

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

Adam D. Hargreaves^a, Long Zhou^b, Josef Christensen^c, Ferdinand Marlétaz^{a,d}, Shiping Liu^b, Fang Li^b, Peter Gildsig Jansen^c, Enrico Spiga^e, Matilde Thye Hansen^c, Signe Vendelbo Horn Pedersen^c, Shameek Biswas^f, Kyle Serikawa^f, Brian A. Fox^f, William R. Taylor^e, John Frederick Mulley^g, Guojie Zhang^{b,h,i,1}, R. Scott Heller^{c,1}, and Peter W. H. Holland^{a,1}

^aDepartment of Zoology, University of Oxford, Oxford, OX1 3PS, United Kingdom; ^bChina National Genebank, BGI-Shenzhen, 518083, Shenzhen, Guangdong, China; ^cNovo Nordisk, Måløv, DK-2760, Denmark; ^dMolecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan; ^eFrancis Crick Institute, London, NW1 1AT, United Kingdom; ^fNovo Nordisk Research Centre, Seattle, WA 98109; ^gSchool of Biological Sciences, Bangor University, Bangor, LL57 2DG, United Kingdom; ^hState Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 650223, Kunming, China; and ⁱCentre for Social Evolution, Department of Biology, University of Copenhagen, DK-2100 Copenhagen, Denmark

Edited by Richard M. Harland, University of California, Berkeley, CA, and approved June 2, 2017 (received for review February 23, 2017)

The sand rat *Psammomys obesus* is a gerbil species native to deserts of North Africa and the Middle East, and is constrained in its ecology because high carbohydrate diets induce obesity and type II diabetes that, in extreme cases, can lead to pancreatic failure and death. We report the sequencing of the sand rat genome and discovery of an unusual, extensive, and mutationally biased GC-rich genomic domain. This highly divergent genomic region encompasses several functionally essential genes, and spans the ParaHox cluster which includes the insulin-regulating homeobox gene *Pdx1*. The sequence of sand rat *Pdx1* has been grossly affected by GC-biased mutation, leading to the highest divergence observed for this gene across the Bilateria. In addition to genomic insights into restricted caloric intake in a desert species, the discovery of a localized chromosomal region subject to elevated mutation suggests that mutational heterogeneity within genomes could influence the course of evolution.

desert rodent | type 2 diabetes | homeobox | *Pdx1* | gene conversion

Arid environments impose extreme physiological demands on animals because of low food and water availability. The sand rat *Psammomys obesus* (Fig. 1A) is a member of the subfamily Gerbillinae, most species of which live in deserts and arid environments (Fig. 1B). *P. obesus* has emerged as a model for research into diet-induced type II diabetes because, if provided with high carbohydrate diets, the majority of individuals become obese and develop classic diabetes symptoms, in the most extreme cases leading to pancreatic failure and death (1–4).

In searching for the molecular basis of this unusual phenotype, attention has been paid to the *Pdx1* homeobox gene, also called *Ipf1*, *Idx1*, *Stf1*, or *Xlox* (5–9), the central and most highly conserved member of the ParaHox gene cluster (10). *Pdx1* is the only member of the Pdx gene family in tetrapods and encodes a homeodomain that has been invariant across their evolution. Mammalian *Pdx1* is expressed in pancreatic beta cells and encodes a homeodomain transcription factor that acts as a transcriptional activator of *insulin* and other pancreatic hormone genes (11, 12). A pivotal role in insulin regulation is also reflected in the association of heterozygous *Pdx1* mutations with maturity-onset diabetes of the young (*MODY4*) and type II diabetes mellitus in humans (13). Contrary to the usual conservation, several studies have reported inability to detect *Pdx1* in multiple gerbil species, including *P. obesus*, by immunocytochemistry, Western blotting, or PCR. However, *Pdx1* is readily detectable in the closely related spiny mouse, *Acomys cahirinus* (Fig. 1B), leading to the hypothesis that the gene has been lost within the Gerbillinae subfamily, contributing to the compromised ability to regulate insulin in the sand rat (14–16). Such a conclusion would raise further questions, because in addition to its adult functions, *Pdx1* is also essential for pancreatic development in the embryo. For example, targeted deletion in mice causes loss of pancreas and anterior duodenum

and is lethal (9, 17). In humans, pancreatic agenesis has been reported in a patient with a homozygous frameshift mutation before the *Pdx1* homeobox and in a compound heterozygous patient with substitution mutations in helices 1 and 2 of the homeodomain (18–20).

