Compromised Function of the Pancreatic Transcription Factor PDX1 in a Lineage of Desert Rodents

Yichen Dai1 · Sonia Trigueros1 · Peter W. H. Holland1

Accepted: 22 March 2021 / Published online: 16 April 2021
© The Author(s) 2021

Abstract

Gerbils are a subfamily of rodents living in arid regions of Asia and Africa. Recent studies have shown that several gerbil species have unusual amino acid changes in the PDX1 protein, a homeodomain transcription factor essential for pancreatic development and β-cell function. These changes were linked to strong GC-bias in the genome that may be caused by GC-biased gene conversion, and it has been hypothesized that this caused accumulation of deleterious changes. Here we use two approaches to examine if the unusual changes are adaptive or deleterious. First, we compare PDX1 protein sequences between 38 rodents to test for association with habitat. We show the PDX1 homeodomain is almost totally conserved in rodents, apart from gerbils, regardless of habitat. Second, we use ectopic gene overexpression and gene editing in cell culture to compare functional properties of PDX1 proteins. We show that the divergent gerbil PDX1 protein inefficiently binds an insulin gene promoter and ineffectively regulates insulin expression in response to high glucose in rat cells. The protein has, however, retained the ability to regulate some other β-cell genes. We suggest that during the evolution of gerbils, the selection-blind process of biased gene conversion pushed fixation of mutations adversely affecting function of a normally conserved homeodomain protein. We argue these changes were not entirely adaptive and may be associated with metabolic disorders in gerbil species on high carbohydrate diets. This unusual pattern of molecular evolution could have had a constraining effect on habitat and diet choice in the gerbil lineage.

Keywords CRISPR/Cas9 · Gene evolution · Insulin · Protein evolution · Sand rat

Introduction

Living in deserts and other arid environments poses particular challenges related to heat, lack of water, and irregular food supply. Desert animals have many physiological adaptations allowing them to cope with these extreme conditions. Amongst desert mammals, rodents are of special interest given their widespread occurrence in desert habitats across six continents and also their small size and large surface-to-mass ratio that may hinder effective temperature regulation and water storage in a harsh environment (Walsberg 2000). Comparative behavioral, morphological, and physiological analysis of desert rodents has revealed many examples of convergent adaptation, for example, nocturnal activity to avoid high heat in daytime, kidney morphology allowing extremely effective water reabsorption, and metabolic changes in response to water deprivation (Walsberg 2000; Takei et al. 2012). In recent years, genome sequencing efforts have revealed the genomic basis of some of these adaptations. Limited plant sources in the desert, many equipped with toxins, have been linked to selective sweeps in bitter taste receptor genes while positive selection affecting genes regulating adipogenesis and lipid metabolism may be associated with metabolic response to dehydration (Tigano et al. 2020).

Here we focus on members of the Gerbillinae subfamily, a lineage consisting predominantly of species inhabiting arid or semi-arid areas of Asia and Africa (Chevret and Dobigny 2005). One example is the sand rat (Psammomys obesus), which feeds primarily on Salicornia plants that have high levels of salt and relatively low caloric content (Schmidt-Nielsen et al. 1964). Morphological changes observed in sand rats include a high medulla to cortex ratio in the kidney, proposed to be an adaptation allowing the animal to produce highly concentrated urine, along with behavior changes such as using...
incisors to strip away salt-laden parts of the *Salicornia* plant before consuming the inner mesophyll tissue (Ojeda et al. 1999). The sand rat also has a strange propensity to develop metabolic abnormalities when given a standard rodent diet in a laboratory environment; some sand rats exhibit elevated blood and urine glucose levels accompanied by rapid weight gain under these conditions (Kalderon et al. 1986), although these metabolic syndromes have not been reported in their natural environment where calorie intake is likely to be lower (Schmidt-Nielsen et al. 1964).

In previous work, we and others have shown that members of the Gerbillinae, including the sand rat, have some highly unusual genomic features (Hargreaves et al. 2017). However, it is unclear whether (or how) these genomic characters relate to life in arid environments. The most striking oddity is that the genomes of sand rat (*P. obesus*) and Mongolian jird (*Meriones unguiculatus*) include several ‘islands’ of extremely high GC-content; these likely arose through localized elevated rates of meiotic recombination causing ‘run away’ GC-biased gene conversion (gBGC) (Hargreaves et al. 2017; Pracana et al. 2020). As a consequence, many protein-coding genes in gerbils have experienced an increase in GC-biased substitutions and in some cases, these nucleotide changes have caused unusual alterations to the encoded protein sequence (Hargreaves et al. 2017; Dai et al. 2020). Some, but not all, of the affected proteins are implicated in dietary physiology and could plausibly affect how gerbils interact with their environment (Dai et al. 2020). The best studied example is the *Pdx1* gene, which encodes a highly conserved homeodomain transcription factor expressed in the vertebrate endocrine pancreas and which regulates, among other targets, *insulin* gene expression in response to food intake (Dai and Holland 2019). Previous work has shown that GC-driven amino acid changes in the sand rat PDX1 homeodomain altered protein stability and could have been selectively deleterious; however, compensatory substitutions have partially restored regulated protein decay (Dai and Holland 2019). In addition, similar amino acid changes are observed in the PDX1 homeodomain of two other gerbil species, the Mongolian jird and Libyan jird (*Meriones libycus*), and the related *Acomys cahirinus* were analyzed in previous work (Hargreaves et al. 2017; Dai et al. 2020). The *Pdx1* gene sequence from the fat-tailed gerbil (*Pachyuromys duprasi*) was obtained from a draft genome assembly generated by the DNA Zoo Consortium from short insert-size PCR-free DNA-Seq data using w2rap-contigger (Clavijo et al. 2017; Dudchenko et al. 2017, 2018). Proteins were aligned using ClustalW in BioEdit v7.2.5 (Hall 1999).

