Weird gene in a weird mammal: Comparative and functional analysis of the highly divergent sand rat *Pdx1* gene

Yichen Dai

(OSS student ID: 1007640)
St. Catherine's College
Department of Zoology

Thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.



University of Oxford

2020

Declaration and Statement of Authorship

I declare that the work presented in this thesis was conducted by myself under the supervision of Prof. Peter W. H. Holland (University of Oxford), except in those instances where the contribution of others has been acknowledged. Any material quoted, referenced, or paraphrased from other sources has been clearly indicated and its origins cited.

Chapter 3 and Chapter 5 have been written up as original scientific research manuscripts with me as the lead author. Chapter 3 has been published in *Molecular Biology and Evolution* and Chapter 5 has been submitted to *BMC Evolutionary Biology*. In both cases, I have drafted the original manuscripts and most of the work done is attributable to me unless otherwise stated. All co-authors on both submitted manuscripts have evaluated the work and contributed to editing the manuscripts prior to submission. A detailed description of author contributions is provided at the start of each data chapter.

Ofuase Dirisu (in2science UK) carried out some experimental work presented in Chapter 3 under my supervision. She constructed one mutant *Pdx1* plasmid and collected protein half-life data for this experimental group. Dr. Rodrigo Pracana (University of Oxford) provided the dataset of 10544 one-to-one orthologues used in Chapter 5 and assisted with protein dissimilarity measurement and PAML analysis used in Chapter 5. I carried out data collection and analysis by myself for the other chapters. Dr. Tatiana Alfonso Perez (University of Oxford) provided the pX8555 and pX8562 plasmids containing the blasticidin and puromycin resistance gene, respectively, used in Chapter 4. Laboratory work was carried out in the laboratory of Prof. Peter Holland using reagents available in the lab. Prof. Peter Holland supervised the development of my work for all data chapters and discussed experimental design of Chapters 2, 3 and 4 with me, along with Dr. Rodrigo Pracana (University of Oxford) in experimental design for Chapter 5. Funding for data collection and data analysis was secured by Prof. Peter Holland through Leverhulme Trust Research Project (Grant No. RPG-2017-321), with the exception of Chapter 3 and 4 where funding for reagents required for protein half-life measurement and CRISPR/Cas9 editing was obtained jointly by me and Prof. Peter Holland through the Elizabeth Hannah Jenkinson Fund. I also acknowledge support from the Rhodes Scholarship.

I hereby confirm that this thesis has not been submitted for any other qualification or degree at this university or any other institution.

Acknowledgements

The work presented in this thesis would not exist without many individuals to whom I own my sincere gratitude. First and foremost, I would like to thank my supervisor, Professor Peter Holland, for his guidance and mentorship throughout my DPhil. As I began stumbling around bioinformatics and rodent phylogenetics, Peter always gave valuable suggestions and guided the progression of my project while leaving me sufficient space to come up with new ideas. He was always keen to discuss new papers with the lab, and he encouraged everyone to share their ideas; in addition, Peter was open to discussing our thoughts on academia – be it the academic publishing process, postdoc requirements, reviewers, …. Peter treated everyone in the lab as fellow scientists and valued input from everyone. He encouraged me to try out experiments that no one knew if they would work or not, and he encouraged me to join conferences and seminars that I was interested in and present my work. Peter used his own actions to show me how a scientist should carry out research, inspire others, and be a good mentor and collaborator. Thanks to Peter, I learnt far more than what is written down in this thesis in the past four years.

I would like to thank Dr. Sonia Trigueros (University of Oxford) for her valuable knowledge on DNA structure and mammalian cell transfection, for her lending us her brilliant microscope that made Chapter 3 possible, and for her positive energy that accompanied me throughout my DPhil. Lots of the bioinformatics work in this thesis would have caused more stumbles and confusion if it were not for help provided by Dr. Rodrigo Pracana (University of Oxford). Dr. Adam Hargreaves (University of Oxford) introduced me to the Linux platform and animal genome assembly, and Dr. Daniel Lunn (University of Oxford) gave me valuable assistance with statistical analysis in Chapter 3. Dr. Thomas Dunwell (University of Oxford) taught me mammalian cell culture, and Dr. Tatiana Alfonso Perez (University of Oxford) shared her protocol for CRISPR/Cas9 editing in mammalian cells and provided two plasmids used in Chapter 4. Dr. Amy Royall (University of Oxford) put in many hours of hard work helping me maintain the CRISPR edited cells while I was away, and Thomas Lewin, Brigid Leung, and Emily Baker were always keen to discuss experimental difficulties and give me much needed suggestions. Dr. John Mulley and Dr. Thomas Brekke (Bangor University) also gave valuable advice and ideas for discussion about the gerbil project. Many thanks to my assessors during my transfer of status and confirmation, Dr. Sebastian Shimeld, Prof. Gil McVean, and Prof. Aris Katzourakis (University of Oxford). I also need to thank everyone that I discussed my project with in the Department of Zoology and at the different seminars and conferences, and thank you, Sharon Cornwell, for all the visits to the open plan office and the biscuits.

I have greatly appreciated the support and companionship of friends outside the lab. Thank you for all the trips to Taberu, for listening to me rant, for the hotpot dinners and boardgame nights, and thank you for sharing the ups and downs of the DPhil with me.

Finally, I owe my parents a massive thank you. Thank you for supporting my decision to pursue a DPhil degree in Oxford, thank you for always being there for me, and thank you for your unconditional love.

Table of Contents

List of Figures iv

Thesis Abstract 1

Chapter 1: General introduction 4

1. GC content in animal genomes and factors affecting it 5

    1.1 GC content varies within the genome and between species 5
    1.2 The effect of mutation bias on GC content 7
    1.3 The effect of selection on GC content 8
    1.4 The effect of biased gene conversion (gBGC) on GC content 9
        1.4.1 Mechanisms of recombination and gBGC 9
        1.4.2 Direct evidence supporting presence of gBGC 11
        1.4.3 Indirect evidence supporting presence of gBGC 11
        1.4.4 The impact of gBGC on GC content and genome evolution 12

2. Sand rat phylogeny, physiology, and GC composition bias 12

    2.1 Phylogeny 12
    2.2 Physiology 12
        2.2.1 Insulin-regulated blood glucose homeostasis 12
        2.2.2 Type 2 diabetes 15
        2.2.3 The sand rat and other gerbils as a type 2 diabetes model 16
    2.3 GC composition bias: the sand rat genome has extremely GC-rich regions 18

3. Overview of pancreatic duodenum homeobox 1 (*Pdx1*) 21

    3.1 The ParaHox gene cluster: *Gsx1*, *Pdx1*, *Cdx2* 21
    3.2 *Pdx1* initiates vertebrate pancreatic development 21
    3.3 *Pdx1* regulates vertebrate adult β-cell function 22
    3.4 PDX1 variants can cause metabolic disease in humans 23

4. Project aims and contributions 24

Chapter 2: Highly divergent *Pdx1* is only observed in the gerbil lineage 26

Statement of Author Contributions 27

1. Abstract 28

2. Introduction 29

3. Materials and methods 31

    3.1 Rodent genome and transcriptome assembly 31
    3.2 PCR verification of the guinea pig *Pdx1* exon 1 sequence 31

4. Results 37

    4.1 Assembly of rodent raw sequencing data 37
    4.2 Alignment of rodent PDX1 hexapeptide and homeodomain sequence 38

4.3 PCR and Sanger sequencing verification of guinea pig *Pdx1* exon 1 39

5. Discussion 41

Chapter 3: Loss of a key PDX1 ubiquitination site is linked to GC composition bias 43

Statement of Author Contributions 44

1. Abstract 45

2. Introduction 46

3. Materials and methods 47

3.1 Plasmid construction and *in vitro* mutagenesis 47
3.2 Cell culture and transfection 47
3.3 SNAP-staining and fluorescence recording 47
3.4 MG132 treatment 47
3.5 Statistical analysis 47

4. Results 48

4.1 Altered stability of gerbil PDX1 protein 48
4.2 Sand rat PDX1 protein degradation is regulated by the UPS 50
4.3 PDX1 ubiquitination sites suggest compensatory molecular evolution 50

5. Discussion 53

6. Supplementary information 55

Chapter 4: Sand rat PDX1 inefficiently regulates insulin expression in rat cells 62

Statement of Author Contributions 63

1. Abstract 64

2. Introduction 65

3. Materials and methods 67

3.1 Construction of plasmid constructs for ectopic overexpression 67
3.2 Mammalian cell culture and transfection 67
3.3 Glucose treatment, total RNA extraction and qPCR 67
3.4 Construction of plasmid constructs for CRISPR/Cas9 gene knock-in 68
3.5 Mammalian cell transfection and selection for CRISPR/Cas9 gene editing 69
3.6 gDNA and cDNA preparation followed by PCR 70
3.7 Protein extraction and Western blot 70
3.8 Protein immunostaining in fixed mammalian cells 71
3.9 Chromatin immunoprecipitation and qPCR 71

4. Results 72

4.1 Sand rat PDX1 inefficiently regulates expression of several key β-cell genes 72

4.2 Establishment of rat β-cells producing a rat-sand rat fusion PDX1 protein 74
4.3 INS-1 cells cultured from single cells show abnormal response to glucose 82
4.4 Sand rat PDX1 has weak binding affinity to the rat insulin I promoter 83

5. Discussion 86

Chapter 5: GC composition bias impacts multiple genes in the gerbil genome 89

Statement of Author Contributions 90

1. Abstract 91

2. Introduction 92

3. Materials and methods 93

3.1 Liver transcriptome sequencing using MinION 93
3.2 Protein dissimilarity calculation and ranking 93
3.3 GC-skew analysis and correlation with high GC peaks 93
3.4 Analysis of protein functions 94

4. Results 95

4.1 Many gerbil proteins are extremely divergent compared to their homologues 95
4.2 Aberrantly divergent gerbil proteins are associated with islands of high GC 97
4.3 Phenotypic implications of aberrantly divergent gerbil genes 100
4.4 MinION sequencing did not yield further information on GC-rich genes 106

5. Discussion 110

6. Supplementary information 114

Chapter 6: Discussion 120

1. The cause of GC composition bias in gerbils 121

2. Sand rat GC composition bias and physiology 123

3. The impact of GC composition bias on evolution 125

Concluding remarks 127

Appendix A: List of abbreviations used 128

Appendix B: Additional files 130

Appendix C: Resulting publications 131

C1 The interaction of natural selection and GC skew may drive the fast evolution of a sand rat homeobox gene 133
C2 Divergent genes in gerbils: origin, evolution and phenotypic relevance 141

References 184

List of Figures

1. General introduction
1.1 Two modes of GC3 content distribution 6
1.2 Mechanisms of recombination and GC-biased gene conversion 10
1.3 Sand rat phylogenetic position 13
1.4 Insulin secretion in human pancreatic β-cells and insulin response in somatic cells 14
1.5 Sand rat genes located in an extreme GC-rich region 19
1.6 Comparison of the ParaHox cluster in amphioxus and mammals 20
2. Highly divergent *Pdx1* is only observed in the gerbil lineage
2.1 Alignment of rodent PDX1 hexapeptide and homeodomain regions 36
2.2 Alignment of human, guinea pig, mouse, and sand rat PDX1 protein 39
2.3 Alignment of PDX1 protein including the updated guinea pig sequence 40
3. Loss of a key PDX1 ubiquitination site is linked to GC composition bias
3.1 Alignment of PDX1 hexapeptide and homeodomain across vertebrates 48
3.2 Measurement of protein stability 49
3.3 Effect of inhibiting UPS on PDX1 protein stability 50
3.4 Protein stability in wild type and mutant proteins 51
3.5 Summary of the possible evolutionary pathway proposed 52
4. Sand rat PDX1 inefficiently regulates insulin expression in rat cells
4.1 Relative gene expression levels 48 hours following transfection 72
4.2 Five essential components of the donor sequence 76
4.3 PCR results from cell line screening 77
4.4 Outcomes from CRISPR/Cas9 mediated gene knock-in 79
4.5 Verification of successful editing at the RNA and protein level 80
4.6 Immunofluorescence staining of different edited cell lines 81
4.7 Relative *Ins2* gene expression levels in multiple edited cell lines 82
4.8 Alignment of the insulin promoter in multiple rodent species 83
4.9 Relative amount of *Ins1* promoter in each ChIP sample 84
5. GC composition bias impacts multiple genes in the gerbil genome
5.1 Phylogenetic relationships and patterns of protein divergence 96
5.2 Dissimilarity difference ranking for 9771 sand rat proteins 97
5.3 Divergent sand rat proteins are frequently encoded by genes in GC-rich islands 98
5.4 Mutational patterns at synonymous sites for genes from four rodent species 100
5.5 Alignment of protein domains in PDX1, INSR, MEDAG, and SPP1 proteins 103
5.6 Distribution of good quality sand rat liver transcriptome reads 106
5.7 Alignment of the sand rat *Insr* gene sequenced using Illumina and Nanopore 107
5.8 Distribution of good quality Mongolian jird liver transcriptome reads 108

Thesis Abstract

Deleterious mutations in highly conserved genes are expected to be removed by natural selection. However, when these genes are trapped in dynamic regions of the genome, for example mutational hotspots, it becomes difficult to predict how the conflict between nucleotide changes and natural selection will be resolved. In this thesis, I investigate highly divergent genes in the fat sand rat (*Psammomys obesus*) associated with GC-skew. To understand how the tug-of-war between GC-skew and natural selection is resolved, I focus on a highly representative example of an extremely conserved gene trapped in an extreme GC-rich region of the sand rat genome: pancreatic duodenum homeobox 1 (*Pdx1*).

In Chapter 2, I hypothesize that the highly divergent sand rat PDX1 protein may be a result of adaptive selection, and that similar amino acid changes may be observed in other arid-living rodents. Using genomic and transcriptomic data gathered from 33 rodent species, I show that there is near 100% conservation of the PDX1 homeodomain sequence in non-gerbil species, including 15 rodent species that reside in arid or semi-arid habitats. The highly divergent *Pdx1* gene observed in the sand rat is only found in three gerbil species, suggesting that adaptive or relaxed selection in an arid environment may not be the key factor affecting *Pdx1* gene evolution in this lineage. Instead, other factors such as GC composition bias may be the main factor driving *Pdx1* gene divergence in the gerbil subfamily.

In Chapter 3, I hypothesize that natural selection may have only managed to purge extremely deleterious mutations while tolerating the existence of mildly deleterious mutations in these genes. Using mammalian cell transfection and pulse-chase labelling, I reveal loss of a conserved ubiquitination site and compensation from a unique, gerbil-specific ubiquitination site. This provides evidence for loss of a conserved amino acid with important function associated with GC-skew and links the presence of a gerbil-unique lysine residue with a specific function.

In Chapter 4, I use qPCR to demonstrate significantly altered regulatory effects of the sand rat PDX1 protein on key target genes in rat pancreatic cells, *Ins1* and *Ins2*. I also present establishment of CRISPR/Cas9 edited knock-in cell lines with the aim of teasing apart whether these observed differences were caused by changes in sand rat PDX1 binding site recognition. I conclude that the highly divergent sand rat PDX1 has likely lost conserved functional residues and has significantly reduced ability to initiate proper biological response to glucose stimulation in a rat cell environment. At the same time, results in Chapter 3 show the presence of gerbil-specific residues with specific function, indicating possible compensation by natural selection.

I use genome wide comparison methods in Chapter 5 to demonstrate that extreme GC-skew has affected multiple highly conserved genes with assorted function. *Pdx1* is not the only aberrantly divergent gene in the gerbil lineage, and these aberrantly divergent genes are linked to GC-skew and high GC regions in the gerbil genome. Four divergent genes, *Pdx1*, *Insr*, *Medag* and *Spp1* are likely associated with obesity and abnormal metabolism observed in gerbils on a standard laboratory diet.

Overall, this research provides further characterization of GC-rich genome regions in the gerbil lineage and functional analysis of a highly divergent sand rat gene, *Pdx1*, that is affected by GC-skew but also has significant functional importance. This work suggests that when extreme changes in genome nucleotide content occurs, natural selection may be incapable of removing some maladaptive mutations in highly conserved genes that are otherwise abolished in other lineages.

Chapter 1:

General introduction

- ❖ GC content in animal genomes and factors affecting it
- ❖ Sand rat phylogeny, physiology, and GC composition bias
- ❖ Overview of pancreatic duodenum homeobox 1 (*Pdx1*)
- ❖ Project aims and contributions

General introduction

Genetic diversity is the basis of evolution and the mechanisms generating and maintaining this diversity has been a much-debated topic in molecular evolution for decades. Changes to genetic material, for example point mutations, have the potential to permit rapid evolution depending on the site at which it occurs, and bigger changes such as transposon movement, tandem gene duplication, chromosome structural changes and whole genome duplication provide ample opportunities for new phenotypes to appear and for species to evolve[1–5]. In this research, I focus on variation within the DNA molecule – the ratio of GC base pairs to AT pairs and its impact on gene evolution.

1. GC content in animal genomes and factors affecting it

Variation in DNA base content within the genome has been widely acknowledged since the 1970s, starting with the observation that digested genomic fragments would settle at different positions in a density gradient[6,7]. The term "isochore" was soon coined to describe regions of the genome with relatively homogenous GC levels, although it gradually fell out of favour as sequencing technology allowed better characterization of GC content distribution in animal genomes[8–10]. Here, I provide a general overview of GC content distribution in different animal genomes, followed by brief summaries of proposed mechanisms causing variation in GC content.

1.1 GC content varies within the genome and between species

The average GC content of animal genomes fall within a range of 35 – 45%, with well-known exceptions including the medusozoan *Hydra magnipapillata* which has a genome with 29% GC and members of the order Hymenoptera, for example the honeybee *Apis mellifera* (33% GC) and the fire ant *Solenopsis invicta* (35% GC)[11–15]. Aside from simply comparing overall average GC content, some papers have focused on analysing GC content variation within the genome; GC level fluctuation across the genome has been described in multiple bacterial, plant and animal genomes[11,16–18].

Several trends have been observed by comparing GC content distribution between different genomes. First, the level of GC variance can vary between different species. In general, when analysing the genome in windows of fixed size (for example 20 kb), a histogram can be plotted with the x-axis displaying average GC% within the window and the y-axis showing the number of windows corresponding to different GC%. Interestingly, two types of GC content distribution have been observed in animal genomes. Tatarinova et al. characterized GC content at the third codon position (GC3) in different species, and they observed that the 12 species analysed could be classified as either "bimodal" or "unimodal" based on the distribution of GC3 content in their genome[18]. Bimodal GC3 distribution describes a broad, high variance distribution in which the distribution is usually right skewed toward high GC% (Fig. 1.1A). For example, the human genome has bimodal GC3 distribution, with 10% of human genes possessing around 75% GC3 and 8% of genes possessing around 45% GC3[18]. In contrast, unimodal GC3 distribution describes a narrow, low variance distribution in which GC3% is closely clustered around the mean (Fig. 1.1B). The zebrafish is a good example of an animal

with unimodal GC3% distribution, with more than 20% of its genes possessing 55% GC3[18]. Even when GC content analysis includes non-coding regions of the genome, a similar trend can be observed. Analysis of the first draft of the human genome showed that when average GC% was calculated over 20 kb regions, the distribution spanned 30 – 68% GC with 41% as the genome-wide average, indicating a wide distribution of GC% in the genome[19]. Similarly, genome-wide analysis of the chicken, crocodile and alligator genomes showed patterns indicating high variation of GC%[11,20]. In contrast, animals like the lizard had similar genome average GC% content compared to the human, but the variation of GC% was far lower in the green anole lizard (*Anolis carolinensis*) genome compared to the human genome (standard deviation in lizard genome equals 0.038 while in human genome standard deviation is 0.064)[20]. Other animals such as the frog, zebrafish and *Ciona intestinalis* also have genomes with clustered distribution of GC% similar to the lizard[20–22].

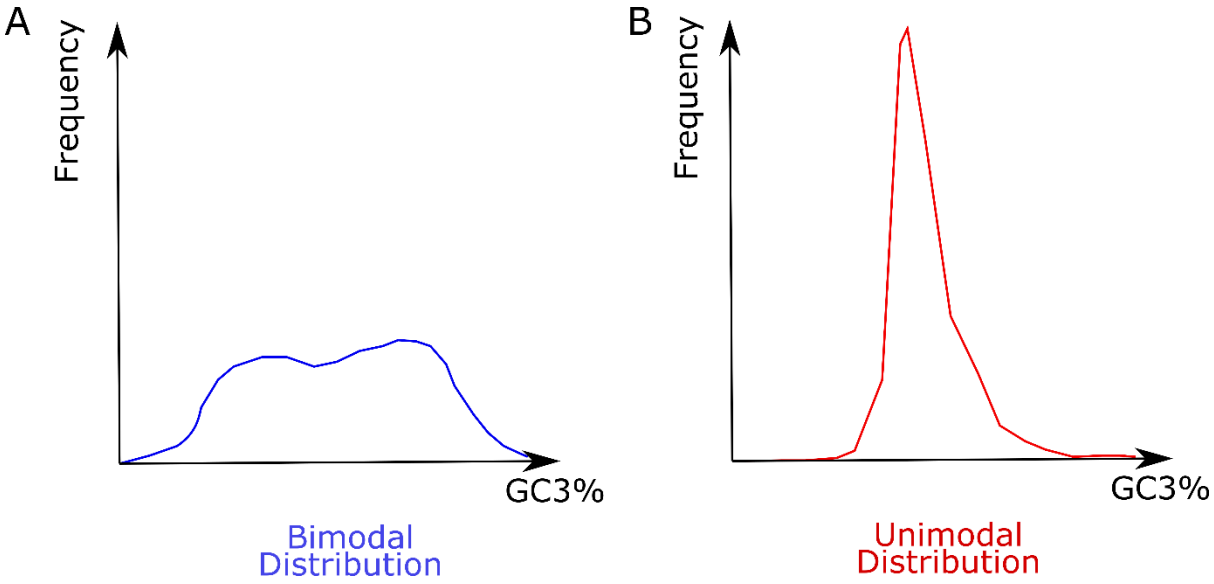


Fig. 1.1 Two modes of GC3 content distribution. This figure is reconstructed from Tatarinova et al. Figure 1[18]. A) genomes with bimodal distribution have a wide spread of GC3% and high levels of variation. B) genomes with unimodal distribution have clustered distribution of GC3% around a certain value and low levels of variation.

Second, very large (>1 Mb) continuous isochores are relatively rare in animal genomes, and these isochores tend to be GC-poor. Cohen et al. defines isochores as chunks of the genome that have a lower GC variability than the chromosome it is located on, without specifying a specific size range for such regions[23]. Therefore, they report only 67% of the human genome can be considered isochores and are >100 kb in size, with only 13% of these regions larger than 1 Mb[23]. However, Clay and Bernardi argue that this definition is problematic as chromosomes have large differences in homogeneity and therefore the threshold used for defining each isochore ends up being different[24]. In addition, Costantini et al. point out that when using regions below 100 kb in size, contribution from different genome structures lead to an increase in variation and the analysis becomes too fragmented, hence they analyse the genome in windows no less than 100 kb, leading them to conclude that 96.4% of the human genome is GC homogenic and >200 kb in size[25]. Although it remains to be concluded what the smallest acceptable size for defining isochores is, both papers agree that the portion of isochores above

1 Mb is relatively low in the human genome, and most of these very large isochores have low GC content[23,25]. This is also true in other species, with vertebrates such as the chicken possessing similar patterns in variation of GC homogeneity and invertebrates such as the honeybee and fruit fly possessing even more fragmented isochores[12].

Many questions have been asked about how isochores arise, whether they are maintained, and why they exist. One approach to answering these questions is probing the biological mechanisms that can alter GC content, which I discuss in the following three sections.

1.2 The effect of mutation bias on GC content

Mutation bias is thought to be an important factor affecting GC content, but the level of influence it has on the genome remains unclear. In this section, I only discuss the effects of point mutations on GC content, which are caused by mechanisms that differ from indels and transpositions[26].

In animals, point mutations lower GC content at the affected base (and therefore are considered to be biased towards AT pairs) due to the strong effect of spontaneous demethylation occurring at methylated cytosine bases, which results in replacement of a cytosine residue with thymine[27–29]. Recent sequencing efforts have shown that demethylation of cytosine is the top cause of *de novo* mutation (DNM) in male germ cells, and even when cytosine is not methylated, a cytosine to thymine base change is the most frequently observed type of base substitution in female germ cells that have not been methylated[30]. Similarly, Arbeithuber et al. report an approximately 11 times higher chance of a GC to AT substitution compared to an AT to GC substitution near crossover sites (locations at which exchange of genetic material between two chromosomes occur) in human sperm cells[27].

However, can point mutations affect GC content distribution at the whole genome level? Early analysis of the human genome suggested that the rate of mutation may be a key factor affecting GC distribution, as CpG sites in the human genome showed a negative correlation between the rate of neutral substitution (assumed to reflect the rate of point mutation) and GC content. In contrast, non-CpG sites showed a much weaker correlation between neutral substitution rate and GC content[31]. This may be explained by AT-bias caused by spontaneous demethylation – at non-CpG sites, cytosine bases are not methylated and therefore point mutations in this area should not be biased towards any of the four nucleotides, hence extremely weak correlation between point mutation rate and GC content. With the advance of DNA sequencing technology in recent years, it has finally become possible to find and characterize DNMs in animals using family trio studies. Smith et al. performed a comprehensive test of whether DNM could actually explain genome-wide GC content variance using three independent human trio datasets, and they concluded that mutation rate variance has very little impact on GC variance[32]. Overall, they report a very weak correlation between the rate of DNM and GC content, which may be caused by discrepancies between the three datasets analysed – separate analysis of each dataset showed a positive correlation between DNM and GC content in one dataset while the second dataset had a negative correlation[32]. Smith et al. suggested that this may be caused by differences in sequencing technology and/or analysis pipeline used to identify the DNMs in

each study[32]. With the development of sequencing techniques and analysis methods, the impact of mutation bias on GC content distribution remains to be further analysed.

1.3 The effect of selection on GC content

It has also been argued that selection may be a force behind GC content variation between species and within the genome. First, early studies on GC content reported wide distribution of GC% throughout the genomes of endotherms (birds and mammals) while exotherms (fish and snakes) seemed to have relatively stable GC% throughout the genome. These observations sparked the hypothesis that high GC content was an adaptation to higher body temperatures[17]. However, this idea is now out of favour following the report of high GC% variance crocodile and alligator genomes, along with bimodal GC3% distribution observed in some plant species such as corn and rice[18,20]. To the best of my knowledge, there are no widely accepted hypotheses that link natural selection and genomic GC content in animals and plants. For bacteria, experiments report a consistent average genomic GC content in bacteria sampled from similar environments, with closely related bacterial species in different environments exhibiting larger than expected GC content differences compared to unrelated species in the same environment[33]. It has been accepted that bacterial genomic GC content is under selection by environmental factors, although the exact factor remains unknown. Supported by newly published archaea genomes, temperature was proposed as the key factor affecting microbe genomic GC content. However, later studies found no correlation between genomic GC content and optimal growth temperature for archaea – instead, high concentrations of potassium ions were proposed to be utilized by archaea to maintain genome stability under high temperature, not increased GC content[34].

At the same time, many theories have proposed that selection is a key factor behind GC content variation within the genome. Analysis of gene function and GC content in both plant and animal species has shown correlation between housekeeping function and high GC content[35–37]. There are also strong correlations between high GC at the third codon position (GC3), low intron numbers and short protein product length[36,38]. At the same time, human genes are clustered in GC-rich regions and GC-rich regions in yeast have a more active chromatin conformation compared to GC-poor regions[39,40]. These correlations seemed to suggest that having higher GC% was advantageous for gene expression, and one hypothesis proposed was that having different GC% for genes with different expression profiles would allow the cell to synchronize codon usage and allocate tRNA resources[41]. For example, fungi genes with similar expression profiles had similar levels of selection pressure on codon usage[41]. However, evidence showing no significant correlation between actual gene expression levels and GC content in mammals cast doubt on this hypothesis[42]. In addition, recent work on synonymous codon usage in mammals showed that codon usage in genes expressed under similar conditions did not necessarily match the tRNA pool of the cells they are expressed in, casting doubt on whether codon usage was selected to match tRNA availability[43]. Despite this, there are potential advantages for high GC content, for example, it can increase the possibility of DNA methylation occurring in the area and can allow the cell to better fine-tune gene expression[44]. This is supported by analysis of plant genomes reporting strong correlation between high GC3 genes and regulation via DNA methylation[18]. In addition, analysis of human, *Drosophila* and

C. elegans genomes showed that exons have higher GC content than introns, which may be correlated to nucleosome modulation and chromatin packaging[45].

In summary, natural selection may be an important factor affecting GC content in the genome, and higher GC content in coding regions may have been selected for due to potential advantages related to gene regulation. However, more evidence supporting this hypothesis is needed, and the impact of selection on genome-wide GC content still requires further analysis.

1.4 The effect of biased gene conversion (gBGC) on GC content

Both Smith et al. and Pouyet et al. support an alternative mechanism other than mutation bias and selection as the main player affecting genome GC distribution: biased gene conversion (gBGC)[32,43]. Since 2000, increasing amounts of literature have been published in support of gBGC as the major force behind genome level GC variation.

1.4.1 Mechanisms of recombination and gBGC

gBGC occurs during recombination, a process during which DNA exchange occurs between homologous chromosomes[46]. gBGC has been observed in mitotic recombination involving ectopic or nonallelic homologous regions[47] and in meiotic recombination, a process essential for proper segregation of chromosomes and genetic variation[48]. Recombination is initiated by double strand breaks (DSB) which cause DNA to unwind and recruit DNA repair enzymes that remove several bases in the 5' direction, leaving single strand DNA near the DSB site[49] (Fig. 1.2). One single strand locates the homologous chromosome and 'invades' the matching strands on the homologous chromosome, forming a structure named the displacement-loop (D-loop)[49] (Fig. 1.2). If the second strand is not captured, synthesis-dependent strand-annealing occurs and the outcome is always a 'noncrossover' event[49] (Fig. 1.2). However, if the second single strand is captured by the D-loop, double Holliday junctions are formed and both 'crossovers' and noncrossovers are possible in this situation[49] (Fig. 1.2). Crossovers are defined by reciprocal DNA exchange between the two homologous chromosomes, while noncrossovers describe situations where reciprocal DNA exchange does not take place and only one chromosome exits the process with altered genetic information[49,50] (Fig. 1.2).

gBGC occurs when recombination happens between homologous chromosomes with heterozygous sites containing adenosine (A) or thymine (T) on one chromosome and cytosine (C) or guanine (G) on the second chromosome[46]. Mismatches occur when one DNA strand (e.g. carrying A) invades the other DNA double helix (e.g. carrying GC) (Fig. 1.2), triggering DNA mismatch repair mechanisms that 'choose' which strand to use as the repair template. Gene conversion occurs when one strand (known as the donor strand) is used as a template to repair the second strand (the reciprocal strand), resulting in replacement of a region of the reciprocal DNA strand with a homologous region on the donor DNA strand[51]. gBGC, as the name implies, describes a biased gene conversion process where the strand containing G or C is preferentially selected as the donor strand, resulting in replacement of the A or T nucleotide on the reciprocal strand with a G or C nucleotide to form a nucleotide pair with the C or G nucleotide on the donor strand[46] (Fig. 1.2). gBGC can happen during both crossovers and noncrossovers[52].

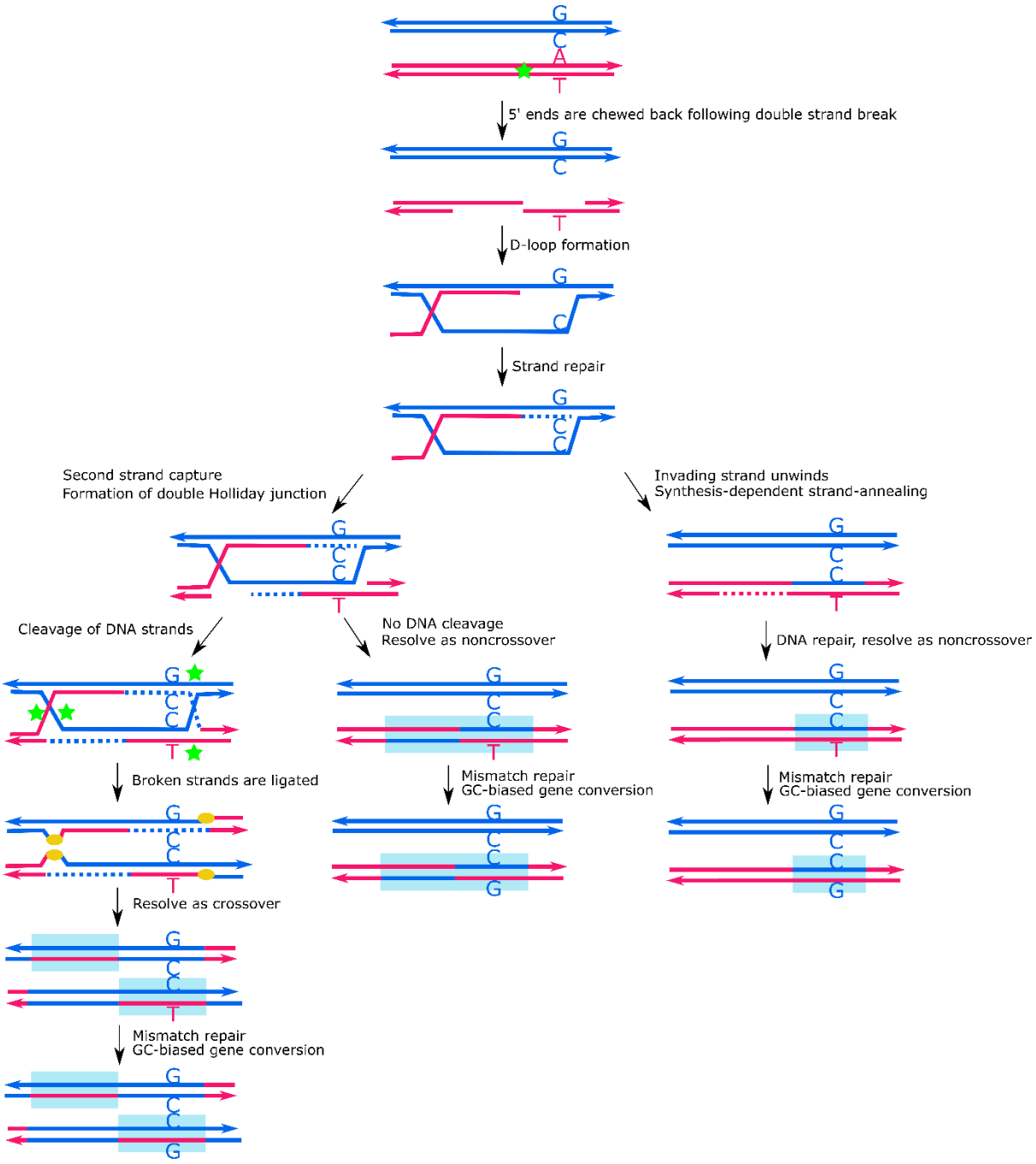


Fig. 1.2 Mechanisms of recombination and GC-biased gene conversion. The two homologous DNA chromosomes are depicted in different colours and nucleotides at one heterozygous site is shown. Three models have been proposed to explain the different outcomes of recombination. In brief, a double strand break initiates the recombination process and DNA repair mechanisms form a D-loop which can be resolved by either synthesis-dependent strand-annealing (SDSA) or formation of a double Holliday junction. The two different pathways have different outcomes, with SDSA always resolving as a noncrossover and double Holliday junctions either resolving as a crossover or a noncrossover. Regions that exit recombination with strand exchange are highlighted in a blue rectangle. gBGC can happen in all three pathways. This figure is modified from Duret & Galtier Figure 1[46].

1.4.2 Direct evidence supporting presence of gBGC

One of the first papers that provided direct evidence of gBGC occurrence identified crossover and noncrossover events in yeast spores and measured the frequency of AT to GC change near these sites[52]. The authors reported presence of GC-bias in yeast at both crossover and noncrossover sites, with GC bases having 1.3% higher occurrence than the expected value (null hypothesis assumes no bias)[52]. For humans, analysis of crossover and noncrossover events in the male germline was performed by Odenthal-Hesse et al. in their study focusing on six identified autosomal recombination hotspots, and the authors reported significant GC bias at noncrossover sites but not crossover sites (71% chance of GC allele transmission at noncrossover sites compared to 51% chance of GC allele transmission at crossovers)[53]. Recently, with the improvement of sequencing technology and decreases in sequencing costs, deep sequencing of multiple human genomes became possible. With the abundance of high coverage human genomic data, comparison of parent and child genome sequences has been made possible and researchers have been able to identify crossover and noncrossover events with higher resolution. For example, Williams et al. analysed 34 three-generation pedigrees and observed strong GC bias at noncrossover sites with GC alleles being selected over AT alleles 68% of the time[54]. Similarly, Halldorsson et al. analysed three-generation pedigrees from Iceland and reported an average GC bias of 67.6% at noncrossover sites with slightly higher GC bias in females compared to males (69.1% in females compared to 64.4% in males)[55]. At the same time, mice intercrossing followed by deep sequencing of more than 100 offspring has led to the detection of thousands of crossover and noncrossover events across the mouse genome and an estimate of 68% of noncrossover events converting an AT nucleotide pair to a GC pair[56]. In conclusion, these studies provide direct evidence of gBGC presence in yeast, human and mouse, with human and mouse possessing very similar rates of gBGC.

1.4.3 Indirect evidence supporting presence of gBGC

Although a limited number of studies have presented direct evidence of gBGC occurring in yeast, mice and humans, multiple lines of indirect evidence support the widespread presence of gBGC in all eukaryotes. In species in which the outcome of recombination has not been investigated at the nucleotide level, gBGC occurrence can be indirectly detected by testing the correlation between GC content and recombination rate. In species that have active gBGC, GC content is expected to have a strong positive correlation with recombination rate as opposed to a negative correlation which is predicted to arise from AT-biased base substitutions near recombination sites (the process of recombination is also highly mutagenic)[27,46]. Even in species that do not have recombination rate measurements, Pessia et al. demonstrate that the presence of gBGC in non-vertebrate eukaryotes can be estimated[57]. In their paper, they use chromosome length as an indicator of likely recombination rate, and they report gBGC in 12 species including the mushroom *Coprinopsis cinerea*, the slime mold *Dictyostelium discoideum*, the parasite *Leishmania major* and the unicellular ciliate *Paramecium tetraurelia*[57]. The authors conclude that as chromosome size is only a rough proxy for recombination rate, such that they did not find a significant correlation between mouse local GC content and chromosome length, the strong correlation they find for the eukaryotes reported here indicate widespread presence of gBGC throughout all eukaryotes[57].

1.4.4 The impact of gBGC on GC content and genome evolution

gBGC requires recombination to take place, and there are known recombination hotspots distributed throughout the genome, creating a potential for some regions to have higher chances of undergoing gBGC than other regions and therefore accumulating higher GC levels in these regions. However, the exact extent that gBGC impacts genome GC content also depends on factors such as the amount of GC in the genome of the most recent common ancestor[58,59]. Overall, the bias causing gBGC does not need to be extremely strong to result in the observed effects on GC distribution within the genome. Even though studies show there is only a slight tendency towards using the GC-allele to resolve the mismatch (e.g. estimated 50.6% in yeast), calculations considering effective population size have predicted gBGC to be sufficient to cause the level GC variation observed in the species mentioned above, with gBGC having stronger effect with larger population size[52,56,60,61].

It remains to be answered whether there is an advantage to having gBGC in the genome, or whether it is simply remnant of a biased DNA repair system that arose early in eukaryote evolution. Potential advantages of GC base pairs over AT have been discussed in the previous section, and one additional explanation for the existence of gBGC has been proposed to be counteraction of AT-biased mutation occurring during recombination[27]. In conclusion, although the exact biological mechanisms behind gBGC remains unclear, it has been established that gBGC is a key factor influencing GC content distribution in eukaryotes.

2. Sand rat phylogeny, physiology, and GC composition bias

2.1 Phylogeny

The fat sand rat (*Psammomys obesus*) is a rodent belonging to the Gerbillinae subfamily and is widely distributed in desert areas around North Africa and the Middle East[62]. To estimate the phylogenetic relationship between Gerbillinae and other rodents, and to further resolve phylogenetic relationship within the gerbil subfamily, Chevret and Dobigny have sequenced gerbil 12S rRNA and cytochrome *b* genes and used gene alignments to estimate the phylogenetic analysis and divergence time estimation[63]. Their results suggest that Gerbillinae may have split from their closest sister group, the Acomyinae subfamily (sometimes referred to as the Deomyinae subfamily, contains species such as the spiny mouse (*Acomys cahirinus*)), around 16 – 18 Myr ago, and the Gerbillinae+Acomyinae subfamily is estimated to have split from the Murinae subfamily (contains the mice and rats) 18 – 22 Myr ago[63] (Fig. 1.3A). In addition, the Gerbillinae subfamily is estimated to include 14 genera, with the sand rat closely related to the Mongolian jird (*Meriones unguiculatus*) (Fig. 1.3B)[63].

2.2 Physiology

2.2.1 Insulin-regulated blood glucose homeostasis

Maintaining a stable level of blood glucose is essential for survival, and in humans under fasting conditions, blood glucose levels are kept within a narrow range of 5.6–6.9 mmol/L[64]. A key factor contributing to blood glucose homeostasis is insulin, a peptide hormone consisting of 51 amino acids[65]. Insulin is produced by pancreatic β-cells, specialized cells located in the

pancreas that produces and secretes insulin in response to changes in blood glucose concentration (Fig 1.4). Aside from glucose, changes in blood fatty acid levels and some amino acids can also induce changes in β-cell insulin production and secretion[65]. In this section, I focus on regulation of β-cell insulin production and secretion mediated by glucose.

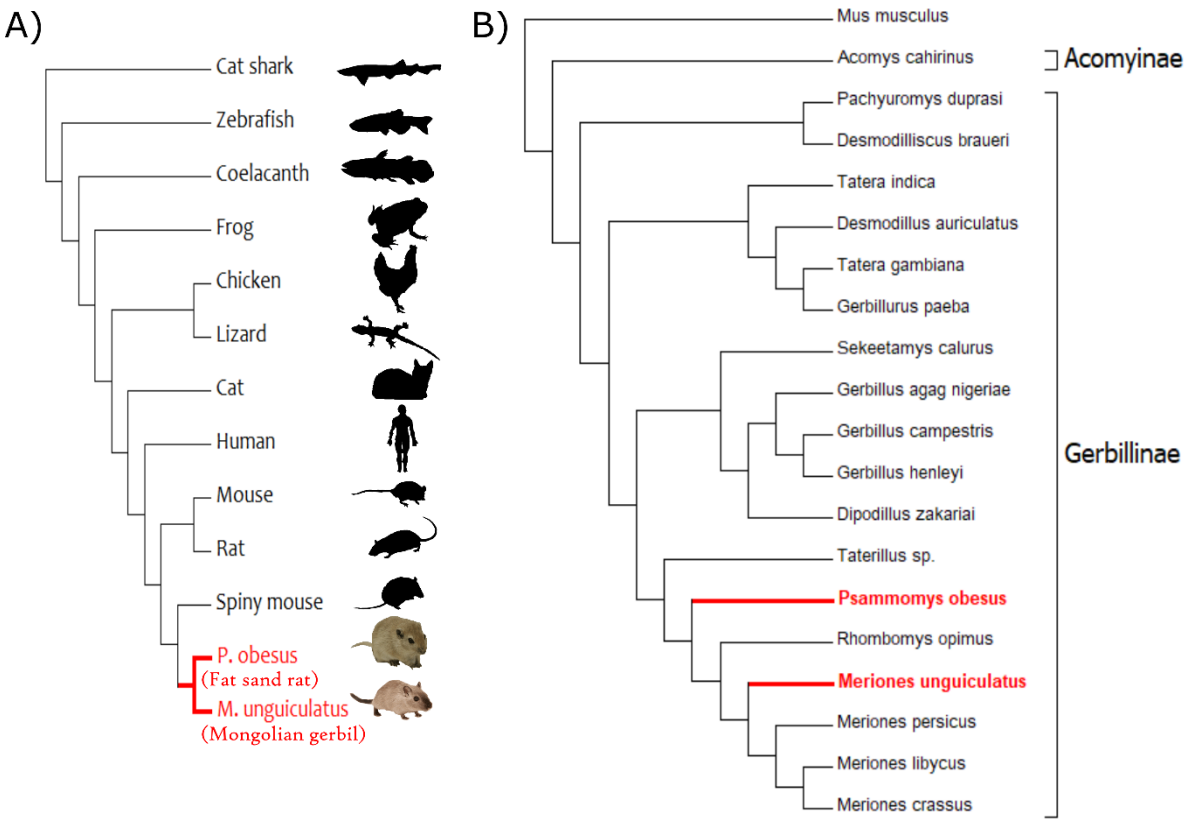


Fig. 1.3 Sand rat phylogenetic position. A) shows the likely phylogenetic position of the fat sand rat and a close relative, the Mongolian jird, compared to other vertebrates. B) shows a more detailed position of the sand rat and Mongolian jird in the Gerbillinae subfamily, with branch positions modified from Chevret and Dobigny[63]. Branch lengths do not represent phylogenetic distance.

The transcription of insulin is strictly regulated in the β-cell, and comparison of vertebrate insulin promoter sequences has shown that all species investigated possess a number of regulatory sequences such as A, C, E, Z and CRE regulatory elements, all of which serve as binding sites allowing key β-cell transcription factors to control insulin gene expression. PDX1 is an example of a highly conserved β-cell transcription factor which binds to A elements in the insulin promoter and regulates insulin expression in response to changes in blood glucose levels. Two regulatory pathways have been proposed to explain the upregulation of insulin gene expression mediated by PDX1 following β-cell exposure to high glucose. The first pathway requires secretion of insulin, which binds to insulin receptors present on the β-cell membrane and triggers the insulin response pathway in β-cells, leading to PDX1 phosphorylation and altered interaction between PDX1 and its cofactors. The second pathway is triggered by glucose entering the β-cell and triggers activation of PDX1 kinase, leading to PDX1 phosphorylation. Phosphorylation of PDX1 has been proposed to promote interaction

with histone acetylase p300 to promote histone acetylation near the insulin promoter, thus stimulating insulin gene expression[66] (Fig. 1.4). Following gene expression, insulin mRNA is first translated into preproinsulin, a 110 amino acid peptide chain that needs to be modified in the endocytomic reticulum (ER) and Golgi apparatus to produce the mature insulin molecule consisting of two peptide chains linked by disulphide bonds[67].

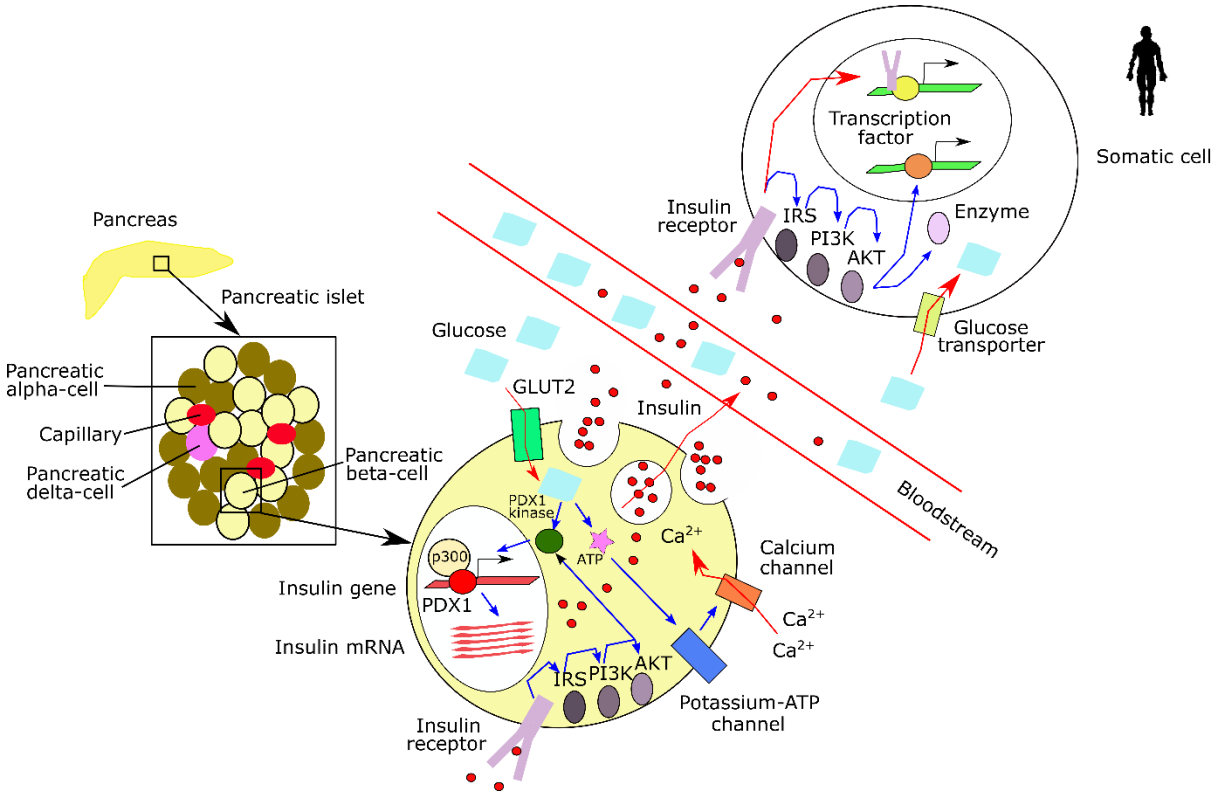


Fig 1.4 Insulin secretion in human pancreatic β-cells and insulin response in somatic cells. β-cells are clustered in pancreatic islets connected to a capillary network. Key components of the insulin production, secretion, and response pathways are shown, with red arrows indicating transport of molecules and blue arrows indicating biochemical reactions. When blood glucose levels rise, glucose can activate PDX1 kinase through a series of biochemical reactions or through the insulin receptor pathway. Phosphorylated transcription factor PDX1 interacts with histone acetylase p300 to upregulate insulin gene expression. Insulin secretion is stimulated by glucose molecules that enter β-cells via glucose transporter GLUT2 and promote ATP production, triggering potassium-ATP channel closure followed by calcium channel opening. Calcium ion influx leads to release of storage granules containing insulin, and insulin travels throughout the body via the bloodstream. Insulin binding to insulin receptors located on somatic cells can regulate cell metabolism by increasing glucose uptake. This can be achieved via a series of biochemical reactions initiated by insulin receptor resulting in altered enzyme activity and gene expression, or via insulin receptor internalization and direct regulation of gene expression. Mature insulin molecules are stored in storage granules in β-cells to await the signal for release; the average mouse β-cell contains approximately 13,000 insulin storage granules which take up approximately 10% of the total β-cell volume[68].

