

1 **Genome sequence of a diabetes-prone desert rodent reveals**
2 **a mutation hotspot around the ParaHox gene cluster**

3
4 Adam D Hargreaves¹, Long Zhou², Josef Christensen³, Ferdinand Marlétaz^{1,4}, Shiping Liu², Fang Li²,
5 Peter Gildsig Jansen³, Enrico Spiga⁵, Matilde Thye Hansen³, Signe Vendelbo Horn Pedersen³, Shameek
6 Biswas⁶, Kyle Serikawa⁶, Brian A Fox⁶, William R Taylor⁵, John F Mulley⁷,
7 Guojie Zhang^{2,8,9*}, R Scott Heller^{3*} and Peter W H Holland^{1*}

8 *Joint corresponding authors

- 9 1- Department of Zoology, University of Oxford, Oxford, OX1 3PS, UK
10 2- China National Genebank, BGI-Shenzhen, 518083, Shenzhen, Guangdong, China
11 3- Novo Nordisk, Måløv, Denmark
12 4- Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University,
13 Onna, Okinawa 904-0495, Japan.
14 5- Francis Crick Institute, London, UK
15 6- Novo Nordisk Research Centre, Seattle, USA
16 7- School of Biological Sciences, Bangor University, UK
17 8- State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology,
18 Chinese Academy of Sciences, 650223, Kunming, China
19 9- Centre for Social Evolution, Department of Biology, Universitetsparken 15, University of
20 Copenhagen, DK-2100 Copenhagen, Denmark

21
22 **The sand rat *Psammomys obesus* is a gerbil native to deserts of North Africa and the Middle East¹.**
23 **Sand rats survive with low caloric intake and when given high carbohydrate diets can become obese**
24 **and develop type II diabetes² which, in extreme cases, leads to pancreatic failure and death^{3,4}.**
25 **Previous studies have reported inability to detect the *Pdx1* gene or protein in gerbils⁵⁻⁷, suggesting**
26 **that absence of this key insulin-regulating homeobox gene might underlie diabetes susceptibility.**
27 **Here we report sequencing of the sand rat genome and discovery of an extensive, mutationally-**
28 **biased GC-rich genomic domain encompassing many essential genes, including the elusive *Pdx1*. The**
29 **sequence of *Pdx1* has been grossly affected by GC-biased mutation leading to the highest divergence**
30 **observed in the animal kingdom. In addition to molecular insights into restricted caloric intake in a**
31 **desert species, the discovery that specific chromosomal regions can be subject to elevated mutation**
32 **rate has widespread significance to evolution.**

33 Linking molecular change to phenotypic change is a central goal of evolutionary biology. Adaptation to
34 arid environments is particularly interesting because of the extreme physiological demands imposed by
35 low food and water availability. The sand rat *Psammomys obesus* (Fig. 1a) is a member of the subfamily
36 Gerbillinae, most species of which live in deserts and arid environments (Fig. 1b). *P. obesus* has emerged
37 as a model for research into diet-induced type II diabetes because, if provided with high carbohydrate
38 diets, the majority of individuals become obese and develop classic diabetes symptoms, in the most
39 extreme cases leading to pancreatic failure and death^{2,3}.

40 In searching for the molecular basis of this unusual phenotype, attention has been paid to the *Pdx1*
41 homeobox gene, also called *Ipf1*, *Idx1*, *Stf1* or *Xlox*⁸⁻¹², the central and most highly conserved member
42 of the ParaHox gene cluster¹³. *Pdx1* is the only member of the Pdx gene family in tetrapods, and encodes
43 a homeodomain that has been invariant across their evolution. Mammalian *Pdx1* is expressed in
44 pancreatic beta-cells¹⁴ and encodes a homeodomain transcription factor that acts as a transcriptional
45 activator of *insulin* and other pancreatic hormone genes^{15,16}. A pivotal role in insulin regulation is also
46 reflected in the association of heterozygous *Pdx1* mutations with maturity-onset diabetes of the young
47 (*MODY4*) and type II diabetes mellitus in humans¹⁶. Contrary to the usual conservation, several studies
48 have reported inability to detect *Pdx1* in gerbils, including *P. obesus*, by immunocytochemistry, Western
49 blotting or PCR⁵⁻⁷, leading to the hypothesis that the gene has been lost, compromising ability to
50 regulate insulin. Such a conclusion would raise further questions, since in addition to its adult functions,
51 *Pdx1* is also essential for pancreatic development in the embryo. For example, targeted deletion in mice
52 causes loss of pancreas and anterior duodenum and is lethal^{8,17}. In humans, pancreatic agenesis has
53 been reported in a patient with a homozygous frameshift mutation before the *Pdx1* homeobox, and in
54 a compound heterozygous patient with substitution mutations in helices 1 and 2 of the homeodomain¹⁸⁻
55 ²⁰.

