation hotspot and other randomly selected genomic regions in the sand rat genome. A significantly strong "W to S" mutation bias was observed compared to other regions in the sand rat genome.

			S to W	W to S	Chisq test p-value
Coding region	Mouse to Gerbil	genes in mutation hotspot	2,361	5,571	< 2.2e-16
		genes picked randomly	7,596	8,419	
	Rat to Gerbil	genes in mutation hotspot	2,414	5,320	< 2.2e-16
		genes picked randomly	8,581	9,249	
Intron region	Mouse to Gerbil	genes in mutation hotspot	43,817	64,272	< 2.2e-16
		genes picked randomly	254,189	260,502	
	Rat to Gerbil	genes in mutation hotspot	45,370	60,488	< 2.2e-16
		genes picked randomly	267,180	259,304	

Mutational bias analysis

We also calculated the *dS* for genes in the region of biased mutation and also genes from other genomic locations which were picked randomly using PAML (64) (version 4.8) with the free-ratios model. For each gene, the protein sequences from sand rat and three other species of rodent (mouse, rat, Guinea pig) were aligned using PRANK (65) followed by removing any poorly aligned sites using Gblocks (66). The alignments were concatenated for 26 genes from the mutation hotspot region and 100 randomly selected genes, respectively, and passed to PAML for *dS* calculation. The *dS* trees are presented in Figure 1d and demonstrate that the mutational bias is significantly skewed towards G/C bases in the 26 genes in the mutation hotspot region in sand rat compared to the syntenic region in 3 other species of rodent and additionally when compared to 100 other randomly selected genes outside of this region.

Analysis of vertebrate *Pdx1* dN/dS

The Pdx1 protein sequences from sand rat and 15 other species were aligned using PRANK (65) followed by removing sites containing gaps and missing data using Gblocks (66). There are a total of 234 positions in the final CDS dataset. Ancestral states were inferred using the Maximum Parsimony method (67) in MEGA7 (68). We calculated synonymous mutation number and non-synonymous mutation number between the child nodes and ancestral nodes (Figure S10).

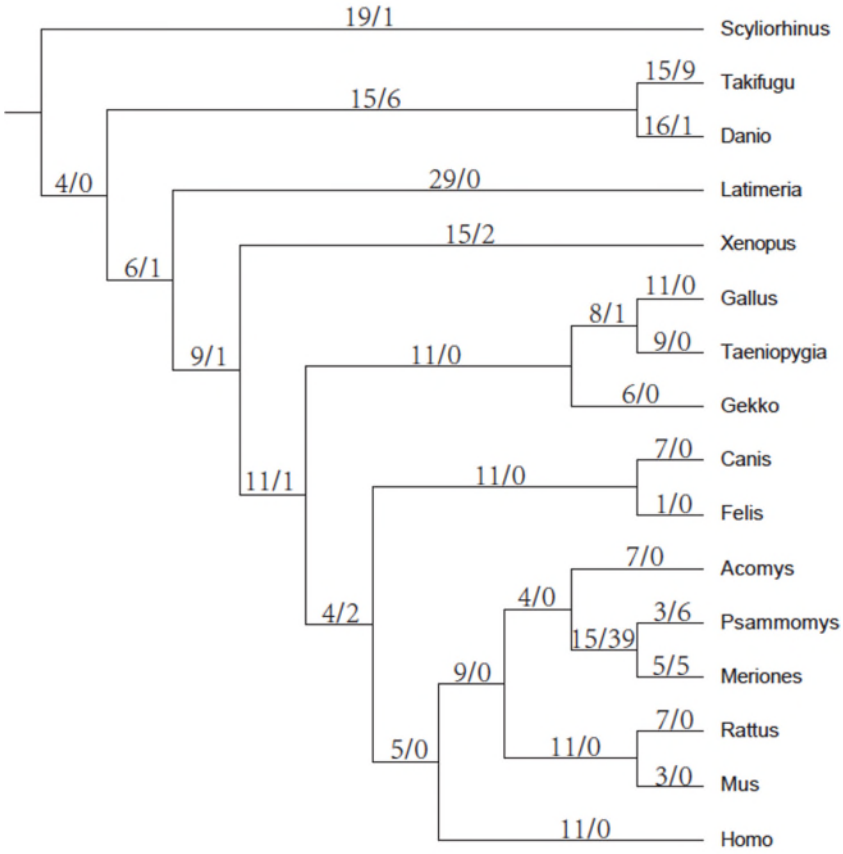


Figure S10. Synonymous mutation number/Non-synonymous mutation numbers were calculated for the child nodes and ancestral nodes in species tree. Ancestral states were inferred using the Maximum Parsimony method (67). The gene has accumulated significantly higher number of non-synonymous substitutions in the common ancestor of *Psammomys* and *Meriones*.