Results

To resolve the conundrum of a putatively absent “essential” gene, we sequenced the *P. obesus* genome by using a standard shotgun strategy (Illumina), using a combination of short and long insert libraries, initially at 85.5× coverage (*SI Appendix, SI Materials and Methods, section 1*). This assembly lacked a *Pdx1* gene, supporting the prevailing hypothesis of a loss of the *Pdx1* gene in gerbils. However, a synteny comparison between *P. obesus* and other mammals delineated a contiguous block of 88 genes (*SI Appendix, Fig. S2*) missing from the assembly including several genes essential

Significance

A core question in evolutionary biology is how mutation and selection adapt and constrain species to specialized habitats. We sequenced the genome of the sand rat, a desert rodent susceptible to nutritionally induced diabetes, and discovered an unusual chromosome region skewed toward G and C nucleotides. This region includes the *Pdx1* homeobox gene, a transcriptional activator of *insulin*, which has undergone massive sequence change, likely contributing to diabetes and adaptation to low caloric intake. Our results imply that mutation rate varies within a genome and that hotspots of high mutation rate may influence ecological adaptation and constraint. In addition, we caution that divergent regions can be omitted by conventional short-read sequencing approaches, a consideration for existing and future genome sequencing projects.

Author contributions: A.D.H., L.Z., W.R.T., G.Z., R.S.H., and P.W.H.H. designed research; A.D.H., L.Z., J.C., S.L., F.L., E.S., M.T.H., S.V.H.P., S.B., K.S., B.A.F., W.R.T., J.F.M., R.S.H., and P.W.H.H. performed research; F.M., W.R.T., and G.Z. contributed new reagents/analytic tools; A.D.H., L.Z., F.M., P.G.J., E.S., S.B., K.S., B.A.F., J.F.M., G.Z., R.S.H., and P.W.H.H. analyzed data; and A.D.H., L.Z., E.S., G.Z., R.S.H., and P.W.H.H. wrote the paper.

Conflict of interest statement: J.C., P.G.J., M.T.H., S.V.H.P., S.B., K.S., B.A.F., and R.S.H. are current or former employees of Novo Nordisk.

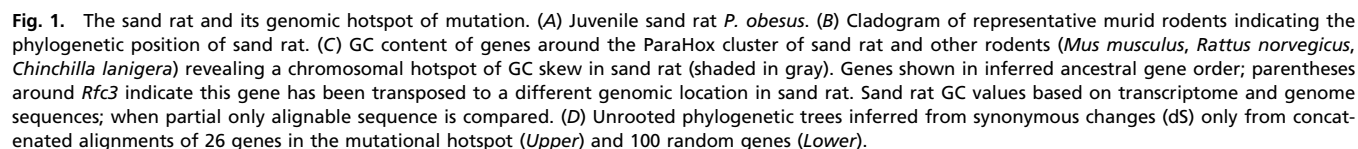
This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the National Center for Biotechnology Information Short Read Archive (accession nos. [SRA502705](#), [SRR5084169](#), [SRR5084170](#), [SRR5092818](#), [SRR5092819](#), [SRR5092820](#), and [SRR5429486](#)) and DDBJ/ENA/GenBank (accession nos. [NESX000000000](#) and [NESX010000000](#)).

¹To whom correspondence may be addressed. Email: peter.holland@zoo.ox.ac.uk, zhanggj@genomics.cn, or richardscottheller@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1702930114/-DCSupplemental.



(*SI Appendix, SI Materials and Methods*, section 1.5). This approach gave a refined assembly with a total size of 2.38 Gb and a scaffold N50 of 10.4 Mb (Table 1 and *SI Appendix, SI Materials and Methods*, sections 1, 3, 4, and 6), including much of the dark DNA region in several scaffolds, and containing genes syntenic to a region of chromosome 12 in rat and a region of chromosome 5 and the subtelomeric region of chromosome 8 in mouse. Analysis indicates that the region was initially omitted by standard genome assembly methods because of lower read coverage of GC regions coupled with short sequence read lengths. Comparison of GC content between species demonstrates that sand rat genes are elevated in GC content across this chromosomal region, syntenic to 12 Mb of the rat genome (Fig. 1C and *SI Appendix, SI Materials and Methods*, section 9). This large region encompasses a 250-kb repeat scaffold containing the sand rat ParaHox cluster and its well-characterized genomic neighbors. We inferred a high W (weak, A/T) to S (strong, G/C) allelic mutation rate in this region of the