### Overexpression and Gene Editing Plasmids

For ectopic overexpression, mouse *Pdx1* coding sequence was cloned into the pSF-CMV-Ub-daGFP-Ascl vector (OXGENE #OG244) using digestion sites EcoRV and XhoI. Sand rat *Pdx1* coding sequence was ‘mousified’ to lower average GC% to a level comparable to mouse *Pdx1* and cloned into pSF-CMV-Ub-daGFP-Ascl between XbaI and XhoI (Fig. S1, Online Resource 1). Clones were verified by Sanger sequencing. For CRISPR/Cas9 gene editing, guide RNA sequences targeted the start and end of rat *Pdx1* exon 2 (Table S2a, Online Resource 1) and were inserted separately into plasmid pSpCas9(BB)-2A-GFP (pX458) (Addgene #48138) (Ran et al. 2013) at the BbsI digestion site. Donor sequences were designed and cloned into pUC19 (NEB #N3041). To construct donor sequences, genomic DNA was extracted from rat INS-1 cells and PCR

### Materials and Methods

#### Rodent Genome and Transcriptome Assembly

Rodent genome and transcriptome sequencing data were sourced from the NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) with usage permission granted by data depositors. Reads were downloaded from SRA, quality was assessed using fastqc v0.11.5 (Andrews 2010), and paired reads separated using sickle v1.33 (Joshi and Fass 2011). Trinity v2.2.0 was used for transcriptome assembly with default settings (Grabherr et al. 2011). For genome assembly, ABYSS v2.0.1 was used (Simpson et al. 2009) with the optimal k-mer value estimated using kmer-eme v1.7016 (Chikhi and Medvedev 2013). Quality of the assembled genomes and transcriptome was assessed using assembly-stats v1.0.0 (Sanger Institute UK 2014). tBLASTn was used to search for *Pdx1* in each assembly, using mouse PDX1 protein as query. For rodents with an annotated *Pdx1* gene sequence on NCBI GenBank, the full coding sequence was downloaded (Table S1, Online Resource 1). *Pdx1* genes from three gerbil species, *P. obesus*, *M. unguiculatus* and *M. libycus*, and the related *Acomys cahirinus* were analyzed in previous work (Hargreaves et al. 2017; Dai et al. 2020). The *Pdx1* gene sequence from the fat-tailed gerbil (*Pachyuromys duprasi*) was obtained from a draft genome assembly generated by the DNA Zoo Consortium from short insert-size PCR-free DNA-Seq data using w2rap-contigger (Clavijo et al. 2017; Dudchenko et al. 2017, 2018). Proteins were aligned using ClustalW in BioEdit v7.2.5 (Hall 1999).
was used to amplify ~800 bp of intron 5’ to rat Pdx1 exon 2 and ~800 bp of 3’ UTR region downstream of rat Pdx1 exon 2. PCR was used to remove the stop codon and add a linker sequence (Ser-Gly-Ser-Gly-Ser-Gly), a V5 peptide tag sequence and P2A sequence to the 3’ end of both rat Pdx1 exon 2 coding sequence and ‘mousified’ sand rat Pdx1 exon 2 coding sequence. Protospacer adjacent motif (PAM) sites, required for Cas9 induced DNA cleavage, were altered by in vitro mutagenesis to avoid repeated cutting by Cas9; at the upstream cut site the PAM site sequence was altered from CACAGG to CCCAGG and at the downstream cut site from upstream cut site the PAM site sequence was altered from CTGAGG to CCGAGG.

Glucose Treatment and QPCR

Cell culture medium was replaced with high or low glucose medium 48 h after electroporation. High glucose medium (25 mM) was prepared by adding stock glucose solution (ThermoFisher #A2494001) to culture medium containing 11.1 mM glucose; low glucose medium (0.5 mM) was prepared by adding stock glucose solution to glucose-free RPMI-1640 medium (ThermoFisher #11879020). Both media contained the same concentration of supplements as above. Total RNA extraction used the RNeasy Plus Mini kit (Qiagen #74134); 500,000 cells from each sample were lysed in Buffer RLT Plus containing 10 ul/ml β-mercaptoethanol. RNA quality was assessed using a NanoDrop 1000 Spectrophotometer (ThermoFisher) and all samples were treated with DNase I (Roche #04716728001). Quantitative PCR (qPCR) used the Luna Universal One-Step RT-qPCR Kit (NEB #E3005); for each reaction, total volume was 20 ul with a final concentration of 1X Luna Universal One-Step Reaction Mix, 1X Luna WarmStart RT Enzyme Mix, 0.4 uM of each primer (Table S2b, Online Resource 1) and 20 ng of DNA-free total RNA.

Antibiotic Selection for CRISPR/Cas9 Gene Editing

For CRISPR/Cas9 gene editing, rat INS-1 832/13 cells were separated into six groups: each group was transfected with a different combination of donor plasmids: (1) rat Pdx1 puromycin resistance plasmid, (2) rat Pdx1 blasticidin resistance, (3) 1:1 ratio mix of rat Pdx1 puromycin resistance and blasticidin resistance plasmids, (4) sand rat Pdx1 puromycin resistance plasmid, (5) sand rat Pdx1 blasticidin resistance, and (6) 1:1 ratio mix of sand rat Pdx1 puromycin resistance and blasticidin resistance plasmids. The use of two resistance genes has been proposed as a means to increase probability of achieving bi-allelic gene targeting (Supharattanasitthi et al. 2019). Prior to CRISPR/Cas9 gene editing, wild type cells were assessed to determine optimal antibiotic concentrations that were neither too toxic nor ineffective, enabling efficient antibiotic selection of gene edited cells. Transfection efficiency was assessed 48 h after electroporation by observing EGFP fluorescence. Live cells from each group were detached, separated into three portions, and seeded onto P15 plates. After 48 h in standard culture medium, cells were cultured in medium containing either puromycin (0.13, 0.15, or 0.2 ug/ml), blasticidin (0.5, 0.6 or 0.7 ug/ml), or both (0.07 ug/ml puromycin with 0.3 ug/ml blasticidin, 0.08 ug/ml puromycin with 0.35 ug/ml blasticidin, or 0.09 ug/ml puromycin with 0.4 ug/ml blasticidin) according to which donor constructs were transfected. After 18 days of antibiotic treatment, single-cell colonies were picked from the P15 plates using a sterile pipette tip and transferred to 24-well plates. Cells were maintained in
antibiotic-containing medium and allowed to grow until ~1 million in number.