Functional constraint

Evidence for functional constraint on the sand rat Pdx1 coding region is given by the finding that two assembled transcripts from the same animal reveal polymorphism in the 3' untranslated region but not the coding sequence (Figure S11).

```
CLUSTAL O(1.2.3) multiple sequence alignment


SANDRAT_k51_1890640     ATGGACAGAGAGGCCGAGCCCTTCTTCGAGGCCTCCTGGGCGTTCCCGGGGCCCGAGTTC
SANDRAT_k55_1486975     ATGGACAGAGAGGCCGAGCCCTTCTTCGAGGCCTCCTGGGCGTTCCCGGGGCCCGAGTTC
                        ************************************************************

SANDRAT_k51_1890640     GCGGCCCCCGCCTCCTCCTGCCTGTTCGAGGGTGGGGGCGGGCAGCCTCCCCCCCACGCT
SANDRAT_k55_1486975     GCGGCCCCCGCCTCCTCCTGCCTGTTCGAGGGTGGGGGCGGGCAGCCTCCCCCCCACGCT
                        ************************************************************

SANDRAT_k51_1890640     CCTCCCCACGCTCCTCCCCACCTCGCCCCGTGCTCCCTGGACCCCACCGGCCTCCAGCCG
SANDRAT_k55_1486975     CCTCCCCACGCTCCTCCCCACCTCGCCCCGTGCTCCCTGGACCCCACCGGCCTCCAGCCG
                        ************************************************************

SANDRAT_k51_1890640     CCCCAGCCCGGGGTCCCCCCGCCGCCACCCGGGGGCCCCGACCAACCGCCCTTTGCCTGG
SANDRAT_k55_1486975     CCCCAGCCCGGGGTCCCCCCGCCGCCACCCGGGGGCCCCGACCAACCGCCCTTTGCCTGG
                        ************************************************************

SANDRAT_k51_1890640     ATGAAGAGCAGCAAAGGCCAAGCCTGGAGCGGCCAGTGGGCAGCCCCCGGCCGCAGGACT
SANDRAT_k55_1486975     ATGAAGAGCAGCAAAGGCCAAGCCTGGAGCGGCCAGTGGGCAGCCCCCGGCCGCAGGACT
                        ************************************************************

SANDRAT_k51_1890640     CGAACCCTGTACACGCGGGCGCAGCGGCTGGAGCTGGAGAAGGAATTCCTCTTCAGCCGC
SANDRAT_k55_1486975     CGAACCCTGTACACGCGGGCGCAGCGGCTGGAGCTGGAGAAGGAATTCCTCTTCAGCCGC
                        ************************************************************

SANDRAT_k51_1890640     TACGTCGCGCGGCCGCGGCGCGTGGAGCTCGCGCGGGCGCTGAACCTCACCGAGAAGCAC
SANDRAT_k55_1486975     TACGTCGCGCGGCCGCGGCGCGTGGAGCTCGCGCGGGCGCTGAACCTCACCGAGAAGCAC
                        ************************************************************

SANDRAT_k51_1890640     GTGAAGGTCTGGTTCCAGAACCGCCGCATGCGCTGGAAGAGGGAGGAGTCCGCGCGGGGG
SANDRAT_k55_1486975     GTGAAGGTCTGGTTCCAGAACCGCCGCATGCGCTGGAAGAGGGAGGAGTCCGCGCGGGGG
                        ************************************************************

SANDRAT_k51_1890640     AGGACGGCGCCCCGGGAGGACGGAGGAGCGGGAGGCTCCCCGCCACCGTCCTCCTCCTCC
SANDRAT_k55_1486975     AGGACGGCGCCCCGGGAGGACGGAGGAGCGGGAGGCTCCCCGCCACCGTCCTCCTCCTCC
                        ************************************************************

SANDRAT_k51_1890640     TCCTCCTCCTCCTCCTCCTCCTCCGCGGCCCGGGATGCTCCTCCTCTTCTTCTCCCTCCT
SANDRAT_k55_1486975     TCC---------TCCTCCTCCTCCGTGGCCCGGGATGCTCCTCCTCTTCTTCTCCCTCCT
                        ***         ************* ***********************************


SANDRAT_k51_1890640     CCTCCTCCTCCTACCGGGGGACTGCGGTGAGGGGGTCGGTGACTCCTCCTCCTCTTCCCC
SANDRAT_k55_1486975     CCTCCTCCTCCTACGGGGGGACTGCGGTGAGGGGGTCGGGGACTCCTCCCTCTCCTCCTC
                        ************** ********************** *********  *** *** * 

SANDRAT_k51_1890640     CTCCTGCTCCTCCTCCTCCTCCTCCCCCTCCTCTTTCTCCTCCT---CCCACTCCTCCTC
SANDRAT_k55_1486975     CTCCTCCTCCCACTCCTCCCCCTCTTCCTCCTCCTCCTCTCACTCCTCCCCTTCTTCCTC
                        ***** ****  ******* ****  ******* * ***   **   ***  ** *****

SANDRAT_k51_1890640     CTCCTCCTCCTCCTCCCACTCCTCCTCCTCCTCCTCCCCCTCCTCCCC------------
SANDRAT_k55_1486975     CTCCTCCTCCTCCTCTCACTCCTCCCCTTCTTCCTCCTCCTCCTCCTCCCACTCCTC
                        *************** ********* * ** ****** ******** *

SANDRAT_k51_1890640     --------------------------------
SANDRAT_k55_1486975     CCCCTCTTCCTCCTCCTTCTCCTCCTCCTCC
```