β-cells are clustered in pancreatic islets, structures that contain groups of pancreatic cells dedicated to releasing hormones such as glucagon (released by α-cells), insulin (released by β-cells) and somatostatin (released by δ-cells). To allow these hormone-releasing cells to quickly detect fluctuations in blood metabolite levels, pancreatic islets are connected to a large network of blood vessels passing through the pancreas[65] (Fig. 1.4). When blood glucose levels rise, β-cells quickly detect this change via an influx of glucose molecules that are shuttled into the cell by the highly affinitive glucose transporter GLUT2. Glucose in the β-cell is quickly converted to pyruvate and used to stimulate NADH and ATP production in the mitochondria[69,70] (Fig. 1.4). A surge in ATP production subsequently induces closure of ATP-sensitive potassium channels on the plasma membrane, leading to depolarization and Ca^{2+} influx, which is the key trigger for insulin exocytosis[71,72].

Once released into the bloodstream, insulin circulates through the body and binds to the insulin receptor (INSR), a receptor molecule that is located on the cell membrane of various somatic cell types, including brain neurons and certain types of T cells involved in immune responses[73] (Fig 1.4). Key tissues that mediate blood glucose homeostasis are liver, skeletal muscle, and white adipose tissue, and the *Insr* gene is expressed at high levels in these tissues[74]. INSR is a large transmembrane protein that consists of more than 1,500 amino acids that make up different protein domains. Key domains include the insulin binding domain, which is located outside the plasma membrane, and the tyrosine kinase domain, which is in the cytoplasm[75]. Once insulin molecules bind to INSR, the receptor triggers a series of reactions that alter cell metabolism. The first type of reaction involves phosphorylation of different insulin receptor substrates (IRS) by the tyrosine kinase domain of INSR, followed by translation of this signal into serine/threonine phosphorylation mediated by enzymes phosphoinositide 3-kinase (PI3K) and protein kinase B (PKB, also referred to as AKT)[76,77] (Fig 1.4). This can directly alter cell metabolism by changing enzyme activity, for example phosphorylation of glycogen phosphorylase can inactivate the enzyme and halt glycogen consumption in skeletal muscle[78]. This can also alter cell metabolism through controlling gene expression, for example phosphorylation of the transcription factor FOXO1 can upregulate genes contributing to glucose storage[76,77]. The second way that insulin moderates cell metabolism through INSR is by direct internalization of INSR (Fig 1.4). Through mechanisms not yet elucidated, insulin binding to INSR can stimulate transport of INSR to the nucleus and direct interaction of INSR with transcription enzyme RNA polymerase II and coregulator protein HCF-1[79]. This process can also alter gene expression and subsequently cause changes to cell metabolism.

2.2.2 Type 2 diabetes

Diabetes is a major chronic disease of extreme significance to global health. The defining symptom of diabetes is hyperglycemia, which describes abnormally elevated blood glucose levels before and after caloric intake[64]. Research into the pathology of diabetes has shown that dysregulation of blood glucose levels in patients is secondary to abnormalities in insulin regulation[64]. Depending on the cause of abnormal insulin regulation and the timing at which it is discovered, diabetes can be classified into different categories (e.g. type 1 diabetes, type 2 diabetes, gestational diabetes etc.). Here, I only discuss type 2 diabetes (T2D), diabetes caused by progressive decline of β-cell insulin secretion and often accompanied by insulin resistance

in tissues such as skeletal muscle, liver, and adipose tissue[64]. Insulin resistance is defined as insufficient response to insulin stimulation, and a patient is diagnosed as insulin resistant if higher than normal levels of circulating insulin is required to induce somatic cell uptake of glucose. Unresponsive somatic cells increase the risk of T2D by inefficiently taking up excess glucose, leading to chronic high blood glucose which constantly stimulates β-cell insulin production and secretion, eventually resulting in cell stress and apoptosis[80].

Aside from insulin resistance, many factors have been associated with increased risk of T2D, for example genetic factors such as *PDX1* gene variants have been linked to decreased insulin gene expression and insufficient insulin secretion[81] and other factors such as obesity have been identified as significant contributors to T2D disease progression[82]. Some of these factors have been connected to insufficient insulin secretion, the defining symptom of T2D, and other factors linked to insulin resistance, a symptom often found in T2D cases. Aside from directly affecting previously mentioned components of the insulin expression and secretion pathway, multiple factors can cause insufficient insulin secretion by lowering the number of healthy pancreatic β-cells[83]. A decrease in β-cell number can be caused by two factors: insufficient β-cell proliferation or increased β-cell death[84]. Aside from physical trauma (e.g. surgical removal of pancreatic tissue), insufficient expression of the transcription factor PDX1 has been linked to increased susceptibility of the β-cell to oxidative stress caused by insulin production and decreased β-cell proliferative capability[85], and obesity has been connected to an increase in the level of circulating non-esterified fatty acids which are detrimental for β-cell survival[83,86]. Many factors are also linked to increased risk of insulin resistance, a symptom that can contribute to T2D development. Insulin resistance can be caused by genetic factors that affect components of the insulin signalling pathway, for example mutations in the human *INSR* gene have been linked to defective insulin binding and loss of insulin receptor tyrosine kinase action, and mutations affecting the downstream protein kinase AKT2 causes insulin resistance in both mouse models and human patients[87–89]. At the same time, other factors such as puberty, pregnancy, aging and obesity can also affect insulin sensitivity and can lead to temporary or permanent insulin resistance[83,90,91].

In conclusion, T2D is defined by progressive decrease of insulin secretion, and various genetic and environmental factors increase T2D risk by directly affecting β-cell function, proliferation, and apoptosis. In addition, factors that cause insulin resistance indirectly increase T2D risk by causing chronic hyperglycemia leading to β-cell stress.

2.2.3 The sand rat and other gerbils as a type 2 diabetes model

As a desert animal, the sand rat mainly feeds on succulent, salt-rich plants from the genus *Salicornia*, which contain salt levels that are higher than that of sea water[92]. They were first reported to develop T2D by Schmidt-Nielsen et al., who brought adult sand rats captured in Egypt into the laboratory[92]. 12 lab-reared sand rats fed standard laboratory chow supplemented with vegetables exhibited elevated urine glucose within 2 months, with 11 out of 12 sand rats showing marked pancreatic β-cell abnormalities after 6 – 7 months[92]. In contrast, the 5 sand rats on a vegetable only diet did not exhibit such symptoms, and maintained an average body weight half of that of the chow-eating sand rats[92]. Schmidt-Nielsen et al. also confirmed that diabetic symptoms were not observed in 40 wild sand rats which had their blood and urine

glucose levels measured at the trapping site[92]. From this point on, sand rats began to be used as a non-transgenic animal model of diet-induced T2D, with multiple other groups reporting a varying percentages of wild-trapped or lab-reared individuals exhibiting symptoms of T2D after being fed a laboratory chow diet[93,94].

The progression of T2D has been well-characterized in sand rats[95]. Different sand rat individuals have varying levels of resistance to T2D when maintained on the same diet, and this phenotype has been selected for to establish two inbred colonies named diabetes prone (DP) and diabetes resistant (DR)[92,93,96]. However, even members of the same colony can progress at different speeds to T2D. Kalderon et al. classified 109 DP sand rats fed on a chow and salt bush plant into four groups with distinct physiological characteristics[96]. After being fed the chow diet for the same period of time, 32% of the animals maintained normal blood glucose, blood insulin levels and body weight (group A), while 29% of the animals showed elevated insulin and body weight but not glucose (group B), 36% had elevated blood glucose and insulin but maintained normal body weight (group C), and only 6% showed elevated blood glucose but normal levels of insulin, along with severe obesity (group D)[96]. Further analysis showed that these different groups represented different stages of T2D development, with group A representing healthy individuals, group B showing insulin resistance, group C showing insulin resistance along with dysregulation of blood glucose, and group D representing individuals with insulin resistance, blood glucose dysregulation and β-cell failure[97]. Bar-On et al. also showed that as long as a strict ration of food is implemented, group C animals could have their symptoms reverted, while group D animals could not be saved due to severe irreversible loss of β-cells[97].

In addition, DP sand rat individuals have a predisposition to diet-induced T2D that can be observed prior to high calorie diet consumption: an innate insulin resistance and inability of β-cells to produce sufficient levels of insulin[98,99]. First of all, insulin resistance is not only observed in diabetic sand rats, but also in non-diabetic, low energy consuming sand rats[98]. In contrast to laboratory rats (*Rattus norvegicus*) which respond strongly to intravenous injection of human insulin, healthy DP sand rats do not show any significant decrease in blood glucose and in healthy DR sand rats the decrease is moderate[98]. In addition, by using the metabolite malonyl-CoA as an indicator of "shifting gear" between glucose and fatty acid metabolism, skeletal muscle cells of healthy DP sand rats show no response to human insulin while DR sand rats and albino rats respond rapidly[98]. Not only are DP sand rats likely to be inherently insulin resistant, they cannot produce enough insulin to meet the demands required to metabolize a high energy diet, and this effect is not due to β-cell apoptosis in the first two stages of disease progression[99]. Kaiser et al. report a decrease in sand rat β-cell mass within the first 2 days of starting a high energy diet, followed by a subsequent rise in β-cell mass up until the end stage of T2D progression[99]. This implies that DP sand rats have issues with maintaining sufficient insulin production under hyperglycaemic conditions, and this problem is not caused by decreased β-cell proliferation[99].

The sand rat is also not the only member of the Gerbillinae subfamily that exhibits a predisposition to developing diet-induced T2D. Boquist reported spontaneous occurrence of obesity in lab reared *Meriones unguiculatus* (Mongolian jird) as early as 1972, with

approximately 10% of the animals maintained on standard lab chow observed to be obese by around 9 months of age[100]. Mongolian jird colonies have also been selected and bred for their susceptibility to T2D, and an inbred line of spontaneously obese, insulin resistant and leptin resistant Mongolian jird has been established by Li et al.[101].

2.3 GC composition bias: the sand rat genome has extremely GC-rich regions

Whole-genome sequencing of the sand rat genome has revealed the presence of an extremely GC-rich region that is estimated to be syntenic to approximately 10 Mb of the rat genome[102] (Fig. 1.5). Comparison of the initial sand rat genome assembly to other mammalian genomes showed that 88 genes were missing as one continuous gene region[102]. Some of these genes are considered crucial for normal vertebrate development and physiology, including the *Pdx1* gene and *Insr* gene mentioned in the previous section. Transcriptome sequencing of liver, pancreatic and duodenum tissue revealed the presence of some "missing" genes and showed these genes were abnormally high in GC content which could explain their absence from the initial assembly which was constructed using Illumina short read sequencing[102]. Their presence in the genome was then proven through isolation of GC-rich genomic DNA using a caesium-chloride gradient and ultracentrifugation, followed by Illumina sequencing of only GC-rich DNA[102]. In conclusion, both genomic and transcriptomic data indicates the presence of a genomic region influenced by strong GC composition bias in the sand rat genome[102].

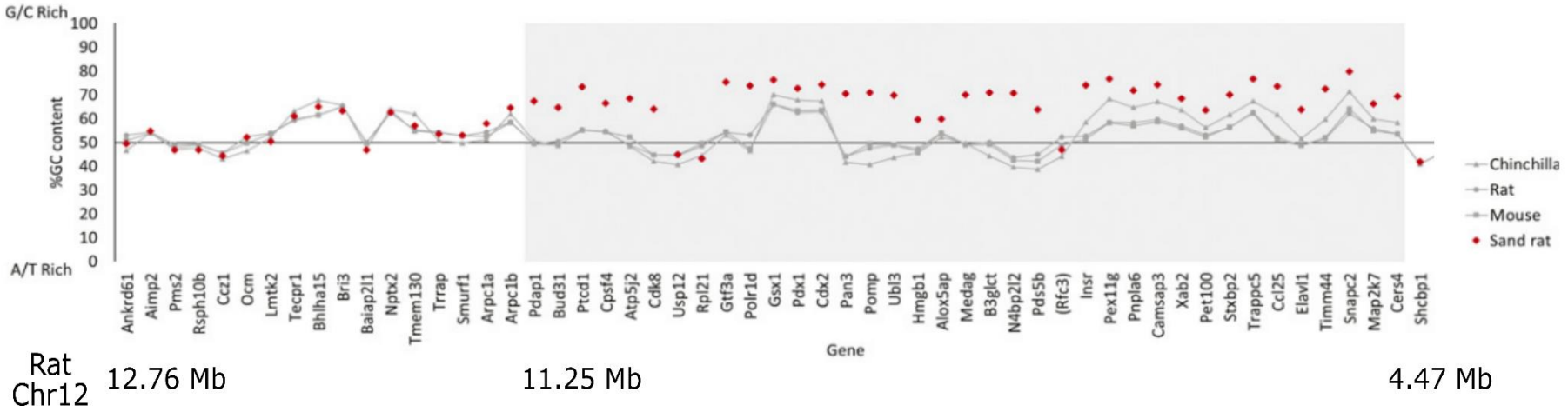


Fig. 1.5 Sand rat genes located in an extreme GC-rich region. This figure is derived from Hargreaves et al.[102]. GC-rich genes fall within the grey box, and sand rat genes are shown as red dots while their rat, mouse, and chinchilla orthologs are shown as grey dots. The gene order shown here is derived from the rat chromosome assembly, and all genes shown here are located on rat chromosome 12 (except *Rfc3* and *Shcbp1*).

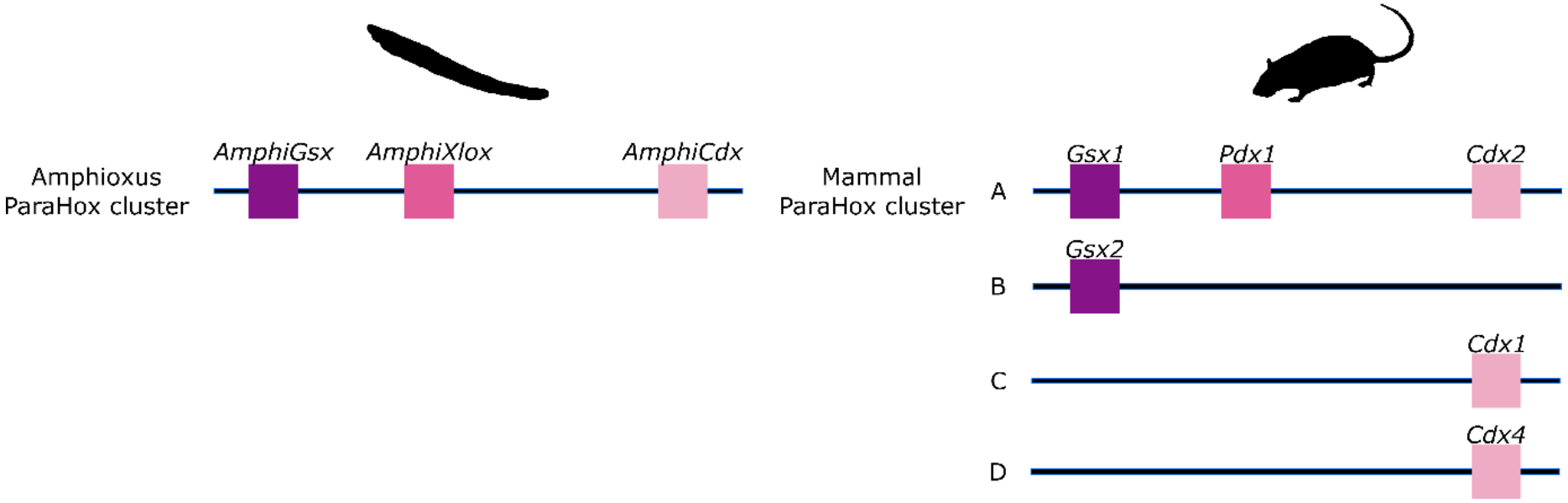


Fig. 1.6 Comparison of the ParaHox cluster in amphioxus and mammals. Due to two rounds of whole genome duplication in the vertebrate lineage, mammals have four ParaHox clusters in their genome. However, only one remains intact with all three ParaHox genes present in the same orientation as the amphioxus ParaHox cluster.

3. Overview of pancreatic duodenum homeobox 1 (*Pdx1*)

3.1 The ParaHox gene cluster: *Gsx1*, *Pdx1*, *Cdx2*

Homeobox genes are genes that encode a protein with a 60 amino acid structure named the homeodomain, which consists of three helix structures allowing the protein to recognize and bind specific DNA sequences[103]. Most homeodomain-containing proteins function as transcriptional factors, and the Hox gene cluster is one of the most well-known gene clusters made up of homeobox genes that are essential for control of animal body plan development[104]. Work on a cephalochordate, the amphioxus *Branchiostoma florida*, was the first to show that three additional homeobox genes form the ParaHox cluster, a likely evolutionary sister gene cluster to the Hox cluster[105]. Like the Hox genes, ParaHox genes are expressed during amphioxus development and the order in which they are located in corresponds to the tissue in which they are expressed in during development (spatial co-linearity): *AmphiGsx* is expressed in the area equivalent to part of the vertebrate brain, *AmphiXlox* in the central gut area, and *AmphiCdx* in the posterior area[105]. Analysis of the human and mouse genomes show that the vertebrate orthologs of these genes, *Gsx*, *Pdx1* and *Cdx* also form gene clusters that have experienced two rounds of duplication (along with the rest of the genome), which after gene loss resulted in only one intact cluster containing all three genes in mammals[106] (Fig. 1.6). Like amphioxus, the two mammalian *Gsx* genes are expressed in the brain, with *Gsx1* shown to be essential for pituitary gland development and *Gsx2* involved in both forebrain and hindbrain development. Mammalian *Pdx1* is expressed in the pancreas and duodenum, and the three mammalian *Cdx* genes, *Cdx1*, *Cdx2* and *Cdx4*, are expressed in the developing intestine[107–113].

3.2 *Pdx1* initiates vertebrate pancreatic development

Pdx1 stands for pancreatic duodenum homeobox 1, a name which has now replaced the historical use of *XlHbox 8*, *Ipf1*, *Idx1*, *Stf1*, *IUF1*, and *Gsf*. Wright et al. first reported expression of a gene named *XlHbox 8* (*Xenopus laevis* homeobox 8) in the endoderm of early *Xenopus* embryos and this expression became restricted to regions of the anterior digestive tract in older embryos[109]. In the same year, isolation of protein factors binding to the human insulin promoter lead to identification of a protein factor named IUF1 (Insulin Upstream Factor 1), which was later discovered to consist of two protein components of different sizes[114,115]. Similar experiments carried out in rat and human pancreatic islets revealed a single protein factor binding at the same site as IUF1, and the first group to report this named this factor GSF (Glucose sensitive factor) given its sensitivity to glucose[116,117]. Two years after discovery of *XlHbox 8*, Ohlsson et al. reported expression of a mouse gene in the developing dorsal and ventral pancreatic buds and named it *Ipf1* (Insulin promoter factor 1), while a separate group described the same mouse orthologue expressed in islet cells and duodenum tissue and named it *Idx1* (Islet/duodenum homeobox 1)[118,119]. In rats, the orthologue was named *Stf1* (Somatostatin transactivating factor 1) given the protein's ability to activate somatostatin expression both *in vivo* and *in vitro*[120]. In 1996, *Pdx1* officially replaced all previous nomenclature, and it was named based on both its expression pattern in the developing embryo and the phenotype exhibited by animal models[121].

The mouse PDX1 protein is a 284 amino acid homeodomain protein that functions as a transcription factor. The initiation of *Pdx1* expression coincides with the initiation of pancreas bud development induced by suppression of endodermal sonic hedgehog (SHH) signalling by notochord factors activin-βB and FGF2[122,123]. *Pdx1* expression is detectable at E8.5 in the mouse embryo, which corresponds to approximately 4 weeks in the human embryo[122,124]. As a master regulator of pancreatic development, *Pdx1* initiates the expression of multiple downstream genes that go on to further coordinate the proliferation of pancreatic cells and the differentiation of acinar cells (cells that secrete digestive enzymes) and endocrine cells (cells that secrete hormones such as insulin)[125,126]. When pancreatic development is complete (E17.5 in the mouse), *Pdx1* expression in the pancreas becomes confined to β-cells[122,125].

The expression pattern of *Pdx1* has indicated its importance in both pancreatic development and endocrine cell differentiation. Functional analysis performed by Jonsson et al. was the first to show that PDX1 was necessary for initiation of pancreatic development, as *Pdx1*$^{-/-}$ mice died shortly after birth due to lack of a pancreas, and Offield et al. independently confirmed these findings with two different mice models completely lacking the *Pdx1* homeobox sequence[110,121]. In contrast, mice with only one functional copy of *Pdx1* managed to develop a pancreas and showed no obvious abnormalities after birth, showing that only one functional copy of *Pdx1* is sufficient for pancreas development[121]. However, a recent study by Zhu et al. shows that *Pdx1*$^{+/-}$ human embryonic stem cells (ESCs) produce 65% less insulin positive cells when induced to differentiate, suggesting that both copies of *Pdx1* may be required for proper endocrine cell differentiation during development[125].

3.3 *Pdx1* regulates vertebrate adult β-cell function

Pdx1 continues to be expressed in adulthood, however its expression in the pancreas becomes restricted to β-cells[97]. It still remains essential for maintaining normal metabolism, as shown using *Pdx1*$^{+/-}$ mouse models and mature β-cell-specific *Pdx1* deletion models which both exhibit glucose intolerance accompanied with insufficient insulin secretion[127–129].

As a master regulator of the adult β-cell, PDX1 maintains normal β-cell function in many ways. First of all, *Pdx1* is essential for β-cell identity, as shown by experiments forcing expression of *Pdx1* in perinatal α-cells to create a β-cell-like transcriptional profile, and experiments knocking out both *Pdx1* alleles in mature β-cells resulting in transition of transcriptional activity towards an α-cell-like state[128,130]. PDX1 maintains β-cell identity mainly through activating expression of itself, insulin and the β-cell specific transcription factor MAFA, while at the same time repressing α-cell identity through suppressing expression of glucagon and the α-cell specific transcription factor MAFB[128,131].

Secondly, PDX1 is essential for insulin production. PDX1 regulation of insulin expression is perhaps the best studied PDX1 function in adult β-cells, and PDX1 has been shown to directly bind to the insulin promoter in human, rat, and mouse islet cells[132–134]. PDX1 recognizes A elements, AT-rich regions that contain a core TAAT sequence (the exception is the A2 element which contains AAAT), and the most important PDX1 binding sites in the insulin promoter are A3 and A4[135]. In addition, PDX1 can upregulate or downregulate insulin expression depending on glucose concentration in the surrounding environment[114,136]. Early experiments using

cultured cells showed that glucose concentration could affect the phosphorylation and dephosphorylation of PDX1, resulting in changes to its binding activity to the insulin promoter[136]. Later, it was shown that phosphorylated/dephosphorylated PDX1 interacted with different cofactors to form different protein complexes that either strongly activated insulin expression or suppressed insulin expression[137,138].

Thirdly, *Pdx1*$^{+/-}$ mouse models have shown that PDX1 is also essential for insulin secretion in response to glucose[129]. PDX1 regulates expression of *Slc2a2*, a gene that encodes the key glucose transporter GLUT2 which facilitates glucose entrance into the β-cell[129]. In addition, PDX1 directly controls expression of a key mitochondrial regulator, mitochondrial transcription factor A (TFAM) and is essential in ensuring the mitochondria can produce sufficient levels of ATP to trigger closure of potassium-ATP channels[129,139].

Finally, PDX1 is also essential for β-cell maintenance. Mass production of insulin can result in ER stress, and if not quickly alleviated, can result in β-cell apotosis[140]. *Pdx1*$^{+/-}$ mice show signs of ER stress and increased β-cell apoptosis, and T2D patients with insufficient *Pdx1* expression also show ER stress and β-cell apoptosis[127,140]. This is likely due to PDX1-mediated transcriptional regulation of key ER function related enzymes, for example PDIA6, a key mediator of disulphide bond formation, and key apoptosis-related genes *Noxa* (also known as *PMAIP1* in humans) and *Puma* (also known as *BBC3* in humans)[141,142].

3.4 *PDX1* variants can cause metabolic disease in humans

The importance of *PDX1* in humans can be inferred from reports of *PDX1* variants correlated to maturity-onset diabetes of the young (MODY), a type of diabetes caused by mutations in a single gene with patients diagnosed before the age of 25[143,144]. Currently, there are a total of 13 subtypes of MODY, each caused by mutations in different genes[145]. MODY4 refers to MODY caused by *PDX1* variants, which occurs at a rare (<1%) prevalence out of all MODY cases and can be due to truncation or amino acid substitution in any region of the PDX1 protein[145].

The most serious consequences are all linked to mutations that affect the PDX1 homeodomain. *PDX1* variant Pro63fsdelC causes loss of the entire homeodomain sequence due to a frameshift mutation and leads to premature truncation of the protein[146]. The patient carrying the Pro63fsdelC variant on both alleles was born without a pancreas, while both parents heterozygous for this *PDX1* variant did not exhibit this phenotype[146]. At the same time, point mutations affecting single amino acids in the homeodomain can lead to equally drastic results. For example, PDX1 variant E164K has an altered amino acid in helix 1 of the homeodomain resulting in drastic decrease of protein half-life, and PDX1 variant E178K with an altered amino acid in helix 2 also has a much shorter half-life[147]. Individuals heterozygous for these two variants may be at risk of developing glucose intolerance as they age, or be at risk for gestational diabetes, while the individual carrying the E178K variant on one allele and E164K on the second was born without a pancreas[147]. The only reported case of a mutation affecting the homeodomain but not causing pancreas agenesis is PDX1 variant R197H, which has an altered amino acid in the third helix of the homeodomain[81]. This mutation led to drastic (<50%) decreases in mutant PDX1 binding affinity to the human insulin promoter and a <50%

reduction in the ability to activate insulin expression under high glucose conditions[81]. This point mutation is highly penetrant and is known to cause MODY4 in heterozygous individuals[81].

In contrast, known PDX1 variants C18R, P33T and D76N have less drastic consequences as they affect the N-terminal transactivation domain[81,148]. Functional assays showed a moderate decline (70 – 80%) in C18R, P33T and D76N PDX1 binding affinity to the human insulin promoter with correlation to the MODY4 phenotype, along with mis-regulation of key genes involved in β-cell development and function[149]. However, all mutations did not exhibit high levels of penetrance as there were individuals between 25 – 30 years of age carrying these mutations and did not exhibit any diabetic symptoms at the time of the study[81,148]. Given the serious consequences of mutations affecting the homeodomain, the most common types of missense PDX1 variants discovered in clinical studies are C18 and P33T[149].

4. Project aims and contributions

The total amount of AT pairs does not necessarily equal that of GC pairs, and the ratio of AT:GC pairs is not always evenly distributed throughout the genome. So far, all eukaryotic genomes show a certain degree of GC heterogeneity. Mutation bias, selection and gBGC are all forces that could potentially contribute to this heterogeneity and affect the outcome of evolution. The sand rat provides an interesting case of GC-skew in a localized region (or regions) of the sand rat genome that affects an ultra-conserved, highly important vertebrate gene. In this thesis, I define "GC-skew" as biased accumulation of GC base pairs over AT base pairs leading to increased genomic GC content, which differs from "GC-skew" used by some papers to describe asymmetric distribution of guanine and cytosine between complementary DNA strands, a phenomenon proposed to be associated with transcription stability[150]. Overall, this project aims to use the sand rat as a model for exploring possible conflict between GC-biased changes in nucleotide composition and natural selection, with focus on the *Pdx1* gene.

In Chapter 2, I ask if other members of the Rodentia order possess a divergent *Pdx1* gene like that observed in sand rats. I search for *Pdx1* in members of the Rodentia order and I show that the sand rat and two close relatives, the Mongolian gerbil and the Libyan jird (*Meriones libycus*) have highly divergent *Pdx1* genes that are extremely different from all other rodent *Pdx1* genes investigated. In this chapter, I downloaded sequencing data from NCBI and assembled multiple rodent genomes and transcriptomes to compile the PDX1 homeodomain sequence of 33 rodent species which span all rodent suborders. Out of all investigated species, only the 3 gerbil species had a PDX1 homeodomain that was highly divergent and GC-rich. I propose that the extreme GC compositional bias reported in sand rats may affect all gerbil species, and runaway GC-skew may have started after the divergence of gerbils from the murines.

In Chapter 3, I ask whether GC compositional bias has altered the biochemical property of sand rat PDX1, especially PDX1 protein half-life, as altered PDX1 half-life has been shown to cause disease in humans. I tagged sand rat and mouse PDX1 and expressed them separately in rat INS-1 cells, and I measured the rate of protein degradation using a fluorescent pulse-chase method at the single cell level. I used drug treatment and *in vitro* mutagenesis to further explore the biological pathways involved in the PDX1 degradation process. In this chapter, I show loss of a key PDX1 ubiquitination site in the sand rat protein linked to GC compositional bias, and

I provide evidence for possible compensation through selection of a new gerbil-specific ubiquitination site. This work provides evidence linking biochemical changes to GC-skew and suggesting presence of compensation due to selection.

In Chapter 4, I ask if GC-skew has affected the biological function of sand rat PDX1, with focus on PDX1 regulated insulin gene expression. First, I overexpressed sand rat and mouse PDX1 separately in rat INS-1 cells followed by treatment with either glucose-rich or glucose-poor medium. I show that compared to mouse PDX1, sand rat PDX1 overexpression does not induce the same expression level changes on rat insulin expression, and that sand rat PDX1 is not responsive to glucose changes in a rat β-cell environment. Second, I replaced the second exon of rat *Pdx1* with the second exon of sand rat *Pdx1* using CRISPR/Cas9 editing to generate rat INS-1 cells expressing fusion rat-sand rat PDX1 proteins at endogenous levels. Using chromatin immunoprecipitation followed by qPCR, I show that compared to the endogenous rat protein, sand rat PDX1 has lower binding efficiency to the rat *Ins1* promoter. These results suggest that amino acid changes linked to GC-skew in sand rat PDX1 have significantly affected its ability to bind to target promoter sequences and likely also changed its interaction with cofactor proteins regulating target gene expression.

In Chapter 5, I ask if gerbils have more aberrantly divergent proteins (proteins with an unexpected amount of amino acid divergence) compared to other rodents (e.g. mouse and rat) and whether aberrantly divergent gerbil proteins are associated with GC-skew and high GC regions in the gerbil genome. I performed a genome-wide analysis of protein-coding sand rat genes and tested the correlation between aberrant protein divergence, GC-skew and abnormal metabolism observed in sand rats. Alignments for sand rat genes that have 1-to-1 orthologues to 11 other rodent species with human as the outgroup were generated and protein alignments were used for protein divergence analysis based on amino acid replacement matrices. Results show that gerbils have higher numbers of aberrantly divergent proteins compared to other members of the mouse-like clade, and these aberrantly divergent proteins are correlated with GC-skew and high GC regions in the gerbil genome. Four of the most divergent proteins, PDX1, INSR, MEDAG and SPP1, are associated with metabolism and may be associated with metabolic dysregulation in gerbils. Finally, to obtain the full-length INSR gene sequence, I performed long-read sequencing of the sand rat and Mongolian jird liver transcriptome. Unfortunately, the full-length gerbil INSR gene sequence was not resolved.

This chapter shows that GC compositional bias in the gerbil genome has affected many proteins, and that gerbils are a good model for studying the tug-of-war between GC-skew and natural selection.

Overall, this dissertation explores the conflict between GC-skew and natural selection by using the sand rat *Pdx1* gene as an example. This works shows that when extreme GC compositional bias occurs in genomic regions, natural selection may not be able to remove all maladaptive mutations in otherwise highly conserved genes trapped in these quagmires.

Chapter 2:

Highly divergent *Pdx1* is only observed in the gerbil lineage

Statement of Author Contributions

The original idea for the project was conceived by discussion between Yichen Dai and Peter W. H. Holland. Adam Hargreaves provided bioinformatics advice in the initial stages of this work. YD performed computational and experimental work. PWHH supervised the work of YD throughout the length of this study.

1. Abstract

The PDX1 transcription factor is highly conserved across the entire vertebrate lineage, with near 100% sequence conservation in the PDX1 homeodomain, the key region that directly interacts with DNA. However, recent sequencing efforts have uncovered the presence of a highly divergent PDX1 protein in a desert rodent, the sand rat (*Psammomys obesus*). The sand rat *Pdx1* gene is trapped in a GC-rich genomic region along with at least 88 genes, which I investigate further in later chapters. Here, I focus on the highly divergent sand rat PDX1 protein and I ask whether excessive amino acid changes in sand rat PDX1 are a characteristic of arid-dwelling rodents, due to adaptive or relaxed selection in these lineages. If yes, this predicts presence of PDX1 proteins with similar amino acid changes in other arid-living rodents. In this brief chapter, I present results from a search for highly divergent *Pdx1* in other rodent species, including multiple arid-living species from different families. I obtained annotated *Pdx1* gene sequences from 16 rodent species available on NCBI, and assembled genome and transcriptome sequences for an additional 19 rodent species using publicly available data in the NCBI Sequence Read Archive (SRA). Combined with rodent sequencing data obtained previously in this research group, PDX1 homeodomain sequences were identified in 33 rodent species. Apart from gerbils, only guinea pigs and kangaroo rats contained changes to normally conserved amino acids in the PDX1 homeodomain. Using Sanger sequencing, I show that the annotated guinea pig *Pdx1* gene on NCBI is likely incorrect and the divergent guinea pig PDX1 homeodomain sequence may be a genome assembly artifact. In neither case does this affect more than 8 amino acids compared to 15 amino acid changes in the sand rat homeodomain. I conclude that extreme divergence in the PDX1 protein is not a general feature in arid-living rodents, and more attention should be given to the impact of GC composition bias on gene evolution in the gerbil lineage.

2. Introduction

As introduced in Chapter 1, the PDX1 transcription factor is highly conserved throughout the vertebrate lineage, with the key DNA-binding homeodomain retaining 100% sequence conservation in nearly all vertebrates, with the exception being teleost fish and lizards. However, even in these species, only 2 to 3 amino acids are different from the conserved sequence. Discovery of the highly divergent sand rat PDX1 protein, with 15 divergent residues out of the 60 amino acids making up the homeodomain, has raised questions around evolution of the *Pdx1* gene in rodents, especially arid-living rodents like the sand rat. It is possible that the highly divergent sand rat *Pdx1* gene may be a result of adaptive selection or relaxed selection in an arid environment. If this is the case, I predict convergent evolution of the *Pdx1* gene in other rodents living in similar environments, which would also raise questions around pancreatic development and function in these species.

To assess whether the *Pdx1* gene is divergent or conserved in different rodent species, availability of sequencing data is crucial. Although the house mouse (*Mus musculus*) and brown rat (*Rattus norvegicus*) are well-studied animal models in the lab possessing high quality annotated genomes, they only represent two closely related rodent species in the Rodentia order. The order Rodentia is massive compared to other orders of mammals, and the 2277 identified rodent species make up approximately half of all placental mammal species on Earth[151]. The closest relatives of rodents are the Lagomorpha order consisting of rabbits and pikas. Rodents are classified into 33 families, and although certain branches of rodent phylogeny remain under debate, rodents can be divided into three main groups: 1) the "mouse-related" clade with the suborder Anomaluromorpha (springhare and flying mice), the suborder Myodonta (mice and voles), the superfamily Geomyoidea (kangaroo rats and pocket gophers) and the family Castoridae (beavers), 2) the "squirrel-related" clade made up of the suborder Sciuriodea (or referred to as Sciuromorpha) (squirrels and mountain beavers) and the family Gliridae (dormice), and 3) the suborder Ctenohystrica (or referred to as Hystricomorpha) (degus and chinchillas)[152].

Rodent evolution has been of interest due to the extremely high speed at which rodents are evolving[151]. Analysis of mitochondrial DNA has shown that rodents are the fastest evolving mammalian lineage, with changes in 7% of mitochondrial DNA per million years compared to 2% in other mammals[153,154]. High fixation rate, rather than mutation rate, is thought to contribute to the high speed of evolution in the rodent lineage and can partially explain why analysis of mammal orthologs reveal large differences between rodents and other mammals[155]. In addition, rodents are good models for studying chromosomal variability and karyotype evolution as many rodent species have extensive chromosomal variability between populations of the same species and between closely related sister species. For example, population studies of the house mouse have discovered 40 different house mice populations around the city of Barcelona possessing different chromosomal arrangements with diploid numbers spanning 27 – 40[156]. In the gerbil subfamily which the sand rat belongs to, the karyotype of different species also varies greatly, with diploid numbers ranging between 34 – 74 and very few species sharing the same karyotype[157,158]. Even for closely related species in the same genus, for example the genus *Gerbillus*, Aniskin et al. reported 33 – 49 chromosomal changes between either of the

four *Gerbillus* species analysed[157]. For the sand rat, it has a diploid number of 48 while its close relatives in the *Meriones* genus, *M. unguiculatus*, *M. libycus* and *M. shawi* all have 44 chromosomes[159–161]. Further research is required to assess whether multiple arid-living rodent lineages have highly divergent *Pdx1* gene, and if so, whether the high fixation rates and extensive variability of chromosome structure (suggesting frequent occurrence of translocation etc.) in rodent lineages are correlated with changes in the *Pdx1* gene.

In this chapter, I ask if *Pdx1* is highly divergent in any rodent species other than the sand rat, and I present results from an extensive search through rodent sequencing data on the public database NCBI. Although many studies on rodent genome evolution exist, only some rodents have had their genome or transcriptome sequenced. In addition, even fewer species have publicly available assembled and annotated genomes or transcriptomes, with most sequenced species possessing only raw read data available on the NCBI Sequence Read Archive (SRA). Therefore, I performed *de novo* genome or transcriptome assembly of 19 rodent species to find the *Pdx1* gene in different rodent species. I show that three gerbil species included in this analysis possessed a highly divergent *Pdx1* gene while 28 rodent species had a 100% conserved PDX1 homeodomain. The only two species with a non-conserved PDX1 homeodomain were the kangaroo rat (*Dipodomys ordii*) and the guinea pig (*Cavia porcellus*). I show Sanger sequencing results for guinea pig *Pdx1* that suggest that the annotated *Pdx1* sequence may be affected by poor sequencing quality or genome assembly errors. In conclusion, gerbils are the only rodent lineage with a highly divergent *Pdx1* gene, and further research on PDX1 function in gerbils has the potential to shed light on the interaction between amino acid changes caused by GC composition bias and natural selection.

3. Materials and methods

3.1 Rodent genome and transcriptome assembly

An extensive search for rodent sequencing data deposited in the Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) was carried out using rodent names in the NCBI Taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9989). Data was downloaded from SRA.

Data quality was assessed with fastqc v0.11.5 and paired reads were separated using sickle v1.33[162,163]. Trinity v2.2.0 was used for transcriptome analysis with default settings[164]. For genome assembly, the most optimal k-mer value was predicted using kmergenie v1.7016 and then used for genome assembly using ABySS v2.0.1[165,166]. Assembly quality was assessed using assembly-stats v1.0.0[167]. All details for the genomes and transcriptomes assembled in this study are listed in Table 2.1. A search for *Pdx1* was carried out using the tBLASTn algorithm with the full mouse PDX1 protein sequence as the query and all parameters set to default. The identity of the top 5 hits was verified by a reciprocal tBLASTn against the human nucleotide collection on NCBI. For rodents with an annotated *Pdx1* gene sequence on NCBI, the full cds was downloaded and used directly for analysis. A complete list of annotated rodent *Pdx1* sequences are in Table 2.2.

In addition, two unpublished rodent genomes for members of the Hystricomorpha lineage, *Pipanacoctomys aureus* and *Tympanoctomys barrerae*, was searched for *Pdx1* (unpublished data by F Marletaz and PWH Holland). In addition, *Pdx1* genes in three gerbil species, *Psammomys obesus*, *Meriones unguiculatus* and *Meriones libycus*, and one closely related species *Acomys cahirinus* were included for alignment analysis. These genes were previously identified in the study published by Hargreaves et al[102].

All rodent *Pdx1* genes identified in this study were aligned using the ClustalW multiple alignment function in BioEdit v7.2.5[168].

3.2 PCR verification of the guinea pig *Pdx1* exon 1 sequence

The Roche DNA Isolation Kit for Cells and Tissues (#11814770001) was used to extract genomic DNA from frozen guinea pig liver tissue stored at -80ºC following the kit protocol. PCR was carried out using the primer sequences listed in Table 2.3, and each PCR reaction consisted of 1X GoTaq Green Flexi Buffer, 2 mM MgCl2, 0.8 mM dNTPs, 0.5 uM forward primer, 0.5 uM reverse primer, 1.25 U GoTaq G2 Flexi DNA Polymerase, 0.5 ug genomic DNA in a total volume of 50 ul. The protocol used for the PCR reaction was 95ºC for 2 minutes, followed by 35 cycles of 95ºC for 15 seconds, 60ºC for 15 seconds, and 72ºC for 30 seconds and a final extension step at 72ºC for 1 minute.

The size of each PCR product was verified using gel electrophoresis and purified using the QIAquick PCR Purification Kit (Qiagen #28104). The purified PCR products were sequenced using Sanger sequencing.

Table 2.1 All genomes and transcriptomes assembled in this study. Raw read data for 3 species from the Hystricomorpha group, 15 from the "mouse-related" clade and 1 from the "squirrel-related" clade was downloaded from SRA (data accessed in October 2016). The transcriptomes were assembled with Trinity v2.2.0 and the genomes were assembled with ABySS v2.0.1[164,165].

Lineage	Species	SRA run #	Data type	Total contig #	N50 (bp)	Contig #	*Pdx1*
Ctenohystrica/ Hystricomorpha	*Cryptomys hottentotus*	SRR2141211	Mixed tissue transcriptome	322,908	570	65,931	Not found
	Georychus capensis	SRR2141216	Mixed tissue transcriptome	245,682	410	67,234	Not found
	Heliophobius argenteocinereus	SRR2141215	Mixed tissue transcriptome	125,401	352	38,843	Not found
"Mouse-related" clade	*Chaetodipus baileyi*	SRR1576060	Kidney transcriptome	86,975	2,028	11,144	Not found
	Rhizomys pruinosus	SRR933768	Brain and liver transcriptome	215,283	1,961	25,136	Complete
	Eospalax fontanierii	SRR3036367	Muscle transcriptome	130,711	791	24,883	Not found
	Spalax carmeli	SRR2016467	Placenta transcriptome	231,671	1,408	28,689	Not found
	Tachyoryctes splendens	SRR2141217	Mixed tissue transcriptome	243,660	724	41,057	Not found
	Ellobius lutescens	SRR3475727	Genome	8,082,007	2,392	279,624	Complete
	Onychomys torridus	ERR968259	Genome	7,536,142	4,512	168,295	Partial
	Peromyscus californicus	SRR545677 SRR545678 SRR545679	Genome	8,236,361	2,366	298,318	Partial
	Peromyscus leucopus	SRR545958 SRR545959	Genome	363,840	1,210	44,960	Complete

		SRR545960					
	Peromyscus polionotus	SRR545671 SRR545672 SRR545673	Genome	6,553,890	2,636	264,284	Partial
	Sigmodon hispidus	SRR2954727	Genome	5,315,264	3,905	185,177	Complete
	Apodemus semotus	SRR3201409	Genome	22,654,712	895	671,723	Partial
	Apodemus sylvaticus	SRR2141382	Genome	27,068,228	293	2,090,123	Partial
	Apodemus uralensis	ERR1101669	Genome	8,983,726	2,876	261,213	Complete
	Mus mattheyi	ERR1101670	Genome	5,112,376	3,367	203,783	Complete
"Squirrel-related" clade	*Spermophilus dauricus*	SRR955327 SRR955328 SRR955330 SRR955331 SRR955334 SRR955335	Genome	59,210	1,093	9,613	Partial

Table 2.2 All annotated *Pdx1* gene sequences in rodent species available on NCBI. A total of 5 species from the Hystricomorpha group, 9 species from the "mouse-related" clade and 2 from the "squirrel-related" are included, and all sequences were downloaded in October 2016. The reference sequences beginning with XM indicate gene sequences predicted by NCBI automated software from genome assemblies.

Lineage	**Species**	**NCBI reference sequence # for annotated *Pdx1***
Ctenohystrica/Hystricomorpha	*Octodon degus*	XM_023705113.1
	Chinchilla lanigera	XM_013509165.1
	Cavia porcellus	XM_003468370.4
	Fukomys damarensis	XM_010605588.2
	Heterocephalus glaber	XM_004865316.1
"Mouse-related" clade	*Dipodomys ordii*	XM_013033503.1

	Jaculus jaculus	XM_004661735.1
	Nannospalax galili	XM_017800993.1
	Mesocricetus auratus	U73854.1
	Cricetulus griseus	XM_016978305.1
	Microtus ochrogaster	XM_005371760.2
	Peromyscus maniculatus	XM_015986228.1
	Mus musculus	NM_008814.3
	Rattus norvegicus	NM_022852.3
"Squirrel-related" clade	*Ictidomys tridecemlineatus*	XM_013361462.2
	Marmota marmota	XM_015483953.1

Table 2.3 PCR primers used for amplification of the guinea pig *Pdx1* sequence. These primers were designed to amplify rodent *Pdx1* exon 1 and they bind to highly conserved regions in this exon. Here, Y refers to a C or T base, R refers to A or G, H refers to A or C or T, and M refers to A or C.

Primer name	**Primer sequence (5' to 3')**
RPFP_1	ATGAAYRGHGAGGAGCAGTACTAYGC
RPFP_2	TGYCTGTAYATGGGYCGYCARCC
RPRP_1	CCTTTCCACGCGTGMGCTTTGG

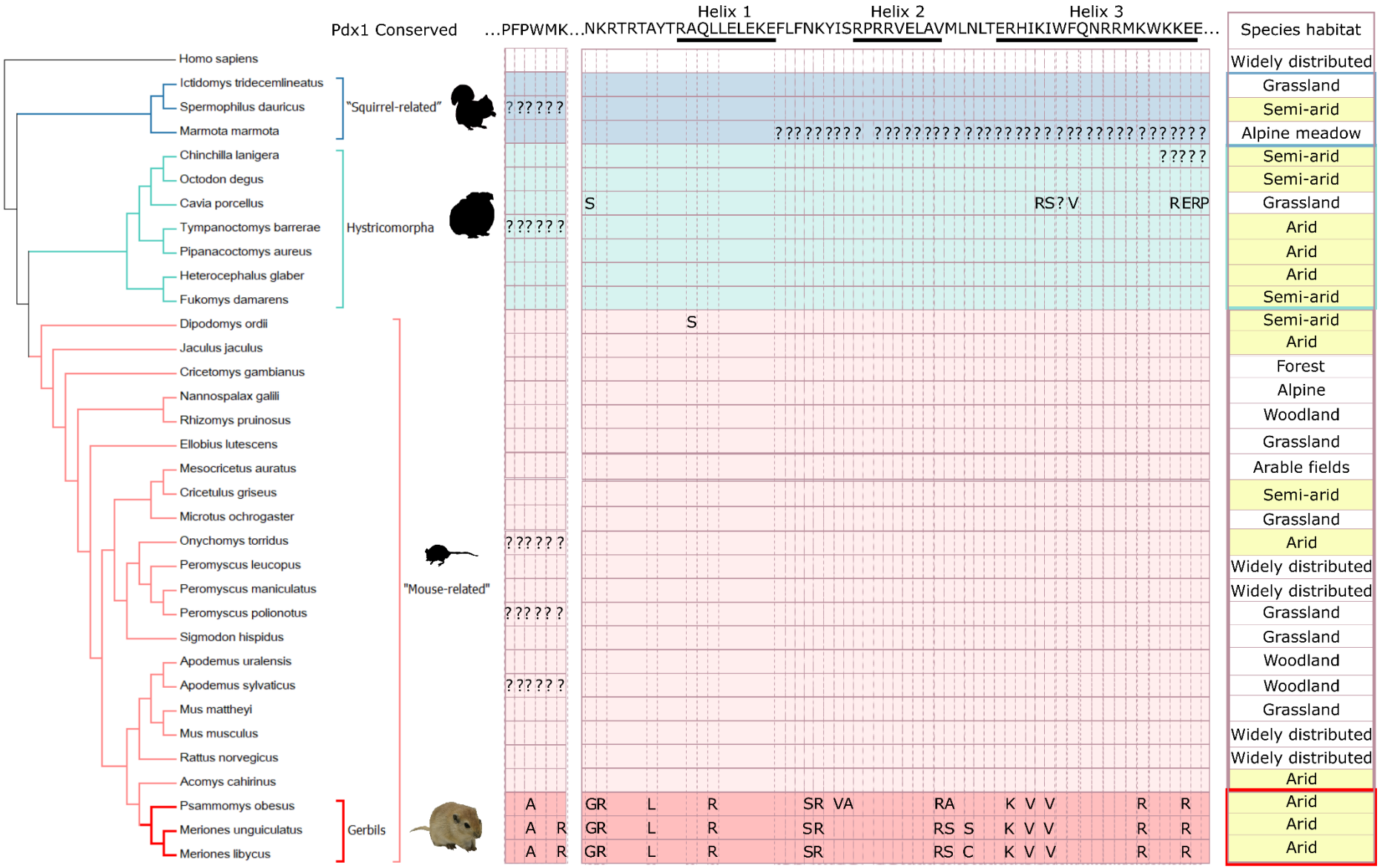

Helix 1
Helix 2
Helix 3
Pdx1 Conserved
...PFPWMK...NKRTRTAYTRAQLLELEKEFLFNKYISRPRRVELAVMLNLTERHIKIWFQNRRMKWKKEE...
Species habitat
Homo sapiens
Widely distributed
Ictidomys tridecemlineatus
Spermophilus dauricus
Marmota marmota
"Squirrel-related"
Grassland
Semi-arid
Alpine meadow
Chinchilla lanigera
Octodon degus
Cavia porcellus
Tympanoctomys barrerae
Pipanacoctomys aureus
Heterocephalus glaber
Fukomys damarens
Hystricomorpha
Semi-arid
Semi-arid
Grassland
Arid
Arid
Arid
Semi-arid
Dipodomys ordii
Jaculus jaculus
Cricetomys gambianus
Nannospalax galili
Rhizomys pruinosus
Ellobius lutescens
Mesocricetus auratus
Cricetulus griseus
Microtus ochrogaster
Onychomys torridus
Peromyscus leucopus
Peromyscus maniculatus
Peromyscus polionotus
Sigmodon hispidus
Apodemus uralensis
Apodemus sylvaticus
Mus mattheyi
Mus musculus
Rattus norvegicus
Acomys cahirinus
"Mouse-related"
Semi-arid
Arid
Forest
Alpine
Woodland
Grassland
Arable fields
Semi-arid
Grassland
Arid
Widely distributed
Widely distributed
Grassland
Grassland
Woodland
Woodland
Grassland
Widely distributed
Widely distributed
Arid
Psammomys obesus
Meriones unguiculatus
Meriones libycus
Gerbils
Arid
Arid
Arid

Fig. 2.1 Alignment of rodent PDX1 hexapeptide and homeodomain regions. The human PDX1 sequence is the outgroup. Unknown or missing amino acid residues are shown with "?" and only residues that are different from the consensus sequence are shown in this alignment. The rodent phylogeny used here is based on the work of Harrison et al. and Blanga-Kanfi et al[151,169]. Rodent habitat is shown next to each corresponding species, with arid and semi-arid environments highlighted in yellow. Rodent habitat information was obtained from the IUCN Red List of Threatened Species (www.iucnredlist.org, accessed July 2020).