**Verification of CRISPR/Cas9 Editing**

GeneJET Genomic DNA Purification Kit (ThermoFisher #K0722) was used to extract genomic DNA from antibiotic resistant INS-1 832/13 rat cell cultures following gene editing. Nested PCR was used to test whether the donor sequence was integrated at the desired site (Table S2c, Online Resource 1). To assess transcript integrity, total RNA was extracted as above and cDNA prepared using the GoScript Reverse Transcription System (Promega #A5000).

One ul cDNA obtained from 80 ng total RNA was used as the template for amplification of the full length Pdx1 gene transcript (Table S2d, Online Resource 1). PCR products were assessed on a 1% agarose gel with 1X TBE buffer, and DNA size was compared to GeneRuler 1 kb DNA Ladder (ThermoScientific #SM0311). To monitor PDX1-V5 protein expression, cells were lysed in RIPA buffer (150 mM NaCl, 1% Triton-X100, 50 mM Tris–HCl pH 7.4, 0.5% sodium deoxycholate, one cOmplete Mini Protease Inhibitor Cocktail tablet (Roche #11836153001) per 10 ml buffer). Forty ug protein was boiled with NuPAGE LDS Sample Buffer (ThermoFisher #NP0007) and 0.2% β-mercaptoethanol, then run on a SDS-PAGE gel (4% stacking, 12% resolving) at 180 V for 45 min with Precision Plus Protein Dual Color Standards (Bio-Rad #1610374) as a reference. Protein was transferred onto a PVDF membrane using semi-wet transfer at 20 V for 40 min. The membrane was blocked with 5% milk powder in PBST, followed by primary antibody binding (1:3000) using monoclonal mouse IgG anti-V5 antibody (Invitrogen #R96025) overnight at 4°C; secondary antibody was goat anti-mouse HRP (Bio-Rad #170–5047; 1:20,000).

To test for PDX1-V5 subcellular location, cells were fixed (1:3000) using monoclonal mouse IgG anti-V5 antibody (Invitrogen #R96025) overnight at 4°C; secondary antibody was goat anti-mouse HRP (Bio-Rad #170–5047; 1:20,000).

**Results**

**Gerbil PDX1 Sequence is Unusual for Arid-living Rodents**

The unusual amino acid sequence of PDX1 previously reported for three gerbil species (P. obesus, M. unguiculatus, and M. libycus) raises the possibility of association with survival in an arid environment (Hargreaves et al. 2017). To test this, we aligned the PDX1 protein sequence of 38 rodent species, including 14 species living in arid or semi-arid habitats representing at least seven evolutionary transitions to these environments (Fig. 1). For 28 species, the Pdx1 gene sequence has been published, made available or annotated (Ohlsson et al. 1993; Rudnick et al. 1994); for ten species we identified the Pdx1 DNA sequence by assembling raw genome or transcriptome data from NCBI SRA (Table 1).
We focus on two key functional regions: the hexapeptide region, which mediates cofactor binding, and the homeodomain, which mediates DNA sequence recognition and binding. We find that 33 rodent species have 100% identity across the conserved hexapeptide and homeodomain sequence, including nine of the 14 arid-living species. Aside from the four gerbil species, the kangaroo rat (*Dipodomys ordii*) is the only rodent that has a PDX1 homeodomain sequence with any change from the ancestral amino acid sequence for mammals. However, kangaroo rat PDX1 protein has only a single altered residue across the homeodomain, in contrast to the 16 altered residues (12 shared) in gerbils (Fig. 1). We conclude that the highly divergent PDX1 protein sequence is specific to the gerbil lineage rather than being a feature common to arid-dwelling rodents.

**Sand Rat PDX1 Inefficiently Regulates Rat *Insulin* Gene Expression**

As a transcription factor, the key function of PDX1 is to correctly regulate the expression of target genes. To test whether the highly divergent sand rat PDX1 protein can regulate the expression of known PDX1 target genes despite radical sequence divergence, we overexpressed the sand rat *Pdx1* gene in a cultured rat pancreatic β-cell line (INS-1 832/13). As controls, we overexpressed mouse *Pdx1* or an empty plasmid. Different culture conditions were used to test if sand rat PDX1 is responsive to elevated glucose because it has been shown previously that mouse or rat PDX1 transactivates an *insulin* promoter effectively only under high glucose conditions (MacFarlane et al. 1994; Marshak et al. 1996; Rosanas-Urgell et al. 2008).

Under high glucose conditions, overexpression of mouse *Pdx1* resulted in a two-fold increase in expression of the two rat *insulin* genes, *Ins1* and *Ins2*, compared to transfection of empty plasmid control (Fig. 2a). In contrast, overexpression of sand rat *Pdx1* did not increase *Ins1* expression, while *Ins2* expression was slightly but significantly decreased (Fig. 2a). We also measured the expression of three other genes known to be directly regulated by PDX1: *MafA*, *Slc2a2* (encoding a glucose transporter GLUT2), and the *Pdx1* gene itself (Fig. 2a). Up-regulation of *Slc2a2* was seen with mouse *Pdx1* overexpression; sand rat had a repressive effect. Down-regulation of *Pdx1* gene expression was only significant with mouse PDX1 protein; *MafA* expression was down-regulated similarly by PDX1 from the two species.