Figure S11. Alignment of two sand rat Pdx1 transcripts. Red coding sequence; bold homeobox; black 3' untranslated region.

12. SUPPLEMENTARY REFERENCES

1. Luo R et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1: 18.

2. Marçais G, Kingsford CA (2011) Fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27: 764-770.

3. H Wickham (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

4. Gregory TR (2016) Animal Genome Size Database. http://www.genomesize.com

5. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2: 231-239.

6. Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University. Available at http://www.bx.psu.edu/~rsharris/lastz/

7. Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0*. 2013-2015 (http://www.repeatmasker.org).

8. Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27: 2957-2963.

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

10. Simpson JT et al (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research* 19: 1117-1123.

11. Camacho C et al (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

12. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841-842.

13. Bradnam KR et al (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2: 10.

14. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210-3212.

15. Lacy PE, Kostianovsky M (1967) Method for the isolation of intact islets of Langerhans from the rat pancreas. *Diabetes* 16: 35-39.

16. Lopez JA, Bohuski E (2007) Total RNA extraction with TRIZOL reagent and purification with QIAGEN RNeasy mini kit. (available at https://wiki.cgb.indiana.edu/download/attachments/22446090/Total_RNA_Extraction_with_TRIZOL_Reagent_10172007.pdf).

17. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120.

18. Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). Available at https://github.com/najoshi/sickle.

19. Robertson G et al (2010) De novo assembly and analysis of RNA-seq data. *Nature methods* 7: 909-912.

20. Grabherr MG et al (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* 29: 644-652.

21. Haas BJ et al (2013) De novo transcript sequence reconstruction from RNA-seq: reference generation and analysis with Trinity. *Nature protocols* 8: 1494-1512. (Available at http://transdecoder.github.io/).

22. Bairoch A et al (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Research* 33: D154-D159.

23. Smit, AFA, Hubley, R, Green P. *RepeatMasker Open-4.0*. 2013-2015 (http://www.repeatmasker.org).

24. Smit, AFA, Hubley, R. *RepeatModeler Open-1.0*. 2008-2015 <http://www.repeatmasker.org>.

25. Goubert C et al (2015) De novo assembly and annotation of the Asian Tiger Mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). *Genome Biology and Evolution* 7: 1192-1205.