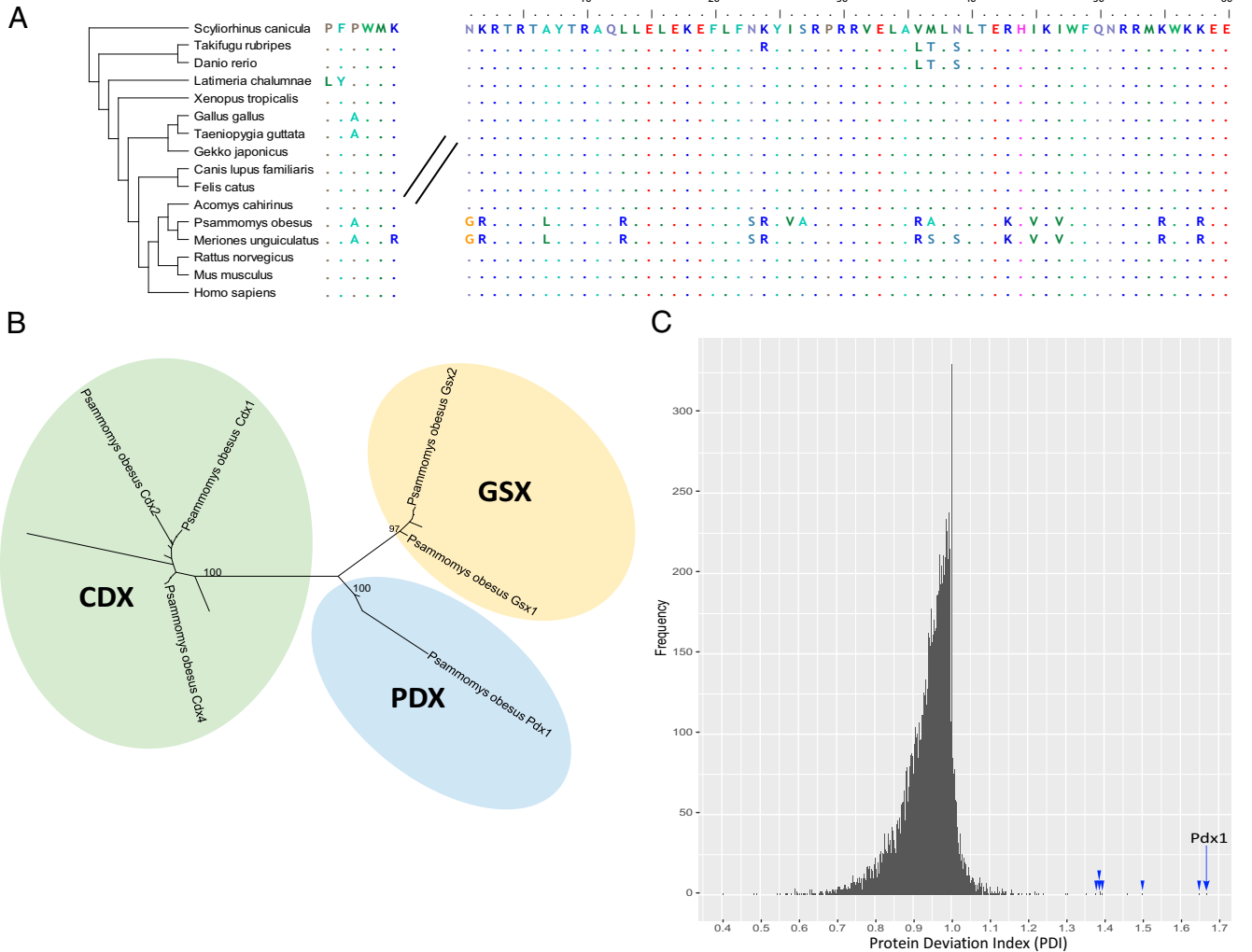


Fig. 2. Molecular divergence of sand rat Pdx1. (A) Alignment of Pdx1 hexapeptide domain and homeodomain sequences across vertebrates. (B) Maximum likelihood tree of ParaHox proteins showing divergent *P. obesus* Pdx1; species included are sand rat, mouse, zebra finch, spotted gar, amphioxus (full tree in [SI Appendix, Fig. S6](#)). (C) Histogram of PDI values for 1:1:1 mammalian orthologs of the sand rat predicted proteins: Pdx1 is marked by an arrow, other genes within the GC-rich region are marked by arrowheads.

likely to be responsible for altered DNA binding affinity and specificity.