Under low glucose conditions, mouse and sand rat PDX1 proteins again elicited distinct effects (Fig. 2b). For example, *Slc2a2* gene expression was activated only by mouse PDX1 protein. An intriguing difference was seen in the response environments highlighted in yellow; habitat information from IUCN Red List of Threatened Species (IUCN 2020). Photograph from J.F. Mulley with permission. Figure created using Inkscape (Inkscape Project 2020)
of the insulin genes, \textit{Ins1} and \textit{Ins2}. Ectopic expression of mouse PDX1 protein resulted in inhibition of \textit{insulin} gene expression under low glucose conditions, in contrast to the activation under high glucose consistent with previous publications (Rosanas-Urgell et al. 2008). Sand rat PDX1 protein, however, gave inhibition or no effect under both glucose concentrations, suggesting that sand rat PDX1 cannot function effectively in the rat glucose-mediated insulin expression pathway.

**Construction of a Rat β-cell Line Producing a Rat-Sand Rat Fusion PDX1 Protein**

Transfection and overexpression of exogenous genes in mammalian cell systems come with some caveats. First, ectopic expression may transcribe the introduced \textit{Pdx1} gene at levels much higher than normal. Second, presence of the endogenous rat PDX1 protein may compete with ectopically expressed PDX1 protein for DNA binding. This is particularly an issue when assaying protein binding affinities. To overcome these issues, we elected to use gene editing methods to alter the endogenous rat \textit{Pdx1} locus using sand rat \textit{Pdx1} gene sequence in a cultured rat pancreatic β-cell line. Specifically, we replaced exon 2 of the endogenous rat \textit{Pdx1} gene, as this includes the complete homeobox sequence encoding the critical homeodomain. Replacement of one exon rather than two was chosen as a balance between technical complexity and effecting an informative change to the encoded protein. We also compensated for the extreme GC-richness of the sand rat sequence (Hargreaves et al. 2017) by altering codon usage, so that the donor exon 2 encodes sand rat protein sequence using codons more compatible with a rat cell line.

The gene editing experimental design is given in Fig. 3. Excision of endogenous rat genomic sequence used guide RNA matching the start and end of rat exon 2, introduced into cells using a plasmid encoding the Cas9 protein and guide RNA to effect double strand breaks (DSBs). Two sand rat donor sequences were also introduced as templates to fill the excised region using the homology directed repair (HDR) pathway which repairs DSBs using a homologous sequence. The donors also contained homologous arms with 100% identity to upstream and downstream genomic regions flanking sand rat \textit{Pdx1} exon 2, a sequence encoding a V5 peptide tag to facilitate downstream analysis, and a drug resistance gene to enable selection (Fig. 3a).

### Table 1

| Lineage | Species | SRA run # | BioProject # | Data type | Total contig # | N50 (bp) | Contig # |
|---------|---------|-----------|--------------|-----------|----------------|----------|----------|
| ‘Mouse-related’ clade | *Cricetomys gambianus* | SRR2069938 | PRJNA285809 (Gingerich et al. 2016) | Genome | 4,704,353 | 12,053 | 54,430 |
| | *Rhizomys pruinosus* | SRR933768 | PRJNA211727 (Lin et al. 2014) | Transcriptome | 215,283 | 1961 | 25,136 |
| | *Elllobius lutescens* | SRR3475727 | PRJNA305123 (Mulugeta et al. 2016) | Genome | 8,082,007 | 2392 | 279,624 |
| | *Onychomys torridus* | ERR968259 | PRJEB8691 (The Wellcome Trust Sanger Institute 2015) | Genome | 7,536,142 | 4512 | 168,295 |
| | *Peromyscus polionotus* | SRR545671, SRR545672, SRR545673 | PRJNA53593 (Bendesky et al. 2017) | Genome | 6,553,890 | 2636 | 264,284 |
| | *Sigmodon hispidus* | SRR2954727 | PRJNA301539 (Beijing Genomics Institute 2015) | Genome | 5,315,264 | 3905 | 185,177 |
| | *Apodemus sylvaticus* | SRR2141382 | PRJNA290427 (University of Liverpool 2016) | Genome | 27,068,228 | 293 | 2,090,123 |
| | *Apodemus uralensis* | ERR1101669 | PRJEB11533 (Neme and Tautz 2016) | Genome | 8,983,726 | 2876 | 261,213 |
| | *Mus mattheyi* | ERR1101670 | PRJEB11533 (Neme and Tautz 2016) | Genome | 5,112,376 | 3367 | 203,783 |
| ‘Squirrel-related’ clade | *Spermophilus dauricus* | SRR955327, SRR955328, SRR955330, SRR955331, SRR955334, SRR955335 | PRJNA215874 (Beijing Genomics Institute 2017) | Genome | 59,210 | 1093 | 9613 |
To increase probability of successful gene editing at both \textit{Pdx1} alleles, two donor templates carrying different selection markers were used simultaneously. To avoid generating large proteins with PDX1 connected to the antibiotic resistance protein, we included a viral ribosome-skipping sequence (referred to as P2A) between the V5 tag and the antibiotic resistance gene (Fig. 3a). If successful gene editing occurs the edited rat cells should produce, from their endogenous \textit{Pdx1} locus, a fusion protein composed of amino acids from endogenous rat \textit{Pdx1} exon 1 and sand rat \textit{Pdx1} exon 2, tagged with V5. Control edited cells were also generated using donor plasmids carrying rat \textit{Pdx1} exon 2, the V5 tag, the ribosome skipping sequence, and antibiotic resistance (Fig. 3b).