26. Keller O, Kollmar M, Stanke M, Waack S (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 15: 757-763.

27. Birney E, Clamp M, Durbin R (2004) GeneWise and GenomeWise. *Genome Research* 14: 988-995.

28. Elsik CG et al (2007) Creating a honey bee consensus gene set. *Genome Biology* 8: R13.

29. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* 25: 1105-1111.

30. Trapnell C et al (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* 28: 511-515.

31. Clarke AG et al (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302: 1960-1963.

32. Kinsella RJ et al (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011: bar030.

33. Katoh K, Misawa K, Kuma K-I, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059-3066.

34. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org).

35. Zhong Y-F, Butts T, Holland PWH (2008) HomeoDB: a database of homeobox gene diversity. *Evolution and Development* 10: 516-518.

36. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 19: 1792-1797.

37. Tamura K et al (2011) MEGA5: Molecular evolutionary genetic analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28: 2731-2739.

38. Gaulton KJ et al (2015) Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature genetics* 47: 1415-1425.

39. Mahajan A et al (2015) Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define and effector transcript at the G6PC2-ABCB11 locus. *PLOS Genetics* 11: e1004876.

40. Prokopenko I et al (2008) Variants in MTNR1B influence fasting glucose levels. *Nature Genetics* 41: 77-81.

41. Yamagata K et al (1996) Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3). *Nature* 384: 455-458.

42. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic acids research* 22: 4673-4680.

43. Longo A, Guanga GP, Rose RB (2007) Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain. *Biochemistry* 46: 2948-2957.

44. Bastidas M, Showalter SA (2013) Thermodynamic and structural determinants of differential Pdx1 binding to elements from the insulin and IAPP promoters. *Journal of Molecular Biology.* 425: 3360-3377.

45. Babin V, Wang D, Rose RB, Sagui C (2013) Binding polymorphism in the DNA bound state of the Pdx1 homeodomain. *Plos Computational Biology* 9: e1003160.

46. Leaver-Fay A et al (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology* 487: 545-74.

47. Abraham MJ et al (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2: 19-25.

48. Cornell WD et al (1995) A Second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* 117: 5179–5197.

49. Hornak V et al (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65: 712–725.

50. Best RB, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix−coil transition of polypeptides. *The Journal of Physical Chemistry B 113:* 9004-9015.

51. Lindorff-Larsen K et al (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78: 1950–1958.

52. Pérez A et al (2007) Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophysical Journal* 92: 3817-3829.

53. Guy AT, Piggot TJ, Khalid S (2012) Single-stranded DNA within nanopores: conformational dynamics and implications for sequencing: a molecular dynamics simulation study. *Biophysical Journal* 103: 1028-1036.

54. Zgarbová M et al (2011) Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of Chemical Theory and Computation 7:* 2886-2902**.**

55. Joung IS, Cheatham TEIII (2008) Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry B 112:* 9020-9041.

56. Bussi G https://github.com/srnas/ff

57. Kollman PA (2000) Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research* 33: 889-897.

58. Miller BR et al (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation 8:* 3314-3321.

59. Humphrey W, Dalke A, Schulten K (1996) VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics* 14: 33-38.

60. Python Software Foundation. Python Language Reference, version 2.7. Available at http://www.python.org

61. McKinnon CM, Docherty K (2001) Pancreatic duodenal homeobox–1, PDX-1, a major regulator of beta cell identity and function. *Diabetologia* 44: 1203-1214.

62. Miller CP, McGehee Jr RE, Habener JF (1994) IDX-1: a new homeodomain transcription factor expressed in rat pancreatic islets and duodenum that transactivates the somatostatin gene. *The EMBO Journal* 13: 1145.

63. Magnuson MA, Jetton TL (1993) Evolutionary conservation of elements in the upstream glucokinase promoter. *Biochemical Society Transactions* 21: 160-163.

64. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586-1591.

65. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *PNAS* 102: 10557-10562.

66. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17: 540-552.

67. Eck RV, Dayhoff MO (1966) Atlas of Protein Sequence and Structure. National Biomedical Research Foundation, Silver Springs, Maryland.

68. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33: 1870-1874.