56 To resolve the conundrum of a putatively absent 'essential' gene, we sequenced the *P. obesus* genome
57 using a standard shotgun strategy (Illumina), using a combination of short and long insert libraries,
58 initially at 85.5X coverage (Supplementary information section 1). This assembly lacked a *Pdx1* gene
59 supporting the prevailing hypothesis of a loss of the *Pdx1* gene in gerbils. However, a synteny
60 comparison between *P. obesus* and other mammals delineated a contiguous block of 88 genes
61 (Supplementary table S2.3.1) missing from the assembly including several genes essential to basic
62 cellular functions, such as *Brca2* and *Cdk8*, in addition to *Pdx1*. This led us to suspect that standard short
63 read sequencing may have given an incomplete genome assembly. To resolve whether this represented
64 a large-scale deletion or an unusual genomic region, we sequenced the transcriptomes of *P. obesus*
65 liver, pancreatic islets and duodenum, which strikingly contained transcripts for many of the missed
66 genes (Supplementary information section 2; Supplementary table S2.3.1). Furthermore, these
67 transcripts show unusually high GC content in most cases, indicating that a large contiguous stretch of
68 elevated GC had either been under-represented in initial sequencing data or had failed to assemble
69 correctly, most likely due to nucleotide compositional bias. We term such cryptic or hidden sequence
70 'dark DNA'. We therefore isolated GC-rich *P. obesus* genomic DNA by Caesium Chloride gradient
71 centrifugation, sequenced this fraction after limited amplification using Illumina MiSeq overlapping
72 paired-end reads, and re-assembled the genome incorporating this sequence data (Supplementary
73 information section 1.5). This gave a refined assembly with a total size of 2.38 Gb and a scaffold N50 of
74 10.4 Mb (Table 1; Supplementary information sections 1,3,4,6), including much of the 'dark DNA' region
75 in several scaffolds, and contains genes syntenic to a region of chromosome 12 in rat and a region of
76 chromosome 5 and the subtelomeric region of chromosome 8 in mouse. Comparison of GC content
77 between species demonstrates that sand rat genes are elevated in GC content across this chromosomal
78 region, syntenic to 12 Mb of the rat genome (Fig. 1c; Supplementary information section 9). This large
79 region encompasses a 250 kb repeat-rich scaffold containing the sand rat ParaHox cluster and its well-
80 characterised genomic neighbours. We inferred a high W (weak, A/T) to S (strong, G/C) allelic mutation
81 rate in this region of the *P. obesus* genome when compared with randomly selected genomic regions or
82 homologous regions in other species of rodent (Fig 1d; Supplementary information section 11,
83 Supplementary tables S11.1 & S11.2). The existence of a localised GC-biased stretch of the *P. obesus*
84 genome is striking and of far-reaching importance, and implies the existence of elevated and biased
85 mutational pressure, acting in one region of the genome.

86

87 The full coding sequence of the *P. obesus Pdx1* gene was deduced from the refined genome and
88 transcriptome assemblies, and the gene was found to be expressed in sand rat pancreatic islets and
89 duodenum (Supplementary information section 7). The 60 amino acid homeodomain of Pdx1 shows
90 100% conservation across other mammals for which data are available; however, in *P. obesus* there are
91 a remarkable 15 amino acid differences in the homeodomain, making this by far the most divergent
92 *Pdx1* gene discovered in the animal kingdom (Fig. 2a). All but one of the amino acid changes are caused
93 by A/T to G/C mutation. The N-terminal and C-terminal regions are also divergent with numerous
94 deletions, although the hexapeptide motif used in heterodimer formation with TALE proteins is
95 conserved (Fig. 2b). Despite its radical divergence, *Pdx1* is the closest homeodomain by blastp and
96 phylogenetic analysis places it as a rodent *Pdx1* on a long branch (Fig. 2c); extensive synteny with the
97 ParaHox region of mouse and rat confirms it is the true and single *Pdx1* ortholog (Supplementary table
98 S7.1). Evidence that the locus is functional includes expression in pancreas and duodenum, and the fact
99 that extensive polymorphism is found in the 3' untranslated region but is very limited in the coding
100 sequence (Supplementary information section 11), indicating that the coding region is under functional
101 constraint despite extensive mutation. Extreme deviation from the expected sequence explains why
102 antibodies and PCR failed to detect *Pdx1* in sand rat⁵⁻⁷.