4. Results

4.1 Assembly of rodent raw sequencing data

To assess whether rodents aside from the sand rat possessed a highly divergent *Pdx1* gene, I obtained the *Pdx1* sequence from as many rodent species as possible. First, *Pdx1* genes from 16 different rodent species were available on the public database NCBI, amongst which the *Pdx1* sequence for three species (house mouse, brown rat, golden hamster *Mesocricetus auratus*) had been verified using PCR and Sanger sequencing. The remaining 13 *Pdx1* sequences were predicted using a standard NCBI pipeline which automatically annotates predicted genes found in assembled genomes uploaded to NCBI. At the same time, 19 rodent species did not have a publicly available *Pdx1* sequence or assembled genome but had raw sequencing data stored in the Sequence Read Archive (SRA) on NCBI. Using SRA data, I assembled 11 rodent genomes using the programme ABySS[165] and 8 rodent transcriptomes using the programme Trinity[164]. Both programmes were selected based on their previous performance in the lab when used to for *de novo* genome/transcriptome assembly. For genome assembly, the most optimal k-mer value (a key factor determining assembly quality) was predicted by the programme kmergenie before the assembly process started[166].

Out of the 8 assembled transcriptomes, only 1 transcriptome constructed from bamboo rat (*Rhizomys pruinosus*) brain and liver RNA contained the *Pdx1* gene. Given that *Pdx1* expression is confined to pancreatic and duodenal tissue, it is likely that the liver tissue used for this RNA sequencing experiment was contaminated. For the other 7 transcriptomes, a tBLASTn query using both the mouse and sand rat PDX1 sequence did not yield potential *Pdx1* hits, as verified by reciprocal BLAST against the human genome. As neither pancreatic nor duodenal tissue were used in the experiments that produced these raw reads, lack of *Pdx1* expression in the tissues sequenced is the most likely explanation for missing *Pdx1* in these 7 assembled transcriptomes.

In contrast to the assembled transcriptomes, all 11 assembled genomes included the *Pdx1* sequence of the respective species. tBLASTn was used to identify potential *Pdx1* sequences in the assembled genomes, and identity of each returned hit was verified using reciprocal BLAST against the human genome. Several *Pdx1* genes obtained from the assembled genomes were incomplete in regions that encode the hexapeptide domain (Fig. 2.1), although all genes had a complete homeodomain region. This may be caused by difficulties sequencing and assembling the exon-intron border, which is located near the sequence coding for the hexapeptide domain. In addition, these assembled genomes were constructed from low coverage reads as the

experiments they came from were not necessarily aimed at *de novo* genome assembly and this may have also led to gaps in the genome assembly.

4.2 Alignment of rodent PDX1 hexapeptide and homeodomain sequence

A total of 33 *Pdx1* genes were obtained as annotated sequences from NCBI, identified using tBLASTn from transcriptomes and genomes that I assembled, or identified using tBLASTn from unpublished transcriptomes and genomes (data by F Marletaz and PWH Holland). To assess if any of these *Pdx1* genes were highly divergent, I generated PDX1 protein alignments including the human PDX1 sequence as the outgroup. Fig. 2.1 shows an alignment of all PDX1 hexapeptide and homeodomain sequences that were identified in this study. A total of 21 rodent species show 100% similarity to the conserved hexapeptide and homeodomain sequence, which is also 100% conserved in the human outgroup. At the same time, the hexapeptide domain of 5 rodent species were missing from their genome assemblies while their PDX1 homeodomain sequence was intact and 100% conserved.

Out of these 33 rodent species, 15 reside in semi-arid or arid habitats. These 15 species are distributed among the three main rodent groups, the "squirrel-related" group, Hystricomorpha, and "mouse-related" group. Among these 15 rodents that inhabit arid environments, 2 non-gerbil species, the long-tailed chinchilla (*Chinchilla lanigera*) and the kangaroo rat (*Dipodomys ordii*), have a PDX1 homeodomain sequence that deviates from the conserved sequence. Both sequences were obtained as annotated genes from NCBI. Given the level of amino acid conservation in other regions of the chinchilla and kangaroo rat PDX1 protein (Fig. 2.1), the *Pdx1* gene annotation is likely in the correct location. However, low coverage of the *Pdx1* gene area and/or poor assembly quality may explain the presence of 5 unknown residues in the chinchilla PDX1 homeodomain. As for the kangaroo rat, it has one residue in the PDX1 homeodomain that differs from the conserved sequence, and no other rodent including gerbils has this divergent residue. The altered site in the kangaroo rat homeodomain is the second residue in helix 1, the first of the three α-helix structures in the PDX1 homeodomain contributing to DNA recognition and binding, and the kangaroo rat divergent residue is serine while the conserved site is alanine. Due to lack of access to kangaroo rat tissue samples, I could not verify if the NCBI sequence reflected a genomic change in kangaroo rats or if the annotated genome had sequencing errors.

Aside from the chinchilla and kangaroo rat, two non-arid dwelling rodents also had non-conserved residues in their homeodomain sequence. These rodents were the alpine marmot (*Marmota marmota*) and the guinea pig (*Cavia porcellus*), with the marmot PDX1 homeodomain containing 40 unknown amino acids and the guinea pig homeodomain containing 8 divergent amino acids and 1 unknown amino acid. Both *Pdx1* gene sequences were obtained from the annotated nucleotide database on NCBI. For the marmot sequence, it is highly likely that this is caused by low sequencing coverage in the *Pdx1* gene area and/or poor assembly quality. The 8 non-conserved residues in the guinea pig PDX1 homeodomain are not observed in any other rodent and are all located in helix 3.

Only the three gerbil species, the sand rat (*Psammomys obesus*), the Mongolian jird (*Meriones unguiculatus*) and the Libyan jird (*Meriones libycus*) have a divergent PDX1 hexapeptide and

homeodomain sequence that shows amino acid differences spread throughout the entire protein domain. The Mongolian jird and Libyan jird's PDX1 sequences are more similar to each other than that of the sand rat, with only these two species possessing an arginine residue instead of lysine in the hexapeptide domain and also possessing either a serine or cysteine in place of the conserved asparagine residue between homeodomain helix 2 and 3. Amongst the three gerbils, sand rat PDX1 is the most divergent, as it has the highest number of divergent residues (not including unknown residues). A total of 16 out of 66 conserved hexapeptide and homeodomain residues are divergent in the sand rat, out of which 14 are also shared by the two other gerbil species.

4.3 PCR and Sanger sequencing verification of guinea pig *Pdx1* exon 1

Aside from the gerbils, the rodent with the highest level of amino acid diversity in the PDX1 homeodomain is the guinea pig. The guinea pig *Pdx1* gene was obtained from an annotation generated by NCBI, and analysis of the full-length guinea pig PDX1 protein shows not only presence of a highly divergent PDX1 homeodomain but also extremely divergent N-terminal regions that are more divergent than the sand rat sequence (Fig. 2.2). Aside from divergent amino acids that are not observed in the other three species, the guinea pig PDX1 N-terminal region also contains several peptides that create large gaps in the alignment. These observations suggested that the divergent guinea pig PDX1 may be a result of assembly errors rather than being reflective of the actual PDX1 sequence in guinea pigs. Therefore, I designed two pair of nested PCR primers for *Pdx1* exon 1, which encodes the PDX1 N-terminal, based on the conserved rodent sequence and tested them on guinea pig gDNA, using mouse gDNA as a control.

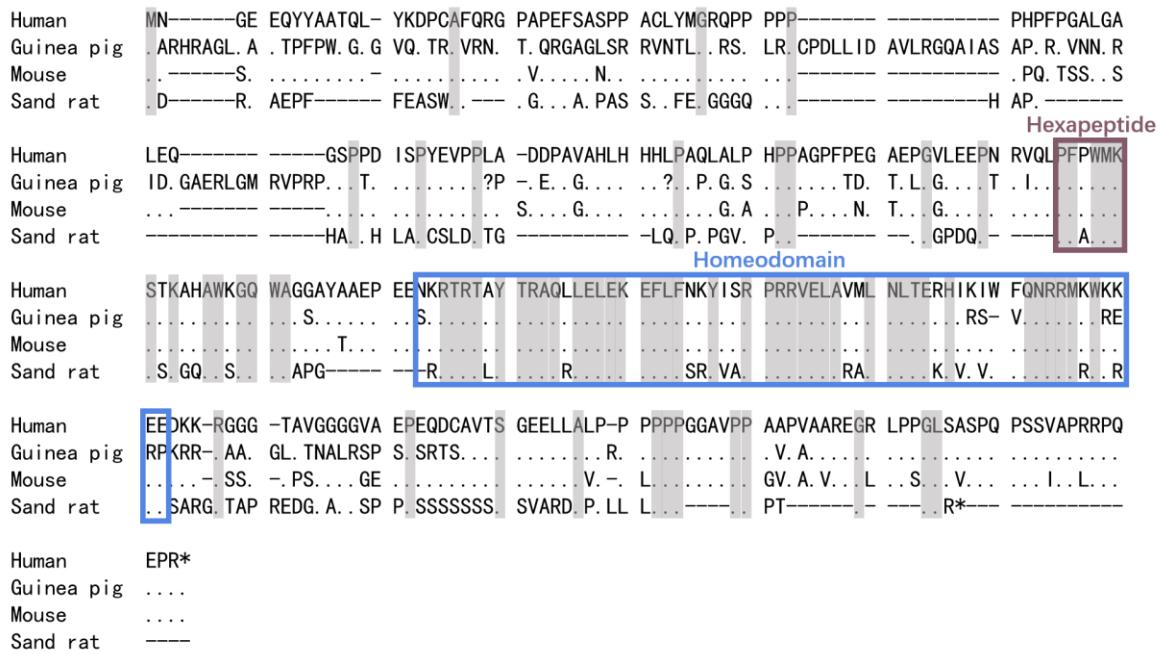


Fig. 2.2 Alignment of human, guinea pig, mouse, and sand rat PDX1 protein. Residues in the rodent sequence that are the same as the corresponding human residues are represented by a dot, while gaps are shown as dashes and unknown residues are question marks. The stop codon

is represented by an asterisk. Sites at which the sequence is 100% conserved are highlighted in grey. The homeodomain region is in blue while the hexapeptide domain is in purple.

Using these primers, I successfully amplified mouse *Pdx1* exon 1 in the control sample, and I also amplified a fragment with similar size to mouse *Pdx1* exon 1 in the guinea pig gDNA sample. Sanger sequencing of this guinea pig DNA fragment returned a sequence with 80% similarity to the mouse PDX1 N-terminal sequence, and reciprocal BLAST against the human transcriptome returned human PDX1 as the top hit. In Fig. 2.3, I show an alignment of human, guinea pig, mouse, and sand rat PDX1 using the Sanger sequencing verified guinea pig N-terminal sequence. More residues conserved between all four species are highlighted, and the guinea pig N-terminal sequence has more residues that are conserved in both the human and mouse sequence. The large discrepancy between the NCBI published guinea pig *Pdx1* sequence and Sanger sequenced *Pdx1* sequence that I generated suggest that the NCBI sequence may have annotated this gene using a genome with poor assembly or sequencing quality in the *Pdx1* region. In conclusion, the large number of divergent residues in the guinea pig PDX1 homeodomain may likely be a result of sequencing and/or assembly errors rather than being reflective of actual divergence of the guinea pig *Pdx1* gene sequence.

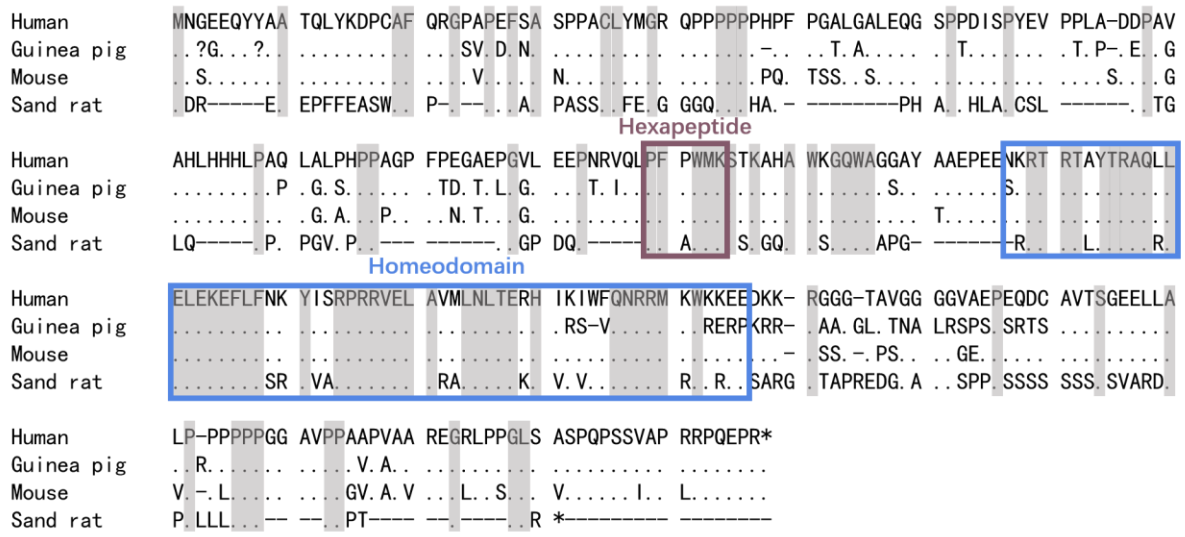


Fig. 2.3 Alignment of PDX1 protein including the updated guinea pig sequence. The guinea pig PDX1 sequence is updated based on Sanger sequencing results. Residues in the rodent sequence that are the same as the corresponding human residues are represented by a dot, while gaps are shown as dashes and unknown residues are question marks. The stop codon is represented by an asterisk. Sites at which the sequence is 100% conserved are highlighted in grey. The homeodomain region is in blue while the hexapeptide domain is in purple.

5. Discussion

Convergent evolution of physiology, morphology, and behaviour has been well-studied in desert rodents, with examples including evolution of kidneys that produce highly concentrated urine, similar limb structure that facilitate burrowing and foraging on rocky or sandy surfaces, and similar rodent-plant interaction patterns[170]. Such comparisons have not been studied in great detail at the molecular level due to the scarcity of high quality desert rodent genomes, but it is expected that desert rodents should have changes in the same molecular pathways that control the processes mentioned above. Discovery of the highly divergent sand rat *Pdx1* gene has led to the hypothesis that this gene may have been an adaptation to desert food sources, and other desert rodents may have convergent changes in the *Pdx1* gene. Indeed, desert rodents in arid regions across the world have similar food options due to limited choice in their habitats; the majority of desert rodents rely on seeds or plant tissue (usually plant tissue with high salt content), with the exception of some carnivorous desert rodents that prey on insects and other arthropods[171].

In this chapter, I show that the highly divergent *Pdx1* gene observed in the sand rat is unique to the gerbil subfamily and may not be present elsewhere in the rodent lineage. 10 non-gerbil desert rodents included in this analysis have 100% conservation of the PDX1 homeodomain sequence, including the spiny mouse (*Acomys cahirinus*). The spiny mouse is the only member of the Acomyinae (also referred to as Deomyinae) subfamily included in this analysis and is the closest relative to gerbils. In addition, 1 non-gerbil desert rodent, the kangaroo rat, has 1 amino acid that differs from the conserved homeodomain sequence. However, this divergent kangaroo rat residue is not present in gerbils, and whether this change has any effect on PDX1 function remains to be assessed. In summary, the results of this chapter strongly rejects the possibility that PDX1 protein divergence may be an adaptation to food sources in a desert environment, and it is very unlikely that the last common ancestor of gerbils and spiny mice already had a highly divergent *Pdx1* gene. Overall, analysis of all 30 non-gerbil species in this chapter shows a very high level of sequence conservation in PDX1 throughout the rest of the rodent lineage. This rejects the previous hypothesis that the highly divergent sand rat *Pdx1* gene may be reflective of relaxed selection in multiple rodent lineages, as extremely high levels of sequence conservation suggest decreased tolerance of non-synonymous mutations affecting the homeodomain.

Regarding the four non-gerbil rodent species that have either unknown residues or non-conserved amino acids in the PDX1 homeodomain, the *Pdx1* gene sequences for these four species are all from NCBI annotated sequences. For the two species with unknown residues in the PDX1 homeodomain (chinchilla and marmot), it is very likely that the genome assembly from which these gene annotations were derived has low sequencing quality or has assembly errors. For the single amino acid discrepancy in the kangaroo rat, I have not been able to verify further whether this was due to a sequencing error given the lack of access to kangaroo rat tissue. However, for the many amino acid discrepancies in the guinea pig, especially in the N-terminal region, I was able to verify the *Pdx1* exon 1 sequence and provide evidence for possible mis-assembly of the guinea pig genome. Although I was not able to directly PCR the *Pdx1* exon 2 sequence, which contains the homeobox, it is highly likely that the amino acid

discrepancies in this area are also due to assembly errors rather than actual *Pdx1* sequence divergence.

In conclusion, the highly divergent *Pdx1* gene can be found in three closely related gerbil species and is unique to the gerbil subfamily. No other rodent, including those found in arid or semi-arid environments, has as many amino acids deviating from the conserved sequence as seen in gerbils. I suggest that rather than being due to positive or relaxed selection, other factors may have played the key part in *Pdx1* gene evolution in the gerbil lineage. Given that the gerbil *Pdx1* gene is located in a GC-rich genomic region, I speculate that an early gerbil ancestor began to accumulate GC content in the genomic region containing *Pdx1*, and for unknown reasons these changes became so pronounced that they resulted in the highly divergent *Pdx1* gene we observe today in the sand rat. However, given the lack of sequencing data in other gerbil species, it is not possible to speculate further about whether this proposed accumulation of GC began to occur in the last common ancestor of all gerbils, or whether this began only in the last common ancestor of sand rats and the two *Meriones* species included in this analysis.

Chapter 3:

Loss of a key PDX1 ubiquitination site is linked to GC composition bias

Statement of Author Contributions

This chapter was submitted as a manuscript formatted as a Research Article for *Molecular Biology and Evolution* (submitted November 2018; Manuscript ID MBE-18-0927). It was accepted for publication on March 26, 2019 after minor revision and first published online on April 9, 2019. The authors on this manuscript are in order: Yichen Dai and Peter W. H. Holland (corresponding author).

The original idea for the project was conceived by discussion between YD, Adam Hargreaves and PWHH. Anna Gloyn and Natasha Ng provided advice on INS-1 cell culture, and Thomas Dunwell provided general mammalian cell culture advice in the initial stages of this work. Sonia Trigueros provided microscopy advice and access to the ZOE Fluorescent Cell Imager (Bio-Rad). Daniel Lunn (University of Oxford Statistical Consultancy Services) provided advice on statistical analysis and modelling. Ofuase Dirisu (in2scienceUK) carried out experimental work included in this chapter under the supervision of YD. She constructed one mutant *Pdx1* plasmid used in this study and helped collect data for one experiment. YD performed all remaining experimental and computational work. PWHH supervised the work of YD throughout the length of this study. Helpful discussions were held with Rodrigo Pracana, Amy Royall, AH, John Mulley, Thomas Brekke, Sebastian Shimeld, and Gil McVean. Both authors discussed the results and revised the manuscript.

The data included in this chapter were also presented (as a contributed talk or poster) at:

- 2019 Society of Molecular Biology and Evolution annual meeting
  - Selected talk at the Walter M. Fitch Symposium
- 33rd International Mammalian Genome Society (IMGS) annual meeting, 2019
  - Selected talk in main conference
  - Winner of Verne Chapman Young Scientist Award
- Manchester's Molecular and Genome Evolution Symposium 2019
  - Selected talk
- 31st International Mammalian Genome Society (IMGS) annual meeting, 2017
  - Awarded *Outstanding Poster* prize
- Euro Evo Devo, organized by European Society for Evolutionary Developmental Biology
  - Awarded *Poster Prize*
- Genetics Society UK Centenary Meeting, 2019
  - Awarded *Poster Prize*

1. Abstract

The sand rat offers an opportunity to study the conflict between extreme GC-skew and natural selection. Large regions in the sand rat genome show accumulation of GC which impact the coding sequence of key genes, for example *Pdx1*. *Pdx1* is highly conserved in all vertebrates and is essential for initiating pancreatic development and maintaining adult islet function. The most conserved region in the PDX1 protein is the homeodomain, the site at which this transcription factor recognizes and binds to DNA, and it is 100% identical between the human, chicken, frog, and shark. However, the sand rat PDX1 homeodomain only has 75% similarity to the conserved vertebrate sequence. In this chapter, I ask whether sand rat PDX1 has a different protein half-life compared to the mouse PDX1 protein, and if so, I ask whether this difference is linked to GC composition bias in the sand rat *Pdx1* gene. I use methods such as mammalian cell transfection, pulse-chase labelling and fluorescence microscopy to directly measure the half-life of tagged PDX1 proteins, and I use drug treatment and *in vitro* mutagenesis to test which pathways and amino acid residues are key factors controlling PDX1 half-life. My results show that GC-skew has indeed lengthened sand rat PDX1 half-life via the loss of AT-rich codons encoding lysine, and I suggest that natural selection may have compensated this loss by selecting a novel gerbil-specific lysine residue. The results of this chapter give molecular insight into the "tug-of-war" between natural selection and possible maladaptive changes caused by localized GC-skew.

2. Introduction

Genome sequencing has revealed uneven distribution of GC content in eukaryote genomes, for example the accumulation of GC at third codon positions and at recombination hotspots[52,58]. gBGC has been shown to be the main cause of GC accumulation in many cases, and this process should be neutral with respect to fitness as it can occur in any position in the genome and can be deleterious, beneficial, or neither[58]. However, in some cases, both theoretical modelling and analysis of empirical data have shown that gBGC mainly results in rapid evolution and fixation of weakly deleterious mutations in coding regions[172,173].

Given the extent of sequence divergence in the sand rat *Pdx1* gene associated with extreme GC-skew, I ask if the observed amino acid changes have altered sand rat PDX1 biochemical property and biological function. I propose that these changes are likely to be deleterious as *Pdx1* is one of the most highly conserved genes in the entire vertebrate lineage[105]. First of all, throughout 420 Myr of vertebrate evolution, *Pdx1* has been under strong purifying selection as 100% sequence conservation is observed in the 60 amino acid long homeodomain in the human, chicken, frog and shark, with only 2 – 4 amino acid differences observed in teleost fish and lizards. Second, *Pdx1* is essential for initiation of vertebrate pancreas development and also crucial for maintaining adult β-cell function, as shown by mouse models and human clinical reports[110,121,147]. It seems unlikely for the changes in sand rat *Pdx1* to be a result of adaptation with significant novel improvements to *Pdx1* function, and given the possible connection between these changes and gBGC, it may be more likely for the changes observed to be a result of fixation of mildly deleterious mutations accompanied by changes selected as compensation.

In this study, I focus on one aspect of protein biochemistry: protein half-life. In the mammalian cell, there are two biological pathways for intracellular protein degradation: the lysosome pathway and the proteasome pathway[174]. For PDX1, one clinical report has shown a connection between human PDX1 mutants E164D and E178K and a decreased protein half-life leading to pancreatic agenesis[147]. It has been shown by the work of multiple groups that PDX1 is degraded by the proteasome pathway, a complex biological pathway that involves the small molecule ubiquitin[175,176]. In brief, proteins bound for degradation via the 26S proteasome are recognized by a E3 ubiquitin-protein ligase and tagged with a ubiquitin molecule[174,177]. Although ubiquitin can bind to the protein via different amino acids, it only binds to the lysine residue in the context of 26S proteasome mediated degradation[174,177]. For PDX1, its degradation is mediated by a protein complex formed by the nuclear factor Pdx1 C terminus–interacting factor-1 (PCIF1) and a CUL3-based ubiquitin ligase, and as the name implies, the PDX1 C-terminal is essential for interaction between PDX1 and the PCIF1-CUL3 complex[176].

In this study, I ask whether a similar change in PDX1 protein half-life can be observed for the highly divergent sand rat PDX1 protein, and whether any possible changes are linked to GC-skew. By using mammalian cell transfection, I express either the mouse or sand rat PDX1 protein tagged with a SNAP tag at the C-terminal. I measure PDX1 half-life using a pulse-chase experiment with a fluorescent chemical specific to the SNAP tag, and I test which amino acids are crucial to controlling protein half-life through *in vitro* mutagenesis.

3. Materials and methods

3.1 Plasmid construction and *in vitro* mutagenesis

Full coding sequences for mouse and sand rat *Pdx1* genes were synthesized by GenScript, with alterations in sand rat codon usage to lower GC content (Supplement S3.1). Mouse *Pdx1* was inserted into pSNAPf vector (New England BioLabs #N9183S) using BamHI and XhoI restriction sites, and sand rat *Pdx1* inserted using NotI and XhoI restriction sites. Stop codons were removed to allow expression of Pdx1-SNAP fusion protein. Successful insertion was verified by Sanger sequencing. *In vitro* mutagenesis was carried out using PCR.

3.2 Cell culture and transfection

INS-1 cells (below 30 passages) were maintained in RPMI-1640 medium, supplemented with 10% fetal bovine serum, 50 mM beta-mercaptoethanol, 1 mM sodium pyruvate, 10 mM HEPES, and 100 U/ml penicillin/streptomycin, at 37℃/5% CO2. For transfection, cells were seeded into 96-well plates at 60,000 cells/well and transfected 5 h later using Lipofectamine 3000 (ThermoFisher #L3000015). For each well, 40 ml DNA-lipid mix with 400 ng plasmid DNA, 0.8 ml P3000 Reagent, and 1.2 ml Lipofectamine was added and incubated overnight.

3.3 SNAP-staining and fluorescence recording

SNAP-Cell TMR-Star dye (New England BioLabs #S9105S) was diluted 1:400 in culture medium. 50 ul was added to each well and incubated at room temperature for 5 min before three washes. Stained cells were incubated in fresh culture medium for 1 h at 37℃, twice, to ensure dissociation of un- bound SNAP-Cell TMR-Star dye prior to analysis. At each timepoint, pictures were taken using a ZOE Fluorescent Cell Imager (Bio-Rad #1450031EDU) with fixed settings: brightfield gain = 18, exposure = 300, LED = 40, contrast = 0; red channel gain = 40, exposure = 500, LED = 50, contrast = 0. Image analysis was performed using ImageJ[178].

3.4 MG132 treatment

MG132 stock solution in dimethyl sulfoxide (DMSO) (Sigma-Aldrich #M7449) was diluted to 10 uM in treatment wells at hour -2. Culture medium with 10 uM MG132 was used until hour 4, giving 6 h treatment in total. Controls were treated in parallel using DMSO.

3.5 Statistical analysis

We used nlme 3.1-137[179] in R 3.5.1[180] to model the pattern of protein degradation based on data collected for all protein and treatment types. Square-root transformed data were used to ensure no obvious deviations from homoscedasticity or normality (Supplement S3.2 – 11). The linear mixed effects model was constructed using the interaction between sqrt(Time) and type of species as the fixed effect, and the random effect was cell identity. Different experiments were analysed individually using the same fixed and random effects. Parameters reported are given (Supplement S3.12 – 16). Half-life predictions were made for each individual cell using the model generated and the generic predict function in R[180]; half-lives of mouse and sand rat PDX1 proteins are reported as mean ± standard deviation.

4. Results

4.1 Altered stability of gerbil PDX1 protein

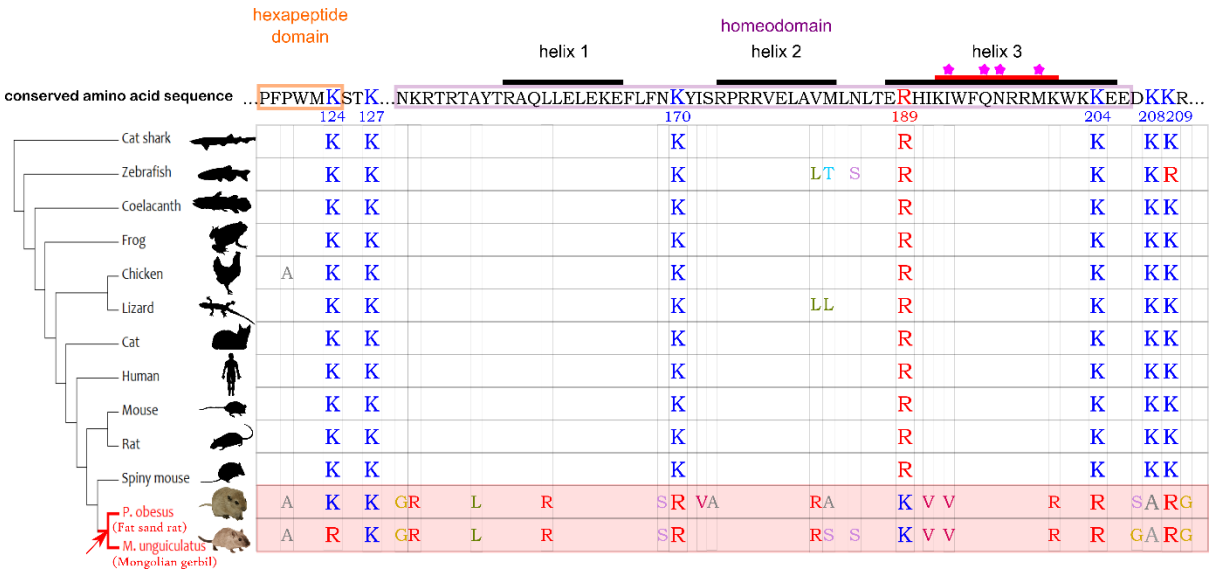


Fig. 3.1 Alignment of PDX1 hexapeptide and homeodomain across vertebrates. The human and mouse amino acid sequence, conserved in most vertebrates, is shown at the top. The three alpha helices are labelled, and key amino acids known to directly interact with DNA bases are highlighted with a star[181]. Lysine/Arginine residues key to the current research are shown for all species. For other residues, only those that deviate from the conserved sequence are shown.

The divergence of sand rat and Mongolian gerbil PDX1 homeobox sequence, compared with the usual extreme conservation (Fig. 3.1), raises the question of whether the gene is non-functional in these species. At the genomic level, previous work has shown that complete loss of gerbil PDX1 function is unlikely given that the *Pdx1* coding sequence is depleted in polymorphisms relative to the 3' untranslated region, suggesting presence of purifying selection[102]. We also note that five of the six hexapeptide residues are unchanged. As for the homeodomain, the amino acid changes in helix 3, which makes close contact with the major groove of DNA[181], are between chemically similar amino acids; there is one chemically significant change (Leu to Arg) in helix 1 (Fig. 3.1). This pattern of change suggests that the gene is functional, and that purifying selection has resisted substitutions of residues that affect interactions with DNA. The non-helix sites that have changed in gerbils are also normally conserved, but in these cases, substitutions driven by GC skew have somehow been tolerated in gerbils only. To test if substitutions are adaptive, we analysed the hexapeptide and homeodomain regions of the 13 species shown in Fig. 3.1 ,and fitted site, branch, and branch-site models using codeml in PAML[182]. We found no support for positive selection occurring within the gerbil lineage (Supplementary S3.17). A two-ratio branch model gave a dN/dS ratio of 0.0197 (dN = 0.1280, dS = 6.0599) on the gerbil lineage, indicating a high rate of synonymous substitution and presence of purifying selection. We suggest the observed amino acid changes in gerbil PDX1 are maladaptive changes.

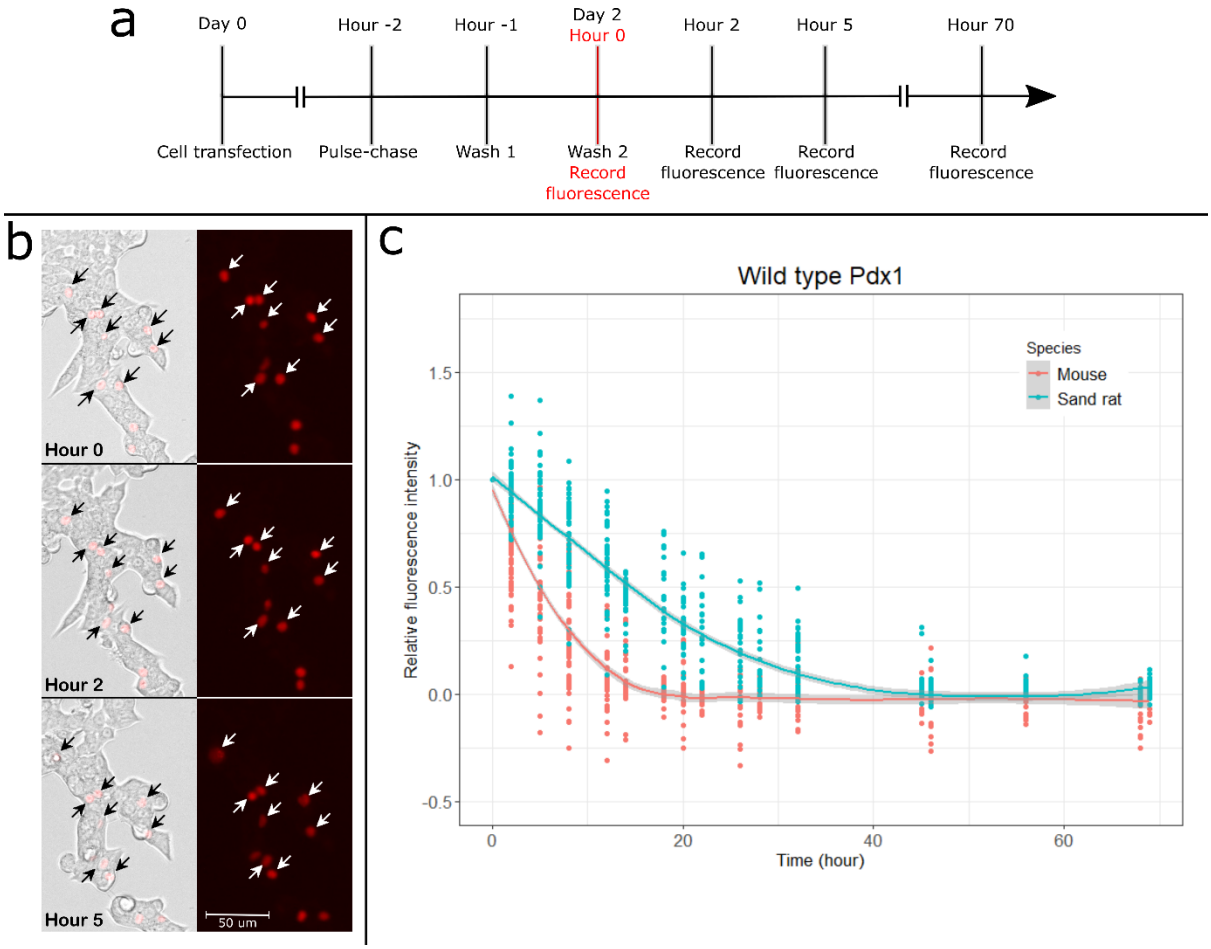


Fig. 3.2 Measurement of protein stability. (a) Experimental design showing timing of cell transfection, pulse-chase, and fluorescence recording. (b) An example of cell tracking over 5 h. The labelled cells are overexpressing mouse PDX1-SNAP. (c) Intensity decline for mouse PDX1 and sand rat PDX1 proteins. Each dot represents an individual cell. The curve shown was generated using the Local Regression (loess) function in the ggplot2 package in R[183].

To assess if PDX1 protein stability has been altered in gerbil evolution, we conducted pulse-chase labelling experiments of ectopically expressed proteins in a rat pancreatic cell line (INS-1). To allow tracking of protein degradation in individual cells, both mouse and sand rat *Pdx1* genes were cloned into pSNAPf vectors, after first lowering the GC content of the sand rat codons ("mousifying") to optimize translation efficiency (Supplementary S3.1). These constructs produce PDX1 fusion proteins with a C-terminal SNAP-tag motif, which will link covalently to a red fluorescent tag (SNAP-Cell TMR-Star) enabling protein stability to be measured after washing out unbound red tag. After transfection into INS-1 cells, pulse-chase was carried out and fluorescence intensity of individual transfected cells recorded at intervals up to 70 h (Fig. 3.2a). Individual cells were identified under brightfield imaging, and cells that exhibited ambiguity in identity were excluded from analysis (Fig. 3.2b). Data from 89 (n = 43 transfected with mouse *Pdx1*, n = 46 transfected with sand rat *Pdx1*) individual cells were used to plot decrease in intensity per cell, revealing a faster rate of intensity decline for mouse PDX1 than for sand rat PDX1 protein (Fig. 3.2c).

To quantify the difference in protein stability, we modelled the decrease of fluorescence intensity using the Linear and Nonlinear Mixed Effects (nlme) Models package in R[179]. This revealed a significant difference between the degradation pattern of mouse PDX1 and sand rat PDX1 proteins (p-value = 0; Supplementary S3.12). Based on this model, we estimate the half-life of mouse PDX1 protein to be 4.7 ± 0.4 h, and the half-life of sand rat PDX1 protein to be 13.7 ± 0.3 h.

4.2 Sand rat PDX1 protein degradation is regulated by the UPS

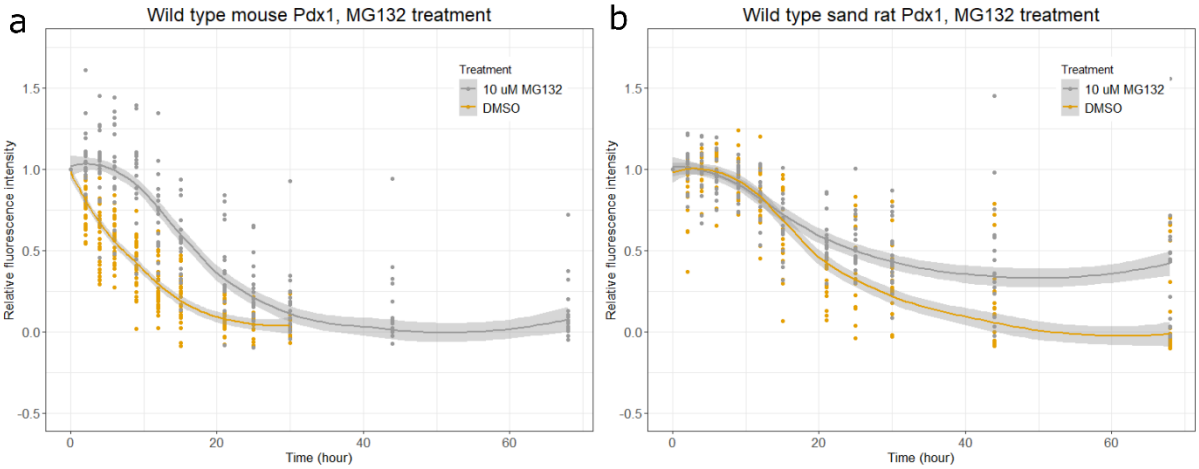


Fig. 3.3 Effect of inhibiting UPS on PDX1 protein stability. Graphs show the decline in fluorescence intensity for (a) mouse PDX1 and (b) sand rat PDX1 in cells treated with DMSO (control) or 10 uM MG132 (inhibitor of UPS), in both cases showing shift of protein decay curve. Statistical modelling confirmed that in both cases the effect is significant (mouse p-value = 1e-04; sand rat p-value = 0.0007); the PDX1 protein half-lives were extended by 10.8 ± 0.9 h (mouse) and 11.3 ± 1.7 h (sand rat).

It has previously been shown that mouse PDX1 protein stability is regulated by the UPS[176]. To test if sand rat PDX1 is regulated by the same biochemical pathway, we treated transfected INS-1 cells with MG132, a reversible inhibitor of the 26S proteasome[184]. MG132 treatment slows the overall rate of mouse PDX1 protein degradation, shifting the best-fit curve to the right (Fig. 3.3a) (p-value = 1e-04, Supplementary S3.13). This is consistent with previous findings that mouse PDX1 protein stability is regulated via UPS. We find that MG132 treatment also slows the rate of sand rat PDX1 degradation pattern (Fig. 3.3b) (p-value = 0.0007, Supplementary S3.14), although less sharply compared with mouse PDX1. Differences in processes prior to proteasome degradation, for example, E3 ligase efficiency in recognizing and tagging the target PDX1 protein, could contribute to subtle differences in how proteins respond to proteasome inhibition[185].

4.3 PDX1 ubiquitination sites suggest compensatory molecular evolution

Although ubiquitin can sometimes be added to the peptide N-terminus or to cysteine residues, it is only added to lysine residues in the context of targeting for degradation in proteasomes[186]. We therefore asked which lysine residues were utilized by the UPS in regulating stability of mouse and sand rat PDX1 proteins. We first used prediction software to deduce which residues

had the highest probability of being a functional ubiquitination site[187–189]. In addition, because ubiquitination acts through covalent addition of a moiety to lysine residues, we excluded key lysine residues in the homeodomain that have been proven to engage in interactions with DNA[181]. This resulted in seven lysine residues of interest: four in the conserved PDX1 protein (mouse, human, and most vertebrates) and three in the divergent sand rat protein.

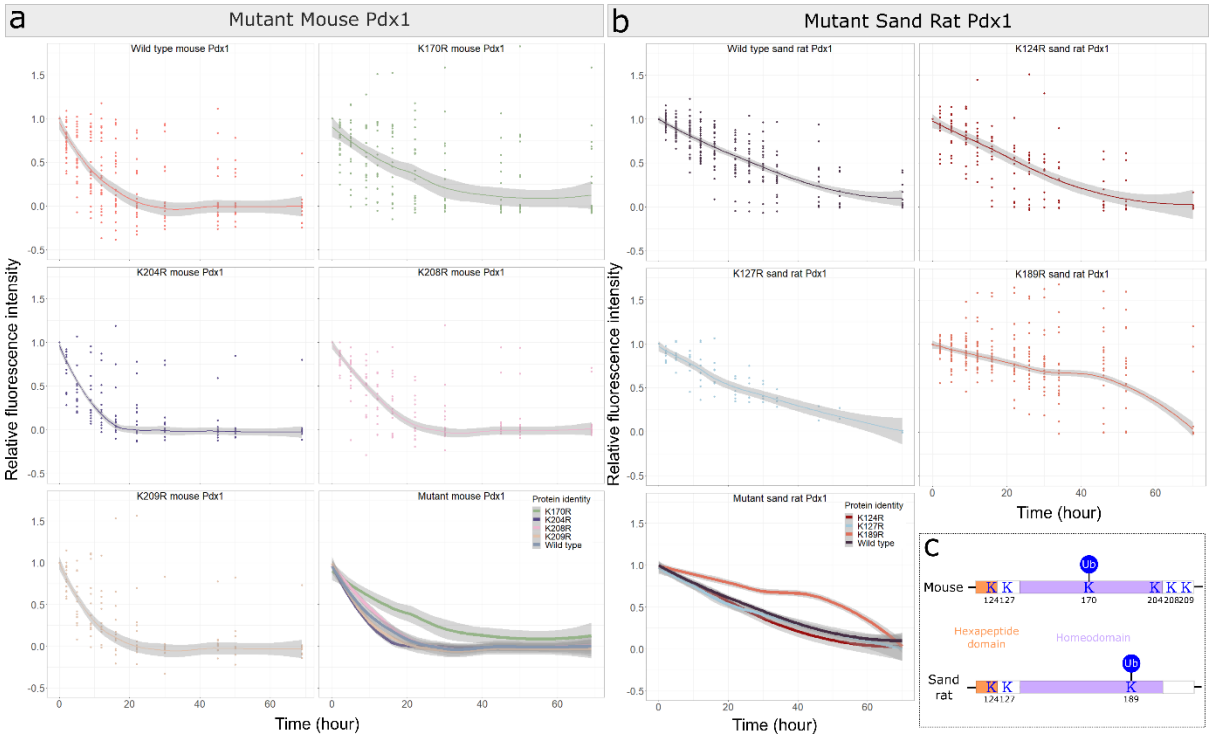


Fig. 3.4 Protein stability in wild type and mutant proteins. (a) Graphs show decline in fluorescence intensity for mouse wild type and individual mutant proteins, with a merged graph in the bottom right corner. (b) Graphs showing decline in fluorescence intensity for sand rat wild type and individual mutant proteins, with a merged graph in the bottom left corner. (c) Summary of ubiquitination site location in mouse and sand rat PDX1.

We used *in vitro* mutagenesis to replace singly each lysine codon of interest with a codon for arginine, an amino acid with similar polarity but which cannot be tagged by ubiquitin. For mouse PDX1, we generated mutant proteins K170R, K204R, K208R, and K209R with C-terminal SNAP-tags (numbering according to mouse PDX1). These constructs were transfected into INS-1 cells and protein stability measured through pulse-chase labelling with SNAP-Cell TMR-Star. As shown in Fig. 3.4a, mutant proteins K204R, K208R, and K209R did not have significantly different half-lives compared with the wild type control (p-value = 0.1487, 0.1341, and 0.5847, respectively; Supplementary S3.15). However, mutant protein K170R showed a very different protein decay pattern, with a much flatter best-fit decay curve and a longer half-life (p-value = 0.0267; Supplementary S3.15). This reveals that lysine residue 170 is the major PDX1 ubiquitination site. We suspect it is not the only ubiquitination site as the protein is still gradually degraded within each cell.

For sand rat PDX1, we generated SNAP-tagged mutant proteins K124R, K127R, and K189R, transfected into INS-1 cells and measured protein stability (numbering according to alignment

with the mouse PDX1 protein). As shown in Fig. 3.4b, mutant proteins K124R and K127R did not have drastically different stability compared with wild type control (P = 0.9649 and 0.5538, respectively; Supplementary S3.16). Only mutant K189R displayed a completely different degradation pattern compared with the other two mutant proteins and the wild type (P = 0.0000; Supplementary S3.16). For K189R, high levels of fluorescence were detected in cells throughout the experimental period, with many cells exhibiting rounded shapes and detaching from the substrate around hour 50. This radical disruption of regulated protein decay indicates that the K189 is the ubiquitination site for sand rat PDX1. Mutation of this site abolishes PDX1 degradation via the UPS pathway.

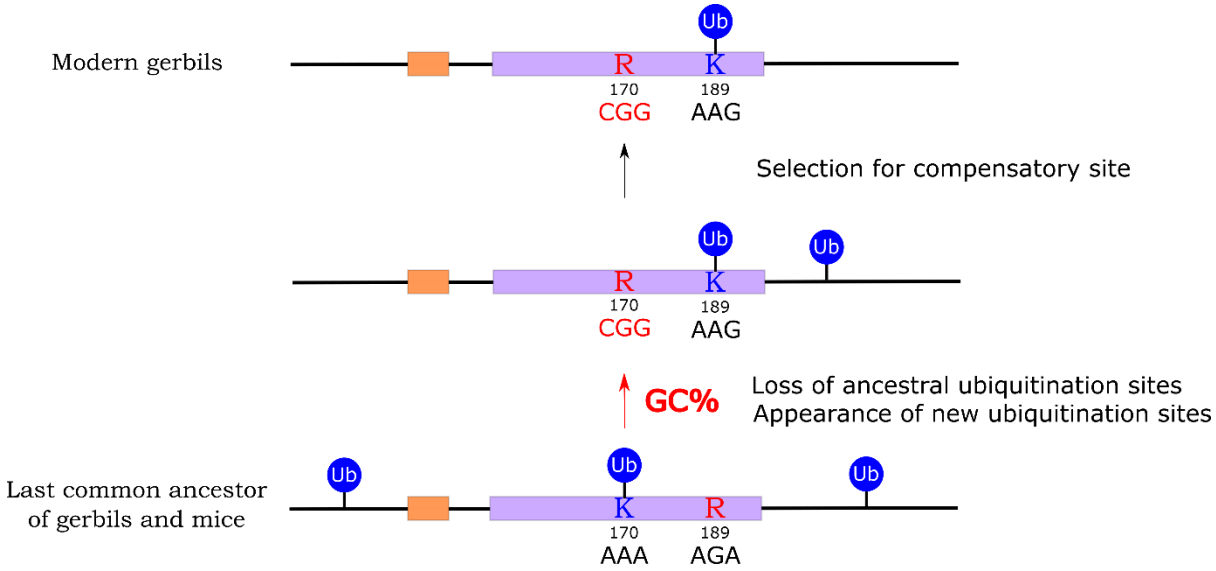


Fig. 3.5 Summary of the possible evolutionary pathway proposed. GC composition bias may be associated with emergence of a compensatory ubiquitination site and loss of conserved sites in gerbil PDX1.