Discussion

We show that an unusual genomic region of biased mutation arose in the evolutionary lineage of the sand rat. One consequence of this hotspot of mutation was the generation of GC-bias in the *Pdx1* gene of *P. obesus*; this process forced modification of the Pdx1 protein sequence, likely affecting its ability to regulate transcription of *insulin* and other pancreatic genes. The sand rat Pdx1 hexapeptide, which mediates cofactor interactions (25), is intact, which may explain why pancreatic development proceeds permitting viable sand rat embryogenesis. We suggest mutation-driven changes have played a role in constraining or adapting the sand rat, and possibly other gerbil species, to arid environments and low caloric intake. Biased gene conversion is a known mechanism that causes GC-biased mutation (21, 26); hence, we suggest this mechanism, driven by elevated localized recombination, is generating a hotspot of skewed base composition. The genomic region we describe here was not detected by standard short-read sequencing approaches, known to be sensitive to nucleotide composition (27). These issues may be

circumvented through the use of third generation sequencing technologies offering substantially longer read lengths and reduced nucleotide bias. The possibility remains that other such dark DNA regions could be widespread features of animal genomes, thus far largely overlooked in comparative animal genomics. Indeed, GC-rich genes are also missing from the chicken genome assembly (28, 29). Hotspots of mutation could drive rapid evolutionary change at the molecular level, and it will be important to decipher to what extent such hotspots have constrained and influenced evolutionary adaptation across the animal kingdom.

Materials and Methods

Sand Rat Genome Sequencing. All animal procedures were carried out in accordance with the regulations specified under Protection of Animals Act by the authority in Denmark, European Union, and Novo Nordisk A/S, or the Animals (Scientific Procedures) Act 1986, U.K., and Bangor University Animal Welfare and Ethical Review Board. Sand rat genome sequencing libraries were constructed from a male *P. obesus* obtained from Hadassah Medical School, Israel. We prepared and sequenced multiple short- and long-insert DNA libraries and sequenced them on an Illumina HiSeq 2000. We also isolated sand rat DNA enriched for GC content through cesium chloride gradient centrifugation, prepared GC-rich DNA libraries, and