103

104 These findings indicate that GC-biased mutation has driven radical changes in an otherwise highly
105 conserved homeobox gene; these changes could be maladaptive and constrain the physiological
106 capability of the sand rat, or adaptive enhancing ability to live in arid regions. To test if the extent of
107 sequence divergence is unusual for sand rat proteins, we calculated a 'protein weirdness index' (PWI)
108 (Supplementary information section 5) for all 1:1 mammalian orthologs by dividing mouse-human
109 protein sequence identity by mouse-sand rat sequence identity (Fig. 2d). This is distinct from identifying
110 the fastest evolving proteins, and specifically identifies proteins that have undergone uncharacteristic
111 divergence in sand rat. We find the majority of sand rat proteins are highly similar to mouse or human
112 (mode PWI = 1.0); in contrast, *Pdx1* is unusually divergent (mouse-sand rat 54.82%, mouse-human
113 91.37%; PWI = 1.67). To test if other genes implicated in glucose metabolism or pancreatic function are
114 also divergent, we compiled a list of 45 candidates from human studies including all genes implicated
115 in monogenic diabetes²¹ and genes for which coding sequence variants have been strongly associated
116 with T2D²². Of the 33 genes with clear 1:1:1 orthologs between human, mouse and sand rat, 32 lie
117 between position 225 and 10,195 in our PWI ranking, indicating that they are not unusually divergent
118 in sand rat. *Pdx1* is ranked 1st and is the most unusually divergent protein identified in the sand rat
119 predicted proteome (Supplementary information section 8; Supplementary table S5.1).

120 The mutations fixed in sand rat *Pdx1* gene do not cause frameshifts or truncations in known domains,
121 and molecular modelling reveals that the sand rat *Pdx1* homeodomain has the ability to form all three
122 helices required for DNA binding (Fig. 3a). To examine if these mutations have resulted in subtle effects
123 on the stability of DNA binding we deployed molecular dynamics simulations with atomistic
124 representation of *Pdx1* homeodomains, DNA target and solvent. From the post-processing of the
125 molecular dynamics simulations we estimated the enthalpy of binding between sand rat and mouse (or
126 other mammal) *Pdx1* and monomer DNA binding sites using the MM-PBSA (Molecular Mechanics
127 Poisson Boltzmann Surface Area) method (Supplementary information section 10). Target DNA
128 sequences used were core *Pdx1*-binding sites of the mouse *insulin* A1 promoter and its sand rat
129 ortholog. From 200 ns molecular dynamics simulations the enthalpy of binding for protein-DNA
130 interaction was calculated to be lower for sand rat than for mouse *Pdx1* (mean -140 kcal/mol vs. mean
131 -122 kcal/mol), indicative of sand rat *Pdx1* binding DNA more 'tightly' than is normal for the mammalian

132 Pdx1 protein (Fig. 3b). One amino acid change was responsible for much of the difference: a Leu to Arg
133 substitution in alpha helix 1 (homeodomain position 13), leading to the positive side chain of Arg making
134 a new indirect contact with the phosphate backbone of DNA. A second substitution, Val to Arg in alpha
135 helix 2 (homeodomain position 36), makes a smaller contribution (Fig. 3c). We also detect modifications
136 to specific base interactions, with sand rat residues Met54 and Arg58 making new contacts to A and T
137 bases within the TAAT core. Hence, stronger DNA binding is most likely driven by increased contacts
138 with the backbone of DNA, coupled with decreased sequence-specificity of DNA interaction. These
139 results suggest that sand rat Pdx1 is suboptimal in DNA-binding affinity and specificity.