Amino acid 170, which we identify as the major functional ubiquitination site in the mouse PDX1 protein, has been conserved across 420 My of jawed vertebrate evolution (Fig. 3.1). This K170 site is invariant across vertebrates except some teleost species and the gerbil species discussed here. In sand rat and Mongolian gerbil, the codon has undergone a unique, GC skew associated mutation to arginine (Fig. 3.1). In contrast, degradation of sand rat PDX1 is regulated through ubiquitination of K189, a site which across the rest of jawed vertebrate evolution is otherwise conserved as arginine. These results reveal a unique shift in the location of the ubiquitination site in PDX1 along the evolutionary lineage leading to gerbils (Fig. 3.4c). I propose that under extreme GC skew, the ultra-conserved lysine residue 170, possibly along with other ubiquitination sites, was removed by mutation and as a compensatory mechanism natural selection favoured an otherwise maladaptive mutation elsewhere in the protein sequence (Fig. 3.5). The fixation of the “new” lysine 189 is consistent with natural selection acting in the face of strong GC skew.

5. Discussion

The sand rat and the Mongolian gerbil are both desert-dwelling rodents, the former native to North Africa and the Middle East[92] and the latter to Mongolia and China[190]. Both are members of the Gerbillinae subfamily that diverged from the Murinae sub-family, including mice and rats, 18 ± 1.8 Ma[63]. The sand rat has been of particular interest in studies of diabetes since both wild-caught and laboratory-reared animals can exhibit insulin resistance, increased body weight and symptoms of type 2 diabetes after being fed a high calorie diet[92,191]. It had been hypothesized that these characteristics are associated with loss of the *Pdx1* gene in the gerbil subfamily[192], but recent work revealed that rather than missing the gene, sand rats and Mongolian gerbils possess a highly divergent *Pdx1*[102].

Two features make the sequence divergence of gerbil *Pdx1* particularly intriguing. First, the results of Chapter 2 show that the divergence of this normally highly conserved protein is an unique feature of gerbils (Fig. 2.1), with nothing similar seen in other vertebrates, including arid-adapted rodents such as the closely related spiny mouse (*Acomys cahirinus*). Second, the divergence of the gene is clearly associated with a strongly pronounced GC skew that encompasses several genes in a large genomic region[102]. In this chapter, my working hypothesis to explain these findings is that focused and extreme GC-biased gene conversion (gBGC) has driven accumulation of GC nucleotides throughout this region, generating maladaptive substitutions. If correct, this hypothesis predicts that natural selection would act to fix compensatory mutations, both within and outside the GC skewed region, to restore biological function.

Before considering the existence of compensatory mutations, I first ask if amino acid changes in gerbil PDX1 are indeed maladaptive. As shown in both Chapter 2 and Fig. 3.1 in this chapter consistent with multiple lines of evidence from comparative analysis, large changes to the homeodomain similar to that observed in gerbil have not been tolerated across the rest of vertebrate evolution or indeed across the Bilateria. In addition, I do not find support for positive selection acting on the hexapeptide region and homeodomain in the gerbil lineage. Direct tests of function, such as gene deletion or experimental manipulation in gerbils, would not answer the question if compensatory changes have occurred elsewhere. Gene replacement in mouse would not give a clear answer for the same reason. Instead, I took a biochemical approach in this chapter and asked whether protein stability was altered, as reported for some human variants associated with disruption of pancreatic function. Using a cell culture expression system, I found that sand rat PDX1 does have a different half-life to mouse PDX1, but surprisingly the change was in the opposite direction from the original expectation. Although two different single amino acid changes in human PDX1 cause decreased half-life, the radically altered sand rat protein has a longer half-life. Perhaps more importantly, the degradation of sand rat PDX1 protein follows a steady curve, indicative of actively regulated degradation rather than unregulated decay, and I obtained experimental verification that protein stability is regulated by the UPS. This regulated degradation is another indication that the sand rat PDX1 gene encodes a functional protein under selection rather than being a non-functional pseudogene. Although these assays were conducted in a heterologous system (a rat cell line), I argue that this is not a confounding factor since the key mediators of PDX1 degradation, PDX1

C-terminal inhibiting factor 1 (PCIF1) and the scaffolding protein CUL3[176], are highly conserved between rat, mouse, and sand rat (>97% protein similarity between all three species).

To investigate candidate maladaptive mutations, I mapped functional sites responsible for UPS-regulated degradation of PDX1. Mutagenesis experiments on mouse PDX1 revealed that one highly conserved lysine, K170, is the critical residue for regulated degradation by UPS. Interestingly, this residue has mutated to an arginine in gerbils and could no longer be used as a site for ubiquitination. Loss of this key ubiquitination site is not seen in any other tetrapod lineage and is likely to have been driven by the gBGC phenomenon described above. However, a new ubiquitinated lysine site has evolved in gerbil PDX1 proteins at a different site in the same protein domain, and it is this new site that enables UPS-regulated degradation. I propose that this has shown that maladaptive change to a key conserved amino acid residue can be compensated for by emergence of a new site fulfilling a similar function. Without further functional tests, I cannot state that the sand rat PDX1 protein is as efficient or "optimal" as the conserved or ancestral sequence present before the emergence of the hotpot of GC skew. Indeed, amino acid changes at other residues may affect other biochemical properties. However, the results of this chapter support the hypothesis that GC skew gradually altered the sequence of critical genes during gerbil evolution while natural selection favoured substitutions elsewhere in these proteins to compensate for loss of critical biochemical function (Fig. 3.5). Changes to sand rat PDX1 half-life caused by alteration of key ubiquitination sites has given an insight into the power of natural selection even in the face of extreme and unusual nucleotide bias.

6. Supplementary information

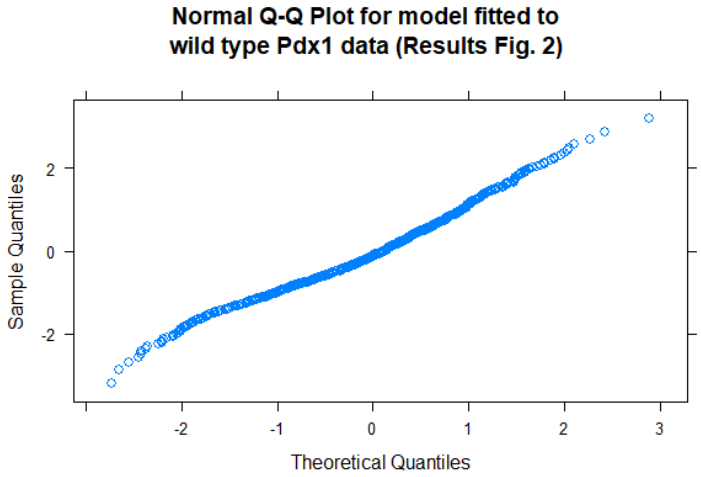
>Sand_rat_original_Pdx1_cds

ATGGACAGAGAGGCCGAGCCCTTCTTCGAGGCCTCCTGGGCGTTCCCGGGGCCCGAGTTCGCGGCCCCCGCCTCC
TCCTGCCTGTTCGAGGGTGGGGGCGGGCAGCCTCCCCCCCACGCTCCTCCCCACGCTCCTCCCCACCTCGCCCCG
TGCTCCCTGGACCCCACCGGCCTCCAGCCGCCCCAGCCCGGGGTCCCCCCGCCGCCACCCGGGGGCCCCGACCAA
CCGCCCTTTGCCTGGATGAAGAGCAGCAAAGGCCAAGCCTGGAGCGGCCAGTGGGCAGCCCCCGGCCGCAGGACT
CGAACCCTGTACACGCGGGCGCAGCGGCTGGAGCTGGAGAAGGAATTCCTCTTCAGCCGCTACGTCGCGCGGCCG
CGGCGCGTGGAGCTCGCGCGGGCGCTGAACCTCACCGAGAAGCACGTGAAGGTCTGGTTCCAGAACCGCCGCATG
CGCTGGAAGAGGGAGGAGTCCGCGCGGGGGAGGACGGCGCCCCGGGAGGACGGAGGAGCGGGAGGCTCCCCGCCA
CCGTCCTCCTCCTCCTCCTCCTCCTCCTCCGTGGCCCGGGATGCTCCTCCTCTTCTTCTCCCTCCTCCTCCTCCT
CCTACGGGGGGACTGCGGTGA

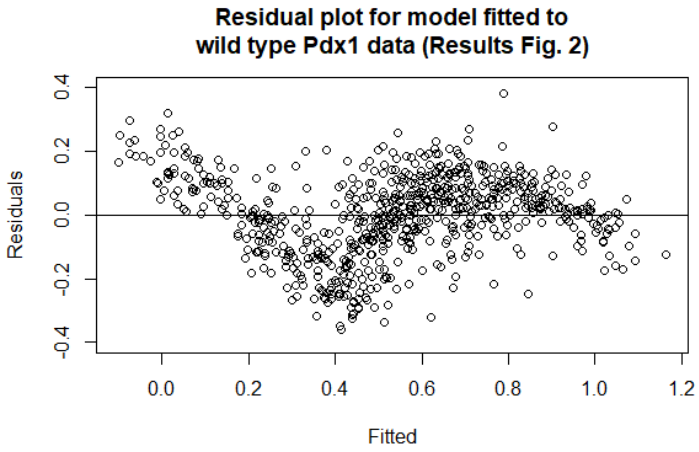
>Sand_rat_mousified_Pdx1_cds

ATGGACAGAGAAGCTGAACCTTTCTTCGAAGCAAGTTGGGCATTCCCTGGTCCAGAGTTCGCGGCTCCCGCCTCC
AGTTGCCTGTTCGAGGGTGGAGGTGGACAGCCTCCACCTCACGCTCCTCCACACGCTCCTCCACACCTCGCCCCA
TGCTCCCTGGACCCAACCGGCCTCCAGCCACCTCAGCCAGGAGTCCCTCCGCCACCACCTGGAGGCCCTGACCAA
CCTCCTTTTGCCTGGATGAAAAGCAGCAAAGGCCAAGCCTGGAGCGGCCAGTGGGCAGCCCCTGGCCGCAGGACT
CGAACACTGTACACTCGGGCGCAGCGGCTGGAGCTGGAGAAGGAATTCTTATTTAGCCGCTACGTCGCGCGGCCT
CGGCGCGTGGAGCTCGCACGGGCGTTGAACTTGACCGAGAAGCACGTGAAAGTCTGGTTCCAAAACCGTCGCATG
CGCTGGAAAAGGGAGGAATCCGCGCGGGGAAGGACGGCGCCTCGGGAAGACGGAGGAGCGGGAGGTTCTCCTCCA
CCTAGTAGTTCAAGCTCTAGTTCCAGCTCCGTGGCCCGGGATGCTCCTCCACTTCTTCTCCCTCCACCACCTCCG
CCAACGGGTGGACTGCGATGA

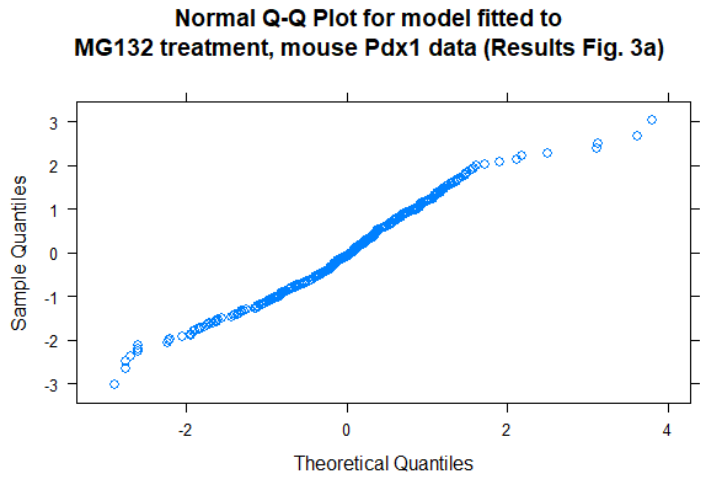
Supplement S3.1 Coding sequence of sand rat *Pdx1* and "mousified" sand rat *Pdx1*. The original sand rat coding sequence is 72.9% GC while the "mousified" coding sequence has 63.0% GC. Both sequences are translated into the same amino acid sequence.



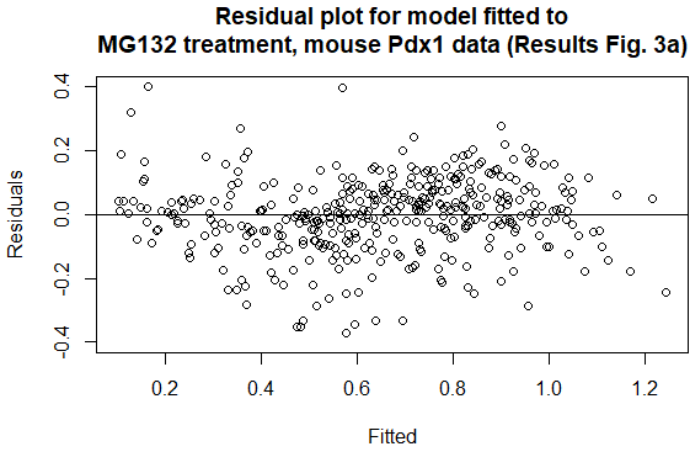
Supplement S3.2 To analyse the normality of the square-root transformed data, Quantile-Quantile Plots were generated using the generic qqnorm function in R[180]. Axes as in Supplementary S3.2.



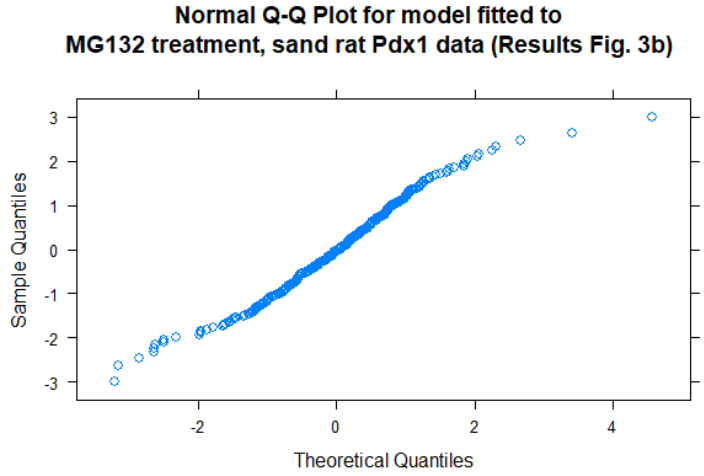
Supplement S3.3 To analyse the homoscedasticity of the square-root transformed data, residual plots were generated using the generic plot function in R[180].



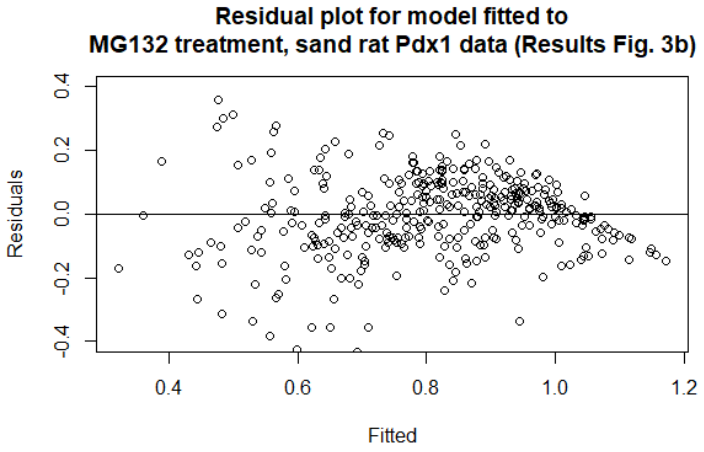
Supplement S3.4 To analyse the normality of the square-root transformed data, Quantile-Quantile Plots were generated using the generic qqnorm function in R[180]. Axes as in Supplementary S3.2.



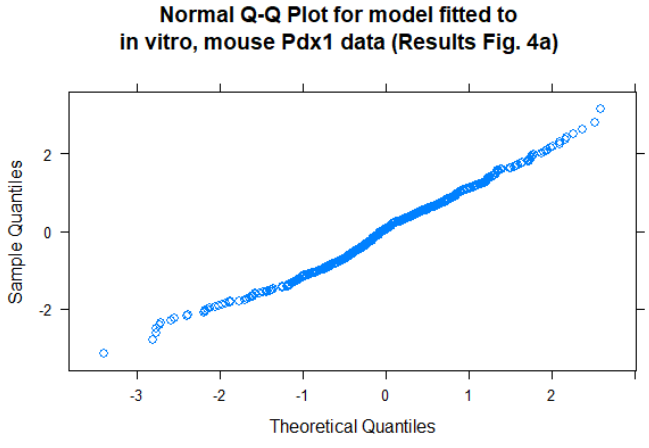
Supplement S3.5 To analyse the homoscedasticity of the square-root transformed data, residual plots were generated using the generic plot function in R[180].



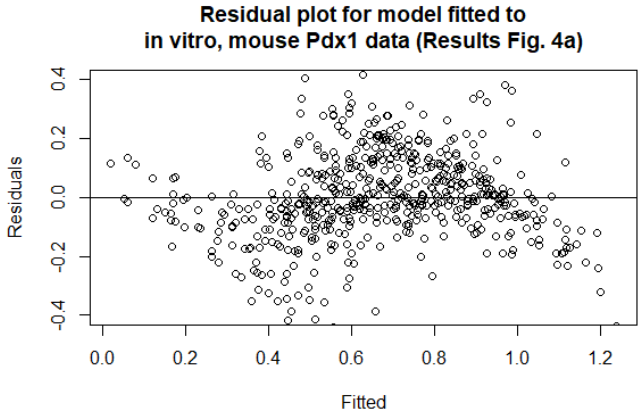
Supplement S3.6 To analyse the normality of the square-root transformed data, Quantile-Quantile Plots were generated using the generic qqnorm function in R[180]. Axes as in Supplementary S3.2.



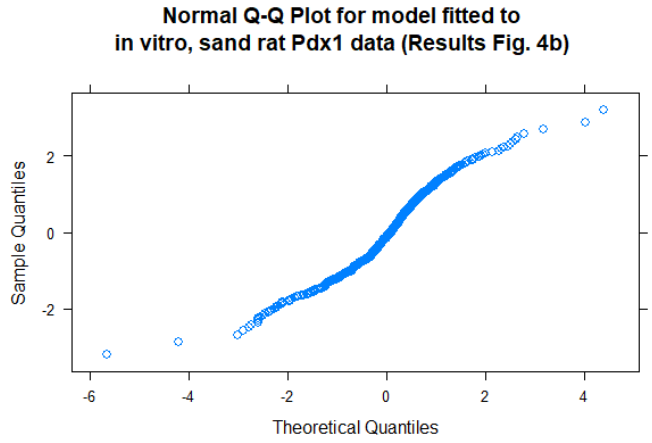
Supplement S3.7 To analyse the homoscedasticity of the square-root transformed data, residual plots were generated using the generic plot function in R[180].



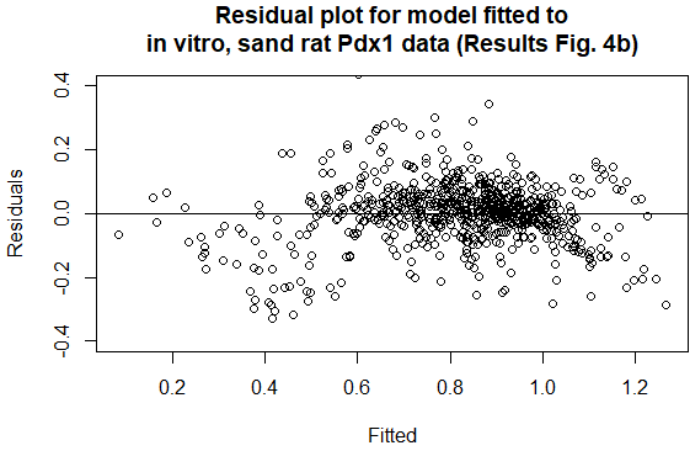
Supplement S3.8 To analyse the normality of the square-root transformed data, Quantile-Quantile Plots were generated using the generic qqnorm function in R[180]. Axes as in Supplementary S3.2.



Supplement S3.9 To analyse the homoscedasticity of the square-root transformed data, residual plots were generated using the generic plot function in R[180].



Supplement S3.10 To analyse the normality of the square-root transformed data, Quantile-Quantile Plots were generated using the generic qqnorm function in R[180]. Axes as in Supplementary S3.2.



Supplement S3.11 To analyse the homoscedasticity of the square-root transformed data, residual plots were generated using the generic plot function in R[180].

Fixed effects	**df**	**t Value**	**p Value**
Intercept	625	38.32785	0
Type: Sand rat	87	12.22414	0
Sqrt(Time)	625	-23.24835	0
Type: Sand rat × Sqrt(Time)	625	-5.24269	0

Supplement S3.12 Parameters reported for the linear mixed effects model generated using data shown in Fig. 3.2. Table shows degrees of freedom, t value, and p value of fixed effects sqrt(Time), type of species, and their interaction. Half-life calculations were obtained using this model.

Fixed effects	**df**	**t Value**	**p Value**
Intercept	350	38.85098	0e+00
Type: Treatment	44	3.67540	6e-04
Sqrt(Time)	350	-22.66775	0e+00
Type: Treatment × Sqrt(Time)	350	3.96114	1e-04

Supplement S3.13 Parameters reported for the linear mixed effects model generated using data shown in Fig. 3.3a. Table shows degrees of freedom, t value, and p value of fixed effects sqrt(Time), type of species, and their interaction. Half-life calculations were obtained using this model.

Fixed effects	**df**	**t Value**	**p Value**
Intercept	325	36.56956	0.0000
Type: Treatment	33	-0.98462	0.3320
Sqrt(Time)	325	-16.39190	0.0000
Type: Treatment × Sqrt(Time)	325	3.40934	0.0007

Supplement S3.14 Parameters reported for the linear mixed effects model generated using data shown in Fig. 3.3b. Table shows degrees of freedom, t value, and p value of fixed effects sqrt(Time), type of species, and their interaction. Half-life calculations were obtained using this model.

Fixed effects	**df**	**t Value**	**p Value**
Intercept	480	18.953975	0.0000
Type: K170R	89	-0.153405	0.8784
Type: K204R	89	-0.444863	0.6575
Type: K208R	89	1.037924	0.3021
Type: K209R	89	0.336021	0.7376
Sqrt(Time)	480	-9.084160	0.0000
Type: K170R × Sqrt(Time)	480	2.223133	0.0267
Type: K204R × Sqrt(Time)	480	-1.446534	0.1487
Type: K208R × Sqrt(Time)	480	-1.500572	0.1341
Type: K209R × Sqrt(Time)	480	-0.546916	0.5847

Supplement S3.15 Parameters reported for the linear mixed effects model generated using data shown in Fig. 3.4a. Table shows degrees of freedom, t value, and p value of fixed effects sqrt(Time), type of treatment, and their interaction. Half-life calculations were obtained using this model.

Fixed effects	**df**	**t Value**	**p Value**
Intercept	626	45.08517	0.0000
Type: K124R	84	-1.01502	0.3130
Type: K127R	84	-0.67135	0.5038
Type: K189R	84	-3.28740	0.0015
Sqrt(Time)	626	-21.83059	0.0000
Type: K124R × Sqrt(Time)	626	0.04408	0.9649
Type: K127R × Sqrt(Time)	626	0.59238	0.5538
Type: K189R × Sqrt(Time)	626	11.01595	0.0000

Supplement S3.16 Parameters reported for the linear mixed effects model generated using data shown in Fig. 3.4a. Table shows degrees of freedom, t value, and p value of fixed effects sqrt(Time), type of treatment, and their interaction. Half-life calculations were obtained using this model.

Model	**Parameters specified**	**lnL**	**ntime**	**np**
Site models				
M0 (one-ratio)	NSsites = 0, model = 0	-1034.558186	23	25
M1a (nearly neutral)	NSsites = 1, model = 0	-1034.558846	23	26
M2a (positive selection)	NSsites = 2, model = 0	-1037.646198	23	28
M7 (beta)	NSsites = 7, model = 0	-1025.473872	23	26
M8 (beta and ω)	NSsites = 8, model = 0	-1025.474532	23	28
Branch models				
Two-ratio model	NSsites = 0, model = 2	-1034.061948	23	26
Branch-site models				
Model A1 (null model)	NSsites = 2, model = 2, fixomega = 1, omega = 1	-1031.077597	23	27
Model A (alternative model)	NSsites = 2, model = 2, fixomega = 0	-1031.077597	23	28

Supplement S3.17 Parameters reported for PAML analysis (version 4.9)[182] generated using sequence alignments of *Pdx1* hexapeptide region and homeodomain from the 13 vertebrate species shown in Fig. 3.1. The tree used for analysis is also shown in Figure 3.1. First, we tested for positive selection on sites in the hexapeptide region and homeodomain by comparing a null model (M0, one ω for all sites) to a nearly neutral model (M1a, two classes of sites $0 < \omega < 1$ and $\omega = 1$), and comparing M1a to positive selection (M2a, three classes of sites $0 < \omega < 1$, $\omega = 1$, $\omega > 1$). We also compared the beta model (M7, multiple categories $0 < \omega < 1$ and $\omega = 1$) to beta with positive selection (M8, multiple categories $0 < \omega < 1$, $\omega = 1$, plus $\omega > 1$). Models M2a and M8 did not significantly outperform models M1a and M7. Second, we tested for presence of positive selection on the gerbil lineage. We compared the null model (M0) to a two-ratio branch model (gerbil lineage as foreground); the latter did not significantly

outperform the null. Third, we used branch-site models to test for positive selection on the gerbil lineage at different sites within *Pdx1*. We compared a branch-site null model (A1, allowing $0 < \omega < 1$, $\omega = 1$) to a branch-site alternative model (A, allowing $0 < \omega < 1$, $\omega = 1$, and $\omega > 1$); the latter did not significantly outperform the null model. Significance was calculated using the Likelihood Ratio Test (LRT)[193]. The analysis showed that models allowing positive selection did not significantly outperform models not allowing positive selection.

Chapter 4:

Sand rat PDX1 inefficiently regulates insulin expression in rat cells

Statement of Author Contributions

The original idea for the project was conceived by discussion between Yichen Dai, Thomas Dunwell and Peter W. H. Holland. TD provided general mammalian cell culture advice in the initial stages of this work and Tatiana Alfonso Perez provided advice on CRISPR/Cas9 editing in mammalian cells. TAP also provided pX8555 plasmids containing the blasticidin resistance gene and pX8562 plasmids containing puromycin resistance gene, from which each gene was cloned using PCR. Sonia Trigueros provided microscopy advice and access to the ZOE Fluorescent Cell Imager (Bio-Rad). Amy Royall and Matias De Vas in the Ferrer group at Imperial College London provided advice on chromatin immunoprecipitation in mammalian cells. YD performed all experimental and computational analysis. PWHH supervised the work of YD throughout the length of this study. Helpful discussions were held with Adam Hargreaves, AR, and Tom Lewin.

1. Abstract

Large regions in the sand rat genome show accumulation of GC which impact the coding sequence of key genes, for example *Pdx1*. The most conserved region in the PDX1 protein is the homeodomain, the site at which this transcription factor recognizes and binds to DNA, and it is 100% identical between the human, chicken, frog, and shark. However, the sand rat PDX1 homeodomain is only 75% conserved. One of the key target genes of the PDX1 protein is known to be *insulin*, and the PDX1 homeodomain region is known to bind to the insulin promoter in multiple vertebrate species. In this chapter, I test whether sand rat PDX1 is capable of regulating rat *insulin* gene expression and whether any functional differences exist in comparison to the mouse or rat PDX1 protein. I use methods including mammalian cell transfection, RNA extraction and qPCR to compare PDX1 regulation of *insulin* gene expression in an overexpression system, and I use CRISPR/Cas9 gene editing to create an endogenous expression system allowing me to perform V5-tag mediated chromatin immunoprecipitation followed by qPCR to compare PDX1 binding to the rat insulin promoter. My results show that sand rat PDX1 is less responsive to changes in glucose concentration and cannot induce *insulin* gene expression changes that are comparable to the mouse PDX1 protein. This may be caused by failure of sand rat PDX1 to correctly recognize and bind to the rat insulin promoter. At the same time, the predicted sand rat insulin promoter PDX1-binding region only differs from the conserved rodent sequence by one nucleotide, suggesting that the amino acid changes in the sand rat PDX1 homeodomain may have led to compromised ability to regulate *insulin* gene expression. The results of this chapter provide evidence comparing sand rat PDX1, mouse PDX1 and rat PDX1, suggesting that the amino acid changes caused by GC-skew are likely maladaptive and contributing to the diabetes-prone phenotype observed in some sand rats.

2. Introduction

Aside from its central role in pancreatic development during embryogenesis, PDX1 is crucial for maintaining β-cell function in adulthood[85,110,118,136]. One of the key functions of PDX1 is direct regulation of insulin gene expression, which can be reflected by the protein's earlier name Insulin Upstream Factor-1 (IUF-1)[114,118]. As introduced in Chapter 1, experiments using human, rat, and mouse islet cells have shown direct binding of PDX1 to the insulin promoter, and the PDX1 binding site in the insulin promoter has been confirmed to be the A elements, AT-rich regions that have a core TAAT sequence (with the exception of the A2 element which contains AAAT) [132–134]. TAAT is the essential recognition site of the PDX1 homeodomain. In addition, although the PDX1 protein interacts with all A elements along the insulin promoter, the key elements for PDX1-mediated insulin gene expression are elements A3 and A4[135]. Like other transcription factors, PDX1 regulation of insulin gene expression is most efficient in presence of other transcription cofactors. Previous studies have shown that PDX1 interacts with the basic-helix-loop-helix (bHLH) protein E47 and Beta2/NeuroD to regulate insulin gene expression in human and rodent islet cells[194,195]

PDX1 binding to the insulin promoter is a tightly controlled process, and PDX1 can activate or suppress insulin expression under different conditions. Glucose concentration is the most well-studied factor affecting the outcome of PDX1 binding to the insulin promoter, with high glucose conditions (20 mM or above) leading to more than 2-fold increases in PDX1 binding affinity compared to low glucose conditions (3 mM or below)[116,136]. Change in glucose concentration has been shown to rapidly alter the sumoylation profile of the PDX1 protein, and given changes in kinase/phosphorylase activity, it has also been long speculated that change in PDX1 phosphorylation can also be a method through which glucose can alter PDX1 activity at the insulin promoter[196,197]. Studies have also confirmed that PDX1 interacts differently with other proteins under different glucose conditions. PDX1 recruit histone acetyltransferase p300 and its close homologue CBP to the insulin promoter under high glucose conditions, leading to acetylation of H4 at the promoter and increased transcription activity[138]. Under low glucose conditions, biochemical modifications on PDX1 change and it prefers interaction with histone deacetylases Hdac-1 and Hdac-2, which in turn deacetylate H4 and decrease transcription activity at the insulin promoter[137].

Given its central role in mediating glucose stimulated insulin expression, PDX1 mis-regulation and mutation can lead to abnormal insulin expression and become a significant contributor to diabetes development[135]. Under chronic hyperglycaemic conditions, *Pdx1* gene expression is severely decreased and in turn, insulin expression becomes insufficient[198]. In addition, mutant versions of PDX1 associated with type II diabetes, C18R, D76N and R197H, cannot fully activate the insulin promoter under high glucose conditions[81].

In this chapter, I ask whether the highly divergent sand rat PDX1 protein is functionally different from the conserved PDX1 protein, and I focus on the interaction between PDX1 and the insulin promoter. First, I use an ectopic overexpression system in a rat pancreatic cell line to test whether sand rat PDX1 activation of insulin expression is comparable to mouse and rat PDX1 under different glucose conditions. Second, I use CRISPR/Cas9 mediated gene editing

to compare sand rat and rat PDX1 proteins expressed at endogenous levels, focusing on the interaction between PDX1 and the rat insulin I gene promoter.

3. Materials and methods

3.1 Construction of plasmid constructs for ectopic overexpression

The full cds for mouse *Pdx1* was cloned into the pSF-CMV-Ub-daGFP-Ascl (OXGENE #OG244, now discontinued) vector using digestion sites EcoRV and XhoI. The full cds for sand rat *Pdx1* was first "mousified" to ensure the average GC% was comparable to the mouse *Pdx1* gene and then cloned into the pSF-CMV-Ub-daGFP-Ascl vector (referred to as pSF) using digestion sites XbaI and XhoI. Sanger sequencing was carried out to verify the success of each cloning experiment. The exact DNA sequence for each gene can be found in the supplementary section of this chapter.

3.2 Mammalian cell culture and transfection

The rat secondary pancreatic cell line INS-1 832/13 was used for all cell culture experiments in this chapter. Cells were cultured in RPMI-1640 medium (ThermoFisher #11875093) supplemented with 10% fetal bovine serum (ThermoFisher #10500064) with a final concentration of 100 U/ml penicillin/streptomycin (ThermoFisher #15140122), 500 uM β-mercaptoethanol (ThermoFisher #31350010), 1 mM sodium pyruvate (ThermoFisher #11360039) and 10 mM HEPES (ThermoFisher #15630056). Cells were maintained in an incubator set to 37℃ with 5% CO_2 and passaged every four days by a ratio of 1:4. Only cells that were passaged less than 30 times were used for the experiments in this chapter.

Lipofection-mediated cell transfection was carried out using the Lipofectamine 3000 kit (ThermoFisher #L3000015) according to the instruction manual. In brief, 800,000 cells in one well of a 6-well plate was transfected 6 hours after being transferred to the well, and a mixture of 7.5 ul Lipofectamine 3000 in 250 ul Opti-MEM medium (ThermoFisher #31985047) and a mixture of 5 ul P3000 Reagent, 2.5 ug plasmid DNA and 250 ul Opti-MEM medium were combined and added to the well. The lipid-DNA mix was incubated with the cells for 16 hours at 37℃ before it was replaced with fresh culture medium. Cells were kept at 37℃ for another 24 hours before being assessed for transfection efficiency.

Electroporation-mediated cell transfection was carried out using the NEPA21 Electroporator (Nepa Gene) according to the instruction manual. In brief, 1 million cells were detached and washed with Opti-MEM medium prior to being mixed with 5 ug plasmid DNA in a NEPA electroporation cuvette with 2 mm gap (Nepa Gene #EC-002S). Electroporation settings were set at poring pulse 200 V, length 5 ms, interval 50 ms, No. = 2, D. Rate 10%, Polarity + and transfer pulse 20 V, length 50 ms, interval 50 ms, No. = 5, D. Rate 40%, Polarity +/-. Electroporated cells were immediately transferred into a 6-well plate with pre-warmed culture medium and incubated at 37℃ for 48 hours before being assessed for transfection efficiency.

3.3 Glucose treatment, total RNA extraction and qPCR

Cells were treated with two different glucose concentrations for 48 hours following electroporation. High glucose medium (25 mM glucose) was prepared by adding 632 ul stock glucose solution (200 g/L, ThermoFisher #A2494001) to 50 ml RPMI-1640 medium while low

glucose medium (0.5 mM glucose) was prepared by adding 22.5 ul stock glucose solution to 50 ml glucose-free RPMI-1640 medium (ThermoFisher #11879020).

Total RNA extraction from mammalian cells was carried out using the RNeasy Plus Mini kit (Qiagen #74134) following the manufacturer's instructions. In brief, no more than 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) according to the recommended protocol.

Quantitative PCR (qPCR) was carried out using the Luna Universal One-Step RT-qPCR Kit (NEB #E3005) according to the manufacturer's instructions. In brief, the total volume was 20 ul for each reaction, with a final concentration of 1X Luna Universal One-Step Reaction Mix, 1X Luna WarmStart RT Enzyme Mix, 0.4 uM of each primer and 20 ng of DNA-free total RNA. Primer sequences used are specified below.

Table 4.1 qPCR primers used in this chapter to measure expression of different rat genes.

Primer name	Target gene	Primer sequence (5' to 3')
q36B4F	*Rplp0* (36B4)	CAGCAGGTGTTTGACAATGGC
q36B4R	*Rplp0* (36B4)	TGAGGCAACAGTCGGGTAGC
FqINS1	*Ins1*	AAACAGCACCTTTGTGGTCCT
RqINS1	*Ins1*	GGGTCCTCCACTTCACGAC
FqINS2	*Ins2*	GCCCAGGCTTTGTCAAACA
RqINS2	*Ins2*	CTCCCCACACACCAGGTAGAG
qPCR-rat Pdx1 pair 1 F	*Pdx1*	TGGGCAGGAGGTGCATACG
qPCR-rat Pdx1 pair 1 R	*Pdx1*	GATGTGTCTCTCACTCAAGTTGAG
qMafA-2-F	*MafA*	CCCACTGGCCATCGAGTACG
qMafA-2-R	*MafA*	CGCAGAAGCTGGGCGAAG
qSlc2a2-2-F	*Slc2a2*	GCACATACGACACCAGACG
qSlc2a2-2-R	*Slc2a2*	TCCGCCTACTGCAAAGCTG

3.4 Construction of plasmid constructs for CRISPR/Cas9 gene knock-in

For CRISPR/Cas9 gene knock-in, two types of plasmids were constructed. The first type of plasmid should contain the *Cas9* gene and the guide RNA sequence. For this experiment, I designed different guide RNA sequences targeting the start and end of rat *Pdx1* gene exon 2, and I selected plasmid pSpCas9(BB)-2A-GFP (pX458) (Addgene #48138) and inserted the guide RNA sequence using the BbsI digestion site. Guide RNA efficiency was tested by transfecting INS-1 cells and performing PCR on the extracted gDNA.

The second type of plasmid contains the donor sequence required for homology-mediated repair and subsequent sequence knock-in. To construct the donor sequence, I first amplified ~800 bp of the intron region flanking the start of rat *Pdx1* exon 2 and ~800 bp of the 3' UTR

region flanking the end of rat *Pdx1* exon 2 from INS-1 gDNA. Then I used standard PCR to amplify the rat *Pdx1* exon 2 sequence from INS-1 gDNA and the sand rat "mousified" *Pdx1* exon 2 sequence from a plasmid, at the same time removing the stop codon and adding a glycine-serine-glycine linker sequence to the 3' end. I also used PCR to add a V5 tag followed by a P2A sequence to allow ribosome skipping. PAM sites used by the selected guide RNAs were altered by *in vitro* mutagenesis to avoid repeated cutting by the Cas9 construct. Finally, I amplified the blasticidin resistance gene and puromycin resistance gene from plasmid constructs. I used the NEBuilder HiFi DNA Assembly Cloning Kit (NEB #E5520S) to "combine" all fragments into one donor plasmid. Following the recommended protocol, I designed PCR primers with overhanging sequences and used the pUC19 plasmid (NEB # N3041) as the donor plasmid backbone. A vector:insert ratio of 1:1 was used for the assembly reaction. All plasmids were verified by Sanger sequencing prior to transfection.

Table 4.2 Guide RNA sequences used to generate cuts at the start and end of rat Pdx1 exon 2. These sequences were ordered as standard DNA oligos which were phosphorylated and annealed *in vitro* before being inserted into the pX458 plasmid.

Oligo name	Cut site	Oligo sequence (5' to 3')
LHA-gRNA-1-1	5 bp upstream of exon 2 start	CACCGTAACCGACGGTGTTGCCCAC
LHA-gRNA-1-2		AAACGTGGGCAACACCGTCGGTTAC
RHA-gRNA-2-1	21 bp upstream of exon 2 end	CACCGCGGGGTTCCTGCGGTCGCAG
RHA-gRNA-2-2		AAACCTGCGACCGCAGGAACCCCGC

3.5 Mammalian cell transfection and selection for CRISPR/Cas9 gene editing

Electroporation-mediated cell transfection was carried out using the NEPA21 Electroporator (Nepa Gene) according to the instruction manual. Compared to the standard protocol described above, transfection of two pX458 plasmids and one/two donor plasmids at the same time was performed using 1 million detached INS-1 cells and 20 ug total plasmid DNA. Other parameters were not altered.

Transfection efficiency was assessed 48 hours following electroporation and cells were dissociated from the 6-well plates. Each well was split into three equal portions and seeded onto different P15 plates. After being cultured in standard culture medium for 48 hours in the P15 plates, cells were cultured in antibiotic-containing medium from this point onwards. Corresponding to the type of donor plasmid added, cells were cultured with different antibiotic combinations. In this section, low puromycin refers to 0.13 ug/ml, medium puromycin refers to 0.15 ug/ml, high puromycin refers to 0.2 ug/ml; low blasticidin refers to 0.5 ug/ml, medium blasticidin refers to 0.6 ug/ml, high blasticidin refers to 0.7 ug/ml; low mixed refers to 0.07 ug/ml puromycin and 0.3 ug/ml blasticidin, medium mixed refers to 0.08 ug/ml puromycin and 0.35 ug/ml blasticidin, high mixed refers to 0.09 ug/ml puromycin and 0.4 ug/ml blasticidin. After 18 days of culture with antibiotic-containing medium, 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. At this point, some gDNA would be extracted for PCR assessment of CRISPR/Cas9 editing

success while the remaining cells would be frozen and stored in liquid nitrogen for further assessment.

Table 4.3 PCR primers used to screen for successfully edited cell lines.

Primer name	Target region	Primer sequence (5' to 3')
RPH1F	~800 bp upstream	ACTCCCTAAAGTTGTTCTCCCA
RPH1R	~800 bp downstream	GAACCAGGTTGCCTTCAGTTCC
Hind III-RPH2F	~750 bp upstream	ACCAAGCTTCACCTCCACATACACCACAC
RPH2R-KpnI	~750 bp downstream	ATCGGTACCCCCACAAACAGCCTACAAGC
Seq-MLHA	~100 bp upstream	AGTGGTGGGTCTCTGAAGCAAG
Seq-MRHA	~100 bp downstream	TATACCCAACTGGCTCTTGTGC

3.6 gDNA and cDNA preparation followed by PCR

Genomic DNA was extracted from ~100,000 cells using the GeneJET Genomic DNA Purification Kit (ThermoFisher # K0722) and assessed using a nested PCR protocol. In brief, three pairs of primers were designed to target the knock-in construct and 5% of the reaction using primer pair 1 would be used as the template for the reaction using primer pair 2 and so on. To assess gene expression and transcript integrity, total RNA was extracted from ~500,000 cells using the protocol described above and cDNA was prepared from the RNA sample using the GoScript Reverse Transcription System (Promega #A5000) following the manufacturer's instructions. 1 ul cDNA obtained from 80 ng total RNA was used as the template for amplification of the full length *Pdx1* gene transcript.

3.7 Protein extraction and Western blot

CRISPR edited lines that passed DNA and RNA assessment were verified for correct protein expression. 1 million cells were trypsinized and lysed in RIPA buffer consisting of 150 mM NaCl, 1% Triton-X100, 50 mM Tris-HCl (pH = 7.4), 0.5% sodium deoxycholate and one cOmplete Mini Protease Inhibitor Cocktail tablet (Roche #11836153001). Protein concentration was assessed using the Bradford BSA assay.

SDS-PAGE gels were cast using 30% Acrylamide/Bis Solution (Bio-Rad #1610158), Tris-HCl solution (pH = 8.8 for resolving gel, pH = 6.8 for stacking gel), 10% SDS (w/v), 10% APS (w/v) and TEMED (Bio-Rad #1610800). For each sample, 40 ug total protein was boiled with NuPAGE LDS Sample Buffer (ThermoFisher #NP0007) and 0.2% β-mercaptoethanol, then run at 180V for 45 minutes. Proteins were transferred onto a PVDF membrane using semi-wet transfer at 20V for 40 minutes. Protein presence was verified using Ponceau S solution (Sigma #P7170). Western blot was carried out first by blocking the membrane with 5% milk in PBST solution, followed by primary antibody binding at a 1:3,000 ratio using a monoclonal anti-V5 antibody, mouse IgG (Invitrogen #R96025). After overnight binding at 4℃, secondary antibody binding was carried out using a goat anti-mouse HRP antibody (Bio-Rad #170-5047) at a ratio of 1:20,000.

3.8 Protein immunostaining in fixed mammalian cells

To verify the location of V5-tagged proteins produced by the CRISPR edited cell line, the cells were first fixed using 4% paraformaldehyde in PBS solution and then permeabilized using 0.25% Triton X-100 in PBST. Background blocking was carried out using 1% BSA in PBST solution and primary antibody binding was carried out using a monoclonal anti-V5 antibody, mouse IgG at a dilution ratio of 1:100. The secondary antibody used was goat anti-mouse IgG with Alexa 594 (Abcam #ab150116) at a dilution ratio of 1:1,000. Cell actin was stained using Alexa Fluor 488 Phalloidin (ThermoFisher #A12379) at a dilution of 1:40, and cell DNA was stained using DAPI at a dilution of 1:30,000.

3.9 Chromatin immunoprecipitation and qPCR

25 million cells from each sample were harvested and fixed in 1% PFA-BP solution for 8 minutes at room temperature. After quenching the reaction with glycine solution, the cells were washed and resuspended using ChIP Lysis Buffer (50 mM HEPES-KOH pH = 7.5, 140 mM NaCl, 1 mM EDTA pH = 8, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, one protease inhibitor cocktail tablet) to a concentration of 1 million cells/100 ul. 300 ul resuspended and lysed cells were sonicated using a Bioruptor Pico sonication device (Diagenode #B01060010) for a total of 8 cycles (30 sec ON, 30 sec OFF).

5 ug monoclonal anti-V5 antibody, mouse IgG was bound in advance to 50 ul Dynabeads Protein G for Immunoprecipitation (Invitrogen #10009D). 100 ul sonicated chromatin was stored as the Input Sample while 1 ml was incubated with the bound antibody-bead mixture in ChIP Lysis Buffer overnight at 4℃ and washed twice with RIPA Wash Buffer (50 mM HEPES-KOH pH = 8, 500 mM LiCl, 1 mM EDTA pH = 8, 1% NP-40, 1% sodium deoxycholate, one protease inhibitor cocktail tablet). Eluted DNA was de-crosslinked by incubation at 65℃ for 9 hours, and DNA was purified using PCI, CI, and ethanol precipitation. DNA was resuspended in 10 ul 1X TE solution and quantified using the Qubit dsDNA HS Assay Kit (ThermoFisher #Q32854). The Input Sample was de-crosslinked and purified using the same protocol.

qPCR was carried out using the Luna Universal One-Step RT-qPCR Kit according to the manufacturer's instructions. In brief, the total volume was 20 ul for each reaction with 1 ul of eluted DNA. Input Sample was diluted by a 1:5, 1:25 and 1:125 ratio and assessed using qPCR to provide reference Cq values.

Table 4.4 qPCR primers used to amplify the rat *Ins1* promoter region and *Alb* promoter region.

Primer name	Target region	Primer sequence (5' to 3')
RPINS1F3	*Ins1* promoter	GCCTTACCCTCTCTGGGACA
RPINS1R3	*Ins1* promoter	CAGAAAGCGCTCATTGGACG
RPALBF1	*Alb* promoter	GGCAAACATACGCAAGGGATTT
RPALBR1	*Alb* promoter	AGTTCATTGGTTCCTACCCCATT

4. Results

4.1 Sand rat PDX1 inefficiently regulates expression of several key β-cell genes

As a transcription factor, a key aspect of PDX1 function is its ability to correctly regulate the expression of target genes. To test whether the highly divergent sand rat PDX1 protein can still regulate the expression of known PDX1 target genes, I overexpressed sand rat *Pdx1* in rat β-cells and compared these cells to rat β-cells overexpressing the mouse *Pdx1* gene and rat β-cells transfected with an empty plasmid. As PDX1 is responsive to changes in glucose concentration[137,138], I also cultured these transfected cells under different glucose conditions to test if sand rat PDX1 was responsive.