\textbf{Fig. 2} Relative gene expression levels 48 h following transfection. Expression of each gene in every group shown is relative to that in the corresponding control. The target gene is shown at the top of each panel. Each point represents one technical repeat of the qPCR reaction. \(* = p\text{-value} < 0.05, \** = p\text{-value} < 0.01, \*** = p\text{-value} < 0.001, \textit{****} = p\text{-value} < 0.0001, \text{ns} = \text{not significant. (a) cells cultured in high (25 mM) glucose medium (b) cells cultured in low (0.5 mM) glucose medium. Figure created using ggplot2 package in R (Wickham 2016; R Core Team 2020). Raw data are available in Online Resource 3}
Following transfection and antibiotic selection, we obtained 279 single cell colonies that were analyzed by PCR and gel electrophoresis, plus Sanger sequencing on cells most likely to have successfully integrated the donor sequence (Fig. S3, Online Resource 1). The sand rat donor sequence replaced both wild type alleles in nine out of 101 cell lines, while seven out of 178 cell lines transfected with a rat Pdx1 donor sequence have the correct insert in both wild type Pdx1 alleles (Fig. 3c). Most antibiotic-resistant cell clones did not contain a donor sequence in either Pdx1 allele. Only the 16 gene-edited cell lines were assessed for presence of the PDX1-V5 fusion protein and used for functional analysis. Examples showing expression of the V5-tagged Pdx1 gene and presence of V5-tagged PDX1 protein are included in Online Resource 1 (Figs. S4 and S5).

Sand Rat PDX1 Homeodomain Inefficiently Binds the Rat Insulin 1 Promoter

We used the gene-edited cell lines to compare PDX1 protein binding to an insulin gene promoter, exploiting the V5 peptide tag to enable protein pull-down using chromatin immunoprecipitation (ChIP). We focused on protein-binding at the rat Ins1 promoter as this was unresponsive to sand rat PDX1 in the ectopic expression experiments above, suggesting inefficient binding between sand rat PDX1 homeodomain and the Ins1 promoter. We extracted DNA crosslinked to protein from rat Pdx1-V5 cell line RBM18 and rat-sand rat fusion Pdx1-V5 cell line SBM18; these cell lines have successful biallelic targeting of the Pdx1 locus. Mouse IgG acts as a control for background binding to the mouse antibody, while the Alb

Fig. 3 Experimental design for Pdx1 exon 2 replacement. (a) Design for experimental exon replacement in which rat exon 2 is replaced by sand rat exon 2 linked to a V5 peptide tag, a P2A ribosome skipping sequence, selection markers, and homology arms. Non-exon regions are in black, sand rat Pdx1 exons are in purple with the homeobox region indicated in dark purple. (b) Design for control exon replacement in which rat exon 2 is replaced by rat exon 2 from a donor sequence linked to a V5 peptide tag, a P2A ribosome skipping sequence, selection markers and homology arms. Stars represent cut sites targeted by the Cas9-gRNA complex. Non-exon regions are in black, rat Pdx1 exons are in green with the homeobox region indicated in dark green. (c) Possible outcomes from CRISPR/Cas9 mediated gene exon replacement, showing percentage of clonal cell lines obtained for each outcome. Only cells that have the donor sequence successfully inserted in both alleles are used for downstream analysis. Photograph from J.F. Mulley with permission. Figure created using Inkscape (Inkscape Project 2020)
(albumin) gene promoter also serves as a control with a TAAT motif which PDX1 can recognize but should not bind (Alb is not expressed in pancreatic cells and the promoter region should be inaccessible).

ChIP using anti-V5 antibody shows enrichment of Ins1 promoter over Alb promoter in both rat PDX1 samples and rat-sand rat fusion PDX1 samples, indicating that the sand rat homeodomain can recognize and bind to the rat Ins1 promoter (compare V5 to IgG precipitated groups in Fig. 4). However, the amount of V5-tagged PDX1 transcription factor binding to the Ins1 promoter sequence is far higher for the rat PDX1 gene-edited cells than for the rat-sand rat fusion PDX1 gene-edited cells (compare V5 precipitation groups in Fig. 4). These results suggest that the sand rat homeodomain sequence has a weaker interaction with the rat Ins1 gene promoter.

Discussion

Gerbils reside in arid environments that pose challenges such as extreme heat, lack of water, and limited food choice (Walsberg 2000). Previous work has identified unusual amino acid changes in the PDX1 protein of three gerbil species: the sand rat, Mongolian jird, and Libyan jird (Hargreaves et al. 2017). Some of these substitutions cause changes to PDX1 protein half-life, and we have argued in previous work that many of these amino acid changes were deleterious, fixed in the population through the ‘selection blind’ process of GC-biased gene conversion (gBGC) rather than by natural selection, albeit with some compensatory change fixed by selection (Dai and Holland 2019; Dai et al. 2020). Such a scenario could only be possible if the rates of recombination and gBGC are aberrantly high, as indeed has been shown for gerbils (Pracana et al. 2020). Nonetheless, the possibility remains that several of the amino acid changes could be adaptive to life in an arid habitat, or they could have constrained gerbils to these environments. In this study, we explore if these amino acid changes are associated more broadly with survival in arid habitats, and whether these changes affect the biological activity of PDX1.

One prediction of the adaptive hypothesis is that convergent changes might be seen in other arid-dwelling rodents. However, comparing PDX1 deduced protein sequences between 38 rodent species, spanning at least seven transitions to arid environments, we find radical changes to the PDX1 protein only in the gerbil lineage with nothing comparable seen in other rodents. We find a single amino acid change, alanine to serine, in the PDX1 homeodomain of the kangaroo rat (D. ordii), a rodent that resides in arid environments in the western United States and feeds mainly on seeds (Kaufman and Kaufman 2015). The altered amino acid site is the second residue in helix 1; this alpha helix does not form direct contact with DNA, but may be involved with stabilizing the binding of helices 2 and 3 to the double helix (Longo et al. 2007). Due to lack of access to multiple kangaroo rat tissue samples, we cannot verify if the substitution is a fixed change, an allelic variant, or even a sequencing error.
Regardless, only the four gerbil species have a radically divergent PDX1 hexapeptide and homeodomain sequence with non-conserved residues spread across the entire region (Fig. 1). We suggest that the majority of amino acid changes in the PDX1 protein of gerbils arose through the selection-blind process of gBGC, which can fix deleterious alleles in extreme circumstances.