140

141 We conclude that an unusual genomic region of biased mutation arose in the evolutionary lineage of
142 the sand rat. One consequence of this hotspot of mutation was the generation of GC-bias in the *Pdx1*
143 gene of *P. obesus*; this forced modification of the Pdx1 protein sequence, affecting its ability to regulate
144 insulin gene transcription and most likely transcription of other pancreatic genes. The sand rat Pdx1
145 hexapeptide, which mediates co-factor interactions²³, is intact, which may explain why pancreatic
146 development proceeds permitting viable sand rat embryogenesis. We suggest mutation-driven changes
147 have played a role in constraining or adapting the sand rat, and possibly other gerbil species, to arid
148 environments and low caloric intake. Biased gene conversion is a known mechanism that causes GC-
149 biased mutation^{24,25}; hence we suggest this mechanism, driven by elevated localised recombination, is
150 generating a hotspot of skewed base composition. The genomic region we describe here was not
151 detected by standard sequencing approaches, raising the possibility that other such dark DNA regions
152 could be widespread features of animal genomes, thus far largely overlooked in comparative animal
153 genomics. Indeed, GC-rich genes are also missing from the chicken genome assembly^{26,27}. Hotspots of
154 mutation could drive rapid evolutionary change at the molecular level, and it will be important to
155 decipher to what extent such hotspots have constrained and influenced evolutionary adaptation across
156 the animal kingdom.

157

158 **Author contributions**

159

160 ADH and PWHH conducted GC-rich DNA isolation, sequencing and analysis; LZ, SL, FL and GZ performed
161 genome assembly and annotation; MTH prepared DNA samples; SVHP and ADH extracted RNA samples;
162 KS, SB, BF and ADH performed RNA-seq and assembly; JC performed laboratory investigations
163 underpinning subsequent work; ADH, LZ, PGJ, JFM and FM undertook bioinformatic analyses; ES and
164 WRT ran molecular dynamic simulations; RSH, GZ and PWHH initiated and directed the research; PWHH,
165 ADH, LZ, GZ and RSH drafted the manuscript. All authors approved the final manuscript.

166

167 **Acknowledgements**

168

169 This work was funded principally by the European Research Council under the European Union's
170 Seventh Framework Programme (FP7/2007-2013 ERC grant 268513 awarded to PWHH), a Strategic
171 Priority Research Program of the Chinese Academy of Sciences (XDB13000000 awarded to GZ) and Novo
172 Nordisk A/S (coordinated by RSH). ES and WRT were supported by the Francis Crick Institute under
173 awards: FC001179. The Crick receives its core funding from Cancer Research UK,
174 the UK Medical Research Council, and the Wellcome Trust. We thank Natasha Ng, Gemma Marfany,
175 Thomas Dunwell, Fei Xu, Shan Quah, Anna Gloyn, Christine Hirschberger, Juliane Cohen, Rhys Morgan,
176 Lorna Witty, Monica Martinez Alonso and Thomas Brekke for assistance and advice, and the Oxford
177 Genomics Centre for GC-rich sequencing.