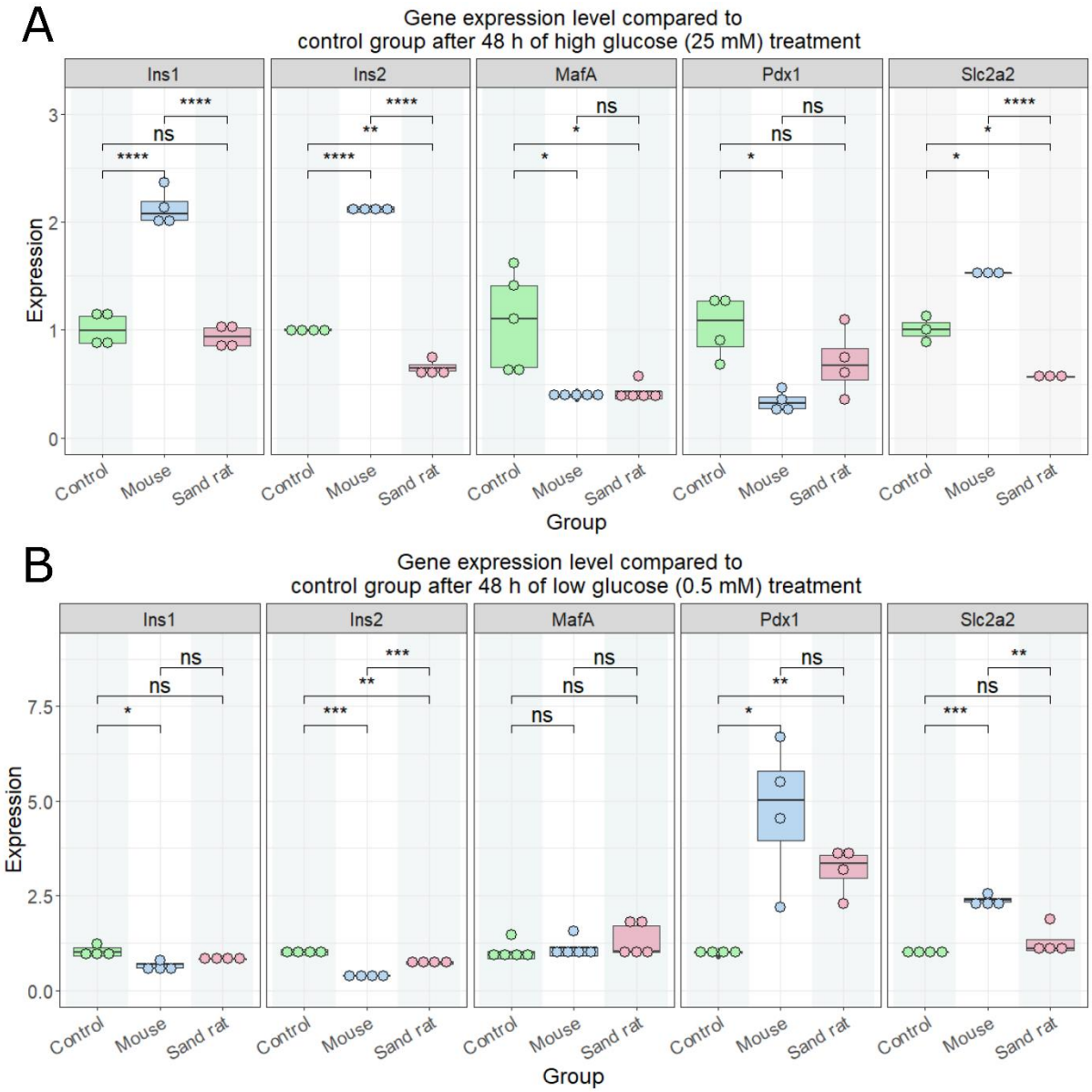


Fig. 4.1 Relative gene expression levels 48 hours following transfection. Expression of each gene in every group shown is relative to that in the corresponding control group. The target gene is shown at the top of each panel, with each group shown in a different colour. Each point represents one technical repeat of the qPCR reaction, and the RNA used for these experiments

are from two independent biological repeats. One asterisk represents a p-value < 0.05, two represents p-value < 0.01, three represents p-value < 0.001, and four represents p-value < 0.0001. 'ns' stands for not significant. (A) cells were cultured in high (25 mM) glucose medium (B) cells were cultured in low (0.5 mM) glucose medium.

First, under high glucose conditions, PDX1 is expected to stimulate insulin expression[138]. This is supported by the 2-fold increase in rat *Ins1* and *Ins2* expression in the mouse *Pdx1* overexpression group compared to the empty plasmid control (Fig. 4.1A). In comparison, the effect of sand rat PDX1 on the *Ins1* gene was not significantly different to the control group, while sand rat PDX1 significantly reduced *Ins2* expression. Aside from insulin, I also measured the expression of other genes known to be directly regulated by PDX1, which includes the *Pdx1* gene itself. Overexpression of mouse PDX1 under high glucose conditions significantly suppressed rat *Pdx1* gene expression compared to the control, showing that mouse PDX1 can regulate expression of the rat *Pdx1* gene. Sand rat PDX1 also caused a weaker decrease of rat *Pdx1* expression, although this effect was not significant compared to the control group. As for the rat *MafA* gene, a key transcription factor gene regulated by PDX1, both mouse and sand rat PDX1 was able to significantly downregulate *MafA* expression compared to the control. This suggests that mouse and sand rat PDX1 interaction with the rat *MafA* promoter and cofactors is likely similar. In contrast, mouse PDX1 significantly upregulated rat *Slc2a2* gene expression compared to the control, which is a gene expressing the key glucose transporter GLUT2, while sand rat PDX1 significantly downregulated *Slc2a2* expression.

I also compared the expression of these rat genes under low glucose conditions (0.5 mM glucose) under which PDX1 should strongly suppress insulin expression[137]. This was supported by the mouse *Pdx1* overexpression group, which significantly lowered *Ins1* and *Ins2* expression compared to the empty plasmid control (Fig. 4.1B). However, for the cells overexpressing sand rat *Pdx1*, *Ins1* and *Ins2* expression levels under low glucose conditions barely differ from that under high glucose conditions. It is likely that sand rat PDX1 is not as sensitive as mouse PDX1 to glucose changes in the rat cell system, or that sand rat PDX1 cannot function properly in the rat glucose-mediated insulin expression pathway. However, both mouse and sand rat PDX1 regulated rat *Pdx1* and *MafA* expression differently under different glucose conditions. Under low glucose levels, both mouse and sand rat PDX1 significantly upregulated rat *Pdx1* expression by 3 – 5-fold, contrasting strongly with the downregulation of *Pdx1* expression observed under high glucose conditions. Neither mouse PDX1 nor sand rat PDX1 significantly altered rat *MafA* expression under low glucose conditions while both PDX1 proteins significantly downregulated *MafA* under high glucose conditions. This suggests that the glucose-mediated insulin expression pathway likely involves different components compared to glucose-mediated regulation of *Pdx1* and *MafA* expression. Finally, the effect of mouse PDX1 on rat *Slc2a2* gene expression did not change under different glucose conditions. In contrast, sand rat PDX1 did not significantly alter *Slc2a2* expression under low glucose conditions while it slightly downregulated *Slc2a2* expression under high glucose conditions.

In summary, these qPCR results suggest three main differences between mouse PDX1 and the highly divergent sand rat PDX1. First, sand rat PDX1 is likely to have a very different role compared to mouse PDX1 in the glucose-mediated insulin expression pathway. Second, self-

regulation of *Pdx1* expression should be still conserved in sand rats, and sand rat PDX1 can still respond to glucose and mediate *Pdx1* expression in a similar way compared to mouse PDX1. Third, amino acid changes in sand rat PDX1 is likely to have affected its regulation of other key β-cell genes such as *MafA* and *Slc2a2*, which might also contribute to the diabetes-prone phenotype observed in some sand rat individuals.

4.2 Establishment of rat β-cells producing a rat-sand rat fusion PDX1 protein

Although it is relatively easy to overexpress exogenous genes in a mammalian cell system, one obvious downfall is this system expresses the gene of interest at levels often more than a hundred times higher than normal. In addition, presence of the endogenous rat *Pdx1* gene complicates interpretation of target gene expression as it is possible that exogenous PDX1 induced the observed changes in gene expression through altering endogenous *Pdx1* expression and/or by competing with the endogenous PDX1 protein for binding sites. Therefore, to compare conserved PDX1 function with sand rat PDX1 function in a system more similar to a normal β-cell system, I decided to use gene editing methods to replace the endogenous rat *Pdx1* gene with a "mousified" sand rat *Pdx1* gene which has similar GC composition compared to the mouse *Pdx1* gene but still encodes for the sand rat PDX1 protein. In this chapter, I targeted only one exon in the endogenous rat *Pdx1* gene. Rat *Pdx1* has two exons and editing both exons at the same time would require either removing the ~3 kb intron sequence or two rounds of CRISPR/Cas9 editing. I decided to focus on exon 2 as this exon encodes the PDX1 homeodomain, the most important structure for PDX1 function.

CRISPR/Cas9 editing relies on Cas9-gRNA, a protein-RNA complex that identifies target sites using guide RNA (gRNA) complementary to target DNA and generates a double-strand break (DSB) in the target using the Cas9 enzyme. To replace rat *Pdx1* exon 2 with sand rat exon 2, I designed gRNA that guide removal of the entire rat exon 2 and constructed four "donor sequences" to "fill" the excised region (Fig. 4.2). Donor sequences take advantage of the homology directed repair (HDR) pathway which repairs DSB using a homologous sequence, allowing gene knock-in. This contrasts with the non-homologous end-joining (NHEJ) repair pathway, which repairs DSB without a homologous sequence, causing indels leading to gene knock-out. The donor sequences consist of five essential components, the first being the *Pdx1* exon 2 sequence. I constructed donors containing either an endogenous rat or "mousified" sand rat *Pdx1* exon 2 and separately transfected them into different groups to account for technical bias (Fig. 4.2). In theory, cells transfected with rat *Pdx1* exon 2 should express a rat PDX1 protein with no amino acid differences from the endogenous protein, and these cells should not have large differences compared to wild type cells. The second essential component in donor sequences is homologous arms, 800 bp regions that have 100% similarity to the upstream and downstream regions flanking endogenous *Pdx1* exon 2 and direct HDR-mediated insertion of the replacement sequence (Fig. 4.2). The third donor sequence component is a sequence encoding the V5 tag, a small chain of amino acids that facilitates protein tracking and identification (Fig. 4.2). The fourth component is an antibiotic resistance gene allowing antibiotic selection of cells that have integrated the donor sequence (Fig. 4.2). To increase the chance of obtaining cells that contain donor sequences in both *Pdx1* loci, I used a combination of two antibiotics (puromycin and blasticidin) to select cells that have a different antibiotic in

each locus. Finally, to avoid generating large fusion constructs with PDX1 connected to the antibiotic resistance protein, I included a viral ribosome-skipping sequence (referred to as P2A) between the V5 tag and the antibiotic resistance gene (Fig. 4.2). Presence of P2A would allow the edited *Pdx1* gene and antibiotic resistance gene to be transcribed on the same mRNA strand but translated as different peptides.

Following transfection of INS-1 cells with the Cas9 plasmid and different combinations of donor plasmids, I obtained 279 single cell colonies by selecting for successfully transfected cells in antibiotic-containing medium. To screen for cells that carry the donor sequence, I used nested PCR to amplify the genomic region around the endogenous *Pdx1* exon 2 loci, with the first pair of primers designed to bind to upstream and downstream regions not included in the donor plasmid while the second and third pairs were designed to bind to sites within the donor plasmid. PCR screening for donor sequence integration is depicted in Fig. 4.3A. I chose to perform three rounds of nested PCR due to difficulties in designing suitable primers for the rat *Pdx1* exon 2 region, and as shown in Fig. 4.3A, products from the third round of PCR was analysed using gel electrophoresis to guide sample selection for Sanger sequencing. Due to addition of components such as the V5 sequence, P2A sequence, and antibiotic resistance gene, the donor sequence is 500 – 800 bp longer than the wild type sequence. Therefore, it is possible to distinguish cell lines with successful integration of the donor sequence on an agarose gel (Fig. 4.3B). Cell lines with insertion of the same donor sequence have a bright band around 1.2 kb or 1.5 kb depending on the antibiotic resistance gene inserted, while cell lines with the wild type sequence should have a bright band around 0.7 kb (Fig. 4.3B). Homozygous insertions of the donor sequence should produce a brighter and cleaner band around 1.2 or 1.5 kb compared to heterozygous insertions, and sometimes homozygous samples do not have any bands smaller than 1.2 kb on the gel (Fig. 4.3B). Alleles with an exon 2 knock-out produce a band smaller than the wild type (Fig. 4.3B).

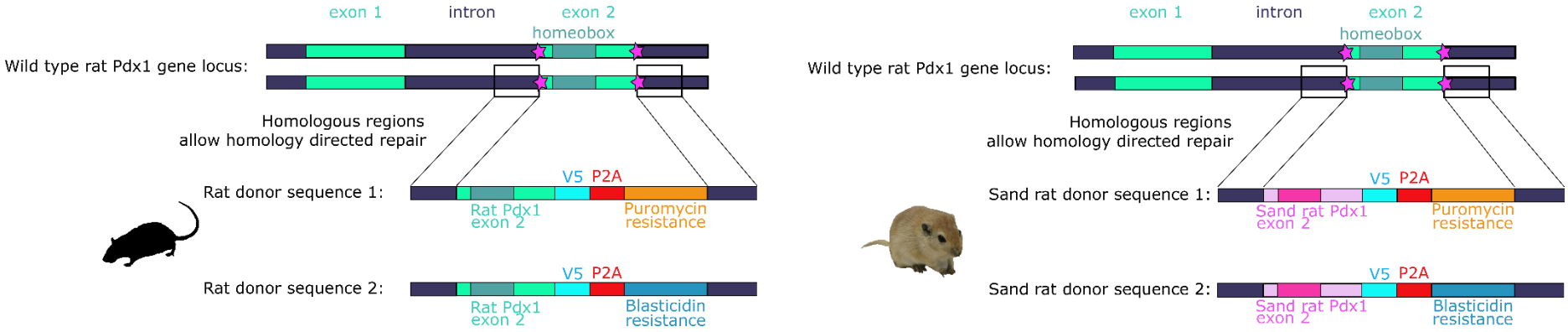


Fig. 4.2 Five essential components of the donor sequence. Pink stars in the wild type rat *Pdx1* gene locus represent cut sites targeted by the Cas9-gRNA complex. Non-exon regions are in black while rat *Pdx1* exons are in green, with the rat homeobox region in deep green. All four donor sequences used this chapter are shown here, with two donor sequences carrying the rat *Pdx1* exon sequence and two donor sequences carrying the sand rat *Pdx1* exon sequence. Each donor sequence component represented by a different colour. The homology of each homologous arm in the donor sequence to the wild type locus is indicated.

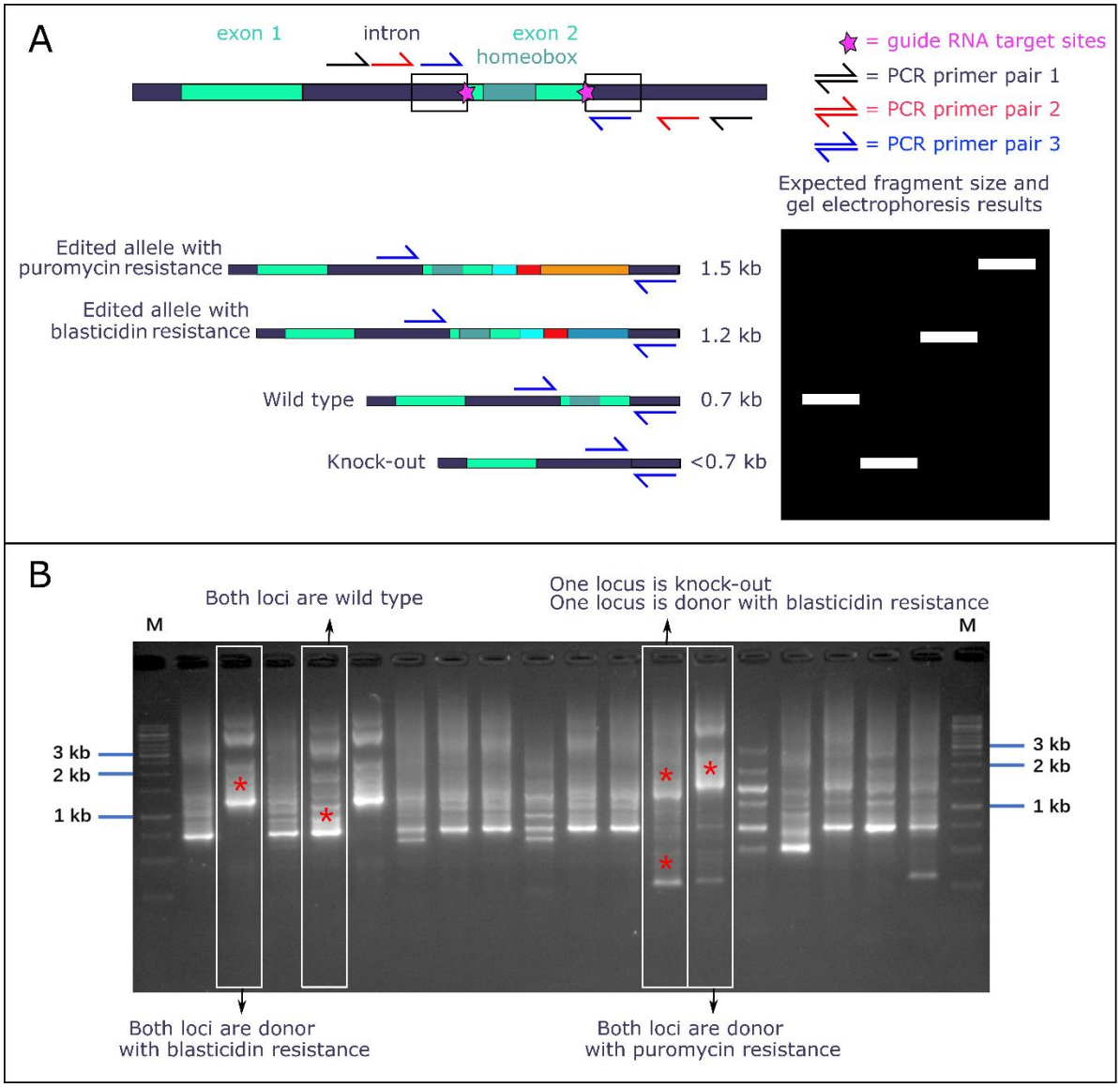


Fig. 4.3 PCR results from cell line screening. A) shows the location of each primer pair. Two rectangular boxes represent regions in the wild type locus that has matches homologous arms in the donor sequence. Products of the third round of PCR were analysed using agarose gel electrophoresis, and four different PCR products was expected. Alleles that have a *Pdx1* knock-out should have a fragment less than 0.7 kb, and alleles with the wild type sequence should have a fragment of 0.7 kb. For alleles that have integrated the rat donor sequence, if the donor contains a blasticidin resistance gene, it should have a 1.2 kb fragment, and if the donor contains a puromycin resistance gene, it should have a 1.5 kb fragment. B) is an example of gel electrophoresis results obtained from 18 different cell lines transfected with the rat donor sequence. A different sample was loaded into each well and the four highlighted samples were verified with Sanger sequencing. Bands on the gel that correspond to fragments amplified from the *Pdx1* allele in each of these four samples are marked with a red asterisk.

Due to the cost of high-fidelity PCR reagents and Sanger sequencing, I did not sequence all 297 cell colonies that survived antibiotic treatment and instead focused sequencing efforts on cell lines that were most likely to contain the donor sequence, as shown by gel electrophoresis

results (Fig. 4.3B). I performed Sanger sequencing for a total of 76 cell lines that were likely to contain an insert of the donor sequence at the correct loci, and as summarized in Fig. 4.4, 7 out of 178 cells transfected with a rat *Pdx1* donor sequence have the donor sequence in both wild type *Pdx1* loci. Out of these 7 cell lines, 4 lines exhibit resistance to blasticidin and 3 lines show resistance to both blasticidin and puromycin (Fig. 4.4). In addition, 1 line integrated the donor sequence into one locus, with the other locus unaffected by the editing process, and 1 line integrated one donor sequence and lost the endogenous exon in the other locus (Fig. 4.4). For cells transfected with the sand rat *Pdx1* donor sequence, 9 out of 101 cell lines had the donor sequence successfully integrated into both wild type *Pdx1* loci. Amongst these 9 homozygous cell lines, 6 are resistant to only one type of antibiotic while 3 are resistant to both antibiotics (Fig. 4.4). In addition, 2 cell lines are heterozygous, with one donor sequence successfully replacing the wild type sequence in one locus and the second locus unaffected by CRISPR editing (Fig. 4.4). Overall, the rate of obtaining homozygous cell lines (i.e. both wild type loci are replaced by a donor sequence) for both rat and sand rat groups is lower than 10%, with most edited cells not containing a donor sequence in either *Pdx1* loci. Given that most DNA repairs in somatic cells are carried out by NHEJ repair, it is expected that most cell lines have lost the endogenous exon 2 sequence or have indels near the Cas9 cut site. Indeed, amongst the 76 cell line samples sent for Sanger sequencing, I identified 3 cell lines that lost one or both loci and did not integrate the donor sequence in the *Pdx1* locus (Fig. 4.4). In theory, cell lines without successful integration of the antibiotic resistance gene should not survive antibiotic treatment (as confirmed by cells not transfected with a control plasmid). However, given that I only screened the *Pdx1* exon 2 loci, it is possible that the knock-out lines have integrated parts of the donor plasmid in different locations, allowing them to survive antibiotic treatment.

In summary, a total of 7 edited cell lines had been successfully edited in both *Pdx1* exon 2 loci and was confirmed via Sanger sequencing to have the rat Pdx1-V5 donor sequence inserted without errors. In addition, 9 edited cell lines contained the sand rat Pdx1-V5 donor sequence instead of the wild type rat sequence in both loci. Only these 16 cell lines were assessed in the following sections for presence of the PDX1-V5 fusion protein and used for functional analysis.



First allele	Second allele	Example	Rat donor sequence	Sand rat donor sequence
Random insertion of donor sequence in non-Pdx1 loci	Wild type unaffected		Not verified	Not verified
	Knock-out		Not verified	At least 3/101 (3.0%)
One Pdx1 allele replaced by donor	Knock-out		1/178 (0.6%)	0/101(0%)
	Wild type unaffected		1/178 (0.6%)	2/101 (2.0%)
Both alleles replaced by donor	Same antibiotic resistance gene		4/178 (2.2%)	6/101 (6.0%)
	Different antibiotic resistance gene		3/178 (1.7%)	3/101 (3.0%)

Use for functional analysis

Legend

- Rat Pdx1 non-exon regions
- Rat Pdx1 exons
- Rat Pdx1 homeodomain
- Sand rat Pdx1 exon
- Sand rat Pdx1 homeodomain
- V5 sequence
- P2A sequence
- Puromycin resistance gene
- Blasticidin resistance gene

Fig 4.4 Outcomes from CRISPR/Cas9 mediated gene knock-in. Some cells still have two wild type *Pdx1* loci or are missing the wild type sequence, and the donor sequence is likely randomly inserted and expressed in a non-*Pdx1* locus in these cells. In addition, some cells are heterozygous in the *Pdx1* locus with one allele carrying the donor sequence and the second allele carrying the wild type sequence or missing the wild type sequence. Only cells that carry a donor sequence in both loci are used for downstream analysis, and the chances of obtaining cells with both loci edited are low.

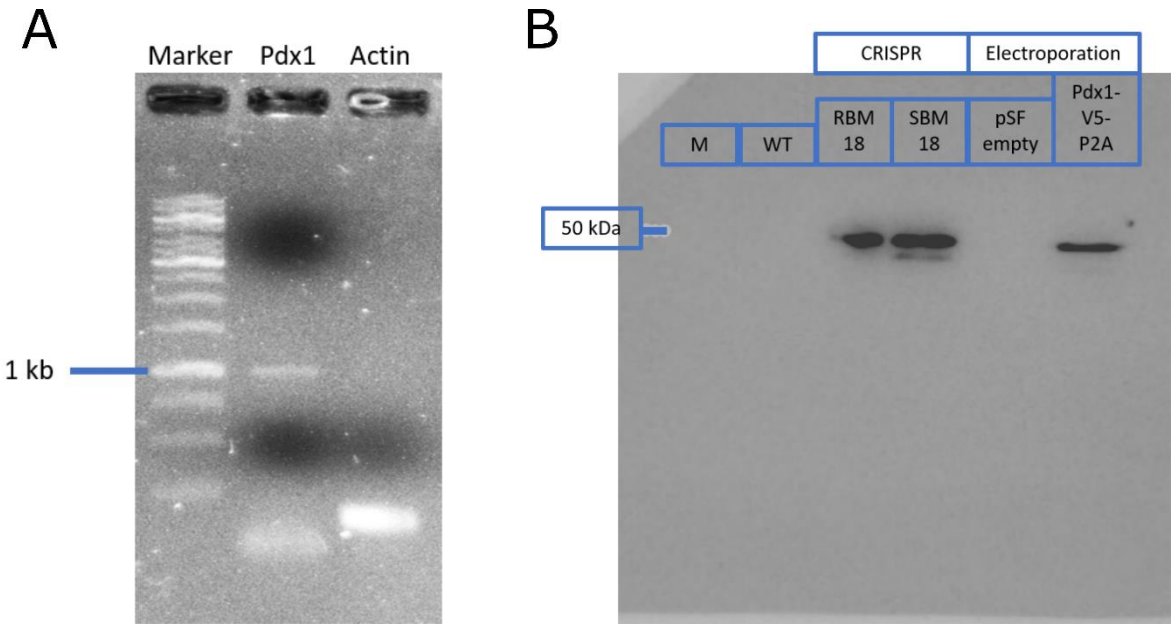


Fig. 4.5 Verification of successful editing at the RNA and protein level. A) Gel electrophoresis showing full-length *Pdx1-V5* amplified from RBM18 cDNA. Primers targeting the rat actin gene were used in the control group. B) Western blot showing presence of the PDX1-V5 fusion protein. "WT" represents wild type INS-1, "RBM18" represents homozygous knock-in rat *Pdx1*-V5 line, and "SBM18" represents a homozygous knock-in sand rat *Pdx1*-V5 line. "pSF empty" and "Pdx1-V5-P2A" are both from wild type cells electroporated with the pSF plasmid. "pSF empty" are transfected with empty pSF plasmids while "Pdx1-V5-P2A" are transfected with pSF containing the full-length rat *Pdx1* cds sequence with the same linker, V5 and P2A sequence used for gene knock-in. "M" is the protein marker. The primary antibody is a monoclonal mouse anti-V5 antibody and the secondary antibody is a goat anti-mouse IgG antibody conjugated with horseradish peroxidase.

To check if cell lines with homozygous knock-in sequences were correctly expressing the fusion PDX1 protein, I prepared cDNA samples from total mRNA extracted from these cell lines. PCR primers recognizing the start of rat *Pdx1* exon 1, a region that was not edited, and the end of the inserted P2A sequence were used, and the expected size of the target region was 1 kb. As shown in Fig. 4.5A, the cell line RMB18 is successfully expressing and splicing the *Pdx1* mRNA strand. To verify protein expression and to check if the P2A sequence successfully induced ribosome skipping, I performed Western blot using total protein extracted from different cell lines. To assess if the fusion protein was produced correctly and if the antibody was working, I prepared wild type cells overexpressing the rat PDX1-V5-P2A fusion protein

as a positive control. As shown in Fig. 4.5B, the rat homozygous knock-in cell line RBM18 has a band about the same size as the positive control. The sand rat homozygous knock-in cell line SBM18 also has a protein of similar size detected by the V5 antibody, although there is a slightly smaller protein also detected. First, this shows that for both edited cell lines, the P2A ribosome skipping sequence is working. If both antibiotic resistance proteins were attached to the PDX1-V5 fusion protein, the resulting construct would not have the same size as the control fusion protein. At the same time, these results indicate that a certain degree of protein degradation or altered post-translation modification may be occurring for rat-sand rat fusion PDX1.

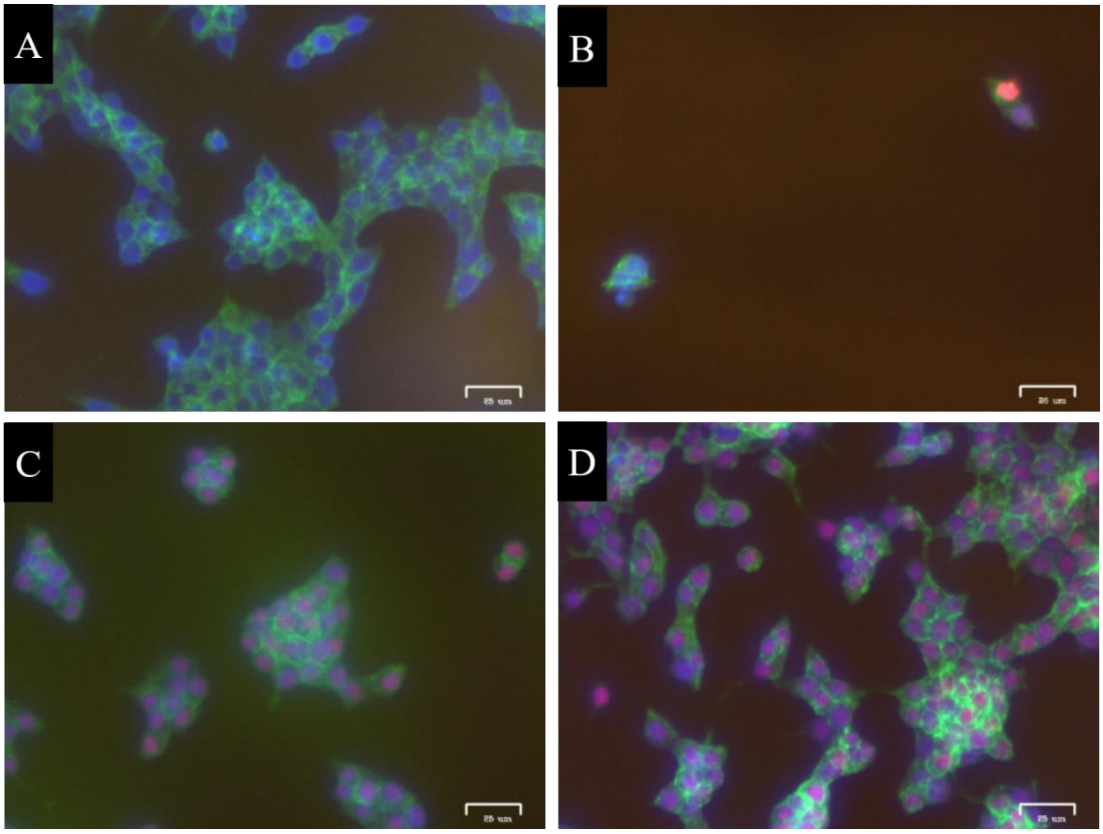


Fig. 4.6 Immunofluorescence staining of different edited cell lines. Blue indicates DAPI staining for DNA, green indicates phallodin staining for actin, red indicates monoclonal mouse anti-V5 antibody recognition of the V5 tag. (A) wild type INS-1 cells as a negative control, (B) wild type cells transfected with pSF-rat Pdx1-V5 as a positive control, (C) cell line RBM18, a homozygous rat Pdx1-V5 knock-in line, (D) cell line SBM18, a homozygous sand rat Pdx1-V5 knock-in line.

Finally, to verify protein subcellular localization, I used immunofluorescence staining to visualize the distribution of the PDX1-V5 fusion protein within the homozygous knock-in cell lines. Fig. 4.6 is an example showing colocalization of the PDX1-V5 protein with nucleus DNA in both knock-in cell lines, with very low levels of background antibody staining in the negative control. At this point, I concluded that I successfully constructed several homozygous knock-in lines of INS-1 cells expressing rat PDX1-V5, and several homozygous knock-in lines of INS-1 cells expression rat-sand rat fusion PDX1-V5.

4.3 INS-1 cells cultured from single cells show abnormal response to glucose

I constructed these homozygous *Pdx1*-V5 knock-in lines in hope of comparing rat and sand rat PDX1 regulation of insulin gene expression following glucose stimulation. However, qPCR analysis showed that these cells may not be the best tool to study glucose-stimulated insulin expression, as shown in Fig. 4.7. For this experiment, I incubated different groups of cells in high glucose (25 mM) culture medium for 48 hours and then extracted total RNA for qPCR analysis. The expression level of *Ins2* in each cell line was normalized to its level of housekeeping gene *36B4*, producing a relative *Ins2* expression level for each sample, and then for each cell line this value was normalized against the relative expression of *Ins2* in the wild type sample. Fig. 4.8 shows that relative expression of *Ins2* is clustered around a value of 1 for the wild type sample while both rat and sand rat knock-in cell lines have a significantly lower expression of *Ins2*. To assess if this difference in *Ins2* expression was caused by errors in gene construct design or the process of establishing these cell lines, I established a wild type cell line which was propagated from a single wild type cell. Aside from going through the process of being separated from its original cell population, this wild type cell line did not experience any additional experimental manipulation. This mimicked the single cell culture procedure that the knock-in cells had to experience, and the single cell wild type line expressed *Ins2* at levels significantly lower than the control wild type cells. This indicated that the INS-1 cell line is likely unsuitable for any processes that require single cell culture or maybe even culturing them at low density. These results suggest that the homozygous edited cell lines are an unsuitable tool to use for comparing glucose-induced insulin expression.

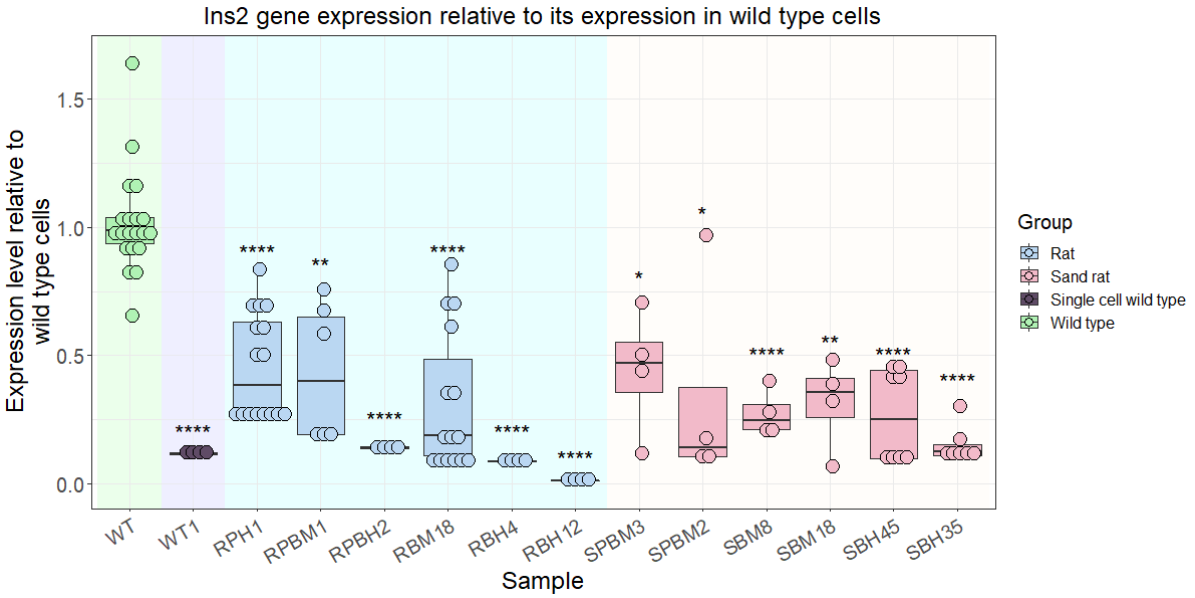


Fig. 4.7 Relative *Ins2* gene expression levels in multiple edited cell lines. The gene expression levels shown are relative to expression levels in wild type cells. Total RNA are extracted from different cell lines maintained in high glucose (25 mM) medium: wild type INS-1, wild type cells propagated from single cells, 6 cell lines with homozygous rat *Pdx1*-V5 alleles and 6 cell lines with homozygous rat-sand rat fusion *Pdx1*-V5 alleles. The data shown here is combined data from 5 different experiments, with each point representing one technical repeat. One asterisk represents p-value < 0.05, two represents p-value < 0.01, three represents p-value <

0.001, and four represents p-value < 0.0001. The unpaired two samples T-test was used to calculate p-values compared to the wild type sample.

4.4 Sand rat PDX1 has weak binding affinity to the rat insulin I promoter

Given that the glucose-induced insulin secretion pathway in the knock-in cell lines were likely disrupted by culturing them from single cells, I decided to use these cell lines to compare a different aspect of PDX1 function: PDX1 binding to the insulin promoter. For the rat-sand rat fusion PDX1 protein, it has the sand rat homeodomain and is expected to have different binding affinities to the insulin promoter compared to the conserved homeodomain sequence[102]. First, I compared the predicted PDX1 binding sites in the insulin promoter of multiple rodents, using the human as the outgroup. Fig. 4.8 compares the key PDX1 binding site sequence in the insulin promoter of several rodent species, with A3 known as the most conserved across species and also the most important PDX1 binding site out of 4 – 5 sites along the insulin promoter[132]. Gerbils have an A3 sequence that is the same as the murid sequence (mouse, rat, and Algerian mouse). Given that key amino acids that directly interact with the TAAT sequence remains conserved in sand rat PDX1[181], it is possible that sand rat PDX1 will still be able to recognize the rat insulin promoter. However, I hypothesize that sand rat PDX1 binding affinity to the rat insulin promoter is different compared to the rat PDX1 protein given the divergent amino acids present in the sand rat homeodomain.

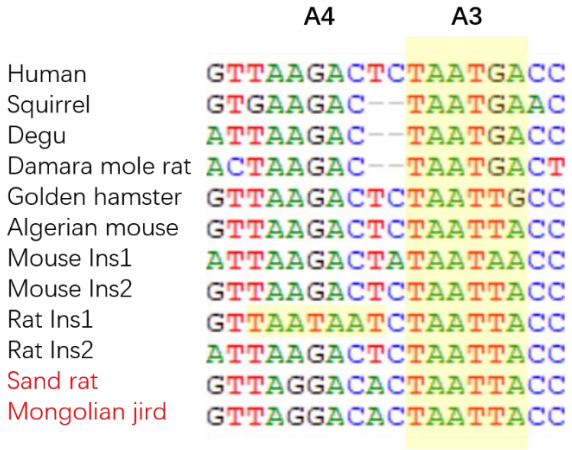


Fig. 4.8 Alignment of the insulin promoter in multiple rodent species. The human insulin promoter is included as the outgroup. For mouse and rat, they have two insulin genes (*Ins1* and *Ins2*), both of which are directly regulated by PDX1. The two gerbil species are located at the bottom of the alignment and their names are in red. Two binding spots for PDX1 are highlighted in yellow, of which A3 is the most highly conserved across species[132].

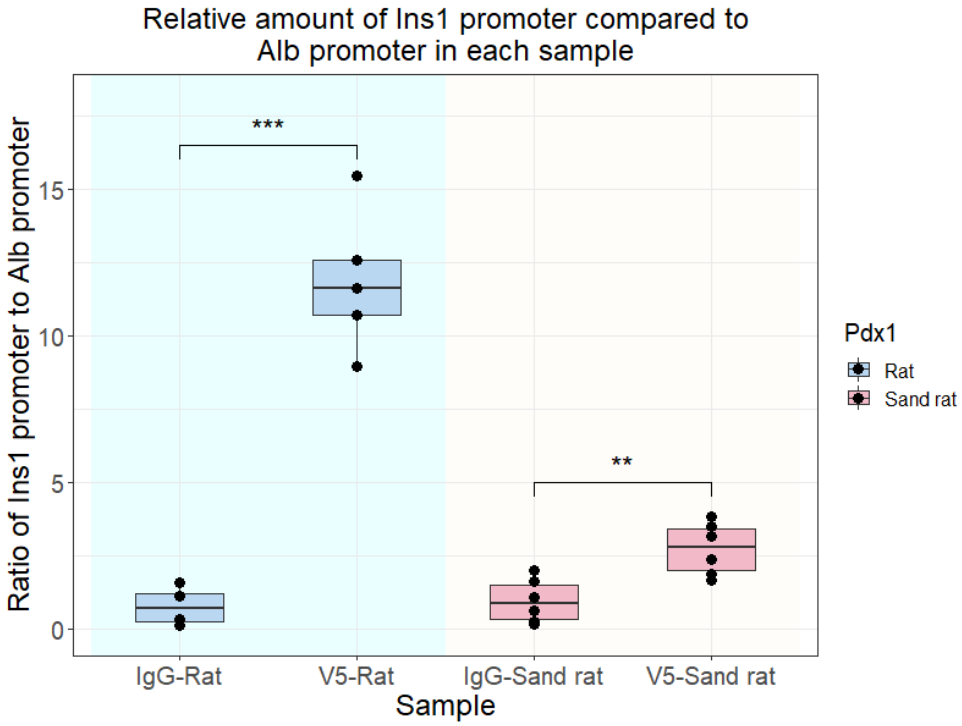


Fig. 4.9 Relative amount of *Ins1* promoter in each ChIP sample. Here, the amount of *Ins1* promoter is shown as relative to the amount of *Alb* promoter in each sample. For rat samples, fixed chromatin was extracted from the rat *Pdx1*-V5 homozygous cell line RBM18. For sand rat samples, fixed chromatin was extracted from the rat-sand rat fusion *Pdx1*-V5 homozygous cell line SBM18. IgG samples were precipitated using mouse IgG while V5 samples were precipitated using a monoclonal mouse anti-V5 antibody. The amount of promoter sequence was measured using qPCR. These measurements were obtained from three biological repeats.

To compare the binding affinity of the rat homeodomain and the sand rat homeodomain, I extracted fixed chromatin from the homozygous rat *Pdx1*-V5 cell line RBM18 and the homozygous rat-sand rat fusion *Pdx1*-V5 cell line SBM18. Proteins bound to genomic DNA are cross-linked by formaldehyde and the genomic DNA is sheared to ~1 kb then incubated with either mouse IgG or a monoclonal mouse anti-V5 antibody. In this experiment, mouse IgG functions as a control for background binding to the mouse antibody. DNA-protein constructs that bind to these antibodies are then precipitated and de-crosslinked to allow purification of the DNA fragments. These fragments are then used for qPCR to measure the relative amount of *Ins1* promoter sequence and *albumin* (*Alb*) promoter sequence. Here, the *Alb* promoter is used as a control as it has a TAAT motif which PDX1 can recognize but should not bind to, as albumin is not expressed in pancreatic cells and the promoter region should be inaccessible[199]. Fig. 4.9 shows the enrichment of *Ins1* promoter over *Alb* promoter in both rat PDX1 samples and rat-sand rat fusion PDX1 samples when incubated with the anti-V5 antibody, indicating that the sand rat homeodomain can indeed recognize and bind to the rat *Ins1* promoter. However, for the rat PDX1 samples, the V5-precipitation group had a much higher level of *Ins1* enrichment compared to the rat-sand rat fusion PDX1 samples, showing that more cells had rat PDX1 bound tightly to the *Ins1* promoter compared to the fusion PDX1. These results suggest that the sand rat homeodomain sequence has a weaker interaction with

the rat insulin promoter, and given the level of sequence conservation between the rat *Ins1* promoter and the sand rat *Ins* promoter, it is likely that the amino acid changes in the sand rat homeodomain has altered its interaction with its target sequences.

5. Discussion

In this chapter, I used ectopic overexpression and gene replacement to compare sand rat and conserved PDX1 protein biological function in a rat pancreatic β-cell background. The results from both ectopic overexpression and gene replacement indicate that the highly divergent sand rat PDX1 protein has altered biological activity in the β-cell, with major changes in its ability to bind to the insulin promoter. I propose that through the course of evolution, natural selection has failed to purge deleterious mutations that affect direct regulation of insulin gene expression by the sand rat PDX1 protein.

PDX1 is known as the master regulator of pancreatic β-cell function because it is required for maintaining β-cell identity and can directly regulate the expression of most essential β-cell proteins, including the signature regulatory molecule insulin, the glucose transporter GLUT2 and the β-cell-specific transcription factor MAFA[114,200,201]. In fact, the total number of target genes regulated by PDX1 has been reported to exceed 4000 in the human islet and 7000 in the mouse islet, including genes that cover nearly all aspects of β-cell metabolism and function[142]. A suitable protein structure is essential for ensuring the PDX1 protein is capable of recognizing its target motif and interacting with cofactor proteins[181,202]. In this discussion, I will focus on analysing sand rat PDX1 recognition of target promoter sequences and its ability to bind to these sequences.

Like other homeodomain proteins, PDX1 has a strong preference for the TAAT motif, and it has been shown to strongly prefer the 5'-(C/T)TAAT(T/G)-3' sequence which is present in the promoter of many β-cell-specific genes including *Ins*, *Slc2a2* and *MafA*[136,203]. Crystal resolution of PDX1 interaction with DNA containing the TAAT sequence has highlighted three amino acids in the homeodomain that interact with nucleotides in this motif: R150 and N196 form direct interactions while I192 forms indirect interactions via van der Waals interaction[181]. For gerbil PDX1, the two key amino acids R150 and N196 remain conserved while I192 is replaced by a valine residue. As I192 interacts with tyrosine through van der Waals interaction, and valine and isoleucine are biochemically similar, it is possible that this residue alteration in gerbils may still allow PDX1 recognition of the key TAAT motif. High levels of A3 sequence conservation in the gerbil insulin promoter also supports the possibility that sand rat PDX1 has retained the ability to recognize the core TAAT motif.

However, it is very likely that the binding affinity of sand rat PDX1 has altered significantly compared to the conserved protein. Mouse PDX1 binding to the core mouse *Ins1* promoter sequence has been compared to sand rat PDX1 binding using computer stimulation, and results predict stronger interaction between the sand rat PDX1 protein and the target DNA sequence due to residue changes that have the potential to increase the number of biochemical interactions between the protein and DNA strand[102]. The results I present in this chapter also suggest changes in sand rat PDX1 binding affinity, although my results show decreased binding affinity of the sand rat PDX1 protein to the rat *Ins1* promoter sequence compared to rat PDX1. I think that these results do not contradict each other, as the A3 box in mouse *Ins1* and rat *Ins1* differ in nucleotide sequences around the core TAAT motif, and it is likely that the residue differences in sand rat PDX1 result in different interactions with these two sequences. In

addition, the computer model may use parameters that deviate from the actual conditions in the ChIP experiment.

The ChIP-qPCR results suggest that weak binding affinity to the rat *Ins1* promoter is a possible explanation for the lack of *Ins1* response to sand rat PDX1 overexpression, as observed in the ectopic overexpression experiment. For genes other than *Ins1*, I did not obtain direct information about sand rat PDX1 binding to their promoters due to the very small amount of material that can be obtained from one ChIP experiment. However, results from the *Pdx1* ectopic expression experiment may provide hints on functional differences between sand rat and conserved PDX1 regulation of other essential β-cell genes. First, based on qPCR results shown in Fig. 4.1, I speculate that sand rat PDX1 may have high affinity for some rat promoter sequences but fail to properly regulate the target gene due to abnormal interaction between sand rat PDX1 and rat cofactor proteins. Examples of such genes include rat *Ins2* and *Slc2a2*, both of which are significantly suppressed when sand rat PDX1 is present, regardless of whether the cell is in a high or low glucose condition. If sand rat PDX1 is weakly interacting with the *Ins2* and *Slc2a2* promoter, it is unlikely that sand rat PDX1 will have a significant effect on the expression of these two genes. I propose that sand rat PDX1 may form stronger interactions with the *Ins2* and *Slc2a2* promoter compared to the endogenous rat PDX1 protein, but sand rat PDX1 fails to interact with the endogenous rat cofactors to correctly regulate insulin expression based on environment glucose content. In the ectopic overexpression experiment, the full coding sequence for sand rat *Pdx1* was overexpressed, and aside from the homeodomain, sand rat PDX1 has highly divergent N-terminal and C-terminal domains. Previous work has shown that PDX1 needs to interact with different cofactors to regulate insulin expression in response to changes in glucose concentration, for example PDX1 interacts with cofactor p300 under high glucose conditions while PDX1 interacts with histone modification enzymes under low glucose conditions[137,138]. These interactions have been further dissected to highlight key PDX1 domains required to maintain these interactions, with the PDX1 N-terminal shown to be essential for interaction with p300 while the C-terminal is more important for interactions with histone modifying enzymes[137,138]. However, these two domains have undergone radical amino acid changes in the sand rat PDX1 protein, and it is very unlikely that sand rat PDX1 will be able to interact properly with these rat cofactors.

The overexpression results also suggest that sand rat PDX1 still has conserved regulation of the rat *Pdx1* and *MafA* genes, both key transcription factors essential for normal β-cell function. First, *Pdx1* is an auto-regulatory gene, as its own protein product can directly bind to its own promoter and regulate its own expression. When a large amount of PDX1 is present, it is expected to suppress *Pdx1* expression, which is observed in cells under high glucose conditions. However, the exact opposite is observed for low glucose conditions. I think that this is because PDX1 also regulates the expression of many cell survival related genes, and under these low glucose conditions, the cells will upregulate *Pdx1* expression to activate survival-related pathways. In general, the very similar effects of mouse and sand rat PDX1 overexpression on rat *Pdx1* expression strongly suggests that sand rat PDX1 still possesses the ability to self-regulate under changing glucose conditions. In addition, MAFA is a transcription factor that also directly regulates insulin expression by binding to the insulin promoter sequence. Under high glucose conditions, when mouse or sand rat PDX1 is overexpressed and therefore

producing an excess amount of PDX1, rat *MafA* expression is downregulated. When PDX1 becomes required for survival under low glucose conditions, rat *MafA* expression is not altered when mouse or sand rat PDX1 is overexpressed. This is mainly because PDX1 and MAFA have overlapping functions regarding regulation of insulin expression but not direct regulation of survival-related genes[131]. The regulatory effect of sand rat PDX1 on these two rat transcription factor genes suggest that sand rat PDX1 may have retained regulatory function of key β-cell transcription factor genes. Although only 5 genes were assessed using this method, it may be possible that the amino acid changes in sand rat PDX1 have deleteriously affected its ability to directly regulate the expression of insulin and other non-regulatory genes such as the glucose transporter gene *Slc2a2* while its ability to regulate expression of other transcription factors remains preserved. It is possible that the presence of other transcription factors such as MAFA has compensated for some losses in certain aspects of sand rat PDX1 function, and natural selection has only purged deleterious mutations that severely alter the capability of sand rat PDX1 to regulate these "replacement" transcription factors. I propose that through the course of evolution, natural selection has purged deleterious mutations that affect sand rat PDX1 regulation of other β-cell transcription factors while allowing mutations that affect direct control of insulin.