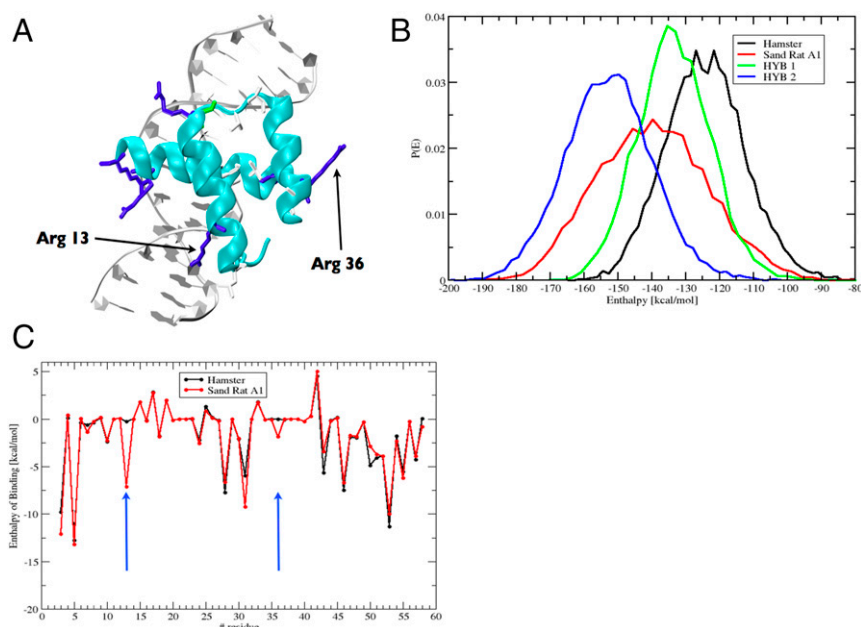


Fig. 3. Molecular modeling of sand rat Pdx1 binding. (A) Molecular model of sand rat Pdx1 homeodomain bound to DNA. The two amino acid changes indicated are the largest contributors to altered enthalpy of binding. (B) Probability distributions of the enthalpy of binding of homeodomain protein–DNA interactions among hamster (normal vertebrate) Pdx1/mouse insulin A1 DNA element (black), sand rat Pdx1/sand rat A1 element (red), hamster Pdx1/sand rat A1 (green), and sand rat Pdx1/mouse A1 (blue) inferred by molecular dynamics simulations and MM-PBSA; sand rat Pdx1 homeodomain has the lowest enthalpy of binding (higher affinity) for each DNA target. (C) Per-site enthalpy of binding comparison between hamster and sand rat Pdx1 revealing contribution of amino acid changes at homeodomain positions 13 and 36 to reduced enthalpy of binding (higher affinity).

sequenced using an Illumina MiSeq. In total, we generated ~398 Gbp of sequencing data, which was assembled by using SOAPdenovo2 (30). Further details are provided in [SI Appendix](#).

Transcriptome Sequencing and Analysis. Total RNA was extracted and purified by using either Qiagen RNeasy column-based methods (pancreatic islets and liver) or TRIreagent (duodenum). All RNA-seq libraries were prepared by using Illumina chemistry. Pancreatic islet libraries were sequenced individually and as pools on the Illumina GAI. RNA-seq libraries for liver and duodenum were sequenced on the HiSeq 2000 (liver) or the HiSeq 4000 (duodenum). The pancreatic islets transcriptome was assembled by using TransABySS (31) using multiple k-mer sizes (41 up to 79, in increments of 2), and the liver and duodenum transcriptomes were assembled by using Trinity (32) ([S/ Appendix](#)).

Gene Prediction and Annotation. We used multiple methods to predict genes in the sand rat genome. Repetitive elements were first masked by using RepeatMasker followed by ab initio gene prediction with AUGUSTUS (33). Homologous proteins from mouse and human were subsequently mapped to the sand rat genome assembly by using TBLASTN, with the aligned sequence being filtered and passed to GeneWise (34) to identify accurate spliced alignments. GLEAN (35) was then used to generate a consensus gene set. These gene models were then further refined by predicting ORFs using genome-guided transcriptome assemblies assembled using TopHat (36) and Cufflinks (37).

Evolutionary Analyses. Using the gene predictions from our sand rat genome assembly and the assembled tissue transcriptome data, we carried out analyses of coding sequence GC content and GC-biased mutation within coding and intronic regions compared with other rodents. We also conducted an analysis to determine the extent of protein divergence within the sand rat predicted proteome compared with mouse and human. Details of these analyses are described in [SI Appendix](#).

Molecular Modeling. We used molecular dynamics simulations to calculate the enthalpy of binding of protein–DNA complexes, namely between the sand rat or mouse *Pdx1* homeodomain and the sand rat or mouse A1 region of the *insulin* promoter, using MM-PBSA analyses (*SI Appendix*).

ACKNOWLEDGMENTS. We thank Natasha Ng, Gemma Marfany, Thomas Dunwell, Fei Xu, Shan Quah, Anna Gloy, Christine Hirschberger, Juliane Cohen, Rhys Morgan, Lorna Witty, Monica Martinez Alonso, and Thomas Brekke for assistance and advice, and the Oxford Genomics Centre for GC-rich sequencing. This work was funded principally by the European Research Council under European Union's Seventh Framework Programme (FP7/2007-2013 ERC Grant 268513 to P.W.H.H.), a Strategic Priority Research Program of the Chinese Academy of Sciences Grant XDB13000000 (to G.Z.), and Novo Nordisk A/S (coordinated by R.S.H.). E.S. and W.R.T. were supported by the Francis Crick Institute under award FC001179. The Crick receives its core funding from Cancer Research UK, the UK Medical Research Council, and the Wellcome Trust.