178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220

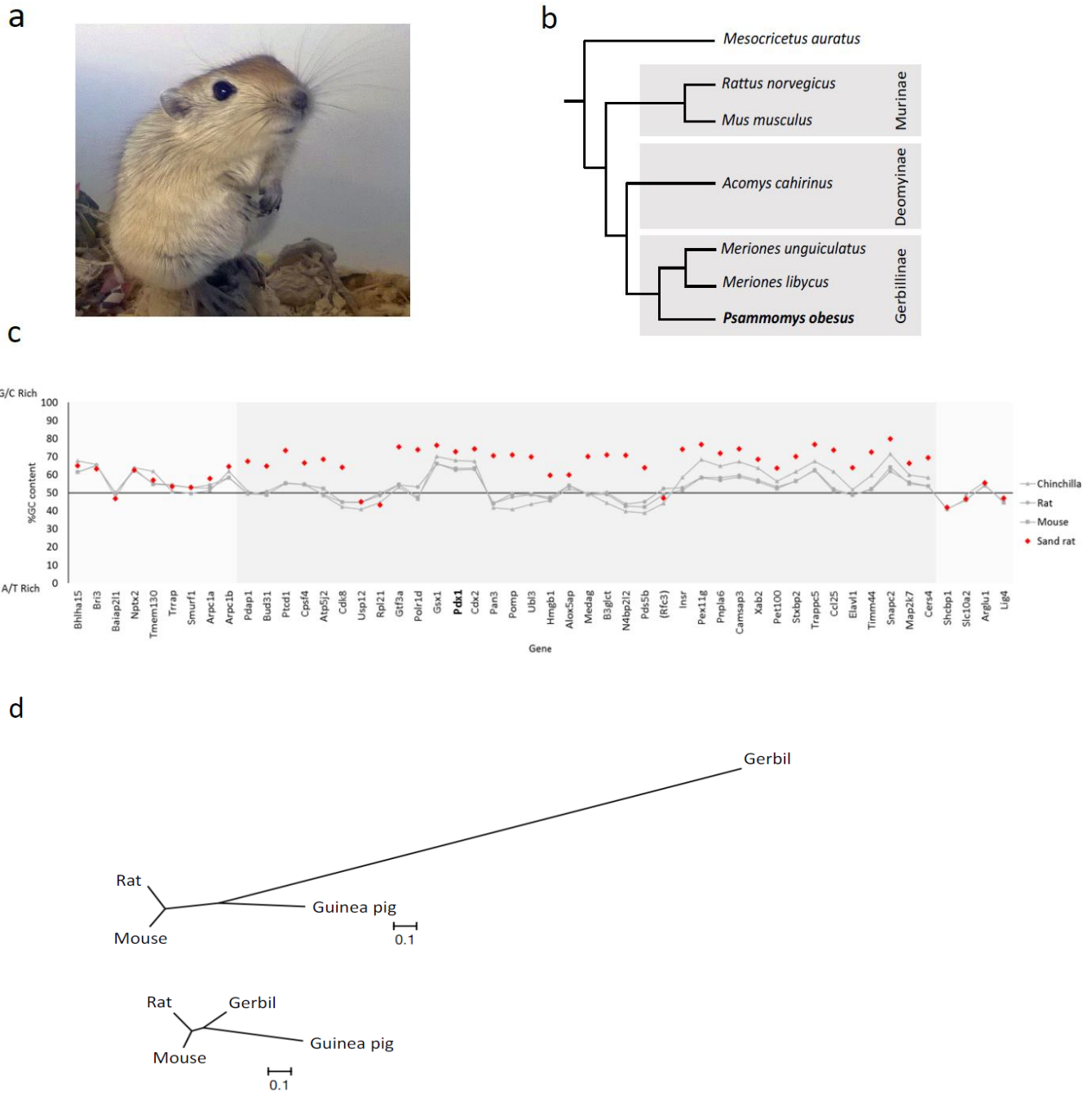
References

1. Kalman, R., Ziz, E., Galila, L. & Shafrir, E. Sand rat. *The laboratory rabbit, guinea pig, hamster, and other rodents* 1171-1190 (Elsevier Inc, 2012).
2. Schmidt-Nielsen, K., Haines, H. B. & Hackel, D. B. Diabetes mellitus in the sand rat induced by standard laboratory diets. *Science* **143**, 689-90 (1964).
3. Kaiser, N., et al. *Psammomys obesus*, a model for environment-gene interactions. *Diabetes* **54** Suppl 2:S137-44 (2005).
4. Bar-On, H., Ben-Sasson, R., Ziv, E., Arar, N. & Shafrir, E. Irreversibility of nutritionally induced NIDDM in *Psammomys obesus* is related to β cell apoptosis. *Pancreas* **18**, 259-265 (1999).
5. Leibowitz, G. et al. IPF1/PDX1 deficiency and β -cell dysfunction in *Psammomys obesus*, an animal with type 2 diabetes. *Diabetes* **50**, 1799-1806 (2001).
6. Vedtofte, L., Bødvarsdóttir, T. B., Karlsen, A. E. & Heller, R. S. Developmental biology of the *Psammomys obesus* pancreas: cloning and expression of the *Neurogenin-3* gene. *J. Histochem. Cytochem.* **55**, 97-104 (2007).
7. Gustavsen, C. R. et al. The morphology of islets of Langerhans is only mildly affected by the lack of Pdx-1 in the pancreas of adult *Meriones jirds*. *Gen. Comp. Endocrinol.* **159**, 241-249 (2008).
8. Offield, M. F. et al. PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* **122**, 983-995 (1996).
9. Ohlsson, H., Karlsson, K. & Edlund, T. IPF1, a homeodomain-containing transactivator of the insulin gene. *EMBO. J.* **12**, 4251-4259 (1993).
10. Miller, C. P., McGehee, R. E. & Habener, J. F. IDX-1: a new homeodomain transcription factor expressed in rat pancreatic islets and duodenum that transactivates the somatostatin gene. *EMBO. J.* **13**, 1145-1156 (1994).
11. Leonard, J. B., Peers, T., Johnson, K., Ferrere, S., Lee, S. & Montminy, M. Characterization of somatostatin transactivating factor-1, a novel homeobox factor that stimulates somatostatin expression in pancreatic islet cells. *Mol. Endocrinol.* **7**, 1275-1283 (1993).
12. Bürglin, T. R. A comprehensive classification of homeobox genes In: *A Guidebook to Homeobox Genes*. Duboule, D. editor. 25-71. (Oxford University Press, Oxford 1994).
13. Brooke, N. M., Garcia-Fernández, J. & Holland, P. W. H. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* **392**, 920-922 (1998).
14. Servitja, J. M. & Ferrer, J. Transcriptional networks controlling pancreatic development and beta cell function. *Diabetologia* **47**, 597-613 (2004).