Finally, the results of CRISPR/Cas9 mediated gene editing show that the biological properties of the cell type used needs to be considered carefully. Pancreatic β-cells have different functions in the islet, with a small population of cells "leading" the response to elevated glucose and most cells "following" this response[204]. The proportion of these different cells in the islet population has also been shown to change in mice that have diabetes, suggesting that overall response to glucose can alter based on the composition of the β-cell population[204]. The rat secondary cell line used in this experiment originated from a pancreatic tumour and was selected for its strong response to glucose[205]. Given that these cells are not primary cells, I did not expect them to depend on population heterogeneity to mediate their response to insulin. However, my results in this chapter show that the process of undergoing single cell culture severely compromises their response to glucose. In addition, although this cell line is known to be less responsive to glucose after multiple passages, the wild type cells that I compared these edited cells to were maintained in parallel to the edited cells. I would caution against using gene editing methods that require isolation of single cells in this cell line.

Chapter 5:

GC composition bias impacts multiple genes in the gerbil genome

Statement of Author Contributions

This chapter was submitted as a manuscript formatted as a Research Article for *BMC Evolutionary Biology* (submitted June 2020; Manuscript ID EVOB-D-20-00133). The authors on this manuscript are in order: Yichen Dai, Rodrigo Pracana, Peter W.H. Holland (corresponding author).

The original idea for the project was conceived by discussion between YD, RP and PWHH. Adam Hargreaves provided advice regarding Nanopore sequencing and data analysis. RP performed orthologue alignment and analysis and provided R code used for protein dissimilarity analysis. In addition, RP provided advice regarding use of Linux and R. YD performed all remaining experimental and computational work. PWHH supervised the work of YD throughout the length of this study. Helpful discussions were held with RP, Amy Royall, Thomas Lewin, Sonia Trigueros, AH, John Mulley, Thomas Breekke, Sebastian Shimeld, and Aris Katzourakis. All authors discussed the results and revised the manuscript.

1. Abstract

Two gerbil species, sand rat (*Psammomys obesus*) and Mongolian jird (*Meriones unguiculatus*), become obese and show signs of metabolic dysregulation when maintained on standard laboratory diets. Genetic factors behind this condition remains unknown. In addition, unusual regions of high GC content are scattered across the sand rat genome, most likely generated by extreme and localized biased gene conversion. PDX1 is encoded by a gene in the most extreme GC-rich region, is remarkably divergent compared to the vertebrate conserved protein sequence and as shown in previous chapters, exhibits altered biochemical properties and biological function. In this chapter, I ask if gerbils have proteins in addition to PDX1 that are aberrantly divergent in amino acid sequence, whether they have also become divergent due to GC-biased nucleotide changes, and whether these proteins could plausibly be connected to metabolic dysfunction exhibited by gerbils. I analysed 10554 proteins with 1-to-1 orthologues in human and rodents and identified 50 proteins that accumulated unusually high levels of amino acid change in the gerbil lineage. These 50 aberrantly divergent gerbil proteins are associated with GC-skew at the nucleotide level and with previously defined high GC-regions in the gerbil genome. I also highlight four aberrantly divergent gerbil proteins, PDX1, INSR, MEDAG and SPP1, that may plausibly be associated with dietary metabolism. In conclusion, I show that through the course of gerbil evolution, GC-biased nucleotide substitution, rather than positive selection, has resulted in many aberrantly divergent proteins in the gerbil lineage. Some of these carry putatively deleterious changes that could underpin metabolic and physiological phenotypes observed in some gerbil species. I propose that these animals provide a useful model to study the ‘tug-of-war’ between natural selection and the excessive generation of deleterious substitutions mutations through biased gene conversion.

2. Introduction

Can GC composition bias lead to fixation of deleterious mutations in multiple genes, and can these mutations be maladaptive for survival? In Chapters 2 – 4, I have focused on the highly divergent sand rat *Pdx1* gene, an example of a functionally important gene trapped in a genomic region with high GC content. My results show that GC composition bias in this region may have led to multiple amino acid changes, and GC composition bias appears to be playing a game of tug-of-war with natural selection in the *Pdx1* genomic region. Apart from the sand rat *Pdx1* gene, other examples of genes accumulating fixation of deleterious mutations due to GC composition bias include the *LEP* gene in the avian lineage[58] and the *Fxy* gene in mice[206]. However, there is yet to be experimental evidence linking these genes affected by GC composition bias with maladaptive features. In Chapters 3 and 4, although I show that the sand rat PDX1 protein has significantly altered protein half-life and binding affinity to the rat *Ins1* promoter, more experiments are needed before concluding that this divergent *Pdx1* gene is the cause of physiological disorders in the sand rat. Here, I move away from the *Pdx1* gene and focus on the entire gerbil genome, which is known to contain multiple GC-rich regions affecting over 100 genes[102,207]. I ask if GC composition bias has affected gerbil genes other than *Pdx1*, and I ask whether these affected genes encode highly divergent proteins that are associated with dysfunctional metabolism observed in gerbils.

Multiple studies show that some gerbils are prone to disorders associated with pancreatic dysfunction. For example, some sand rats develop diet-induced type 2 diabetes (T2D) when fed standard laboratory rodent food[92,208] and some Mongolian jirds raised on a laboratory diet also spontaneously become obese and show poor glucose tolerance[100,101]. In humans, T2D is a complex metabolic disease caused by a mixture of genetic and environmental factors[209–211]. In addition, some T2D-related genes are not directly associated with adult β-cell function; for example, mutations in *LEP* can lead to T2D by altering appetite and body weight, while mutations altering the insulin receptor gene *INSR* can cause T2D due to insulin resistance in tissues responding to insulin[87,212].

In this chapter, I first ask how many other proteins, besides PDX1, have become 'unusually altered' in amino acid sequence during the evolution of two gerbils: sand rat (*Psammomys obesus*) and Mongolian jird (*Meriones unguiculatus*). I apply a comparative perspective, searching for genes that have changed in gerbils but are conserved in other mammals, which may include genes that have accumulated potentially deleterious changes. Second, I ask if the genes encoding these unusually altered proteins are associated with GC skew: an excess of changes from AT to GC base pairs. This tests whether unusual protein sequence change is being driven by the process of gBGC, rather than natural selection. Third, I ask if any of these proteins could feasibly be connected to metabolic disorders. Together, these three lines of investigation combine to test whether 'blind' fixation of maladaptive mutations though biased gene conversion had deleterious consequences for the biology of these desert-living rodents.

3. Materials and methods

3.1 Liver transcriptome sequencing using MinION

Total RNA was extracted from 50 – 100 mg sand rat liver tissue using TRIzol Reagent (ThermoFisher #15596026) and total RNA quality was assessed on non-denaturing agarose gel (1% agarose, 1x TBE buffer). mRNA was isolated from purified RNA samples using the polyA Spin mRNA Isolation Kit (NEB #S1560) and treated with the TURBO DNA-free kit (Invitrogen #AM1907) to further remove any contaminating DNA. cDNA libraries were prepared following Nanopore protocol Direct cDNA Sequencing (SQK-DCS108) and sequenced using flow cell FLO-MIN106 on a MinION. The same procedure was used for Mongolian jird liver RNA extraction and sequencing.

3.2 Protein dissimilarity calculation and ranking

Scripts used to perform protein dissimilarity calculation and ranking are accessible via GitHub[213]. In brief, any sites in protein alignments that contained a stop codon, an unknown residue, or a missing residue in one or more species were removed. All sites which had the same amino acid in three outgroup species were then identified. Outgroups were usually human (*Homo sapiens*), guinea pig (*Cavia porcellus*) and squirrel (*Ictidomys tridecemlineatus*), but for 7 alignments these were human, degu (*Octodon degus*) and chinchilla (*Chinchilla lanigera*) due to lack of guinea pig and squirrel sequence, for 376 alignments these were human, guinea pig and jaculus (*Jaculus jaculus*) due to lack of squirrel sequence, and for 270 alignments we used the human, squirrel and degu sequence due to lack of guinea pig sequence. Sites with the same amino acid in all three outgroups were defined as 'conserved positions'; only these sites were used to calculate protein dissimilarity. Proteins with <50 residues meeting this criterion were excluded, reducing the dataset to 9771 genes for sand rat and 10069 for Mongolian jird.

To calculate a Sneath score[214], each site which differed from a 'conserved position' was assigned a value dictated by the Sneath Index, and values summed to produce the Sneath value for the protein in that species. Epstein's Coefficient assigns different values to each amino acid difference; these were summed in the same way[215]. Scores were divided by the length of the parsed alignment to produce adjusted protein dissimilarity values.

3.3 GC-skew analysis and correlation with high GC peaks

To measure GC-skew, synonymous substitution rates for weak to strong changes (dS_{ws}) were divided by dS for strong to weak changes (dS_{sw}), estimated using the species tree and calculated from the murid-gerbil node to the terminal branch tips. Genes with $dS_{ws}/dS_{sw} >> 1$ are under strong GC-skew, and genes with $dS_{ws}/dS_{sw} << 1$ are under strong AT-skew. To test for correlation with position, 10000 random samples of 50 sand rat genes from the 9771 proteins dataset or 50 Mongolian jird genes from the 10069 dataset were selected and chromosomal location in mouse recorded; samplings were compared to locations for the 50 genes with highest adjusted Sneath value. To test for correlation with high GC peaks or islands in the sand rat and Mongolian jird genomes, we used peaks defined previously using sliding window analysis of dSws[207] (coordinates in Supplement S5.1).

3.4 Analysis of protein functions

To test if 'aberrantly divergent' sand rat proteins are enriched for biological functions, we tested for statistical overrepresentation using PANTHER and Fisher's Exact test with calculation of False Discovery Rate as the correction method[216,217]. PANTHER identified mouse orthologue IDs for 9671 of the 9771 proteins of interest (including 49 out of 50 aberrantly divergent proteins).

4. Results

4.1 Many gerbil proteins are extremely divergent compared to their homologues

To detect protein-coding genes that have become unusually divergent in gerbils, I used measures of 'relative divergence' rather than 'absolute divergence' because different types of protein evolve at different rates. High absolute sequence divergence could simply reflect low levels of selective constraint on amino acid sequence, which would result in high sequence divergence in all lineages. In contrast, high relative divergence reflects amino acid sequence changes in one species (or lineage) that is greater than the expectation for that protein.

Comparing predicted proteomes from the genomes of 12 rodents, plus human as an outgroup, gave a set of 10554 genes with one-to-one orthology between all species (Fig. 5.1; Additional file 1). To quantify relative divergence in amino acid sequence, I used two different indices for measuring the biological and chemical relevance of amino acid changes: the Sneath Index and Epstein's Coefficient[214,215]. Using 134 different categories, the Sneath Index fully characterizes the biochemical properties of all 20 natural amino acids, giving a full picture of biochemical differences between amino acid properties[214]. In comparison, other measurements take far less biochemical properties into account, with the Epstein's Coefficient only considering amino acid polarity and side chain size[215], Miyata's distance only comparing volume and polarity[218], and Grantham's distance only analysing amino acid composition, polarity, and molecular volume[219]. When compared to Miyata's distance and Grantham's distance, the Epstein's Coefficient stands out as it is sensitive to the direction of amino acid change: for example, the Epstein's Coefficient will differentiate between a change from phenylalanine to methionine and a change from methionine to phenylalanine, while the other two measures do not consider the direction of amino acid change[215,218,219].

I limited analysis to positions in multiple sequence alignments that had the same residue in three nested outgroup species (usually human, guinea pig and squirrel; asterisks in Fig. 5.1A); these residues should represent the ancestral state for rodents and they are likely functionally conserved. Some proteins had few residues meeting this criterion, reducing the dataset to 9771 sand rat genes and 10069 Mongolian jird genes. For each sequence in two gerbil species and two murine species (mouse and rat), I scored the level of divergence for each residue relative to that of the outgroup species. The sum across all residues gives a 'Sneath value' and 'Epstein's Coefficient' for each protein; these are divided by the protein alignment length to give an 'adjusted Sneath value' or 'adjusted Epstein's Coefficient'.

For most genes, the adjusted Sneath value is similar between species as expected (most dots lie along the x=y axis in Fig. 5.1B-D). For example, there is very high correlation between values for two gerbil species (Pearson correlation coefficient $r = 0.973$, $p < 2.2e\text{-}16$; Fig. 5.1B) and also between two murid rodent species ($r = 0.943$, $p < 2.2e\text{-}16$; Fig. 5.1C). The comparison between gerbil and murid proteins includes several outlier sequences (dots away from the x=y line), although the overall correlation is still high ($r = 0.909$, $p < 2.2e\text{-}16$; Fig. 5.1D). Similar results were obtained using Epstein's coefficient: very high correlation two gerbil species ($r = 0.965$, $p < 2.2e\text{-}16$), between two murid species ($r = 0.924$, $p < 2.2e\text{-}16$), and lower correlation between murids and sand rat ($r = 0.884$, $p < 2.2e\text{-}16$) (Supplement S5.2).

Several proteins lie away from the x=y axis, indicating differences in Sneath values. To identify the proteins that have 'unusually high' sequence divergence in gerbils compared to mice and rats, I calculated the difference between adjusted Sneath values for sand rat and their corresponding murid protein average values and then ranked these 'dissimilarity differences' from largest to smallest. High ranking proteins are abnormally divergent in the sand rat compared to their mouse and rat homologues, while low ranking proteins are divergent in mouse and rat compared to the sand rat. By comparing rank order with dissimilarity difference, we also looked for discontinuities in the distribution that could be used to identify a set of 'unusually' divergent proteins (Fig. 5.2). Most proteins have a dissimilarity difference close to 0, while a small number of proteins differ drastically between the sand rat and murid lineage.

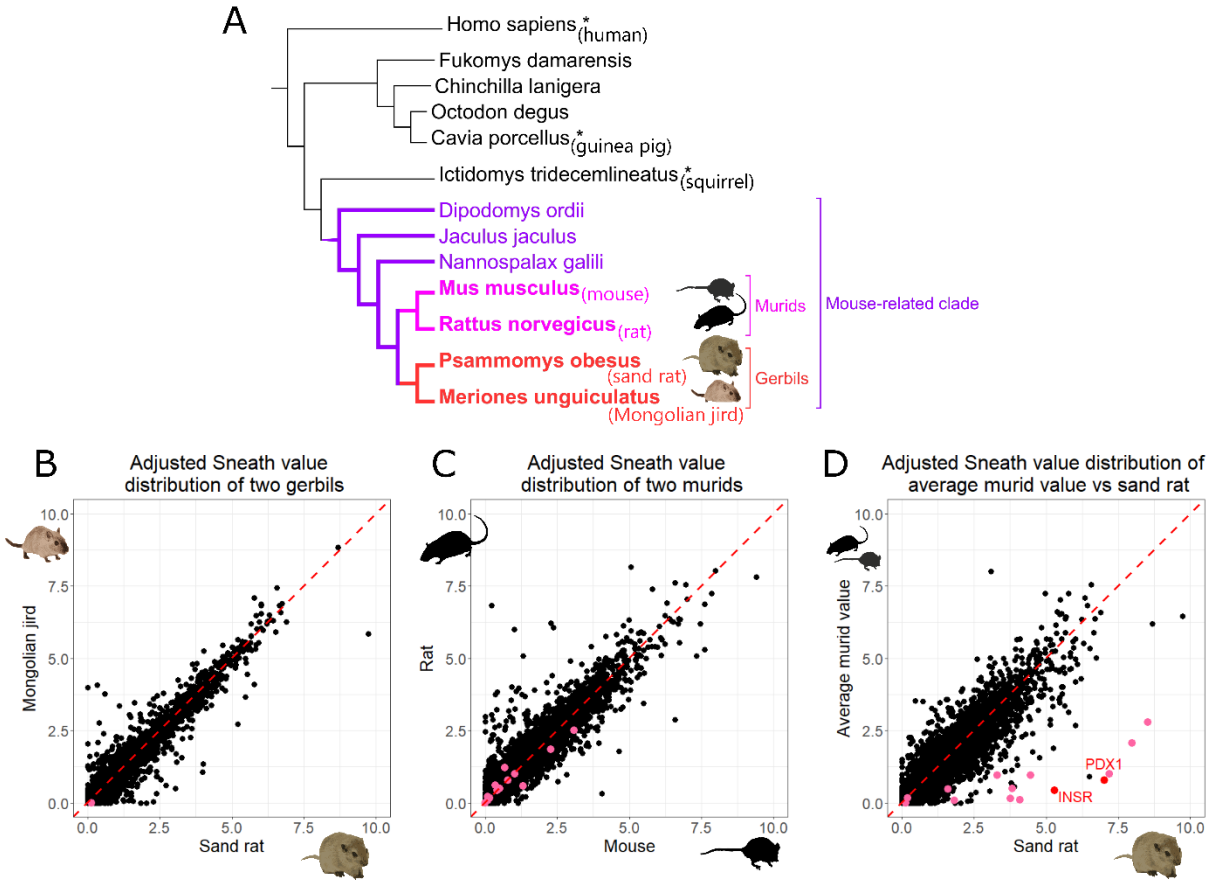


Fig. 5.1 Phylogenetic relationships and patterns of protein divergence. (A) Phylogenetic tree showing evolutionary relationships between the species. Outgroup species used for most proteins are marked with an asterisk, murid species in pink, gerbil species in red. (B–D) Adjusted Sneath values for orthologous proteins compared between (B) two gerbil species, (C) two murid species, (D) the average of two murid species and the sand rat. Each point represents one protein; proteins encoded by genes in the extreme GC-rich region of sand rat are shown in pink or (for PDX1 and INSR) in red. Only one protein encoded within the extreme GC-rich region is present in (B) due to missing genes in the Mongolian jird genome assembly.

Discontinuities in Fig. 5.2 suggest there are 50 'aberrantly divergent' protein coding genes in sand rat with difference values above 1.80, of which 16 are exceptionally unusual with Sneath dissimilarity differences above 3. In contrast, murids have 18 aberrantly divergent protein

coding genes with difference values above 1.80, with just 2 highly unusual genes showing a difference value above 3 (Fig. 5.2; Supplement S5.3).

To further test if these results are driven by unusual evolution in gerbils, and not slow evolution in mouse and rat, I also compared Sneath values for sand rat proteins to the average Sneath values of their homologues across a wider clade encompassing kangaroo rat (*Dipodomys ordii*), jerboa (*Jaculus jaculus*), blind mole rat (*Nannospalax galili*), mouse and rat. This analysis did not change the overall picture, identifying 67 sand rat proteins (dissimilarity difference > 1.80) including 17 exceptionally unusual (dissimilarity difference > 3) (Supplement S5.4).

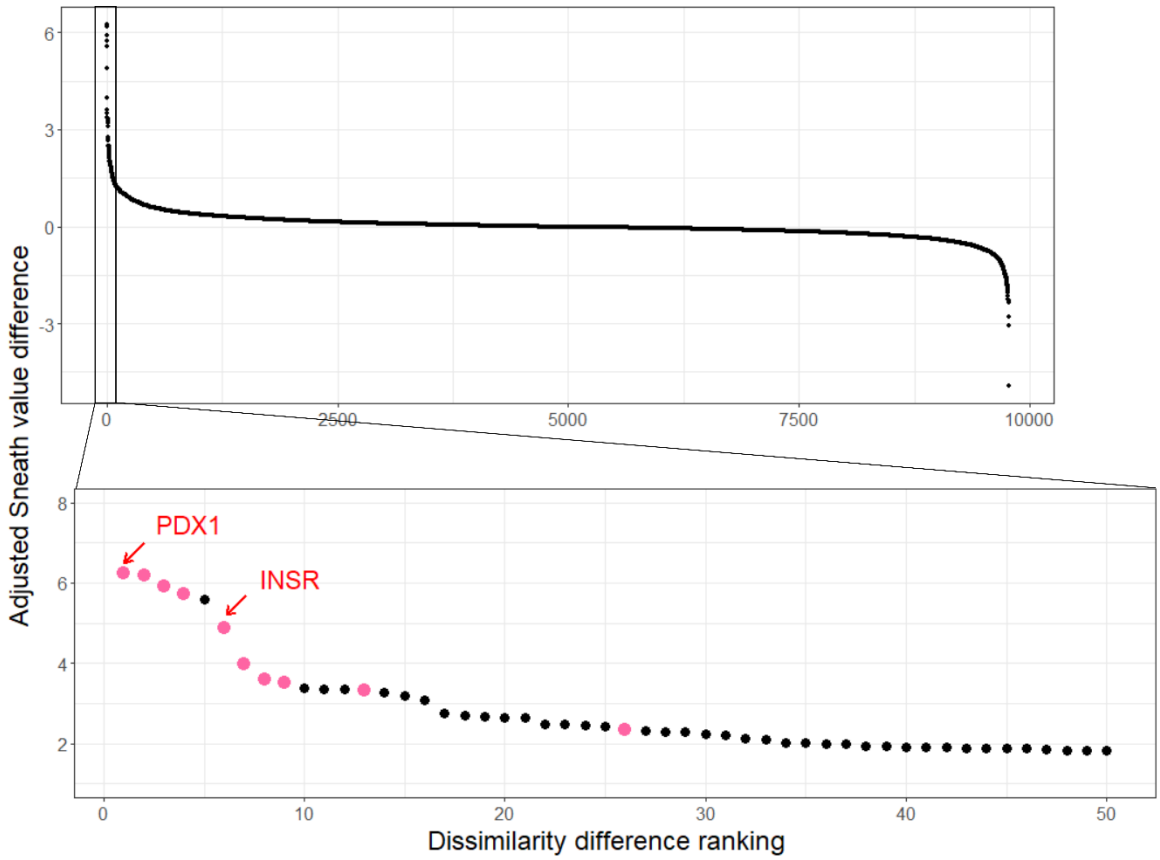


Fig. 5.2 Dissimilarity difference ranking for 9771 sand rat proteins. Dissimilarity difference ranking for each protein is plotted against its difference in adjusted Sneath value compared to the murid homologue. The top 61 ranked proteins are enlarged in the bottom plot with proteins PDX1 (rank 1) and INSR (rank 6) marked with arrows. Amongst these proteins, those encoded by genes in the extreme GC-rich region are shown in pink.

4.2 Aberrantly divergent gerbil proteins are associated with islands of high GC

It has been shown previously that gerbil genomes have islands of GC skew, and for the *Pdx1* gene (located in the island of most extreme GC skew) this is associated with radical protein sequence divergence[102,207]. To test whether excess mutational change and/or GC-skew has driven unusual protein sequence divergence in other gerbil genes, I focused on the 50 genes with aberrantly high relative sequence divergence in sand rats.

First, I asked how many of the aberrantly divergent proteins are encoded by genes located in the most extreme GC-rich region of the sand rat genome (the set of genes including *Pdx1*)[102,207]. The dataset of 9771 orthologous genes includes 14 genes from the extreme GC-rich region in the sand rat genome, out of which 10 are present in the set of 50 most aberrantly divergent protein sequences (Fig. 5.3). Extreme GC-rich genes are significantly enriched in the set of aberrantly divergent proteins (Fisher's Exact Test; p-value = 4.7e-21). These 10 genes are *Pdx1*, *Medag*, *Pex11g*, *Tex45*, *Insr*, *Trappc5*, *Pan3*, *B3glct*, *Pdap1*, and *Cdx2*.

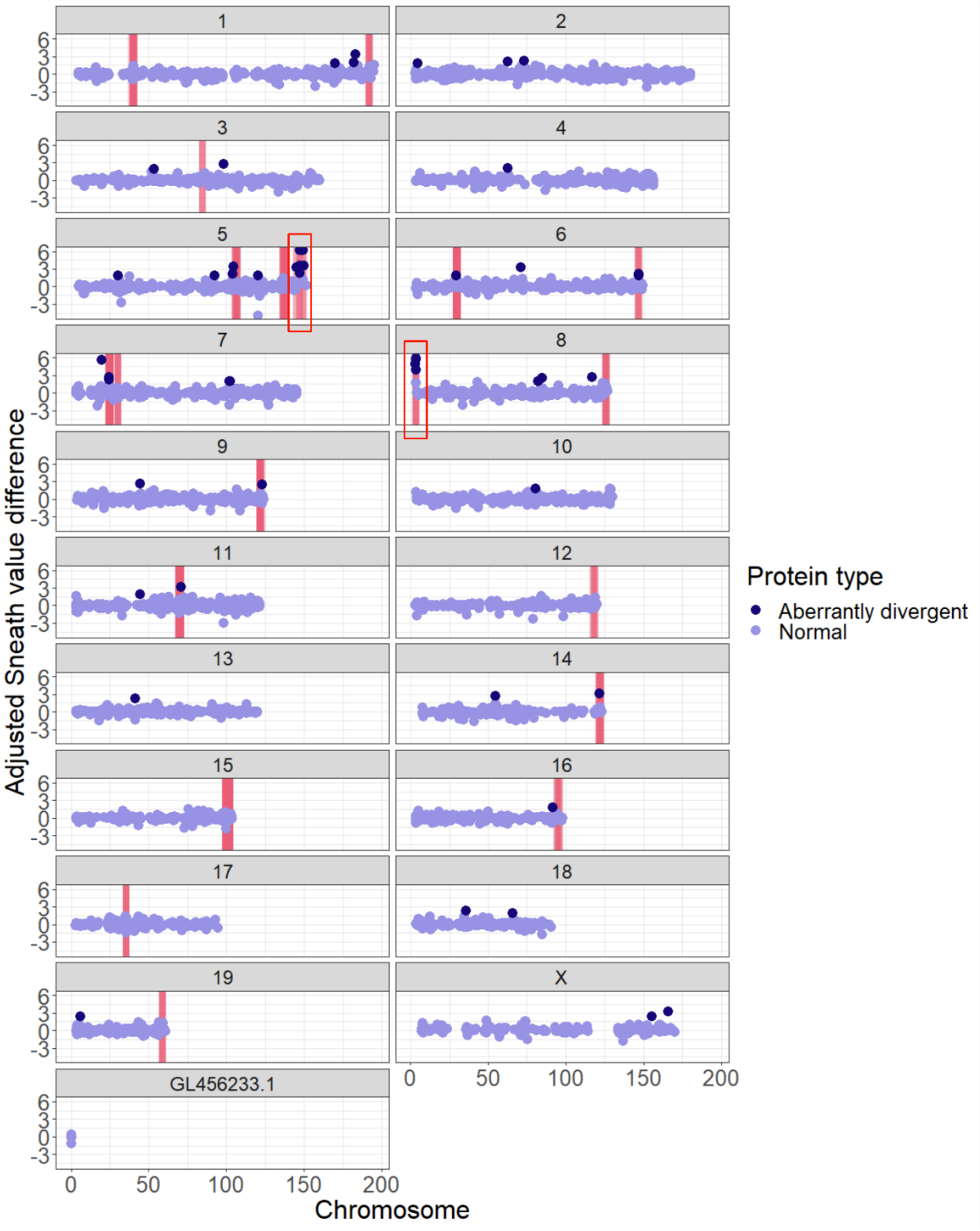


Fig. 5.3 Divergent sand rat proteins are frequently encoded by genes in GC-rich islands. Each panel shows one mouse chromosome (scale in Mb) to which the locations of sand rat orthologues are mapped. All analysed sand rat genes are displayed as dots plotted according to the midpoint position of their corresponding mouse orthologue. The position of each dot on the y-axis shows the difference in adjusted Sneath value between the sand rat and mouse orthologues. Pink lines indicate locations of GC-rich regions identified previously[102,207]. 'Clusters' of aberrantly divergent proteins with more than two proteins mapped to regions less than 1 Mb apart are marked with open red boxes.

I extended this analysis beyond the extreme GC-rich genomic region and tested if other aberrantly divergent genes are associated with other less extreme high GC regions identified previously in the sand rat genome[102,207]. In total, 19 of the 50 most aberrantly divergent proteins are encoded by genes in known high GC regions (including the 10 noted above), compared to 277 out of 9721 non-divergent genes in these regions. This is a significant enrichment of aberrantly divergent proteins encoded within the known GC-rich islands of the sand rat genome (Fisher's Exact Test; p-value = 1.0e-16). Even when the genes located in the extreme high GC region were removed from this analysis, the enrichment is significant (9/40 of divergent proteins in high GC regions, compared to 273/9717 non-divergent proteins; Fisher's Exact Test p-value = 1.5e-06). Using high GC region coordinates previously reported for the Mongolian jird[102,207], we also found a similar enrichment of aberrantly divergent proteins in GC regions (Fisher's Exact Test; p-value = 0.00096) (Supplement S5.5 and S5.6).

The above tests focus on the chromosomal location of each gene and the nucleotide compositional effects occurring in genomic regions. I also analysed genes individually to enable a higher resolution analysis of the patterns of nucleotide substitution underpinning protein divergence. Amino acid substitutions, which occurred to an extreme extent in the aberrantly divergent proteins, are caused by non-synonymous substitutions (dN). However, I focus on synonymous substitution rates (dS) to detect the presence of underlying mutational processes independent of selection on amino acid changes. Specifically, I tested whether aberrantly divergent gerbil proteins are encoded by genes that have experienced GC-skew (an excess of weak to strong mutations) and/or an excess of mutational change in all categories. I calculated dS values for all categories of nucleotide changes: weak-to-strong (dS_{ws} = rate of A or T being altered to C or G), strong-to-weak (dS_{sw}: rate of G/C to A/T), weak-to-weak (dS_{ww}) and strong-to-strong (dS_{ss}). I defined mutational 'outliers' as genes with dS values more than 2.5 times the average dS for that category[207].

Out of the 50 genes encoding aberrantly divergent proteins, 52% are dS_{ws} outliers (26/50) compared to only 4% of the remaining genes (358/9721), showing that dS_{ws} outlier genes are significantly enriched for aberrantly divergent proteins (Fisher's Exact Test; p-value = 6.3e-24). We also find that dS_{sw}, dS_{ww}, or dS_{ss} outliers are significantly enriched for aberrantly divergent proteins (dS_{sw} p-value = 4.4e-30; dS_{ww} p-value = 4.4e-19; dS_{ss} p-value = 8.8e-25) (Supplement S5.7 – 9). These results indicate that aberrantly divergent gerbil proteins are largely encoded by genes that have experienced unusually high levels of mutational change, even at synonymous sites, and especially high levels of weak-to-strong mutations.

To formally test if these elevated mutational effects are associated with GC-skew (excess of change towards G or C nucleotides), I measured the ratio of dS_{ws} to dS_{sw} (high dS_{ws} increases GC content, high dS_{sw} increases AT content). Plotting dS_{ws} for each gene against dS_{sw} shows GC-skew affecting multiple gerbil genes, as previously reported (sand rat dS_{ws}/dS_{sw} mean = 1.81; Mongolian jird dS_{ws}/dS_{sw} mean = 1.99; Fig. 5.4A and 5.4B) (5). The same effect is not seen in murid rodents (mouse dS_{ws}/dS_{sw} mean = 1.20; rat dS_{ws}/dS_{sw} mean = 1.24; Fig. 5.4C and 5.4D). Interestingly, most of the aberrantly divergent proteins in gerbils lie above the x=y line, indicating they are encoded by genes that have strong GC-skew (sand rat aberrantly divergent genes dS_{ws}/dS_{sw} mean = 2.19; Mongolian jird aberrantly divergent genes dS_{ws}/dS_{sw} mean = 1.97; Fig. 5.4A and 5.4B). These results argue that high mutation accumulation and GC-skew are major factors contributing to the evolution of divergent gerbil proteins.

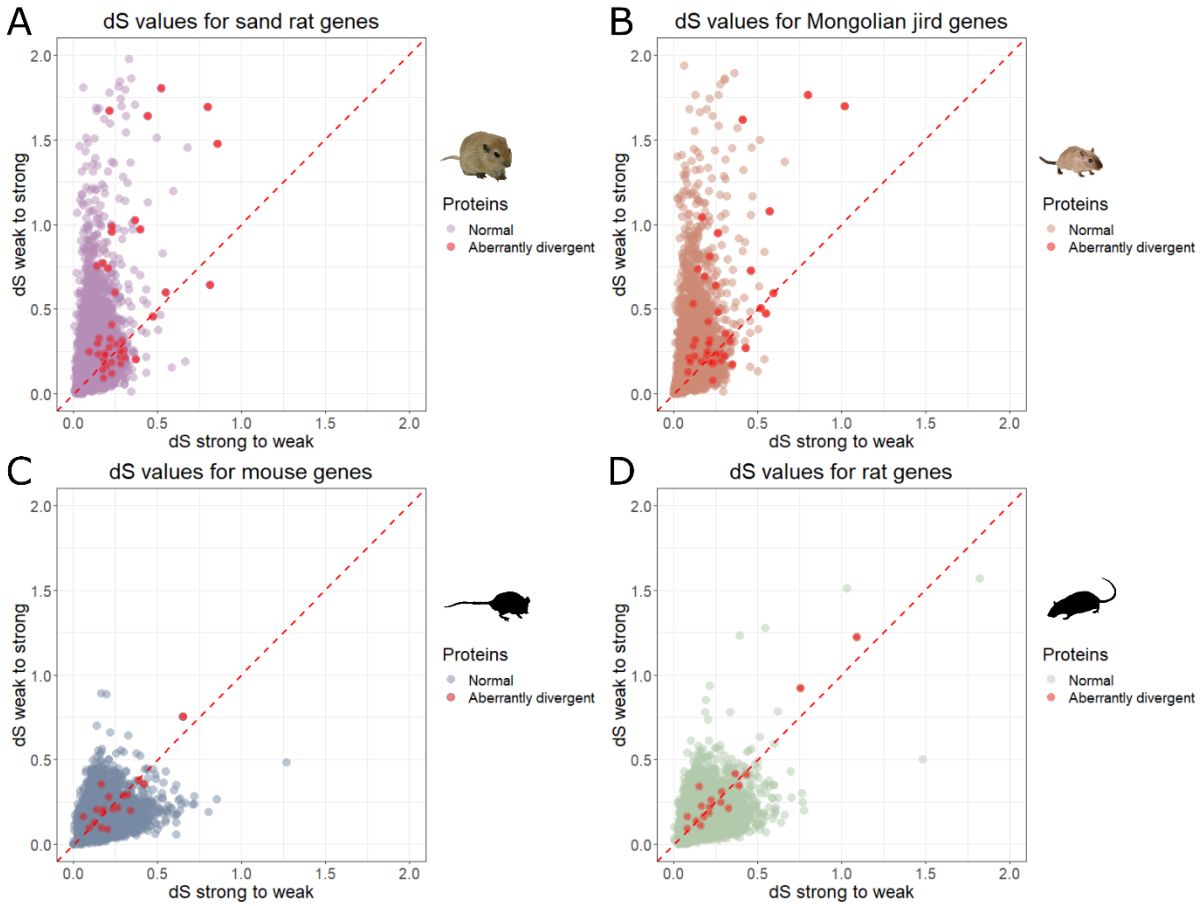


Fig. 5.4 Mutational patterns at synonymous sites for genes from four rodent species. Genes encoding aberrantly divergent proteins are highlighted with red circles. (A) Graph comparing weak-to-strong (dS_{ws}) and strong-to-weak (dS_{sw}) synonymous mutation rates for sand rat genes. The "chimney" shape indicates that many sand rat genes have undergone GC-skew; most of the aberrantly divergent proteins are in this category. Eighteen sand rat genes with dS_{ws} values larger than 2 are omitted from the graph, among which 12 encode aberrantly divergent sand rat proteins. (B) Graph comparing weak-to-strong (dS_{ws}) and strong-to-weak (dS_{sw}) synonymous mutation rates for Mongolian jird genes. The "chimney" shape indicates GC-skew; most aberrantly divergent proteins are in this category. Ten Mongolian jird genes with dS_{ws} values larger than 2 are omitted, among which two encode aberrantly divergent Mongolian jird

proteins. (C) Graph comparing dS_{ws} and dS_{sw} for mouse genes. (D) Graph comparing dS_{ws} and dS_{sw} for rat genes.

4.3 Phenotypic implications of aberrantly divergent gerbil genes

I asked if the evolution of aberrantly divergent proteins, driven by elevated mutational and gBGC processes, could have had phenotypic consequences for gerbils. Specifically, I tested if the evolution of divergent gerbil genes could be associated with the type 2 diabetes-prone phenotype observed in sand rats and Mongolian jirds.

First, I performed Gene Ontology (GO) analyses to test for over-representation of biological functions or molecular pathways in the set of aberrantly divergent genes, and as a control I also analysed the initial set of 9771 genes. There is significant enrichment of a number of biological processes, molecular functions and cellular components in the dataset of 9771 sand rat genes when compared to the mouse genome (Fisher's Exact Test, q-value < 0.05), but overrepresentation was modest with maximum fold-change 2.01. Other functions showed underrepresentation including sensory perception of smell (GO:0007608; 20-fold underrepresentation, q-value = 5.44e-137), likely a result of only using 1-to-1 orthologues causing bias against dynamic gene families with extensive gene duplication and loss. For the aberrant sand rat genes, there was no significant overrepresentation or underrepresentation for biological processes, functions or cellular components compared against the dataset of 9771 sand rat genes (q-value < 0.05). These results suggest that the aberrantly divergent gerbil proteins are associated with a diversity of biological functions, not only dietary metabolism.

To search for possible associations between aberrant protein divergence and propensity to type 2 diabetes, I took a candidate gene approach. Specifically, I asked if any of the proteins showing extreme relative divergence in gerbils have been linked to dietary metabolism in other species. I note that amongst the top 10 most aberrantly divergent sand rat proteins, four are clearly associated with lipid or carbohydrate metabolism. The genes encoding these are all located in the extreme GC-rich region of sand rat genome (Table 5.1). These four proteins are the previously discussed PDX1 transcription factor, plus MEDAG (Mesenteric Estrogen Dependent Adipogenesis), INSR (Insulin Receptor) and SPP1 (Secreted Phosphoprotein 1 or osteopontin). In Fig. 5.4, I show protein sequence alignments of PDX1, INSR, MEDAG and SPP1 to highlight some of the highly unusual amino acid changes observed in sand rat and related gerbil species, compared to other mammals. In the PDX1 protein, there is a high degree of conservation across vertebrates of the hexapeptide domain (a cofactor binding domain) and the homeodomain (DNA-binding and sequence recognition domain), but extreme divergence in three gerbil species (Fig. 5.5A)[102]. At the same time, the three amino acids in the PDX1 homeodomain that have been previously associated with T2D in humans are conserved in the three gerbils (Fig. 5.5A). For INSR, a strong candidate for association with metabolic function, there are many amino acid changes unique to the gerbil lineage throughout the protein, although again nearly all key amino acid sites previously associated with T2D in humans remained conserved (Fig. 5.5B). An exception is K1095Q located in the INSR tyrosine kinase domain; an identical point mutation is associated with insulin resistance and T2D in humans, although the biochemical effect of this missense mutation remains unknown[220]. As for MEDAG and SPP1, I have not found literature linking protein variants and metabolic dysfunction in humans,

and key functional domains in both proteins remains to be elucidated. Regardless of this lack of identified essential residues and key domains, Fig. 5.5C shows that gerbils have 15 out of 56 residues that differ from the conserved sequence in this MEDAG protein region that is well-conserved in mammals. In addition, Fig. 5.5D shows 14 out of 51 non-conserved residues in the sand rat SPP1 protein, of which 6 sites are non-conserved only in gerbils.

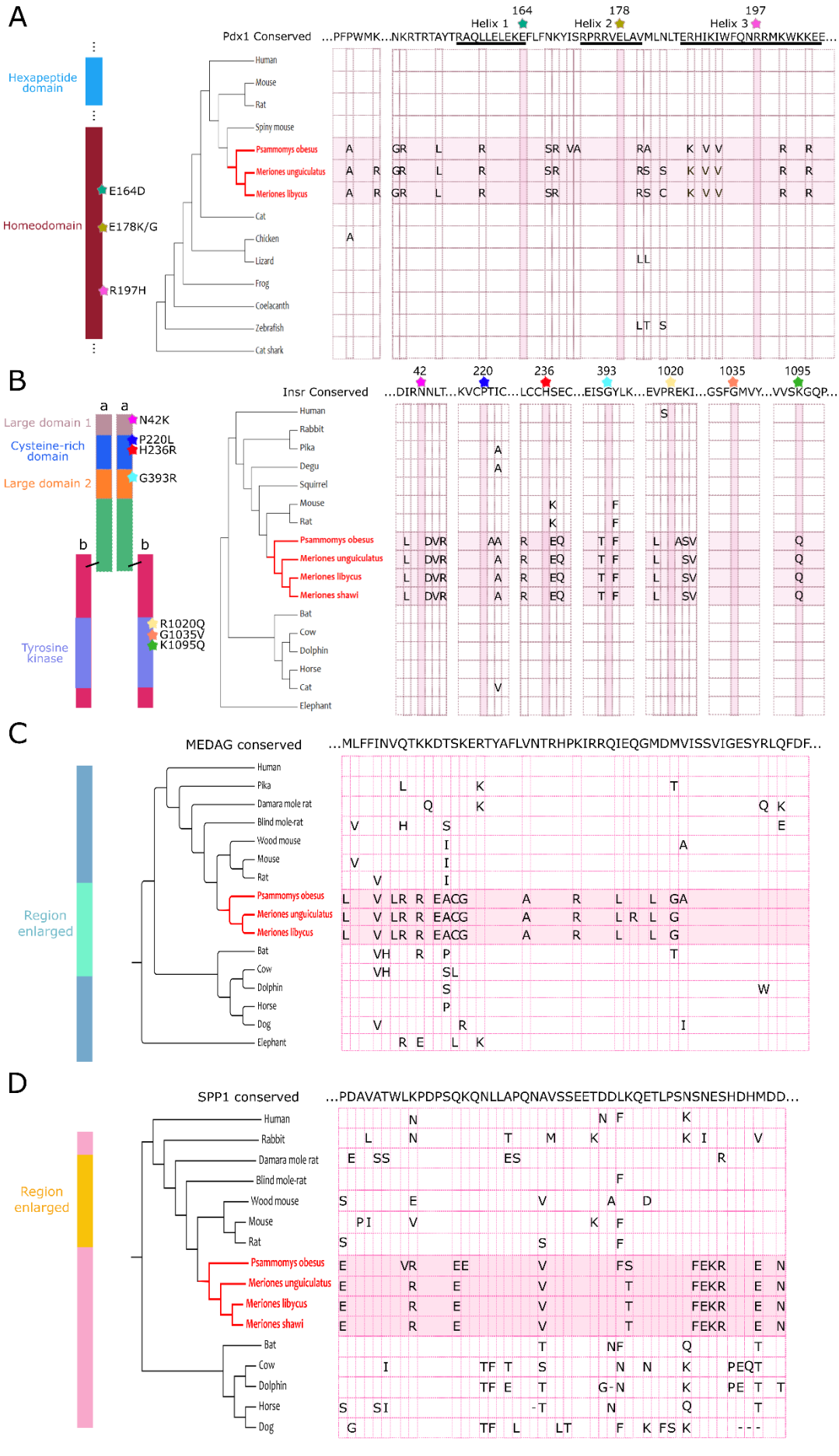

A
Pdx1 Conserved
Helix 1
Helix 2
Helix 3
Hexapeptide domain
Homeodomain
E164D
E178K/G
R197H
Human
Mouse
Rat
Spiny mouse
Psammomys obesus
Meriones unguiculatus
Meriones libycus
Cat
Chicken
Lizard
Frog
Coelacanth
Zebrafish
Cat shark
B
Insr Conserved
Large domain 1
Cysteine-rich domain
Large domain 2
Tyrosine kinase
N42K
P220L
H236R
G393R
R1020Q
G1035V
K1095Q
Rabbit
Pika
Degu
Squirrel
Meriones shawi
Bat
Cow
Dolphin
Horse
Elephant
C
MEDAG conserved
Region enlarged
Damara mole rat
Blind mole-rat
Wood mouse
D
SPP1 conserved
Region enlarged

Fig. 5.5 Alignment of protein domains in PDX1, INSR, MEDAG, and SPP1 proteins. (A) Alignment of the conserved PDX1 hexapeptide and homeodomain sequence from representative vertebrate species. The three gerbil species are shown in red and sites where amino acid substitutions are associated with T2D in humans are marked with a star. (B) Alignment of representative regions in the INSR protein, including both large domains, the cysteine-rich and tyrosine kinase domain. The four gerbil species are shown in red and sites where amino acid substitutions are associated with T2D in humans are marked with a star. (C) Alignment of a representative region in the MEDAG protein. The three gerbil species are shown in red. (D) Alignment of a representative region in the SPP1 protein. The four gerbil species are shown in red.

Rank	Gene	GC-rich region	Sand rat dS_{ws} value	Encoded protein
1	*Pdx1*	Yes	20.09	Pancreatic duodenal homeobox. Transcription factor involved in pancreas development and β-cell function.
2	*Medag*	Yes	3.09	Mesenteric estrogen dependent adipogenesis. Involved in adipocyte differentiation and glucose uptake.
3	*Pex11g*	Yes	4.58	Peroxisomal membrane protein. Involved in growth and division of peroxisomes (organelles partaking in lipid biosynthesis).
4	*Tex45*	Yes	1.48	Testis-expressed protein. Understudied gene expressed in testis.
5	*Cblc*	No	0.60	E3 ubiquitin-protein ligase. Involved in protein degradation; down-regulates receptor-tyrosine kinase (RTK) signaling.
6	*Insr*	Yes	7.08	Insulin receptor. RTK receptor involved in insulin signaling and stimulation of glucose uptake.
7	*Trappc5*	Yes	34.40	Trafficking protein particle complex subunit. Implicated in vesicular transport from ER to Golgi.
8	*Pan3*	Yes	10.66	Subunit of PAN2-PAN3 deadenylase complex. Trim poly(A) tail prior to degradation.
9	*B3glct*	Yes	2.29	Beta 3-glucosyltransferase. Transfers glucose onto fucose on TSR repeat proteins.
10	*Spp1*	No	0.22	Secreted phosphoprotein 1 or osteopontin. Implicated in bone mineralization, immune system, and diet-induced insulin resistance.

Table 5.1 A list of the sand rat genes that encode the most aberrantly divergent sand rat proteins ordered by ranking, whether these genes are in GC-rich regions, dS_{ws} value and a short description of the function of their encoded protein.

4.4 MinION sequencing did not yield further information on GC-rich genes

Illumina sequencing was previously used to assemble gerbil genomes[102,221], however the GC-rich regions remain incomplete, likely due to issues such as GC-rich region fragility and PCR difficulties. Due to this, I have partial or missing information for several essential genes, such as the insulin receptor gene *Insr*, which is missing parts of the 5' and 3' region. To allow better analysis of high GC genes and to facilitate functional studies of these genes, I test if long-read sequencing technology can sequence the sand rat insulin receptor *Insr* in full length. Out of the two most prominent long-read sequencing technologies available, I chose Nanopore over PacBio given that the Nanopore sequencing device, the MinION, is portable and easy to use. As I am aiming to sequence the *Insr* coding region, I chose to sequence the liver transcriptome, a tissue where the *Insr* gene is highly expressed in mammals[79].

For sand rat liver tissue, a total of 294198 reads were sequenced, out of which 124029 were classified as failed reads while 170169 reads were classified as successful reads by the Nanopore developed programme MinKNOW[222]. For the 170169 good quality reads, the average read length is 1508 bp and the average read GC content is 47.14% (Fig. 5.6). All good quality reads have been deposited on the SRA database (accession ID: SRR12284189).

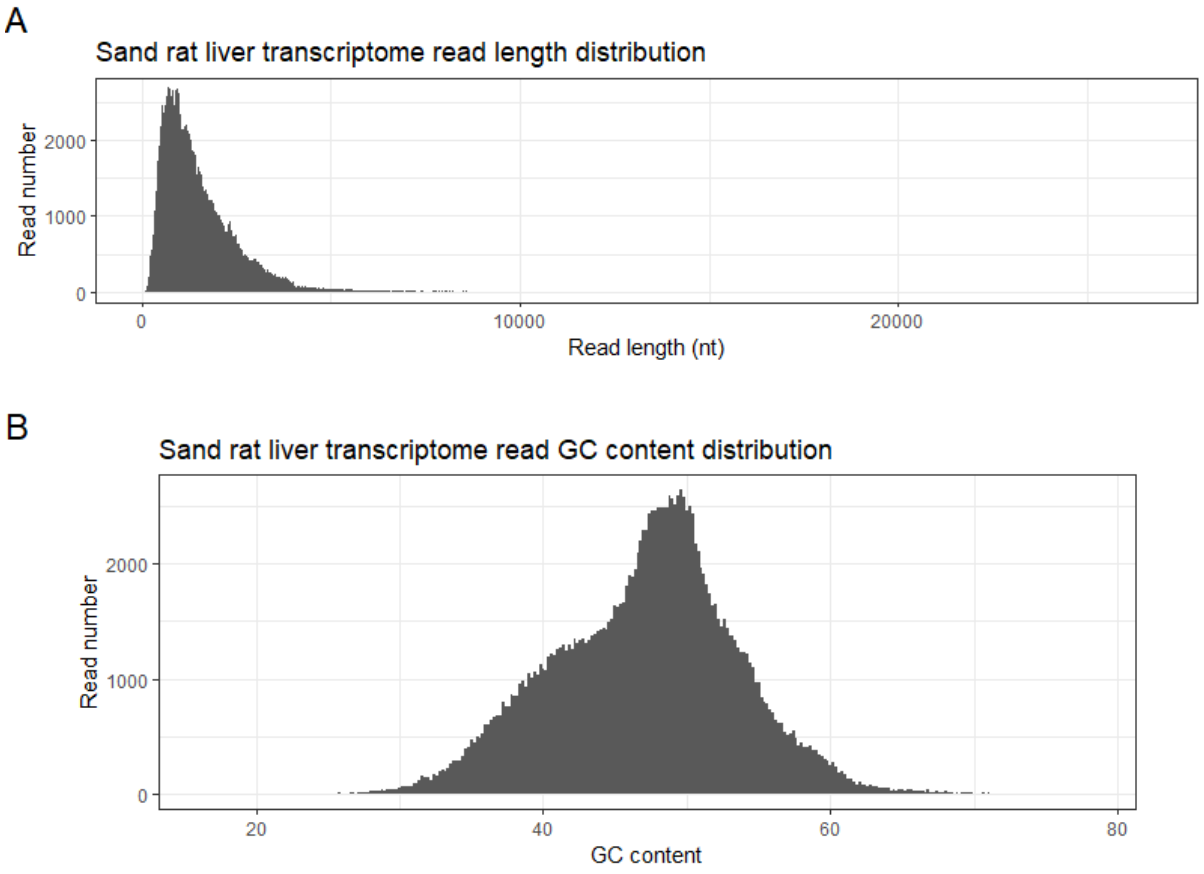


Fig. 5.6 Distribution of good quality sand rat liver transcriptome reads. A) Read length. The shortest read 123 bp and the longest read is 26567 bp. B) Read GC content. The read with the lowest GC content is 16.35% GC while the read with highest GC content is 78.07% GC.