In a previous study, we showed that one of the amino acid changes in gerbil PDX1, in homeodomain helix 3, resulted in gain of a ubiquitination site; this was likely an adaptive change fixed to compensate for loss of conserved ubiquitination sites removed by gBGC (Dai and Holland 2019). Gain of this ubiquitination site enables a degree of cellular control over protein half-life, but does not completely mirror or restore the ancestral condition. How other gerbil-specific amino acids affect PDX1 activity have not been previously assessed. As PDX1 is a direct regulator of insulin gene expression, we hypothesized that DNA-binding to the insulin promoter or activation of transcription could be compromised. In addition, the effect of PDX1 on insulin gene expression is regulated by glucose levels, with PDX1 having a stimulatory effect under high glucose (>20 mM) and a repressive effect under low glucose conditions (<3 mM) (MacFarlane et al. 1994; Mosley and Özcan 2004; Rosanas-Urgell et al. 2008). We wished to know if this complex regulatory ability was retained in PDX1 proteins of the gerbil lineage.

We used two different experimental approaches to address these questions: ectopic expression follow by QPCR, and gene editing followed by chromatin immunoprecipitation. Each has advantages and disadvantages. Ectopic overexpression is a faster experimental process; however, the endogenous rat protein remains present and could potentially compete with the overexpressed protein for binding sites. Using ectopic expression in a rat pancreatic β-cell line, we found evidence that sand rat PDX1 is indeed compromised in its ability to regulate insulin gene expression. In our experiments, sand rat PDX1 protein did not increase transcription of either of the two rat insulin genes, Ins1 or Ins2, under high or low glucose conditions. This is in marked contrast to ectopic expression of mouse PDX1 protein, which was able to differentially regulate Ins1 and Ins2 as expected. We interpret these results to mean that the radical amino acid changes that occurred outside the homeodomain have altered cofactor interactions. For example, the N-terminal of PDX1 is essential for interaction with the coactivator p300, which is involved in PDX1-mediated insulin gene expression (Stanojevic et al. 2004). We also note that sand rat PDX1 cofactors do not seem to have undergone accelerated evolution alongside the divergent sand rat PDX1 protein, as all known sand rat PDX1 cofactors have high amino acid sequence conservation compared to their mouse and rat homologues. In addition, we note that sand rat PDX1 has lost conserved phosphorylation sites in the N- and C-terminal region, which may also affect its interaction with cofactors (An et al. 2010; Frogne et al. 2012).

Using CRISPR/Cas9 gene editing, we removed the second exon of the endogenous rat Pdx1 gene and replaced it with the second exon of sand rat Pdx1, generating cells that produce a fusion protein from both alleles. This hybrid PDX1 protein contains the rat N-terminal region, thought to be responsible for protein–protein interactions between PDX1 and its cofactors (Stanojevic et al. 2004), and the sand rat homeodomain primarily controlling DNA-sequence recognition and binding. The successfully edited cells no longer have the full endogenous gene, so there is no competitor PDX1 protein present. These cells should enable us to test if the sand rat PDX1 homeodomain can bind an insulin promoter efficiently, something we could not assess using ectopic expression. It is important to note, however, that we cannot assess the full repertoire of cofactor interactions with these cells as the fusion protein includes the N-terminal region of PDX1 from rat, not from sand rat. To enable analysis of cofactor interactions, it would have been necessary to replace both exons of the Pdx1 gene, at both alleles, which would have necessitated an even more complex experimental design. Even in the replacement of a single exon, we found
that the majority of cell clones that survived antibiotic treatment did not have successful gene editing and must have integrated donor plasmid DNA at inappropriate sites in the genome; successful gene editing of both alleles was a rare event.

Chromatin immunoprecipitation revealed that, in gene-edited pancreatic β-cells, the fusion PDX1 protein is bound to the Ins1 promoter. This indicates that the highly divergent sand rat PDX1 homeodomain can recognize and bind the Ins1 promoter, despite radical amino acid changes, consistent with in silico prediction results performed previously (Hargreaves et al. 2017). However, this interaction is much weaker than with rat PDX1 protein under the same conditions. This is not due to a species mismatch as the known target sequence in the promoter, the A3 box region, is well-conserved and is the same in gerbils and several murids (mouse, rat, and Algerian mouse; Fig. S6, Online Resource 1). The difference in binding efficiency indicates that amino acid changes in the sand rat homeodomain have disrupted ability to interact with a well-known target sequence. We suggest this could have a severely detrimental impact on sand rat PDX1 function. As for other PDX1 target genes, we could not obtain full-length sand rat promoter sequence due to incomplete sequence upstream of Pdx1, Mafa, and Slc2a2.

Both ectopic gene expression and the gene edited cells provide complementary information concerning the biological activity of the PDX1 transcription factor in sand rat and by implication other gerbils. The gene-edited cells reveal that the unusual amino acid changes in gerbils, which may have arisen through gBGC, have compromised but not abolished the ability of the PDX1 protein to bind promoter sequences upstream of an insulin gene target. However, we could not assess glucose-stimulated insulin gene expression in gene edited cells, because these cells had lost the ability to respond to glucose. Control cells revealed this reflects a phenotypic alteration caused by passing through a single-cell state (Fig. S7, Online Resource 1). Taken together, we suggest that the amino acid changes that accumulated in PDX1 during gerbil evolution compromised the DNA-binding efficiency of the protein and disrupted ability to stimulate rapid insulin production after glucose stimulation. We have not assessed potential functional differences between PDX1 proteins of different gerbil species, as these show only minor protein sequence variation.