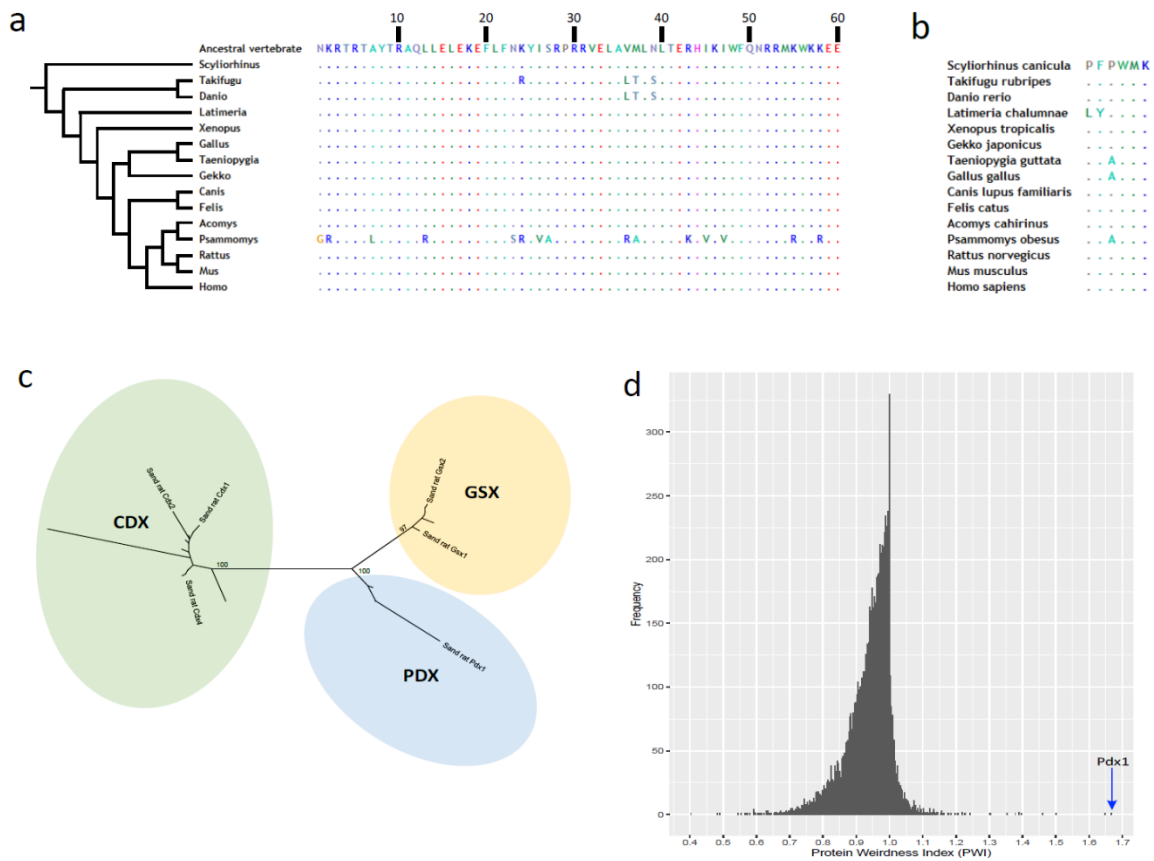
- 221 15. Ashizawa, S., Brunicardi, F. C. & Wang, X. P. PDX-1 and the pancreas. *Pancreas* **28**, 109-120 (2004).
222
- 223 16. Stoffers DA et al. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nat. Genet.* **17**,
224 138–139 (1997).
225
- 226 17. Jonsson, J., Carlsson, L., Edlund, T. & Edlund, H. Insulin-promoter-factor 1 is required for pancreas
227 development in mice. *Nature* **371**, 606-609 (1994).
228
- 229 18. Stoffers, D. A., Zinkin, N. T., Stanojevic, V., Clarke, W. L. & Habener, J. F. Pancreatic agenesis
230 attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nat. Genet.* **15**,
231 106-110 (1997)
232
- 233 19. Thomas, I. H. et al. Neonatal diabetes mellitus with pancreatic agenesis in an infant with
234 homozygous IPF-1 pro63fsX60 mutation. *Pediatr. Diabetes.* **10**, 492-496 (2009).
235
- 236 20. Schwitzgebel, V. M. et al. Agenesis of human pancreas due to decreased half-life of Insulin Promoter
237 Factor 1. *J. Clin. Endocrinol. Metab.* **88**, 4398-4406 (2003)
238
- 239 21. Schwitzgebel, V. M. Many faces of monogenic diabetes. *J. Diabetes. Investig.* **5**, 121-133 (2014).
- 240 22. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47 (2016).
241
- 242 23. Moens, C. B. & Selleri, L. Hox cofactors in vertebrate development. *Dev. Biol.* **291**, 193-206 (2006).
243
- 244 24. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes.
245 *Annu. Rev. Genomics. Hum. Genet.* **10**, 285-311 (2009).
246
- 247 25. Pessia, E. et al. Evidence for widespread GC-biased gene conversion in Eukaryotes. *Genome Biol.*
248 *Evol.* **4**, 675-682 (2012).
249
- 250 26. Hron, T., Pajer, P., Pačes, J., Bartůněk, P. & Elleder, D. Hidden genes in birds. *Genome. Biol.* **16**: 164
251 (2015).
252
- 253 27. Seroussi et al. Identification of the long-sought Leptin in Chicken and Duck: expression pattern of
254 the highly GC-rich avian *Leptin* fits an autocrine/paracrine rather than endocrine function.
255 *Endocrinology* **157**, 737-751 (2015).
256
257
258
259
260
261
262
263
264