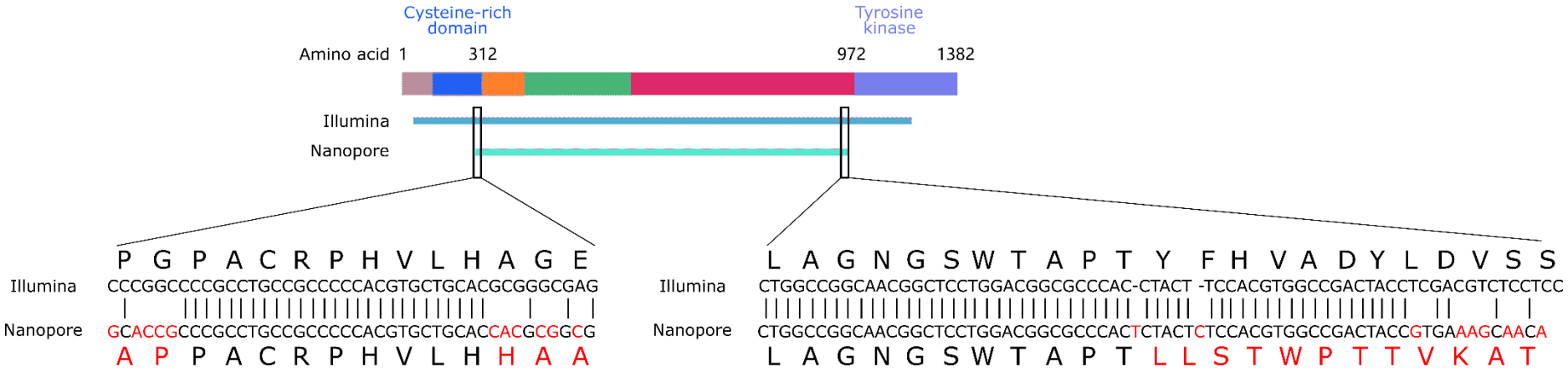


Fig. 5.7 Alignment of the sand rat *Insr* gene sequenced using Illumina and Nanopore. Key INSR protein regions are shown, with human INSR coordinates given. The Illumina sequenced sand rat *Insr* gene is incomplete and missing amino acids near the start and end of the INSR protein. In comparison, the Nanopore sequenced sand rat *Insr* gene is shorter and lacks regions that code for approximately 300 amino acids near the INSR N-terminal and the entire tyrosine kinase domain. As shown in the DNA alignment, the raw Nanopore read contains many nucleotide differences that are likely sequencing errors, along with nucleotide insertions.

However, only one read was identified as sand rat *Insr*. A tBLASTn search using the partial sand rat INSR protein sequence as the query, resulted in one hit in all good quality Nanopore raw reads that was confirmed to be *Insr* using reciprocal BLAST. Searching failed reads did not yield any sand rat *Insr* hits, and neither did searching all reads using the highly conserved INSR tyrosine kinase domain as the query sequence yield any more *Insr* hits. The Illumina-assembled partial sand rat *Insr* sequence is 3882 bp and 73.9% GC in comparison to the 2970 bp, 70% GC sand rat *Insr* read sequenced using Nanopore (Fig. 5.7). For the MinION *Insr* read, 988 – 2970 bp aligns to the 855 – 2758 bp region in the Illumina *Insr* sequence with 89% identity (Fig. 5.7). This relatively low level of sequence similarity is caused by the large number of indels in the Nanopore read, with 6% of the sequence possessing either an insert or deletion (Fig. 5.7). In addition, the first 988 bp of the Nanopore *Insr* read could not be aligned to the Illumina sequence or identified using reciprocal BLAST. The lack of *Insr* reads in the sand rat liver transcriptome is likely caused by difficulties sequencing GC-rich sequences using Nanopore technology, not low expression levels given the Illumina transcriptome.

Similarly, the Mongolian jird liver transcriptome had 594195 good quality reads and 451697 failed reads. The average read length is 1304 bp (Fig. 5.8A) and the average read GC content is 47.05% (Fig. 5.8B). Although the Mongolian jird liver transcriptome yielded more reads than the sand rat transcriptome, none were identified as the Mongolian jird *Insr* gene. This gene is also GC-rich, as confirmed by the partial sequence assembled using Illumina sequencing. All good quality reads have been deposited on the SRA database (accession ID: SRR12284190).

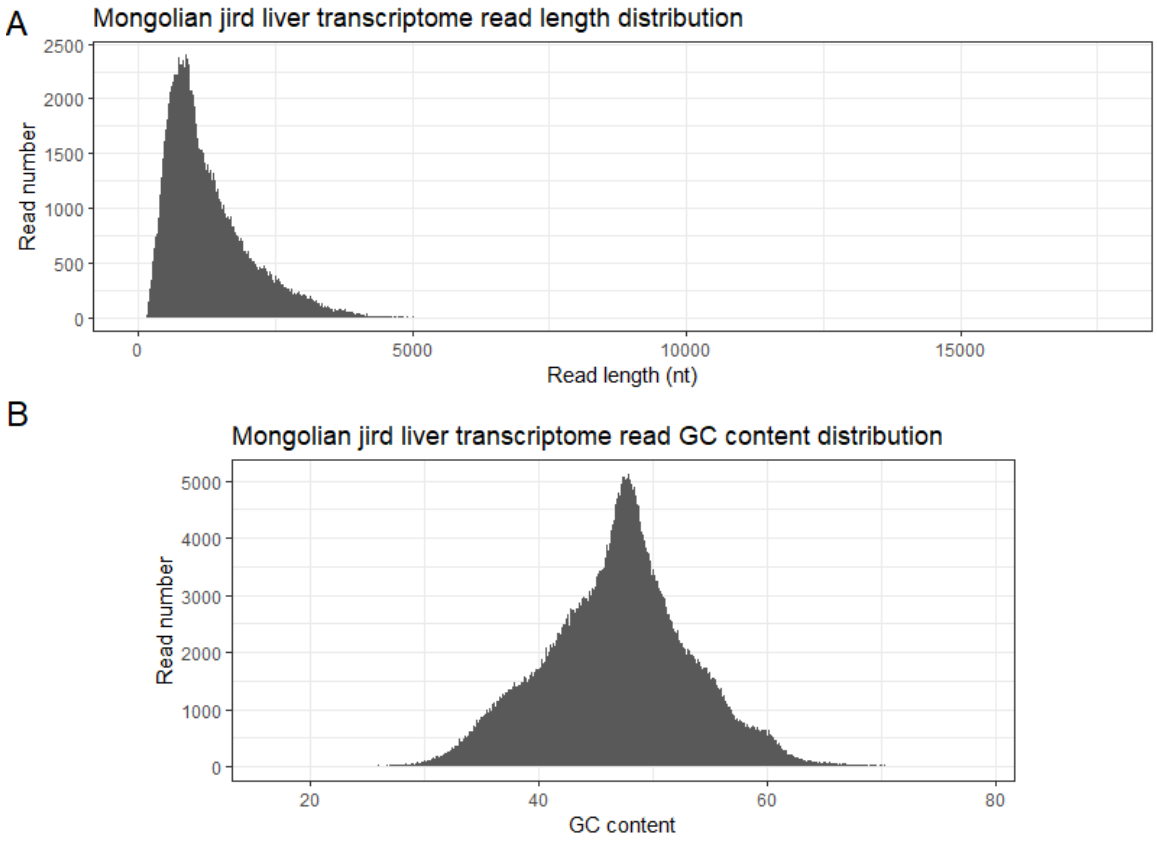


Fig. 5.8 Distribution of good quality Mongolian jird liver transcriptome reads. A) Read length. The shortest read 70 bp and the longest read is 17616 bp. B) Read GC content. The read with the lowest GC content is 16.32% GC while the read with highest GC content is 78.51% GC.

5. Discussion

When unusual or divergent proteins are observed in some species and not others, this is generally thought to result from natural selection acting to adjust or optimize a protein for a new or modified function. The contributions of mutation rate variation and recombination frequency differences between species are easily overlooked. In this chapter, I move away from investigating the *Pdx1* gene and focus on the entire gerbil genome, asking whether recombination-related processes are affecting the evolution of genes in this lineage, including genes that are highly conserved in other mammals[102,207]. I ask if gerbils have an abnormal number of aberrantly divergent genes compared to related rodent species, and test if these genes have become divergent because of association with GC-rich genomic regions and GC-skew, rather than because of positive selection. In addition, I ask whether these aberrant genes could be linked to physiological abnormalities observed in some gerbils.

Aberrantly divergent gerbil proteins have an extremely high number of amino acid changes at sites that are conserved in other rodent lineages or across the vertebrates. One explanation for the presence of these aberrantly divergent proteins is that strong positive selection and adaptive changes occurred in the gerbil lineage. However, I detect signals of positive selection in only 3 out of 50 aberrantly divergent proteins (Additional file 2). These 3 proteins are TEX37 (Testis Expressed 37), SYCP2L (Synaptonemal Complex Protein 2 Like) and IL15 (Interleukin 15); they are not associated with GC-skew nor located in GC-rich regions. TEX37 is predominantly expressed in the testis[223] while SYCP2L is mainly expressed in the ovary[224]. In addition, gene association studies report a correlation between SYCP2L variants and lipid metabolism, although the mechanism remains unknown[225,226]. IL15 is a proinflammatory cytokine that activates T-cells and aberrant activity has been linked to destruction of pancreatic beta-cells in type 1 diabetes (T1D)[227,228]. However, autoimmune destruction of β-cells has not been proposed as a cause of metabolic dysfunction in gerbils. In general, reproductive proteins and immunity-related proteins are under positive selection in many species[229,230]. For the 47 other aberrantly divergent gerbil proteins, the extent of change exceeds other lineages and there is no clear signature of positive selection, making an adaptive hypothesis unlikely.

A second possibility is that these aberrant proteins have lost function and are randomly accumulating amino acid changes. This is again unlikely given that no early stop codons or frameshifts in the coding sequences of these genes are observed. A third hypothesis is that genomic processes have inadvertently caused unusual sequence changes to become fixed in populations, even if maladaptive at the time of origination. For example, GC-biased gene conversion (gBGC) can convert heterozygous sites to homozygous sites, and overall gBGC promotes the fixation of neutral and slightly deleterious weak-to-strong substitutions[46,231]. This hypothesis predicts that aberrantly divergent proteins would be encoded by genes located in GC rich genomic regions, as these are localized regions where gBGC has occurred at high rates. Consistent with this prediction, we find a significant enrichment of aberrantly divergent genes mapping to previously described high GC regions in the gerbil genome[102,207]. Also consistent with this hypothesis, aberrant gerbil genes are enriched for dS_{ws} outliers and show strong GC-skew in gerbils (sand rat average dS_{ws}/dS_{sw} = 2.19) compared to murids (mouse average dS_{ws}/dS_{sw} = 1.20).

I argue, therefore, that aberrantly divergent gerbil genes are likely to have arisen through unusually high levels of nucleotide substitution, including especially high levels of weak-to-strong substitutions resulting in GC-skew. Given the degree of amino acid change observed in these 50 gerbil proteins, the high rates of nucleotide substitution have likely resulted in fixation of slightly deleterious mutations in these proteins, with natural selection only successfully removing individuals with the most severely deleterious amino acid changes. I suggest that gerbil evolution has witnessed a tug-of-war between localized GC-skew in certain genomic regions causing fixation of deleterious mutations and natural selection preserving essential genes trapped in these 'quagmires' of GC-skew.

The two gerbil species focused on in this study have also been reported to become obese and exhibit metabolic abnormalities when maintained on a standard laboratory diet or a high fat diet[92,100,232,233]. For example, sand rat islet cells show more pancreatic β-cell damage compared to rat islets when exposed to high glucose concentrations, and even healthy sand rats on plant-based diets do not have a strong response to human insulin[97,234]. Could these metabolic disorders be related to the accumulation of deleterious mutations in gerbils, driven by excessive biased gene conversion? Chapters 3 and 4 suggest that genetic changes in sand rat *Pdx1* are very likely linked to abnormal regulation of insulin expression, and there is a plausible connection between amino acid changes linked to GC-skew and metabolic dysfunction in gerbils. In this chapter, I highlight four aberrantly divergent gerbil proteins for which the human and mouse orthologues have been associated with dietary metabolism: these proteins are PDX1, INSR, MEDAG and SPP1. For two genes, the human orthologues have been associated directly with type 2 diabetes or other metabolic diseases, and these associations can be tracked down to single amino acid substitutions or other small mutations. In gerbils, there is dramatic amino acid change at multiple conserved sites in these two genes. For the other two genes, functional studies in mouse or human indicate key roles in adipose tissue.

PDX1 is a homeodomain transcription factor essential for vertebrate pancreatic development and normal β-cell function[85,110,121]. Given its essential roles in embryonic development and adult physiology, it is not surprising that PDX1 is highly conserved across vertebrates. Indeed, PDX1 mutations have been linked metabolic disease in humans spanning various degrees of severity. The homeodomain is the mostly highly conserved protein domain in PDX1 and point mutations in this region severely compromise PDX1 function in humans. For example, an individual born with pancreas agenesis was reported to carry a E164D substitution on one allele and a E178K substitution on the other, with both mutations shown to significantly shorten PDX1 protein half-life[147]. In addition, the mutation R197H has >50% decreased binding affinity with the human insulin promoter sequence and is associated with type 2 diabetes[81]. These conditions in humans are associated with single amino acid changes to the homeodomain, yet by contrast the sand rat PDX1 homeodomain has a total of 15 amino acid differences compared to the conserved sequence (Fig. 5.5A). It is the most aberrantly divergent protein out of 9771 analysed proteins in the present study. Although the 15 substitutions do not include changes at the three best-characterized disease-associated sites in humans, it is likely that the combination of large number of substitutions would radically compromise protein function.

INSR, the insulin receptor, is the cell surface receptor for insulin and IGF peptides and initiates the insulin signalling pathway in vertebrates[235]. In sand rat, the INSR protein is the 6th most aberrantly divergent protein, carrying many amino acid substitutions at otherwise conserved sites. Other vertebrates carry small changes to the INSR protein, and these highlight the possible phenotypic consequences of amino acid substitutions. Recent work has uncovered adaptive change in the INSR protein of Mexican cavefish, notably a P211L substitution mutation (equivalent to P220L in human INSR) that has been fixed by positive selection[236]. This mutation reduces binding capability to the insulin peptide and results in insulin resistance in the cavefish, which paradoxically may be a trait beneficial to survival in a cave environment lacking a stable food supply[236]. In humans, deleterious mutations affecting INSR function have also been reported, with over 30 different point mutations described in clinical reports[87]. In gerbils, such as sand rat, we find that although there are a large number of unusual changes, these generally do not match the individual mutations associated with human metabolic conditions, apart from point mutation K1095Q, nor the P220L mutation reported in Mexican cavefish populations. Concerning point mutation K1095Q, this was reported in a small cohort study of diabetes patients with insulin resistance, but the exact biochemical effects of this point mutation are not yet known[220]. In summary, these observations suggest that severely detrimental amino acid changes occurring in PDX1 and INSR of gerbils are likely to have been removed by natural selection, while changes in residues with lower levels of functional importance have been allowed, even though these may have been deleterious at the time of origin. Nonetheless, it is difficult to predict the consequences of all amino acid changes in the gerbil based on these comparisons, especially as there are no reports approaching this dramatic extent of amino acid change in other organisms.

For the two other aberrant proteins putatively associated with dietary metabolism, MEDAG and SPP1, it is difficult to make comparisons to variants associated with human phenotypes. However, the known functions of these proteins suggest plausible association with dietary metabolism. MEDAG is a well-conserved pro-inflammatory protein (91% sequence identity between humans and mice) that promotes adipocyte differentiation and regulates adipocyte glucose uptake[237]. A study comparing islet transcriptomes of diabetic and healthy individuals reported significantly different expression of MEDAG between these two groups[238], and a clinical study has reported association between copy number variation in the MEDAG loci and obesity[239]. SPP1, also known as osteopontin, is a multifunctional protein best known for its role in biomineralization and bone remodelling. However, SPP1 also has roles in in adipose tissue and contributes to adipose tissue inflammation and insulin resistance[240]. Mice fed on a high fat diet, or with genetic predisposition to obesity, show enormous elevation of SPP1 expression levels[241], while blocking SPP1 function by antibodies or genetic mutation reduces adipose inflammation and improves insulin sensitivity[242,243]. In humans, SPP1 expression is also reported to be elevated in adipose tissue of obese individuals[241,244]. Therefore, it is possible that the amino acid changes in these gerbil proteins are causing changes in protein function and could contribute to the unusual dietary physiology of gerbils.

In this chapter, I show there are more than twice the number of aberrantly divergent proteins in gerbils compared to mice and rats, and that to a large extent the genes that encode these aberrant proteins are associated with GC-skew and previously reported GC-rich genomic

regions[102,207]. I propose that a tug-of-war between GC-skew and natural selection has been taking place during gerbil evolution, leaving behind an excess of deleterious mutations in several aberrant gerbil genes, several of which may be partly responsible for the abnormal metabolic phenotypes observed in some gerbil species.

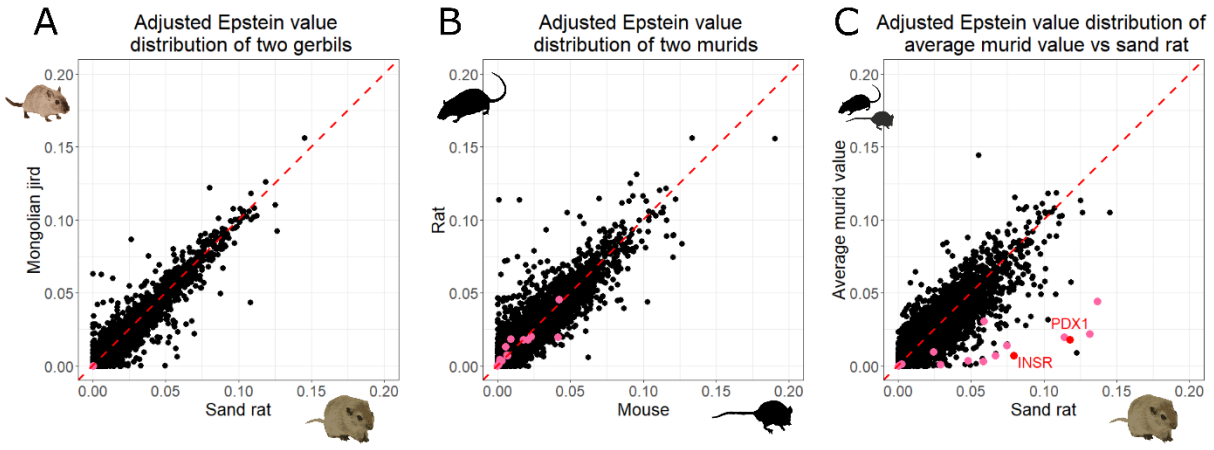
6. Supplementary information

A. High GC regions in the Mongolian jird genome (1 Mbp window, 0.25 Mbp step)			
Chromosome	**Start**	**End**	**Length (bp)**
1	38,735,985	41,485,985	2,750,000
1	191,735,985	192,735,985	1,000,000
3	84,269,840	85,269,840	1,000,000
5	104,917,342	107,167,342	2,250,000
5	136,167,342	137,917,342	1,750,000
5	*145,128,769*	*149,762,599*	*4,500,000*
6	29,118,273	30,368,273	1,250,000
6	146,618,273	147,618,273	1,000,000
7	23,720,729	25,720,729	2,000,000
7	30,970,729	31,970,729	1,000,000
8	*3,150,922*	*3,681,255*	*500,000*
8	125,200,606	126,700,606	1,500,000
9	121,297,555	122,797,555	1,500,000
11	69,541,271	71,291,271	1,750,000
12	117,314,511	119,064,511	1,750,000
14	120,701,122	122,701,122	2,000,000
15	100,271,842	101,521,842	1,250,000
16	94,103,884	96,353,884	2,250,000
18	6,101,319	7,101,319	1,000,000

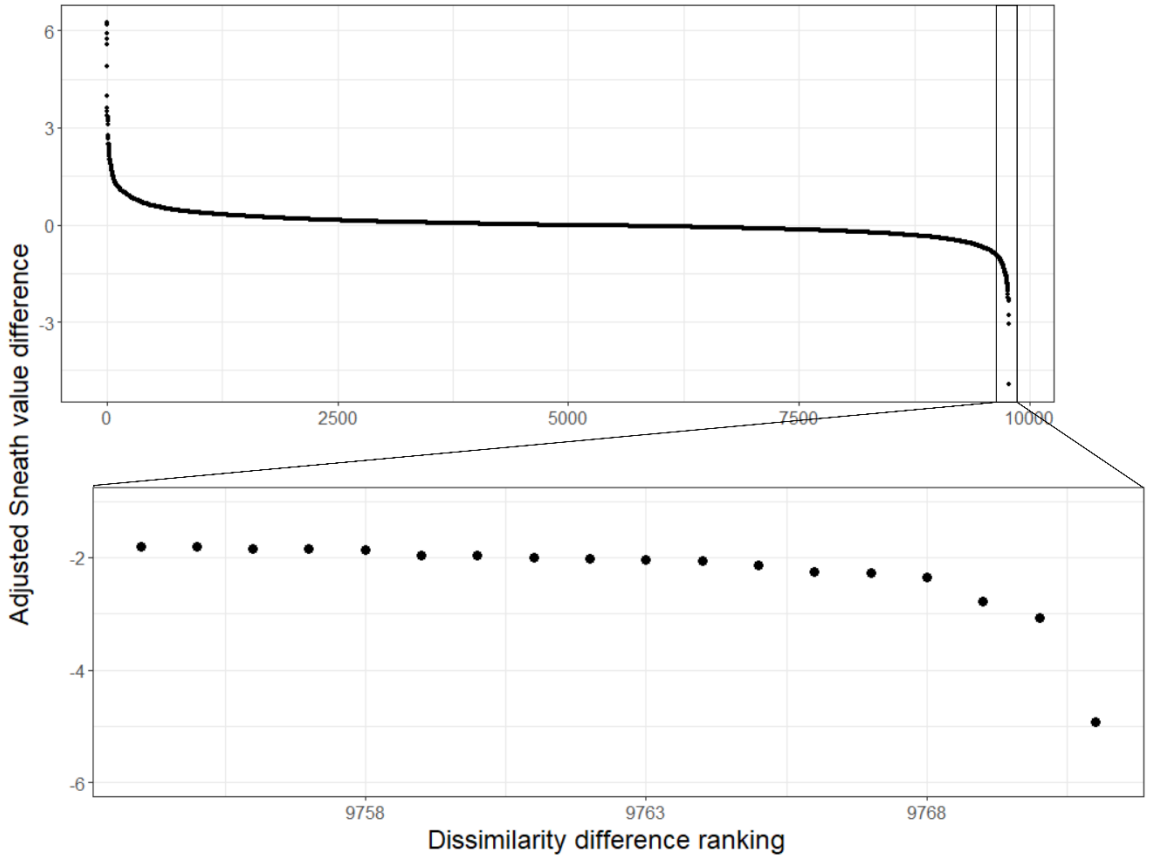
B. High GC regions in the sand rat genome (1 Mbp window, 0.25 Mbp step)			
Chromosome	**Start**	**End**	**Length (bp)**

1	38,735,985	39,735,985	1,000,000
1	39,985,985	41,485,985	1,500,000
1	191,735,985	192,735,985	1,000,000
3	84,269,840	85,269,840	1,000,000
5	104,917,342	107,417,342	2,500,000
5	136,167,342	137,917,342	1,750,000
5	*145,128,769*	*149,762,599*	*4,500,000*
6	29,368,273	30,618,273	1,250,000
6	146,618,273	147,618,273	1,000,000
7	23,720,729	26,220,729	2,500,000
8	*3,150,922*	*3,681,255*	*500,000*
8	125,200,606	126,700,606	1,500,000
9	121,297,555	123,047,555	1,750,000
11	69,541,271	71,291,271	1,750,000
12	117,314,511	119,064,511	1,750,000
14	120,951,122	122,701,122	1,750,000
15	99,771,842	101,521,842	1,750,000
15	101,771,842	102,771,842	1,000,000
16	94,103,884	96,353,884	2,250,000

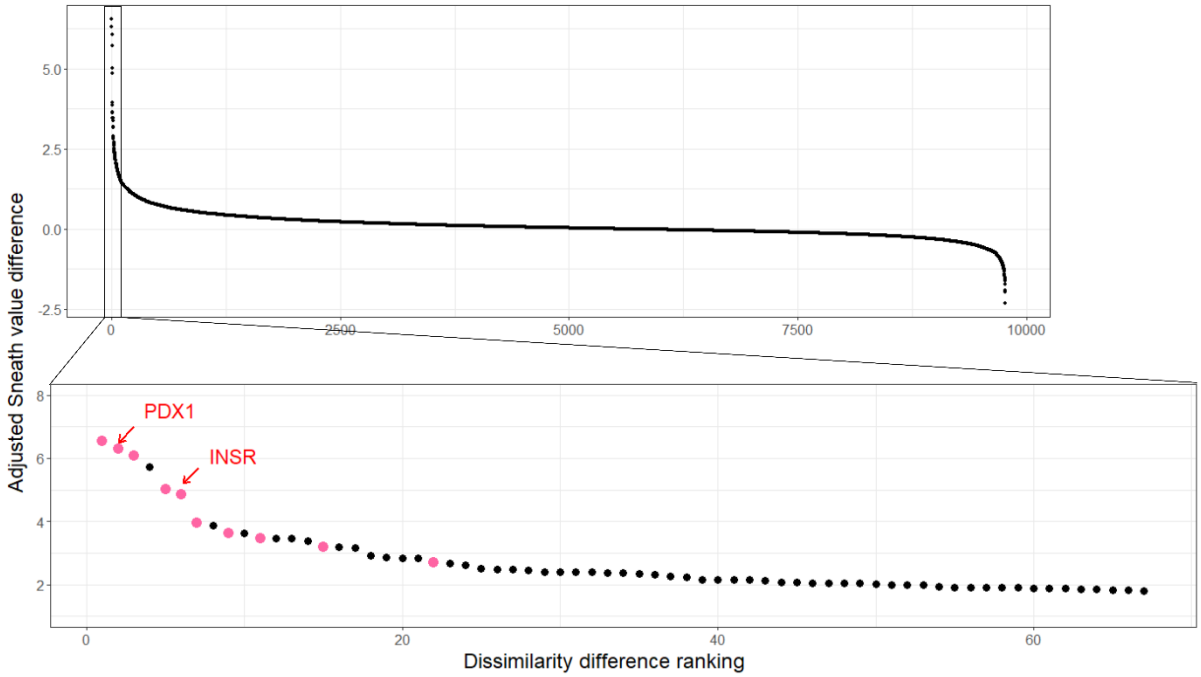
Supplement S5.1 List of high GC regions used for analysis defined using sliding window analysis of average dS_{ws} and direct comparison of gene coding sequence[102,207]. These locations are mapped using mouse chromosome coordinates. Most regions are obtained from dS_{ws} outlier analysis, with the exceptions being regions in italics, which are obtained from gene coding sequence comparison focusing on the ParaHox region and surrounding genes. A) shows high GC regions in the Mongolian jird genome while B) shows high GC regions in the sand rat genome.



Supplement S5.2 Patterns of protein sequence divergence. Adjusted Epstein values orthologous proteins compared between (A) two gerbil species, (B) two murid species, (C) the average of two murid species and the sand rat. Each point represents one protein; proteins encoded by genes in the extreme GC-rich region of sand rat are shown in pink or (for PDX1 and INSR) in red. Only one protein encoded within the extreme GC-rich region is present in (A) due to missing genes in the Mongolian jird genome assembly.



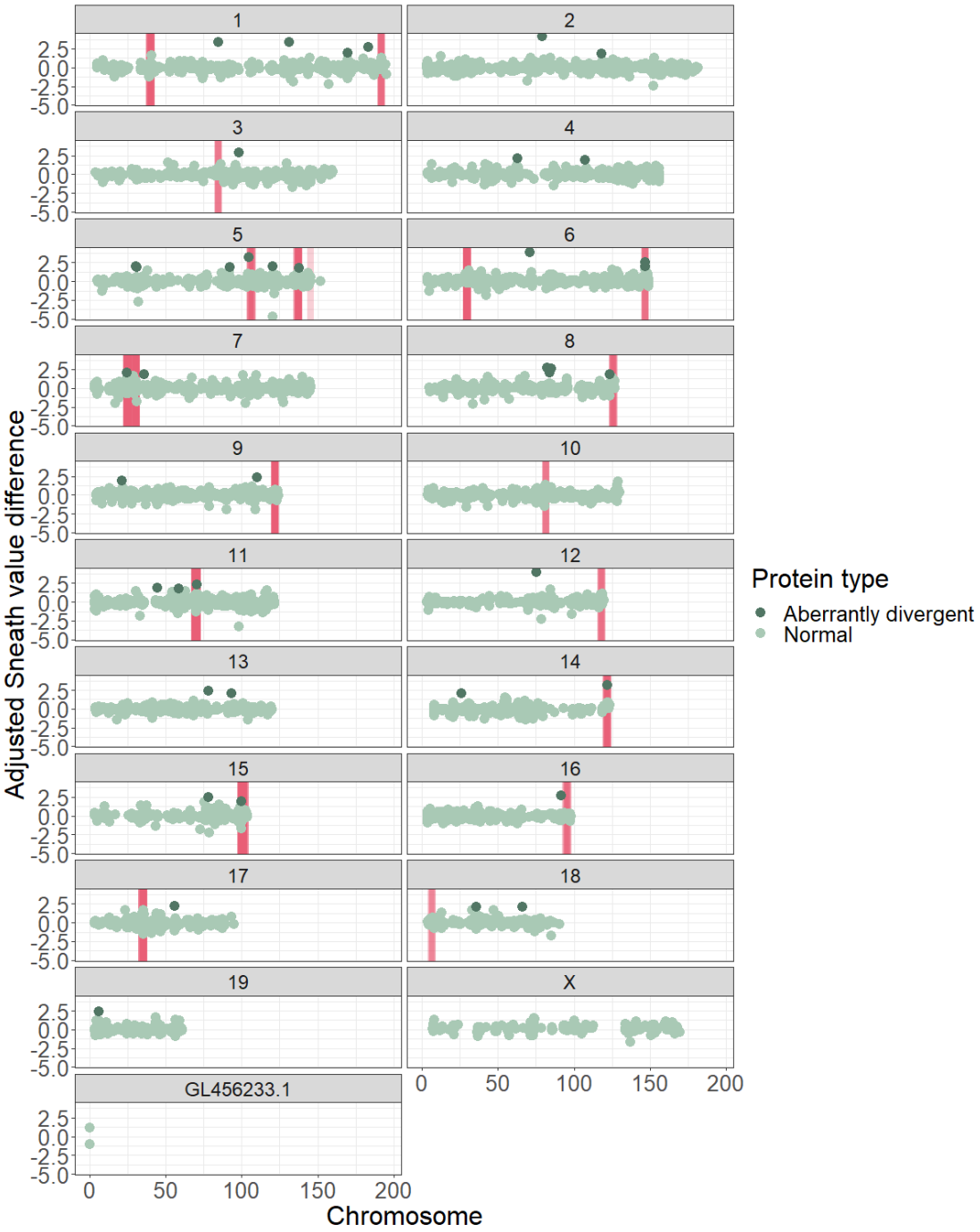
Supplement S5.3 Dissimilarity difference ranking for all 9771 sand rat proteins against the difference in adjusted Sneath value compared to the murid homologue. The last 18 ranked proteins are enlarged in the bottom plot.



Supplement S5.4 Dissimilarity difference ranking for all 9771 sand rat proteins against the difference in adjusted Sneath value compared to the average of five species in the mouse-related clade (kangaroo rat (*Dipodomys ordii*), jerboa (*Jaculus jaculus*), blind mole rat (*Nannospalax galili*), mouse and rat). The top 67 ranked proteins are enlarged in the bottom plot with proteins PDX1 (rank 1) and INSR (rank 6) marked with an arrow. Amongst these proteins, those encoded by genes in the extreme GC-rich region are shown in pink.

Gene location **Protein type**	**High GC**	**Not high GC**	**Total**
Aberrantly divergent	7	43	50
Normal	314	9705	10019
Total	321	9748	10069

Supplement S5.5 Summary of aberrantly divergent Mongolian jird proteins and whether their corresponding genes are in high GC peaks. Fisher's Exact Test reports p-value = 0.00096, estimated odds ratio = 5.03.



Supplement S5.6 Aberrantly divergent Mongolian jird proteins are frequently encoded by genes in GC-rich islands. Each panel shows one mouse chromosome, to which the locations of Mongolian jird orthologues are mapped. All analysed Mongolian jird genes are displayed as dots plotted according to the midpoint position of their corresponding mouse orthologue. The position of each dot on the y-axis shows the difference in adjusted Sneath value between the Mongolian jird and mouse orthologues. Pink lines indicate locations of GC-rich regions identified previously[102,207].

Gene type	dS_{sw} outlier	Not dS_{sw} outlier	Total

Protein type			
Aberrantly divergent	19	31	50
Normal	45	9676	9721
Total	64	9707	9771

Supplement S5.7 Summary of aberrantly divergent sand rat proteins and whether their corresponding genes are dS strong to weak outliers. Fisher's Exact Test reports p-value = 4.4e-30, with an estimated odds ratio of 130.70.

Gene type **Protein type**	**dS_{ww} outlier**	**Not dS_{ww} outlier**	**Total**
Aberrantly divergent	16	34	50
Normal	106	9615	9721
Total	122	9649	9771

Supplement S5.8 Summary of aberrantly divergent sand rat proteins and whether their corresponding genes are dS weak to weak outliers. Fisher's Exact Test reports p-value = 4.4e-19, with an estimated odds ratio of 42.58.

Gene type **Protein type**	**dS_{ss} outlier**	**Not dS_{ss} outlier**	**Total**
Aberrantly divergent	15	35	50
Normal	29	9692	9721
Total	44	9727	9771

Supplement S5.9 Summary of aberrantly divergent sand rat proteins and whether their corresponding genes are dS strong to strong outliers. Fisher's Exact Test reports p-value = 8.8e-25, with an estimated odds ratio of 142.16.

Chapter 6:

Discussion

- ❖ What caused GC composition bias in gerbils?
- ❖ Sand rat GC composition bias and physiology
- ❖ The impact of GC composition bias on evolution
- ❖ Concluding remarks

Discussion

In Chapter 2, I show that a rodent subfamily, the Gerbillinae, includes several species that have an aberrantly divergent *Pdx1* gene. Previous sequencing results have shown that these species have a GC-rich *Pdx1* gene, and for the sand rat, it has multiple GC-rich genomic regions with aberrantly divergent genes[102,207]. Throughout the vertebrate lineage, the *Pdx1* gene is highly conserved; therefore, I ask if the amino acid changes in sand rat PDX1 affect its biochemical properties and its biological function. I show in Chapter 3 that GC composition bias is linked to amino acid substitutions affecting key conserved functional sites in the sand rat PDX1 protein, and during the course of gerbil evolution, natural selection likely compensated for this loss through conservation of an alternative ubiquitination site. In addition, Chapter 4 indicates that amino acid changes linked to GC composition bias has altered sand rat PDX1 protein binding affinity to target promoter sequences, with the sand rat PDX1 conserving its ability to regulate other transcription factors while its amino acid changes have affected its ability to directly control insulin expression and respond to changes in glucose concentration. *Pdx1* is not the only gene that has been affected by both GC composition bias and natural selection; Chapter 5 shows that gerbils have a large number of aberrantly divergent genes that encode proteins with amino acid changes not observed in other rodent species or even in other vertebrate species. Aberrantly divergent genes are strongly associated with GC composition bias, while only a small number of these genes show signs of positive selection, suggesting that these aberrant gerbil genes are a result of a tug-of-war between random GC-skew and natural selection. I propose that the gerbil is a good model for studying the effect of GC composition bias on evolution, and similar clashes between GC composition bias and natural selection may be present in other animals but less extreme. In this chapter, I discuss three questions that are worth further investigation: the cause of GC composition bias in gerbils, sand rat GC composition bias and its possible implications on sand rat physiology, and the impact of GC composition on animal evolution.

1. The cause of GC composition bias in gerbils

The molecular mechanisms behind GC composition bias have been partially elucidated. GC-biased gene conversion (gBGC) has been proposed to be responsible for GC-skew in plants and animals, and many studies provide direct or indirect evidence for the presence of gBGC[52,55,56,245]. Proteins that are responsible for initiating gBGC in mammals are also known. During meiotic recombination, zinc finger protein PR/SET Domain 9 (PRDM9) recognizes specific DNA motifs at recombination hotspots and modifies surrounding histones[56]. These changes in histone markers subsequently recruit the enzyme SPO11, a key member of the topoisomerase 6 complex that creates DSBs[51]. Once a DSB is generated, 5’ ends at the break are chewed back to expose single strand DNA which is bound by strand exchange proteins RAD51 and DMC1[56], the latter used as an immunofluorescence target to visualize recombination sites in mice models. From this point onwards, several proteins have been proposed to be involved in resolving recombination. As mentioned in Chapter 1, the two DNA strands resulting from DSB can either form a Holliday junction or unwind and initiate the SDSA pathway. In yeast, the helicase Srs2 has been shown to promote activation of the SDSA pathway by removing strand exchange proteins from DNA, while the helicase Srs1 has been

linked to formation of the double Holliday junction[51]. In mammals, it is not yet clear whether the Srs2 homologue serves a similar function, while the Srs1 homologue, the BLM helicase, has been shown to play a key part in initiating formation of the double Holliday junction[51]. However, the proteins behind GC nucleotide bias during mismatch resolution, the key step of gBGC, remains unknown. To the best of my knowledge, no genes or proteins have been linked to GC nucleotide preference during DNA repair, and it is also possible that this GC nucleotide preference could have biochemical explanations (e.g. higher binding affinity of guanine or cytosine bases at the DNA repair site). Based on their discovery of the 68% rate of GC-biased gene conversion at noncrossover sites, Li et al. propose the presence of a DNA repair mechanism that is strand-aware, or in other words, capable of selecting a donor strand based on the nucleotide it contains[56]. Currently, the presence of such a repair pathway has not yet been experimentally identified.

Gerbils may be a good animal model for studying the molecular mechanisms of GC composition bias given the prevalence of GC-rich regions likely distributed throughout their genome. First, the sand rat has a good quality genome with several genes in high GC-regions identified, and the breast cancer susceptibility gene *Brca2* is in the same GC-rich region as *Pdx1*. BRCA2 protein is essential for DNA repair, genome maintenance, and is essential in assisting RAD51 binding to single strand DNA at recombination sites. In addition, protein structure resolution has shown that the C-terminal of BRCA2 directly interacts with RAD51 in the RAD51-DNA complex formed at recombination sites. In the sand rat, the *Brca2* gene is extremely GC rich (>70% GC) and is predicted to encode a highly divergent protein; due to low coverage of this gene area, the sand rat *Brca2* gene is partially assembled. This leads to questions around sand rat BRCA2 protein function, and whether the amino acid changes observed in BRCA2 are linked to evolution of high GC regions in the genome. Aside from BRCA2, other sand rat proteins that participate in recombination may provide hints on how GC composition bias affected gerbil genome evolution. Although key proteins such as SPO11, RAD51 and DMC11 are highly conserved between the gerbils, mice and rats, further analysis of all recombination-related proteins might identify more divergent proteins in the gerbil. In addition, PRDM9, the protein that marks recombination hotspots in mammals (except dogs, which have a pseudogene), is predicted to recognize GC-rich DNA motifs in the sand rat. This suggests a possible link between GC-content and recombination hotspots in gerbils, and increased recombination rate in GC-rich areas may contribute to formation of extreme GC-rich regions in the gerbil genome. Finally, aside from having several candidate genes that may be linked to GC composition bias, gerbils have established colonies in the lab that make it possible to study crossover and non-crossover events in animals with a relatively inbred genetic background.

In summary, gerbils are a good model for studying GC content evolution, and they might provide insight into the molecular mechanisms behind GC composition bias. Detailed analysis of sand rat proteins that mediate recombination is needed, and high-quality transcriptomic data from reproductive tissue may facilitate reconstruction of the full-length gerbil *Brca2* gene and other components of the meiotic recombination pathway.

2. Sand rat GC composition bias and physiology

In this thesis, I show that GC composition bias is linked to multiple amino acid changes in the gerbil PDX1 protein, a highly conserved protein that has near 100% in the PDX1 homeodomain throughout the entire vertebrate lineage. In most cases, amino acid divergence is assumed to indicate adaptive selection; however, only 2 – 3 amino acid changes have been tolerated in the PDX1 homeodomain over 420 million years of vertebrate evolution. This extreme conservation, along with gene knock-out experiments in both zebrafish and mouse models showing that the *Pdx1* gene is essential for initiation of pancreatic development[110,121,246], indicate strong selective pressure against non-synonymous amino acid changes in the PDX1 protein. To assess whether amino acid changes in the sand rat PDX1 protein are a desert rodent adaptation, I analysed the PDX1 protein sequence of 14 other desert rodents in 3 different rodent groups. My results show that no other desert rodent species, apart from members of the gerbil subfamily, have similar amino acid changes in the PDX1 protein. The lack of support for convergent evolution in the *Pdx1* gene in desert rodents, along with the presence of a highly divergent *Pdx1* gene in all three gerbil species included in my analysis, cast doubt on whether the 15 amino acid changes in the sand rat PDX1 homeodomain are simply a result of adaptive evolution. Instead, I hypothesize that residue changes observed in the highly divergent gerbil PDX1 protein are maladaptive, and indeed my results show that the sand rat PDX1 protein has altered biochemical properties and cannot correctly regulate expression of the rat insulin genes. In addition, analysis of the sand rat genome has shown that proteins that directly mediate PDX1 degradation and regulation of target gene expression have high levels of sequence similarity to their mouse and rat homologues, further casting doubt on the possibility that sand rat PDX1 has gained adaptive features. I propose that rather than adaptive selection, GC composition bias is the main reason sand rat PDX1 became highly divergent, and I propose that extreme increase of GC content has mainly resulted in slightly maladaptive mutations that have been tolerated by natural selection. In addition, the change in ubiquitination site observed in Chapter 3 suggests that for sand rat PDX1, natural selection is also compensating some of the deleterious mutations caused by GC composition bias via selecting for compensatory sites.

This likely also applies to other gerbil genes with divergent amino acids; my results in Chapter 5 suggest that GC composition bias in the sand rat genome is also linked to many other divergent proteins such as INSR and MEDAG, and these proteins, similar to PDX1, also have amino acid changes in highly conserved protein domains that are only observed in the gerbil lineage. First, higher quality sequenced transcripts of the *Insr* and *Medag* genes are needed before functional comparisons are possible. Currently, both gerbil genes face similar problems: a complete coding sequence for both GC-rich genes are not available. Protocols involving isolation of GC-rich reads or genomic regions prior to sequencing may help solve this problem. Once this issue is overcome, it will be possible to compare the function of the gerbil proteins with mouse or rat proteins in a mammalian cell system. For example, using fibroblast cells or human embryonic kidney (HEK) 293 cells overexpressing gerbil or mouse INSR, it is possible to compare insulin binding efficiency by utilizing insulin molecules tagged to a fluorescent molecule. For MEDAG, a protein that promotes adipocyte differentiation, it may be possible to overexpress gerbil or mouse MEDAG in non-adipocyte cells such as fibroblasts and assess

whether any changes in key adipocyte differentiation genes (e.g. *Pparg*) occur, or to overexpress gerbil or mouse MEDAG in established adipocyte cell models such as 3T3-L1 cells, mouse pre-adipocyte cells that readily differentiate into mature adipocytes, which allow comparison of glucose uptake and fatty acid storage rates. For SPP1, it is secreted into the extracellular matrix and reported to increase expression of inflammatory genes in target cells. It might be possible to produce gerbil and mouse SPP1 in HEK 293 cells and treat different groups of mammalian cells with the purified protein to compare possible differences in the expression of different inflammatory genes. In conclusion, once the complete coding sequence of gerbil *Insr* and *Medag* genes are elucidated, there are multiple methods utilizing mammalian cell culture that can allow functional comparison of gerbil and conserved rodent INSR, MEDAG and SPP1 proteins.

In addition, although GC composition bias is linked to many divergent gerbil proteins, the link between these divergent proteins and gerbil physiology has not yet been fully established. In this thesis, I show evidence of biochemical and functional changes in sand rat PDX1, but the direct link between sand rat PDX1 and sand rat metabolic dysfunction remains to be investigated. Here, I am not implying that amino acid changes in one protein is responsible for the metabolic abnormalities exhibited in gerbils on a standard laboratory diet. Indeed, many sand rat proteins have more divergent amino acids than their mouse and rat homologues, and some of these proteins have important functions related to metabolism and fat storage. Given that experiments have shown multiple abnormalities in sand rats fed a standard laboratory diet, including pancreatic β-cell susceptibility to oxidative stress, insufficient insulin secretion, and persistent insulin resistance, it is likely that multiple sand rat metabolic pathways may have abnormalities.

Future projects probing the link between GC composition bias and metabolic dysfunction in gerbils need to tackle the sequencing issues hindering understanding of the GC-rich genomic regions. Even with the help of long-read sequencing technology such as Nanopore and transcriptome sequencing of tissues where some GC-rich genes are highly expressed, the gerbil genome still has incomplete gene sequences affecting functional comparison of gene products. Hopefully, advances in sequencing technology will help overcome this issue. Apart from this difficulty, sand rats can be good animal models for studying gene regulatory networks, especially how the network adjusts itself in response to amino acid changes affecting key proteins. For metabolic disease pathology, it is worth investigating how sand rats become obese and undergo metabolic dysregulation when given a standard laboratory diet. It may be possible to compare the pathways that are activated in sand rat β-cells and other tissues to pathways that are active in mice and rats when challenged with a diet with slightly increased caloric content. Comparisons with the closely related spiny mice (*Acomys cahirinus*) will also be interesting, as the spiny mouse is also a desert rodent that exhibits spontaneous obesity when fed a standard laboratory diet[247]. The spiny mouse does not possess an aberrantly divergent *Pdx1* gene and exhibits different pathologies compared to the sand rat[248,249], however such comparisons may help elucidate which changes in metabolic pathways are adaptations to desert environments, and which are linked to GC-skew.

3. The impact of GC composition bias on evolution

The impact of mutations and recombination on genome evolution has been increasingly documented and appreciated in recent years. However, the concept of "non-random" mutations and its impact on genome evolution has not been fully appreciated. Random mutations refer to the fact that mutations are "blind" to phenotypic consequence, and mutations do not occur for the sake of inducing phenotypic change. Apart from this aspect, mutations are not random: the rate of mutation at different genomic locations can vary, and certain nucleotide changes happen at higher rates than others. First, mutation hotspots exist in nearly all organisms studied, and high-resolution sequencing of human genomes has allowed researchers to characterize the average mutation rate of different regions in each chromosome. Mutation hotspots can be under positive selection, as shown in *Escherichia coli* that have mutation hotspots predicted to occur via positive selection, with pathogenic *E. coli* strains possessing the highest numbers of genes carrying mutations compared to other strains[250]. In addition, mutation hotspots can contribute to animal evolution, as shown in freshwater stickleback species that have repeatedly lost pelvic fins due to loss of an enhancer region trapped in a mutation hotspot[251]. Second, mutations are AT-skewed due to spontaneous deamination occurring at methylated cytosine bases, converting cytosine to thymine, and nucleotide changes associated with recombination are GC-skewed due to gBGC. Compared to mutation hotspots, the impact of skewed nucleotide substitution, especially skew towards increased GC content, on animal evolution is often overlooked. Genomic regions that become GC-rich have increased potential to gain CpG islands and/or become regions with altered DNA structure or stability. In addition, increased GC content in gene coding regions has the potential to alter codon usage and limit the types of amino acids that can appear in the protein.

Gerbils are a good model for studying the impact of GC composition bias on animal evolution given the presence of high GC regions in its genome and the estimated high level of GC bias required to accumulate such regions. Previous research has identified several GC-rich regions in the gerbil genome that affect hundreds of genes trapped in these regions; are there differences in mutation rates in these GC-rich regions compared to other regions? If yes, does this imply that GC-rich region genes have more gene variants compared to genes in other regions? Most importantly, do GC-rich regions have increased rates of adaptive evolution, and can GC-rich regions be selected for in certain genomic regions containing genes that can benefit from rapid evolution, such as immunity-related genes that participate in an evolutionary arms race with pathogens[252]? The results I present in this thesis show that GC-rich regions can also induce maladaptive changes, and this needs to be considered along with potential benefits of having GC composition bias. Sequencing and comparing the genomes of more gerbil species, especially species that are more distantly related to the sand rat and *Meriones* species, will provide further information on how GC composition bias affected gerbil evolution and whether GC composition bias led to fast evolution and adaptation to the desert environment, or led to accumulation of deleterious mutations affecting gerbil metabolism.