We argue that the amino acid changes observed in the PDX1 protein of sand rat and other gerbils are unlikely the result of adaptation to an arid environment as they are not seen in other desert rodents. This finding is consistent with previous work that has highlighted an extreme amount of gBGC in gerbils, a molecular process that can fix deleterious alleles in a selection-blind manner. We show that the highly derived sand rat PDX1 protein may carry deleterious amino acid changes. The sand rat protein inefficiently binds the rat insulin promoter and ineffectively regulates insulin gene expression in response to glucose, although it has retained ability to regulate expression of some other target genes which may partially compensate. We speculate that these functional changes could have reduced the ability of gerbils to handle carbohydrate-rich diets during their evolution. Thus, while radical amino acid changes in gerbils were not adaptive, they may have been ecologically constraining.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10914-021-09544-x.

Acknowledgements We thank Tatiana Alfonso-Pérez and Francis Barr for plasmids containing antibiotic resistance genes and for assistance with the CRISPR protocol, the Feng Zhang lab for the pX458 vector, Ignacio Maeso and Matias De Vas for advice on chromatin immunoprecipitation, and Amy Royall, Thomas Lewin, and Rodrigo Pracana for advice and helpful discussion. This work was facilitated by researchers giving access to published and unpublished DNA sequencing data; in particular, we thank Thomas Keane for permission to use Onychomys torridus sequence data and the Beijing Genomics Institute, University of Liverpool, and the DNA Zoo consortium for making datasets publicly available.

Authors Contributions YD and PWHH conceived the study. YD, ST and PWHH jointly designed the laboratory experiments. YD performed laboratory experiments and data analysis with advice from ST and PWHH. YD and PWHH drafted the first version of this manuscript. All authors contributed to revision and critical editing of the manuscript and approved the final version.

Funding This work was supported by the Elizabeth Hannah Jenkinson Fund (to Y.D.), the Rhodes Trust (to Y.D.), and a Leverhulme Trust Research Project Grant (RPG-2017-321 to P.W.H.H.).

Data Availability Data generated and methods necessary to replicate analyses are included in this published article and its Online Resource files.

 Declarations

Conflicts of Interest/Competing Interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

 Springer
References

Alfonso-Pérez T, Hayward D, Holder J, Grüneweg U, Barr FA (2019) MAD1-dependent recruitment of CDK1-CCNB1 to kinetochores promotes spindle checkpoint signaling. J Cell Biol 218:1108–1117

An R, da Silva Xavier G, Semplici F, Vakhshouri S, Hao HX, Rutter J, Pagano MA, Meggio F, Pinna LA, Rutter GA (2010) Pancreatic and duodenal homebox 1 (PDX1) phosphorylation at serine-269 is HIPK2-dependent and affects PDX1 subnuclear localization. Biochem Biophys Res Commun 399:155–161

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastq/

Asfari M, Janjic D, Meda P, Li G, Halban PA, Wollheim CB (1992) Establishment of 2-mercaptoethanol-dependent differentiated insulin-secreting cell lines. Endocrinology 130:167–178

Beijing Genomics Institute (2015) Whole genome sequencing of cotton rat. https://db.cngb.org/search/project/CNPhis0000287/

Beijing Genomics Institute (2017) Ground squirrel genome sequencing revealed genetic adaptations for cellular stress in hibernation. https://www.ncbi.nlm.nih.gov/genome/38559

Bendesky A, Kwon Y-M, Lassance J-M, Lewarch CL, Yao S, Peterson BK, He MX, Dulac C, Hoekstra HE (2017) The genetic basis of parental care evolution in monogamous mice. Nature 544:434–439

Chevret P, Dobigny G (2005) Systematics and evolution of the subfamily Gerbillinae (Mammalia, Rodentia, Muridae). Mol Phylogenet Evol 35:674–688

Chikhi R, Medvedev P (2013) Informed and automated k-mer size selection for genome assembly. Bioinformatics 30:31–37

Clavijo BJ, García Accinelli G, Wright J, Heavens D, Barr K, Yanes L, Di-Palma F (2017) W2RAP: a pipeline for high quality, robust assemblies of large complex genomes from short read data. bioRxiv:1–12

Dai Y, Holland PWH (2019) The interaction of natural selection and GC skew may drive the fast evolution of a sand rat homeobox gene. Mol Biol Evol 36:1473–1480

Dai Y, Pracana R, Holland PWH (2020) Divergent genes in gerbils: prevalence, relevance to GC-biased substitution, and phenotypic relevance. BMC Evol Biol 20:1–15

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger E, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL (2017) De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356:92–95

Dudchenko O, Shamim MS, Batra SS, Durand NC, Musial NT, Mostofa R, Pham M, Glenn St Hilaire B, Yao W, Stamenova E, Hoeger E, Nyquist SK, Korchina V, Petch K, Planagan JP, Tomaszewicz A, McAloose D, Pérez Estrada C, Novak BJ, Omer AD, Aiden EL (2018) The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv

Frogne T, Sylvestersen KB, Kubicek S, Nielsen ML, Hecksher-Sørensen J (2012) Pdx1 is posttranslationally modified in vivo and serum 61 is the principal site of phosphorylation. PLoS One 7:e35233

Gingerich TJ, Stumpo DJ, Lai WS, Randall TA, Steppan SJ, Blackshear TB (2016) Emergence and evolution of Zfp36l3. Mol Phylogenet Evol 94:518–530

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Girschik J, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–52

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hargreaves AD, Zhou L, Christensen J, Marlétaz F, Liu S, Li F, Jansen PG, Spiga E, Hansen MT, Pedersen SVH, Biswas S, Serikawa K, Fox BA, Taylor WR, Mulley JF, Zhang G, Heller RS, Holland PWH (2017) Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. Proc Natl Acad Sci USA 114:7677–7682