265 **Table 1. Metrics of sand rat raw genomic sequencing data and final genome assembly.** Coverage was
266 calculated using an estimated genome size of 2.51 Gb based on a k-mer analysis (Supplementary
267 information section 1.3) and is based upon paired-end sequencing data only.

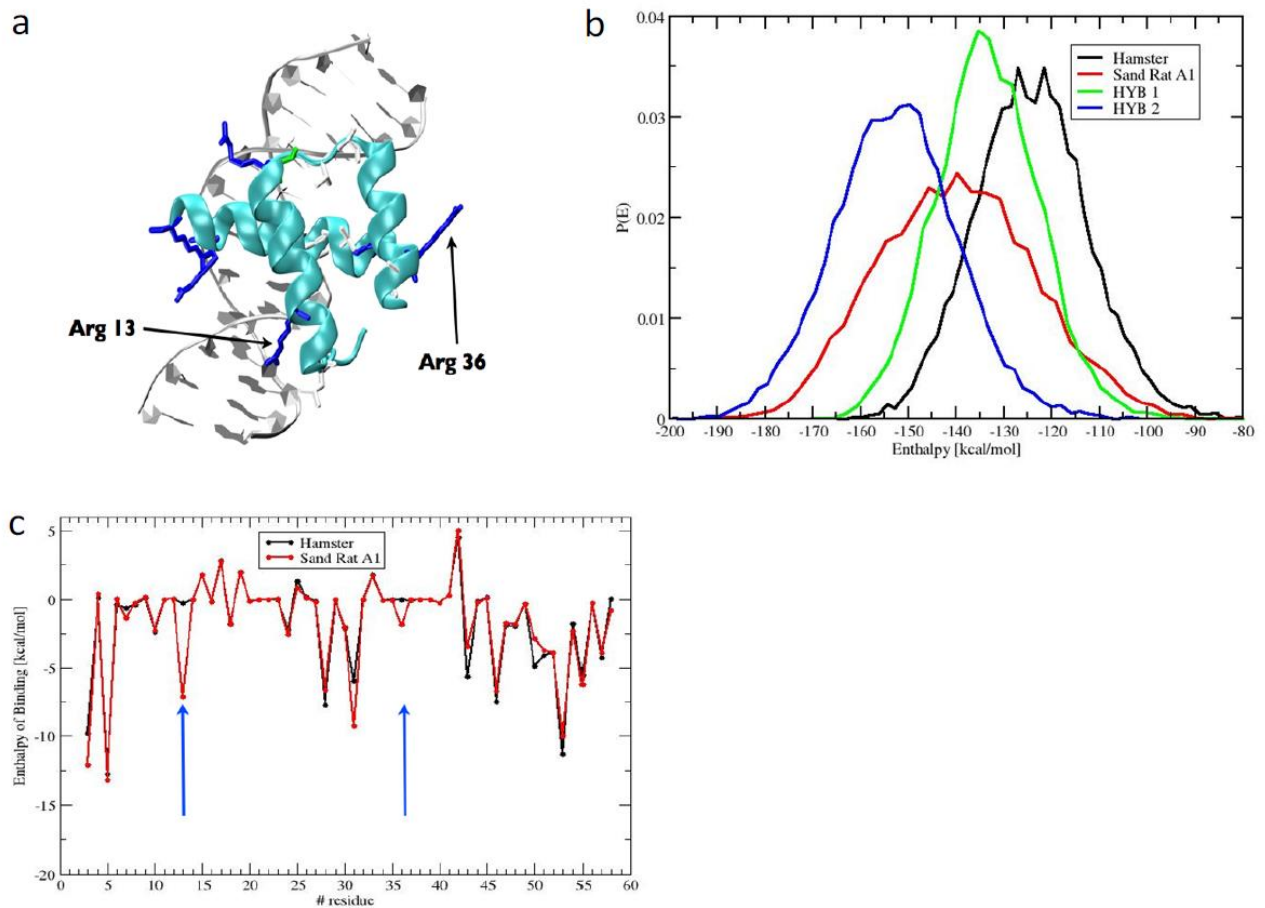
Total number of paired-end reads	724,377,486
Total number of mate-pair reads	1,780,436,140
Total bases sequenced	394,396,928,120
Estimated sequencing coverage (x)	87.6
Number of scaffolds >2 kb	1,737
Total length of assembly (bp)	2,381,209,849
Longest scaffold (bp)	54,616,910
Mean scaffold length (bp)	15,794
Scaffold N50 (bp)	10,461,538
Scaffold L50	63
Contig N50 (bp)	83,904
Percentage of assembly in scaffolds	98.6%