- Schmidt-Nielsen K, Haines HB, Hackel DB (1964) Diabetes mellitus in the sand rat induced by standard laboratory diets. *Science* 143:689–690.
- Bar-On H, Ben-Sasson R, Ziv E, Arar N, Shafir E (1999) Irreversibility of nutritionally induced NIDDM in *Psammomys obesus* is related to β -cell apoptosis. *Pancreas* 18:259–265.
- Kaiser N, et al. (2005) *Psammomys obesus*, a model for environment-gene interactions in type 2 diabetes. *Diabetes* 54:S137–S144.
- Kalman R, Ziv E, Galila L, Shafir E (2012) Sand rat. *The Laboratory Rabbit, Guinea Pig, Hamster, and Other Rodents* (Elsevier, London), pp 1171–1190.
- Ohlsson H, Karlsson K, Edlund T (1993) IPF1, a homeodomain-containing transactivator of the insulin gene. *EMBO J* 12:4251–4259.
- Leonard J, et al. (1993) Characterization of somatostatin transactivating factor-1, a novel homeobox factor that stimulates somatostatin expression in pancreatic islet cells. *Mol Endocrinol* 7:1275–1283.
- Bürglin TR (1994) A comprehensive classification of homeobox genes. *A Guidebook to Homeobox Genes*, ed Duboule D (Oxford Univ Press, Oxford), pp 25–71.
- Miller CP, McGehee RE, Jr, Habener JF (1994) IDX-1: A new homeodomain transcription factor expressed in rat pancreatic islets and duodenum that transactivates the somatostatin gene. *EMBO J* 13:1145–1156.
- Offield MF, et al. (1996) PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* 122:983–995.
- Brooke NM, García-Fernández J, Holland PWH (1998) The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 392:920–922.
- Ashizawa S, Brunicaudi FC, Wang XP (2004) PDX-1 and the pancreas. *Pancreas* 28:109–120.
- Servitja JM, Ferrer J (2004) Transcriptional networks controlling pancreatic development and beta cell function. *Diabetologia* 47:597–613.
- Stoffers DA, Ferrer J, Clarke WL, Habener JF (1997) Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. *Nat Genet* 17:138–139.
- Leibowitz G, et al. (2001) IPF1/PDX1 deficiency and β -cell dysfunction in *Psammomys* obesus, an animal with type 2 diabetes. *Diabetes* 50:1799–1806.

15. Vedtofte L, Bødvarsdóttir TB, Karlén AE, Heller RS (2007) Developmental biology of the *Psammomys obesus* pancreas: Cloning and expression of the *Neurogenin-3* gene. *J Histochem Cytochem* 55:97–104.
16. Gustavsen CR, et al. (2008) 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.
17. Jonsson J, Carlsson L, Edlund T, Edlund H (1994) Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* 371:606–609.
18. Stoffers DA, Zinkin NT, Stanojevic V, Clarke WL, Habener JF (1997) Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nat Genet* 15:106–110.
19. Schwitzgebel VM, et al. (2003) Agenesis of human pancreas due to decreased half-life of insulin promoter factor 1. *J Clin Endocrinol Metab* 88:4398–4406.
20. Thomas IH, et al. (2009) Neonatal diabetes mellitus with pancreatic agenesis in an infant with homozygous IPF-1 Pro63fsX60 mutation. *Pediatr Diabetes* 10:492–496.
21. Pessia E, et al. (2012) Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* 4:675–682.
22. Ratnakumar A, et al. (2010) Detecting positive selection within genomes: The problem of biased gene conversion. *Philos Trans R Soc Lond B Biol Sci* 365:2571–2580.
23. Schwitzgebel VM (2014) Many faces of monogenic diabetes. *J Diabetes Investig* 5:121–133.
24. Fuchsberger C, et al. (2016) The genetic architecture of type 2 diabetes. *Nature* 536:41–47.
25. Moens CB, Selleri L (2006) Hox cofactors in vertebrate development. *Dev Biol* 291: 193–206.
26. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* 10:285–311.
27. Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC (2013) Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS One* 8:e62856.
28. Hron T, Pajer P, Pačes J, Bartůněk P, Elleder D (2015) Hidden genes in birds. *Genome Biol* 16:164.
29. Seroussi E, et al. (2016) Identification of the long-sought Leptin in Chicken and Duck: Expression pattern of the highly GC-rich avian *Leptin* fits an autocrine/paracrine rather than endocrine function. *Endocrinology* 157:737–751.
30. Luo R, et al. (2012) SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18.
31. Robertson G, et al. (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods* 7:909–912.
32. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652.
33. Keller O, Kollmar M, Stanke M, Waack S (2011) A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 27: 757–763.
34. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* 14: 988–995.
35. Elisk CG, et al. (2007) Creating a honey bee consensus gene set. *Genome Biol* 8: R13.
36. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111.
37. Trapnell C, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28:511–515.