Hohmeier HE, Mulder H, Chen GX, Henkel-Rieger R, Prentki M, Newgard CB (2000) Isolation of INS-1-derived cell lines with robust ATP-sensitive K+ channel-dependent and -independent glucose-stimulated insulin secretion. Diabetes 49:424–430

Inkscape Project (2020) Inkscape. https://inkscape.org

IUCN (2020) The IUCN red list of threatened species. Version 2020–2. https://www.iucnredlist.org

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

Kalderon B, Gutman A, Levy E, Shafrir E, Adler JH (1986) Characterization of stages in development of obesity-diabetes syndrome in sand rat (Psammomys obesus). Diabetes 35:717–724

Kaufman DW, Kaufman GA (2015) Ord’s kangaroo rats in north-central Kansas: patterns of body size and reproduction. Trans Kansas Acad Sci 118:251–263

Lin G-H, Wang K, Deng X-G, Nevo E, Zhao F, Su J-P, Guo S-C, Zhang T-Z, Zhao H (2014) Transcriptional sequencing and phylogenomic resolution within Spalacidae (Rodentia). BMC Genomics 15:32

Longo A, Gupta GP, Rose RB (2007) Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain. Biochemistry 46:2948–57

MacFarlane WM, Read ML, Gilligan B, Bujalska I, Docherty EK (1994) Glucose metabolism: the binding activity of the beta-cell transcription factor IUF1 in a phosphorylation-dependent manner. Biochem J 303:625–31

Marshall S, Totary H, Cerami E, Melloul D (1996) Purification of the β-cell glucose-sensitive factor that transactivates the insulin gene differentially in normal and transformed islet cells. Proc Natl Acad Sci USA 93:15057–15062

Matsuoka TA, Ketano H, Stein R, Miyazaki T, Kawamori D, Henderson E, Kojima I, Matsuura M, Hori Y, Yamasaki Y (2007) MafA regulates expression of genes important to islet β-cell function. Mol Endocrinol 21:2764–2774

Mosley AL, Ozcan S (2004) The pancreatic duodenal homeobox-1 protein (Pdx-1) interacts with histone deacetylases Hdac-1 and Hdac-2 on low levels of glucose. J Biol Chem 279:54241–54247

Mulugueta E, Wassenberg E, Sleedens-Linkels E, van IJken WJF, Heard E, Grooteogel JA, Just W, Gribnau J, Baarends WM (2016) Genomes of Elobius species provide insight into the evolutionary dynamics of mammalian sex chromosomes. Genome Res 26:1202–1210

Neme R, Tautz D (2016) Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife 5:e09977

Ohlsson H, Karlsson K, Edlund T (1993) IPF1, a homeodomain-containing transactivator of the insulin gene. EMBO J 12:4251–4259

Ojeda RA, Borghi CE, Diaz GB, Giannoni SM, Mares MA, Braun JK (1999) Evolutionary convergence of the highly adapted desert rodent Typanoctomys barrerae (Octodontidae). J Arid Environ 41:443–452

Pracana R, Hargreaves AD, Mulley JF, Holland PWH (2020) Runaway GC evolution in gerbil genomes. Mol Biol Evol 37:1262–1270

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F (2013) Genome editing using the CRISPR-Cas9 system. Nat Protoc 8:2281–2308

Rosanas-Urgell A, Garcia-Fernández J, Marfany G (2008) ParaHox genes in pancreatic cell cultures: effects on the insulin promoter regulation. Int J Biol Sci 4:48–57

Springer
Rudnick A, Ling TY, Odagiri H, Rutter WJ, German MS (1994) Pancreatic beta cells express a diverse set of homeobox genes. Proc Natl Acad Sci USA 91:12203–12207
Sanger Institute UK (2014) Assembly-stats. https://github.com/sanger-pathogens/assembly-stats
Schmidt-Nielsen K, Haines HB, Hackel DB (1964) Diabetes mellitus in the sand rat induced by standard laboratory diets. Science 143:689–690
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123
Stanojevic V, Habener JF, Thomas MK (2004) Pancreas duodenum homeobox-1 transcriptional activation requires interactions with p300. Endocrinology 145:2918–2928
Supharattanasitthi W, Carlsson E, Sharif U, Paraoan L (2019) CRISPR/Cas9-mediated one step bi-allelic change of genomic DNA in iPSCs and human RPE cells in vitro with dual antibiotic selection. Sci Rep 9:1–7
Takei Y, Bartolo RC, Fujihara H, Ueta Y, Donald JA (2012) Water deprivation induces appetite and alters metabolic strategy in Notomys alexis: unique mechanisms for water production in the desert. Proc R Soc B Biol Sci 279:2599–2608
The Wellcome Trust Sanger Institute (2015) Dissecting behavioural traits of the grasshopper mouse. https://www.ncbi.nlm.nih.gov/sra/?term=ERR968259
Tigano A, Colella JP, MacManes MD (2020) Comparative and population genomics approaches reveal the basis of adaptation to deserts in a small rodent. Mol Ecol 29:1300–1314
University of Liverpool (2016) De novo sequencing of wood mouse (Apodemus sylvaticus): liver. https://gold.jgi.doe.gov/project?id=Gp0143296
Waeber G, Thompson N, Nicod P, Bonny C (1996) Transcriptional activation of the GLUT2 gene by the IPF-1/STF-1/IDX-1 homeobox factor. Mol Endocrinol 10:1327–34
Walsberg GE (2000) Small mammals in hot deserts: some generalizations revisited. Bioscience 50:109–120
Wickham H (2016) Ggplot2: elegant graphics for data analysis. Springer-Verlag, New York
Zhang C, Moriguchi T, Kajihara M, Esaki R, Harada A, Shimohata H, Oishi H, Hamada M, Morito N, Hasegawa K, Kudo T, Engel JD, Yamamoto M, Takahashi S (2005) MafA is a key regulator of glucose-stimulated insulin secretion. Mol Cell Biol 25:4969–76