276 **Figure 1. The sand rat and its genomic hotspot of mutation.** (a) Juvenile sand rat *Psammomys obesus*.
 277 (b) Cladogram of representative murid rodents indicating the phylogenetic position of sand rat. (c) GC
 278 content of genes around the ParaHox cluster of sand rat and other rodents (*Mus musculus*, *Rattus*
 279 *norvegicus*, *Chinchilla lanigera*) revealing a chromosomal hotspot of GC skew in sand rat (shaded in
 280 grey). Genes shown in inferred ancestral gene order; parentheses around *Rfc3* indicate this gene has
 281 been transposed to a different genomic location in sand rat. Sand rat GC values based on transcriptome
 282 and genome sequences; when partial only alignable sequence is compared. (d) Unrooted phylogenetic
 283 trees inferred from synonymous changes (dS) only from concatenated alignments of 26 genes in the
 284 mutational hotspot (top) and 100 random genes (bottom).



285 **Figure 2. Molecular divergence of sand rat Pdx1.** (a) Alignment of Pdx1 homeodomain sequences
286 across vertebrates. (b) Alignment of Pdx1 hexapeptide domain across vertebrates. (c) Maximum
287 likelihood tree of ParaHox proteins showing divergent *Psammomys obesus* Pdx1; species included are
288 sand rat, mouse, zebra finch, spotted gar, amphioxus (full tree Figure S7.1). (d) Histogram of Protein
289 Weirdness Index (PWI) values for 1:1:1 mammalian orthologs of the sand rat predicted proteins: Pdx1
290 is marked by an arrow.



291 **Figure3. Molecular modelling of sand rat Pdx1 binding.** (a) Molecular model of sand rat Pdx1
292 homeodomain bound to DNA. The two amino acid changes indicated are the largest contributors to
293 altered enthalpy of binding. (b) Probability distributions of the enthalpy of binding of homeodomain
294 protein-DNA interactions between hamster (normal vertebrate) Pdx1/hamster insulin A1 DNA element
295 (black), sand rat Pdx1/sand rat A1 element (red), hamster Pdx1/sand rat A1 (green) and sand rat
296 Pdx1/hamster A1 (blue) inferred by molecular dynamics simulations and MM-PBSA; sand rat Pdx1
297 homeodomain has the lowest enthalpy of binding (higher affinity) for each DNA target. (c) Per-site
298 enthalpy of binding comparison between hamster and sand rat Pdx1 revealing contribution of amino
299 acid changes at homeodomain positions 13 and 36 to reduced enthalpy of binding (higher affinity).