Aside from gerbils, the impact of GC composition bias on evolution in general remains to be investigated further. The link between gBGC, a key mechanism leading to increased GC content, and GC content variation throughout the genome has been established[46], but further

research is needed to understand how GC content variation affected animal evolution. Whether GC content variation arose randomly or whether GC content is under selection remains a topic of debate. In addition, the functional consequences of GC composition bias has not been well characterized given that genomic areas with GC composition bias have the potential to generate more gene variants and speed up gene evolution, but at the same time it can generate deleterious mutations in highly conserved genes. It is also possible for GC composition bias to lead to convergent evolution, for example in species that have similar degrees of GC composition bias affecting syntenic regions and generating similar gene variants. Future analysis can benefit from taking GC composition bias into consideration when analysing genome evolution, and more studies are needed to understand how GC composition bias affected evolution in general.

Concluding remarks

This dissertation presents a case of tug-of-war between GC composition bias and natural selection in a desert-dwelling rodent, the sand rat (*Psammomys obesus*). Focusing on a highly conserved gene in vertebrates, pancreatic duodenal homeobox 1 (*Pdx1*), I show that this gene is highly divergent and GC-rich in several gerbil species. In addition, GC composition bias is linked to amino acid changes in the sand rat PDX1 protein which have altered key functional sites, and it is likely that natural selection compensated such changes by retaining mutations at other locations. Based on functional comparisons between conserved and sand rat PDX1, I propose that amino acid changes altering sand rat PDX1 binding to the insulin promoter has not been removed by natural selection while selection has removed mutations that affect sand rat PDX1 regulation of other key pancreatic transcription factors, including itself. Analysis of the gerbil genome shows that gerbils possess many aberrantly divergent genes that are linked to GC composition bias and high GC regions. This thesis uses PDX1 as an example to explore the functional consequences of GC composition bias and shows that the gerbil is a good model for studying evolution under the effect of GC composition bias.

Appendix A: List of abbreviations used

This appendix contains non-S.I. abbreviations, excluding gene and protein names, used in this thesis.

AT	Adenine-thymine
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	Base pair
cDNA	Complementary DNA
cds	Coding sequence
ChIP	Chromatin immunoprecipitation
Cq	Quantification cycle
CRISPR	Clustered regularly interspaced short palindromic repeats
dN	Nonsynonymous distance
DNM	*De novo* mutation
DP	Diabetes prone
DR	Diabetes resistant
dS	Synonymous distance
DSB	Double strand break
ER	Endoplasmic reticulum
ESC	Embryonic stem cell
gBGC	GC biased gene conversion
GC	Guanine-cytosine
GC3	Guanine-cytosine content at the third codon position
gDNA	Genomic DNA
GO	Gene Ontology
gRNA	Guide RNA
HDR	Homology directed repair
HEK	Human embryonic kidney
kb	Kilo base pairs (1000 base pairs)

LRT	Likelihood Ratio Test
Mb	Mega base pairs (1,000,000 base pairs)
MODY	Maturity onset diabetes of the young
mRNA	Messenger RNA
NADH	Nicotinamide adenine dinucleotide in reduced form
NCBI	National Center for Biotechnology Information
NHEJ	Non-homologous end joining
ns	Non-significant
PAM	Protospacer adjacent motif
PCR	Polymerase chain reaction
qPCR	Quantitative polymerase chain reaction
SDSA	Synthesis-dependent strand-annealing
SHH	Sonic hedgehog
SRA	Sequence Read Archive
SS	Strong to strong
SW	Strong to weak
tBLASTn	Translated nucleotide BLAST (using a protein query)
tRNA	Transfer RNA
T2D	Type 2 diabetes
UPS	Ubiquitin-proteasome system
WS	Weak to strong
WW	Weak to weak

Appendix B: Additional files

Additional file 1: Accession IDs for nucleotide sequences and protein sequences used in Chapter 5. The ranking of 9771 1-to-1 orthologues containing a sand rat gene calculated using the Sneath Index and Epstein's Coefficient are provided in separate sheets. The ranking of 10069 1-to-1 orthologues containing a Mongolian jird gene calculated using the Sneath Index and Epstein's Coefficient are also provided.

Additional file 2: DNA alignments for 10554 1-to-1 orthologues used in Chapter 5. Each alignment file is named using its orthogroup ID.

Appendix C: Resulting publications

All publications associated with work presented in this thesis is listed here:

Accepted publications

Dai, Y. and Holland, P.W., 2019. The interaction of natural selection and GC skew may drive the fast evolution of a sand rat homeobox gene. *Molecular biology and evolution*, *36*(7), pp.1473-1480.

Under review

Dai, Y., Pracana, R. and Holland, P.W. Divergent genes in gerbils: origin, evolution and phenotypic relevance.

The Interaction of Natural Selection and GC Skew May Drive the Fast Evolution of a Sand Rat Homeobox Gene

Yichen Dai[1] and Peter W.H. Holland*[,1]
[1]Department of Zoology, University of Oxford, Oxford, United Kingdom
***Corresponding author:** E-mail: peter.holland@zoo.ox.ac.uk.
Associate editor: Bing Su

Abstract

Several processes can lead to strong GC skew in localized genomic regions. In most cases, GC skew should not affect conserved amino acids because natural selection will purge deleterious alleles. However, in the gerbil subfamily of rodents, several conserved genes have undergone radical alteration in association with strong GC skew. An extreme example concerns the highly conserved homeobox gene *Pdx1*, which is uniquely divergent and GC rich in the sand rat *Psammomys obesus* and close relatives. Here, we investigate the antagonistic interplay between very rare amino acid changes driven by GC skew and the force of natural selection. Using ectopic protein expression in cell culture, pulse-chase labeling, in vitro mutagenesis, and drug treatment, we compare properties of mouse and sand rat Pdx1 proteins. We find that amino acid change driven by GC skew resulted in altered protein stability, with a significantly longer protein half-life for sand rat Pdx1. Using a reversible inhibitor of the 26S proteasome, MG132, we find that sand rat and mouse Pdx1 are both degraded through the ubiquitin proteasome pathway. However, in vitro mutagenesis reveals this pathway operates through different amino acid residues. We propose that GC skew caused loss of a key ubiquitination site, conserved through vertebrate evolution, and that sand rat Pdx1 evolved or fixed a new ubiquitination site to compensate. Our results give molecular insight into the power of natural selection in the face of maladaptive changes driven by strong GC skew.

Key words: **mutation, biased gene conversion, genome evolution, protein evolution, Pdx1, homeodomain.**

Introduction

Vertebrate genomes are far from uniform in their proportions of A, C, G, and T nucleotides and are mosaics of more AT-rich and more GC-rich regions (Eyre-Walker and Hurst 2001). The processes that generate these patterns are complex but include variation in the mutability of certain nucleotides (Ehrlich and Wang 1981), differences in the probabilities of each transition and transversion mutation (Petrov and Hartl 1999; Schaibley et al. 2013), and the influence of chromosome pairing at meiosis through both recombination-associated AT skewed point mutation (Arbeithuber et al. 2015) and by GC-biased gene conversion (gBGC; Galtier 2001; Galtier and Duret 2007; Arbeithuber et al. 2015). The latter is considered to be strong force acting to elevate GC content locally. gBGC is a process in which GC pairs are preferred over AT pairs at polymorphic sites and occurs by biased repair of mismatched DNA heteroduplexes formed by strand invasion after pairing of homologous chromosomes; it leads to gradual accumulation of GC in the affected genomic region (Galtier 2001; Galtier and Duret 2007). gBGC has been shown to be widely present in vertebrates and is a prominent effect leading to local accumulation of GC content, especially in intergenic regions and synonymous third codon positions (Figuet et al. 2015).

The predominant localization of GC skew to synonymous sites (Eyre-Walker and Hurst 2001) implies that natural selection is usually effective at removing GC changes that negatively affect protein function. However, an unusual and extreme case has been discovered recently in gerbil genomes whereby very high GC skew is associated with radical coding sequence divergence in localized regions of the genome (Hargreaves et al. 2017). In these animals, several highly conserved proteins have diverged dramatically over a short evolutionary period through amino acid changes almost always associated with changes from AT- to GC-rich codons. This finding allows us a rare insight into the inevitable conflict between GC skewed nucleotide changes and protein function. For example, some highly conserved sites, inferred to be under strong purifying selection, have undergone unique changes in gerbils. These rare cases offer the chance to explore how molecular evolution might compensate for maladaptive functional changes.

The *pancreatic duodenum homeobox 1* gene (*Pdx1*, also referred to as *Xlox*, *Ipf1*, *Idx1*, *Stf1*, *XlHbox8*, or *Lox*) is a member of the ParaHox gene cluster and one of the most highly conserved of all homeobox genes (Wright et al. 1989; Brooke et al. 1998; Ferrier and Holland 2002). The Pdx1 protein is the key transcription factor initiating vertebrate pancreatic development and is also essential for maintaining pancreatic β-cell identity in adulthood (Jonsson et al. 1994; Ahlgren et al. 1996; Ahlgren et al. 1998). There are two key functional domains in the Pdx1 protein: the hexapeptide domain (PFPWMK)

© The Author(s) 2019. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5466080 on 30 October 2019

responsible for dimerization with PBC proteins (Liberzon et al. 2004), and the 60 amino acid homeodomain responsible for binding to DNA and sequence-specific recognition of promoter sequences (Lu et al. 1996). Throughout vertebrate evolution, the Pdx1 homeodomain has been highly conserved, retaining 100% amino acid identity between catshark, coelacanth, *Xenopus*, chicken, cat, mouse, and human, and just two to four amino acid changes seen in teleost fish and *Anolis* lizard. This extreme conservation, for over 420 My, suggests that every residue in the 60 amino acid homeodomain is likely to be under strong purifying selection, not just those encoding the three alpha helices key to homeodomain folding.

The importance of the homeodomain sequence is also indicated by human phenotypes. A single amino acid change in the third alpha helix, R197H, is associated with drastically lowered levels of insulin expression and a propensity for heterozygous carriers to develop type II diabetes (Macfarlane et al. 1999). The most extreme phenotype reported is absence of a pancreas in an infant carrying one *Pdx1* allele with a E164D point mutation in the homeodomain and a second allele with the mutation E174K (Schwitzgebel et al. 2003). These changes did not alter protein localization, binding to the *insulin* gene promoter, or interaction with known Pdx1 cofactors, but both mutant proteins had significantly shortened protein half-life as measured in cell culture assays. Given that previous work has shown that Pdx1 is degraded via the ubiquitin 26S-proteasome pathway (Claiborn et al. 2010), it is likely that these mutations altered ubiquitination sites and/or the interaction between Pdx1 and ubiquitin pathway proteins.

The only known exceptions to the extreme conservation of vertebrate *Pdx1* are rodents of the gerbil subfamily. It was recently shown that *Pdx1* is highly divergent in two gerbil species: *Psammomys obesus* (sand rat) and *Meriones unguiculatus* (Mongolian gerbil) (Hargreaves et al. 2017). In these animals, the normally invariant 60 amino acid homeodomain has experienced 15 and 14 amino acid changes, along with drastic change outside the homeodomain. This implies that radical Pdx1 sequence divergence occurred in the ancestry of gerbils. Although amino acid changes would normally be indicative of either reduced selection pressure or positive adaptive change, neither seems likely in this case. Instead, detailed analysis in *Psammomys obesus* reveals that the changes have been driven by extreme accumulation of GC nucleotides across a very extensive genomic region, possibly by gBGC, and that the GC bombardment of the gene may have driven maladaptive changes to fixation. Similar substitutions are not seen in any other animal lineage. This unusual situation gives an opportunity to examine how natural selection responds to intense mutation accumulation. First, we ask whether amino acid changes, driven by GC skew, have affected protein stability, by analogy to some human mutations. Second, we use a pharmacological inhibitor to test if, despite divergence, sand rat protein stability is regulated by the ubiquitin-proteasome system (UPS). Third, we use in vitro mutagenesis to identify which amino acid residues are targeted for ubiquitination and activation of UPS in mouse and sand rat proteins and uncover possible compensatory mutations. Our data suggest that extreme accumulation of GC changes removed a critical and conserved ubiquitination site, and that this maladaptive change was overcome by emergence of a novel compensatory change. Hence, despite an onslaught of GC skew, some critical protein properties are maintained.

Results

Altered Stability of Gerbil Pdx1 Protein

The divergence of sand rat and Mongolian gerbil *Pdx1* homeobox sequence, compared with the usual extreme conservation, raises the question of whether the gene is nonfunctional in these species. We have previously shown that this is unlikely since the *Pdx1* coding sequence is depleted in polymorphisms relative to the 3′ untranslated region, suggesting presence of purifying selection (Hargreaves et al. 2017). We also note that five of the six hexapeptide residues are unchanged. As for the homeodomain, the amino acid changes in helix 3, which makes close contact with the major groove of DNA, are between chemically similar amino acids; there is one chemically significant change (Leu to Arg) in helix 1 (fig. 1).This pattern of change suggests that the gene is functional, and that purifying selection has resisted substitutions of residues that affect interactions with DNA. The nonhelix sites that have changed in gerbils are also normally conserved, but in these cases, substitutions driven by GC skew have somehow been tolerated in gerbils only. To test if substitutions are adaptive, we analyzed the hexapeptide and homeodomain regions of the 13 species shown in figure 1, and fitted site, branch, and branch-site models using codeml in PAML (Yang 1997). We found no support for positive selection occurring within the gerbil lineage (Supplementary Material online). A two-ratio branch model gave a dN/dS ratio of 0.0197 ($dN = 0.1280$, $dS = 6.0599$) on the gerbil lineage, indicating a high rate of synonymous substitution and presence of purifying selection. We suggest the observed amino acid changes in gerbil Pdx1 are maladaptive changes.

To assess if Pdx1 protein stability has been altered in gerbil evolution, we conducted pulse-chase labeling experiments of ectopically expressed proteins in a rat pancreatic cell line (INS-1). To allow tracking of protein degradation in individual cells, mouse and sand rat *Pdx1* genes were cloned into $pSNAP_f$ vectors, after first lowering the GC content of the sand rat codons ("mousifying") to optimize translation efficiency (Supplementary Material online). These constructs produce Pdx1 fusion proteins with a C-terminal SNAP-tag motif, which will link covalently to a red fluorescent tag (SNAP-Cell TMR-Star) enabling protein stability to be measured after washing out unbound red tag. After transfection into INS-1 cells, pulse-chase was carried out and fluorescence intensity of individual transfected cells recorded at intervals up to 70 h (fig. 2*a*). Individual cells were identified under brightfield imaging, and cells that exhibited ambiguity in identity were excluded from analysis (fig. 2*b*). Data from 89 ($n = 43$ transfected with mouse *Pdx1*, $n = 46$ transfected with sand rat *Pdx1*) individual cells were used to plot decrease in intensity per cell, revealing a faster rate of intensity decline for mouse Pdx1 than for sand rat Pdx1 protein (fig. 2*c*).

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5435962 by guest on 30 October 2019

Interaction of Natural Selection and GC Skew · doi:10.1093/molbev/msz080 MBE

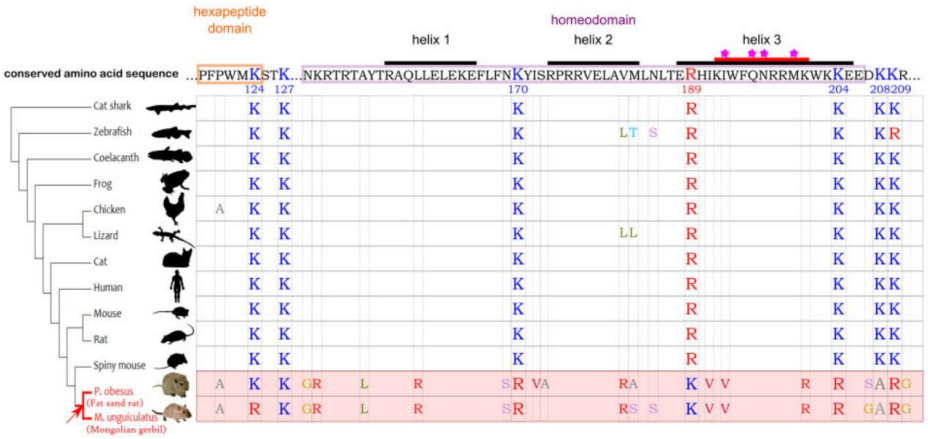


FIG. 1. Alignment of Pdx1 hexapeptide and homeodomain sequence across vertebrates. The human and mouse amino acid sequence, conserved in most vertebrates, is shown at the top. The three alpha helices are labeled, and key amino acids known to directly interact with DNA bases are highlighted with a star (Longo et al. 2007). Lysine/Arginine residues key to the current research are shown for all species. For other residues, only those that deviate from the conserved sequence are shown.

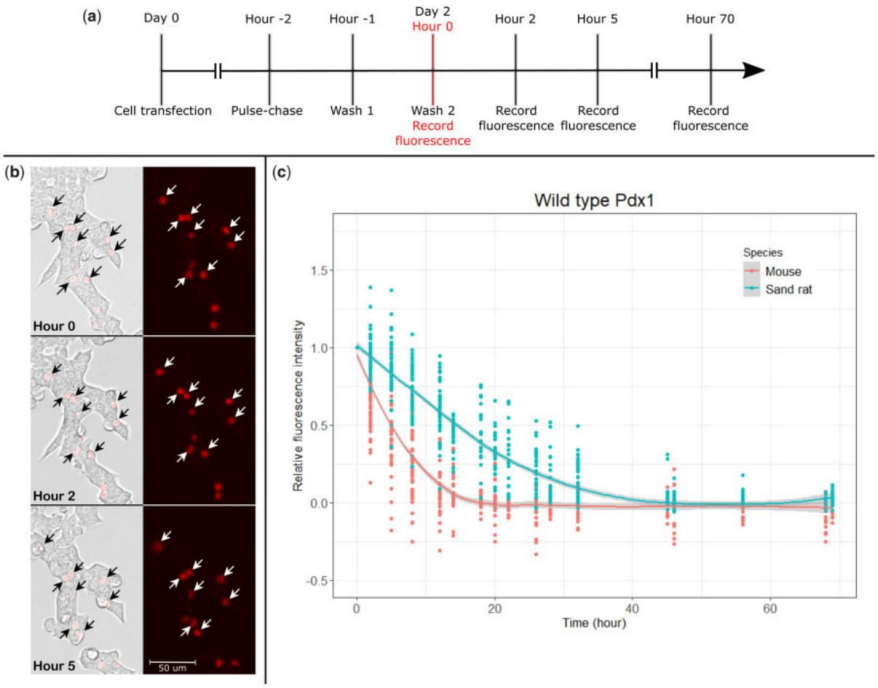


FIG. 2. Measurement of protein stability. (*a*) Experimental design showing timing of cell transfection, pulse-chase, and fluorescence recording. (*b*) An example of cell tracking over 5 h. The labeled cells are overexpressing mouse Pdx1-SNAP. (*c*) Intensity decline for mouse Pdx1 and sand rat Pdx1 proteins. Each dot represents an individual cell. The curve shown was generated using the Local Regression (loess) function in the ggplot2 package in R (Wickham 2016).

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5435962 by guest on 30 October 2019

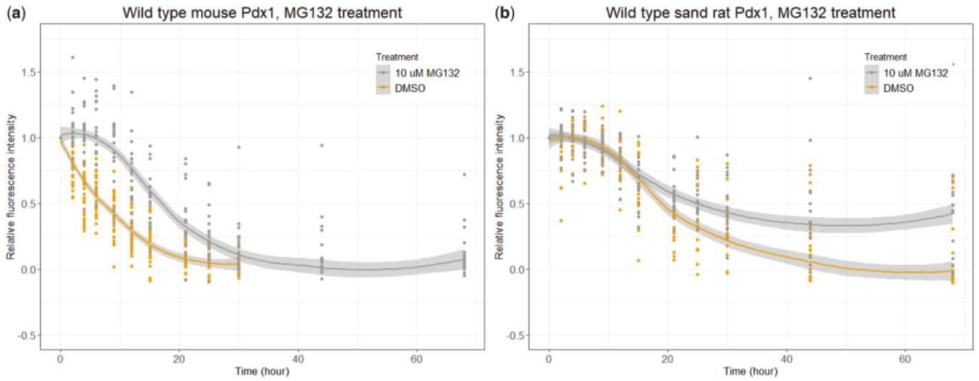


Fig. 3. Effect of inhibiting UPS on Pdx1 protein stability. Graphs show the decline in fluorescence intensity for (*a*) mouse Pdx1 and (*b*) sand rat Pdx1 in cells treated with DMSO (control) or 10-μM MG132 (inhibitor of UPS), in both cases showing shift of protein decay curve. Statistical modeling confirmed that in both cases the effect is significant (mouse $P = 1e\text{-}04$; sand rat $P = 0.0007$); the Pdx1 protein half-lives were extended by 10.8 ± 0.9 h (mouse) and 11.3 ± 1.7 h (sand rat).

To quantify the difference in protein stability, we modeled the decrease of fluorescence intensity using the Linear and Nonlinear Mixed Effects (nlme) Models package (Pinheiro et al. 2018) in R. This revealed a significant difference between the degradation pattern of mouse Pdx1 and sand rat Pdx1 proteins ($P = 0$; Supplementary Material online). Based on this model, we estimate the half-life of mouse Pdx1 protein to be 4.7 ± 0.4 h, and the half-life of sand rat Pdx1 protein to be 13.7 ± 0.3 h.

Sand Rat Pdx1 Protein Degradation Is Regulated by the UPS

It has previously been shown that mouse Pdx1 protein stability is regulated by the UPS (Claiborn et al. 2010). To test if sand rat Pdx1 is regulated by the same biochemical pathway, we treated transfected INS-1 cells with MG132, a reversible inhibitor of the 26S proteasome. MG132 treatment slows the overall rate of mouse Pdx1 protein degradation, shifting the best-fit curve to the right (fig. 3*a*, P = 1e-04; Supplementary Material online). This is consistent with previous findings that mouse Pdx1 protein stability is regulated via UPS. We find that MG132 treatment also slows the rate of sand rat Pdx1 degradation pattern (fig. 3*b*, $P = 0.0007$; Supplementary Material online), although less sharply compared with mouse Pdx1. Differences in processes prior to proteasome degradation, for example, E3 ligase efficiency in recognizing and tagging the target Pdx1 protein, could contribute to subtle differences in how proteins respond to proteasome inhibition.

Identification of Pdx1 Ubiquitination Sites Suggests Compensatory Molecular Evolution

Although ubiquitin can sometimes be added to the peptide N-terminus or to cysteine residues, it is only added to lysine residues in the context of targeting for degradation in proteasomes (Kerscher et al. 2006). We therefore asked which lysine residues were utilized by the UPS in regulating stability of mouse and sand rat Pdx1 proteins. We first used prediction software to deduce which residues had the highest probability of being a functional ubiquitination site (Chen et al. 2013; Gao et al. 2013; Wang et al. 2017). In addition, because ubiquitination acts through covalent addition of a moiety to lysine residues, we excluded key lysine residues in the homeodomain that have been proven to engage in interactions with DNA. This resulted in seven lysine residues of interest: four in the conserved Pdx1 protein (mouse, human, and most vertebrates) and three in the divergent sand rat protein.

We used in vitro mutagenesis to replace singly each lysine codon of interest with a codon for arginine, an amino acid with similar polarity but which cannot be tagged by ubiquitin. For mouse *Pdx1*, we generated mutant proteins K170R, K204R, K208R, and K209R with C-terminal SNAP-tags (numbering according to mouse Pdx1). These constructs were transfected into INS-1 cells and protein stability measured through pulse-chase labeling with SNAP-Cell TMR-Star. As shown in figure 4*a*, mutant proteins K204R, K208R, and K209R did not have significantly different half-lives compared with the wild type control ($P = 0.1487$, 0.1341, and 0.5847 respectively; Supplementary Material online). However, mutant protein K170R showed a very different protein decay pattern, with a much flatter best-fit decay curve and a longer half-life ($P = 0.0267$; Supplementary Material online). This reveals that lysine residue 170 is the major Pdx1 ubiquitination site. We suspect it is not the only ubiquitination site as the protein is still gradually degraded within each cell.

For sand rat Pdx1, we generated SNAP-tagged mutant proteins K124R, K127R, and K189R, transfected into INS-1 cells and measured protein stability (numbering according to alignment with the mouse Pdx1 protein). As shown in figure 4*b*, mutant proteins K124R and K127R did not have drastically different stability compared with wild type control ($P = 0.9649$ and 0.5538, respectively; Supplementary Material

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5435962 by guest on 30 October 2019

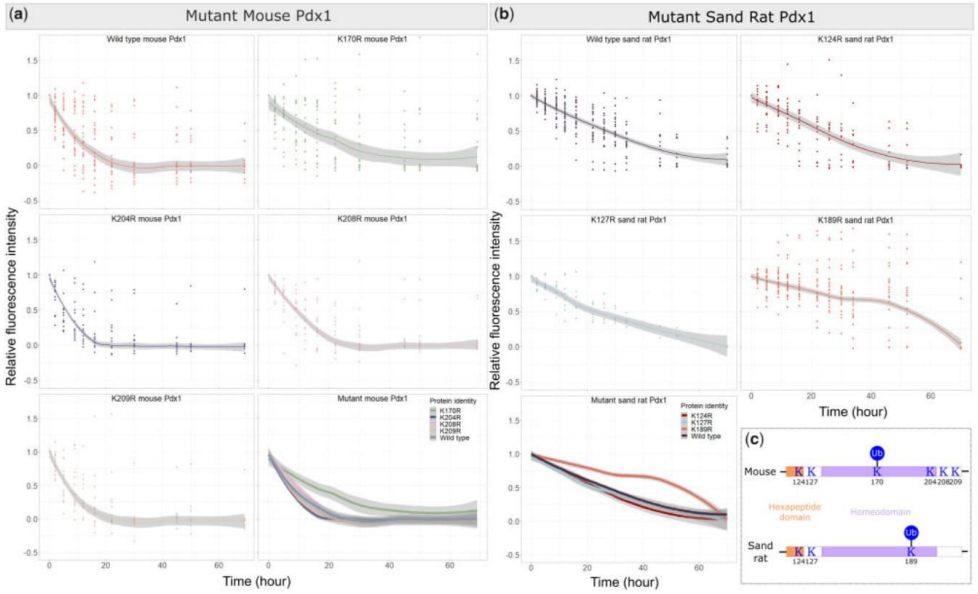


FIG. 4. Protein stability comparison between wild type and mutant proteins. (*a*) Graphs show decline in fluorescence intensity for mouse wild type and individual mutant proteins, with a merged graph in the bottom right corner. (*b*) Graphs showing decline in fluorescence intensity for sand rat wild type and individual mutant proteins, with a merged graph in the bottom left corner. (*c*) Summary of ubiquitination site location in mouse and sand rat Pdx1.

online). Only mutant K189R displayed a completely different degradation pattern compared with the other two mutant proteins and the wild type ($P = 0.0000$; Supplementary Material online). For K189R, high levels of fluorescence were detected in cells throughout the experimental period, with many cells exhibiting rounded shapes and detaching from the substrate around hour 50. This radical disruption of regulated protein decay indicates that the K189 is the ubiquitination site for sand rat Pdx1. Mutation of this site abolishes Pdx1 degradation via the UPS pathway.

Amino acid 170, which we identify as the major functional ubiquitination site in the mouse Pdx1 protein, has been conserved across 420 My of jawed vertebrate evolution. This K170 site is invariant across vertebrates, with the exception of some teleosts and the gerbil species discussed here. In sand rat and Mongolian gerbil, the codon has undergone a unique, GC skew associated mutation to arginine (fig. 1). In contrast, stability and degradation of sand rat Pdx1 is regulated through ubiquitination of K189, a site which across the rest of jawed vertebrate evolution is otherwise conserved as arginine. These results reveal a unique shift in the location of the ubiquitination site in Pdx1 along the evolutionary lineage leading to gerbils (fig. 4*c*). We propose that in the course of evolution, and under extreme GC skew, the ultra-conserved lysine residue 170, possibly along with other ubiquitination sites, was removed by mutation and as a compensatory mechanism natural selection favored an otherwise maladaptive mutation elsewhere in the protein sequence (fig. 5). The fixation of the "new" lysine 189 is consistent with natural selection acting in the face of strong GC skew.

Discussion

The sand rat and the Mongolian gerbil are both desert-dwelling rodents, the former native to North Africa and the Middle East (Schmidt-Nielsen et al. 1964) and the latter to Mongolia and China (Mallon 1985). Both are members of the Gerbillinae subfamily that diverged from the Murinae subfamily, including mice and rats, 18 ± 1.8 Ma (Chevret and Dobigny 2005). The sand rat has been of particular interest in studies of diabetes since both wild-caught and laboratory-reared animals can exhibit elevated fasting blood glucose levels, increased body weight and symptoms of type 2 diabetes after being fed a high calorie diet (Schmidt-Nielsen et al. 1964; Shafrir and Gutman 1993). It had been hypothesized that these characteristics are associated with loss of the *Pdx1* gene in the gerbil subfamily (Leibowitz et al. 2001), but our recent work revealed that rather than missing the gene, sand rats and Mongolian gerbils possess a highly divergent *Pdx1* (Hargreaves et al. 2017).

Two features make the sequence divergence of gerbil Pdx1 particularly intriguing. First, the divergence of this normally highly conserved protein is unique, with nothing similar seen in other vertebrates, including arid-adapted rodents such as the closely related spiny mouse *Acomys cahirinus*. Second, the

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5435962 by guest on 30 October 2019

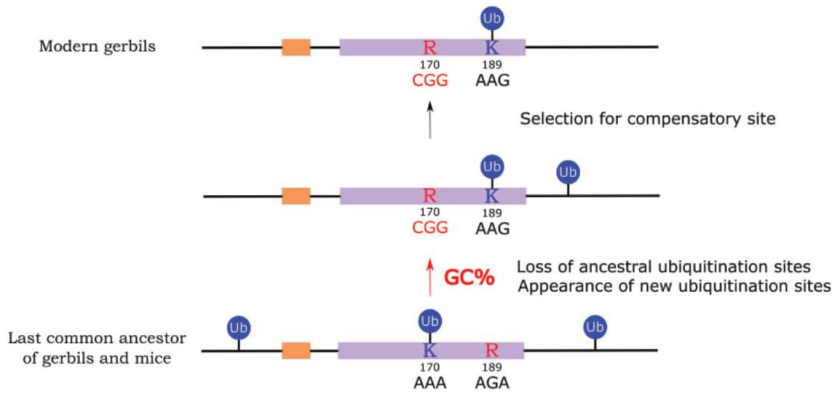


Fig. 5. Summary of the possible evolutionary pathway leading to emergence of a compensatory ubiquitination site and loss of conserved sites in gerbil Pdx1. This process is associated with GC skew.

divergence of the gene is clearly associated with a strongly pronounced GC skew that encompasses several genes in a large genomic region (Hargreaves et al. 2017). Our working hypothesis to explain these findings is that focused and extreme gBGC has driven accumulation of GC nucleotides throughout this region, generating maladaptive substitutions. If correct, this hypothesis predicts that natural selection would act to fix compensatory mutations, both within and outside the GC skewed region, to restore biological function.

Before considering the existence of compensatory mutations, we first ask if amino acid changes in gerbil Pdx1 are indeed maladaptive. The evidence from comparative analysis is consistent with this view since large changes to the homeodomain, such as in gerbils, have not been tolerated across the rest of vertebrate evolution or indeed across the Bilateria. In addition, we do not find support for positive selection acting on the hexapeptide region and homeodomain in the gerbil lineage. Direct tests of function, such as gene deletion or experimental manipulation in gerbils, would not answer the question if compensatory changes have occurred elsewhere. Gene replacement in mouse would not give a clear answer for the same reason. Instead, we took a biochemical approach and asked whether protein stability was altered, as reported for some human variants associated with disruption of pancreatic function. Using a cell culture expression system, we found that sand rat Pdx1 does have a different half-life to mouse Pdx1, but surprisingly the change was in the opposite direction from our original expectation. Although two different single amino acid changes in human Pdx1 cause decreased half-life, the radically altered sand rat protein has a longer half-life. Perhaps more importantly, we noted that degradation of sand rat Pdx1 protein follows a steady curve, indicative of actively regulated degradation rather than unregulated decay, and we obtained experimental verification that protein stability is regulated by the UPS. This regulated degradation is another indication that the sand rat *Pdx1* gene encodes a functional protein under selection rather than being a nonfunctional pseudogene. Although these assays were conducted in a heterologous system (a rat cell line), we argue that this is not a confounding factor since the key mediators of Pdx1 degradation, Pdx1 C-terminal inhibiting factor 1 (Pcif1) and the scaffolding protein Cul3 (Claiborn et al. 2010), are highly conserved between rat, mouse, and sand rat (>97% protein similarity between all three species).

To investigate candidate maladaptive mutations, we mapped functional sites responsible for UPS-regulated degradation of the mouse Pdx1 protein. These mutagenesis experiments revealed that one highly conserved lysine, K170, is the critical residue for regulated degradation by UPS. Interestingly, this residue has mutated to an arginine in gerbils and could no longer be used as a site for ubiquitination. Loss of this key ubiquitination site is not seen in any other tetrapod lineage and is likely to have been driven by the gBGC phenomenon described above. However, a new ubiquitinated lysine site has evolved in gerbil Pdx1 proteins, at a different site in the same protein domain, and it is this new site that enables UPS-regulated degradation. We suggest that the maladaptive change to a key amino acid residue, conserved for over 420 My, is compensated for by emergence of a new site fulfilling a similar function. We cannot state that the sand rat Pdx1 protein is as efficient or "optimal" as the conserved or ancestral sequence present before the emergence of the hotpot of GC skew. Indeed, amino acid changes at other residues may affect other biochemical properties. However, we propose that as GC skew gradually altered the sequence of this critical gene during gerbil evolution, natural selection favored substitutions elsewhere in the protein to compensate for loss of a critical biochemical function (fig. 5). This gives us an insight into the power of natural selection even in the face of extreme and unusual nucleotide bias.

Materials and Methods

Plasmid Construction and In Vitro Mutagenesis

Full coding sequences for mouse and sand rat *Pdx1* genes were synthesized by GenScript, with alterations in sand rat

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5435962 by guest on 30 October 2019

codon usage to lower GC content (Supplementary Material online). Mouse *Pdx1* was inserted into pSNAP$_f$ vector (New England BioLabs) using *Bam*HI and *Xho*I restriction sites, and sand rat *Pdx1* inserted using *Not*I and *Xho*I restriction sites. Stop codons were removed to allow expression of Pdx1-SNAP fusion protein. Successful insertion was verified by Sanger sequencing. In vitro mutagenesis was carried out using polymerase chain reaction.

Cell Culture and Transfection

INS-1 cells (below 30 passages) were maintained in RPMI-1640 medium, supplemented with 10% fetal bovine serum, 50-μM beta-mercaptoethanol, 1-mM sodium pyruvate, 10-mM HEPES, and 100-U/ml penicillin/streptomycin, at 37 °C/5% CO_2 and passaged every 4 days. For transfection, cells were seeded into 96-well plates at 60,000 cells/well and transfected 5 h later using Lipofectamine 3000 (ThermoFisher Scientific). For each well, 40-μl DNA-lipid mix containing 400-ng plasmid DNA, 0.8-μl P3000 Reagent, and 1.2-μl Lipofectamine was added and incubated overnight.

SNAP-Staining and Fluorescence Recording

SNAP-Cell TMR-Star dye (New England BioLabs) was diluted 1:400 in culture medium. Fifty microliters was added to each well and incubated at room temperature for 5 min before three washes. Stained cells were incubated in fresh culture medium for 1 h at 37 °C, twice, to ensure dissociation of unbound SNAP-Cell TMR-Star dye prior to analysis. At each timepoint, pictures were taken using a ZOE Fluorescent Cell Imager (Bio-Rad) with fixed settings: brightfield gain = 18, exposure = 300, LED = 40, contrast = 0; red channel gain = 40, exposure = 500, LED = 50, contrast = 0. Image analysis was performed using ImageJ (Schneider et al. 2012).

MG132 Treatment

Ten-millimolar MG132 stock solution in dimethyl sulfoxide (DMSO, Sigma-Aldrich) was diluted to 10-μM in treatment wells at hour −2. Culture medium containing 10-μM MG132 was used for all medium changes until hour 4, giving a total of 6-h treatment. Controls were treated in parallel using DMSO.

Statistical Analysis

We used nlme 3.1-137 (Pinheiro et al. 2018) in R 3.5.1 (R Core Team 2018) to model the pattern of protein degradation based on data collected for all protein and treatment types. Square-root transformed data were used to ensure no obvious deviations from homoscedasticity or normality (Supplementary Material online). The linear mixed effects model was constructed using the interaction between sqrt(Time) and type of species as the fixed effect, and the random effect was cell identity. Different experiments were analyzed individually using the same fixed and random effects. Parameters reported are given in Supplementary Material online. Half-life predictions were made for each individual cell using the model generated and the generic predict function in R (R Core Team 2018); half-lives of mouse and sand rat Pdx1 proteins are reported as mean ± standard deviation.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Sonia Trigueros for advice on microscopy, Anna Gloyn and Natasha Ng for INS-1 cell culture advice, Daniel Lunn (University of Oxford Statistical Consultancy Services) for advice on data analysis, Rodrigo Pracana, Amy Royall, Adam Hargreaves, John Mulley, Thomas Brekke, Sebastian Shimeld, and Gil McVean and for helpful discussions, Ofuase Dirisu for assistance, and reviewers for useful suggestions. 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.).

References

Ahlgren U, Jonsson J, Edlund H. 1996. The morphogenesis of the pancreatic mesenchyme is uncoupled from that of the pancreatic epithelium in IPF1/PDX1-deficient mice. *Development* 122(5):1409–1416.

Ahlgren U, Jonsson J, Jonsson L, Simu K, Edlund H. 1998. β-Cell-specific inactivation of the mouse Ipf1/Pdx1 gene results in loss of the β-cell phenotype and maturity onset diabetes. *Genes Dev.* 12(12):1763–1768.

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci U S A.* 112(7):2109–2114.

Brooke NM, Garcia-Fernàndez J, Holland PWH. 1998. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 50(2):61–63.

Chen Z, Zhou Y, Song J, Zhang Z. 2013. HCKSAAP-UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim Biophys Acta Proteins Proteomics* 1834(8):1461–1467.

Chevret P, Dobigny G. 2005. Systematics and evolution of the subfamily Gerbillinae (Mammalia, Rodentia, Muridae). *Mol Phylogenet Evol.* 35(3):674–688.

Claiborn KC, Sachdeva MM, Cannon CE, Groff DN, Singer JD, Stoffers DA. 2010. Pcif1 modulates Pdx1 protein stability and pancreatic β cell function and survival in mice. *J Clin Invest.* 120(10):3713–3721.

Ehrlich M, Wang R-H. 1981. 5-Methylcytosine in eukaryotic DNA. *Science* 212(4501):1350.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. *Nat Rev Genet.* 2(7):549–555.

Ferrier DEK, Holland P. 2002. *Ciona intestinalis* ParaHox genes: evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal colinearity. *Mol Phylogenet Evol.* 24(3):412–417.

Figuet E, Ballenghien M, Romiguier J, Galtier N. 2015. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol Evol.* 7(1):240–250.

Galtier N. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Dev Psychobiol.* 39(4):251–256.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet.* 23(6):273–277.

Gao T, Liu Z, Wang Y, Cheng H, Yang Q, Guo A, Ren J, Xue Y. 2013. UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation. *Nucleic Acids Res.* 41(D1):445–451.

Hargreaves AD, Zhou L, Christensen J, Marlétaz F, Liu S, Li F, Jansen PG, Spiga E, Hansen MT, Pedersen SVH, et al. 2017. Genome sequence of

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5435962 by guest on 30 October 2019

a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. *Proc Natl Acad Sci U S A* 114(29):7677–7682.
Jonsson J, Carlsson L, Edlund T, Edlund H. 1994. Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* 371(6498):606–609.
Kerscher O, Felberbaum R, Hochstrasser M. 2006. Modification of proteins by ubiquitin and ubiquitin-like proteins. *Annu Rev Cell Dev Biol*. 22(1):159–180.
Leibowitz G, Ferber S, Apelqvist Å, Edlund H, Gross DJ, Cerasi E, Melloul D, Kaiser N. 2001. IPF1/PDX1 deficiency and β-cell dysfunction in *Psammomys obesus*, an animal with type 2 diabetes. *Diabetes* 50(8):1799–1806.
Liberzon A, Ridner G, Walker MD. 2004. Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1. *Nucleic Acids Res*. 32(1):54–64.
Longo A, Guanga GP, Rose RB. 2007. Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain. *Biochemistry* 46(11):2948–2957.
Lu M, Miller C, Habener JF. 1996. Functional regions of the homeodomain protein IDX-1 required for transactivation of the rat somatostatin gene. *Endocrinology* 137(7):2959–2967.
Macfarlane WM, Frayling TM, Ellard S, Evans JC, Allen LIS, Bulman MP, Ayres S, Shepherd M, Clark P, Millward A. 1999. Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *J Clin Invest*. 104: R33–R39.
Mallon DP. 1985. The mammals of the Mongolian People's Republic. *Mamm Rev*. 15(2):71–102.
Petrov DA, Hartl DL. 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc Natl Acad Sci U S A*. 96(4):1475–1479.
Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team. 2018. nlme: linear and nonlinear mixed effects models. R package version 3.1-137. Available from: https://CRAN.R-project.org/package=nlme
R Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.r-project.org/
Schaibley VM, Zawistowski M, Wegmann D, Ehm MG, Nelson MR, St. Jean PL, Abecasis GR, Novembre J, Zollner S, Li JZ. 2013. The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res*. 23(12):1974–1984.
Schmidt-Nielsen K, Haines HB, Hackel DB. 1964. Diabetes mellitus in the sand rat induced by standard laboratory diets. *Science* 143(3607):689–690.
Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH image to ImageJ: 25 years of image analysis. *Nat Methods*. 9(7):671–675.
Schwitzgebel VM, Mamin A, Brun T, Ritz-Laser B, Zaiko M, Maret A, Jornayvaz FR, Theintz GE, Michielin O, Melloul D, et al. 2003. Agenesis of human pancreas due to decreased half-life of insulin promoter factor 1. *J Clin Endocrinol Metab*. 88(9):4398–4406.
Shafrir E, Gutman A. 1993. *Psammomys obesus* of the Jerusalem colony: a model for nutritionally induced, non-insulin-dependent diabetes. *J Basic Clin Physiol Pharmacol*. 4(1–2):83–100.
Wang JR, Huang WL, Tsai MJ, Hsu KT, Huang HL, Ho SY. 2017. ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. *Bioinformatics* 33(5):661–668.
Wickham H. 2016. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag.
Wright CVE, Schnegelsberg P, Robertis EDE. 1989. XlHbox 8: a novel Xenopus homeo protein restricted to a narrow band of endoderm. *Development* 105(4):787–794.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13(5):555–556.

Downloaded from https://academic.oup.com/mbe/article-abstract/36/7/1473/5435962 by guest on 30 October 2019

1 **Divergent genes in gerbils: origin, evolution and phenotypic relevance**

2

3 Yichen Dai, Rodrigo Pracana and Peter WH Holland *

4 Department of Zoology, University of Oxford, 11a Mansfield Road,

5 Oxford, OX1 3SZ, UK

6

7 *** Correspondence: peter.holland@zoo.ox.ac.uk**

8

Abstract

Background

Two gerbil species, sand rat (*Psammomys obesus*) and Mongolian jird (*Meriones unguiculatus*), become obese and show signs of metabolic dysregulation when maintained on standard laboratory diets. The genetic basis of this condition is unknown. Recently, genome sequencing has uncovered very unusual regions of high guanine and cytosine (GC) content scattered across the sand rat genome, most likely generated by extreme and localized biased gene conversion. A key pancreatic transcription factor PDX1 is encoded by a gene in the most extreme GC-rich region, is remarkably divergent compared to the vertebrate conserved protein sequence and exhibits altered biochemical properties. Here, we ask if gerbils have proteins in addition to PDX1 that are aberrantly divergent in amino acid sequence, whether they have also become divergent due to GC-biased nucleotide changes, and whether these proteins could plausibly be connected to metabolic dysfunction exhibited by gerbils.

Results

We analyzed 10544 proteins with 1-to-1 orthologues in human and rodents and identified 50 proteins that accumulated unusually high levels of amino acid change in the gerbil lineage. We show that these 50 aberrantly divergent gerbil proteins are associated with GC-skew at the nucleotide level and with previously defined high GC-regions in the gerbil genome. We also highlight four aberrantly divergent gerbil proteins, PDX1, INSR, MEDAG and SPP1, that may plausibly be associated with dietary metabolism.

29 **Conclusions**

30 We show that through the course of gerbil evolution, GC-biased nucleotide substitution, rather
31 than positive selection, has resulted in many aberrantly divergent proteins in the gerbil lineage.
32 Some of these carry putatively deleterious changes that could underpin metabolic and
33 physiological phenotypes observed in some gerbil species. We propose that these animals
34 provide a useful model to study the 'tug-of-war' between natural selection and the excessive
35 generation of deleterious substitutions mutations through biased gene conversion.

36 **Keywords:** gBGC, GC-skew, genome evolution, insulin receptor, *Medag*, metabolism,
37 osteopontin, pancreatic duodenal homeobox 1, protein evolution.

38

39 **Background**

40 As animal genomes are sequenced, unexpected features are often discovered, shaping our
41 understanding of molecular evolution and its relation to phenotypic evolution. Amongst
42 mammals, the genomes of gerbils (order Rodentia, subfamily Gerbillinae) have emerged as
43 some of the most unusual discovered to date. Several puzzling features of gerbil genomes are
44 still not fully understood, and these raise important questions about mechanisms underlying
45 molecular evolution and possible constraints or challenges to natural selection.

46 The first indications of unusual genome structure in this group came from analyses of the sand
47 rat (*Psammomys obesus*), a desert-living gerbil from the Middle East and North Africa (1). After
48 initial attempts to clone individual genes gave misleading results (2), genome sequencing
49 revealed presence of an 'island' or 'islands' of remarkably high GC content, within which
50 conserved protein-coding sequences have become radically altered through extensive changes
51 from AT (weak) to GC (strong) base pairs (3). These alterations have caused unusual amino
52 acid substitutions in genes that are conserved across other vertebrate clades, with potential to
53 adversely impact physiology or development (3,4). Analysis of transcriptomes from additional
54 species reveal that the 'GC islands' phenomenon is shared between different gerbil species and
55 occurred in the gerbil evolutionary lineage (5).

56 Variation in GC content within a genome is not in itself unusual. With the increased availability
57 of published genomes, many studies have reported differences between nucleotide composition
58 in regions of the genome and between species (6,7). What marks gerbil genomes as unusual is
59 the extreme nature of the nucleotide compositional bias associated with extensive change to

protein-coding genes. Recent work has revealed that gerbil genomes likely have one large island (~10 Mb) of extreme GC skew, where GC content at non-synonymous sites reaches almost 100%, and several other less extreme islands on other chromosomes (5). In eukaryotes, the location of GC-rich domains is often correlated with chromosomal regions of high recombination, such as telomeric regions and small chromosomes (8,9). The correlation is thought to be driven by GC-biased gene conversion (gBGC) (10,11), a phenomenon that occurs at meiosis when homologous chromosomes contain GC/AT heterozygous sites (8). During strand invasion after meiotic pairing, the mismatch between strands at these sites tends to be repaired using the strand containing a G or C nucleotide in preference to the strand containing A or T (8).

This gBGC process can profoundly affect genome GC composition over generations through fixation of AT to GC (weak to strong) mutations in genomic regions with a high recombination rate (8). This underlying process generates a pattern of GC skewed evolution in these regions, with 'weak to strong' nucleotide substitutions occurring more often than 'strong to weak' nucleotide substitutions. Importantly, in genomic regions where the effect of gBGC is particularly strong, deleterious 'weak to strong' mutations may become fixed in the population despite the action of natural (purifying) selection (12). Suggested examples include coding sequence changes in the *LEP* gene in the avian lineage (13), the *Fxy* gene in mice (14), and the *Pdx1* gene in sand rats and other gerbils (3). The radical changes to the *Pdx1* gene associated with GC skew in the gerbil lineage could have physiological implications. *Pdx1* is a homeobox gene encoding a highly conserved transcription factor essential for pancreatic development and function (15–18). In humans, *PDX1* variants are linked to pancreatic dysfunction of varying

82 severity (19,20), while experiments in mice have shown that *Pdx1* is essential for pancreatic
83 development (15,16). Consistent with evidence from mutations, comparisons between species
84 reveal extreme evolutionary conservation, especially in the homeodomain motif, indicating that
85 each amino acid site is under purifying selection and that few changes are tolerated. It is
86 therefore striking that the 60 amino acid PDX1 homeodomain, normally 100% conserved
87 across mammals, has 15 amino acid changes in the sand rat and 14 amino acid changes in the
88 Mongolian jird (3).

89 Despite the occurrence of these otherwise 'disallowed' mutations, sand rats and other gerbils
90 do develop a pancreas and do secrete insulin from pancreatic β-cells (21,22). There are
91 indications, however, that sand rats and possibly other gerbils are prone to disorders associated
92 with pancreatic dysfunction. For example, at least some sand rats develop diet-induced type 2
93 diabetes (T2D) when fed standard laboratory rodent food, and physiological studies have shown
94 a predisposition towards insulin resistance and stress-induced β-cell apoptosis in T2D-prone
95 individuals (23,24). In addition, some Mongolian jirds raised on a laboratory diet have also
96 been observed to spontaneously become obese and develop poor glucose tolerance (25,26).

97 We must be cautious before concluding that changes to *Pdx1*, driven by GC-skew, are the cause
98 of physiological disorders in sand rats and other gerbils. In humans, T2D is a complex metabolic
99 disease caused by a mixture of genetic and environmental factors, with more than 98% of T2D
100 cases estimated to be caused by coding/regulatory sequence mutations in more than one gene
101 (27–29). In addition, some T2D-related genes are not clearly associated with adult β-cell
102 function; for example, mutations in the *LEP* gene can lead to T2D by altering appetite and body

103 weight, while mutations altering the insulin receptor gene *INSR* can cause T2D due to insulin
104 resistance in tissues responding to insulin (30,31).

105 In this study we first ask how many other proteins, besides PDX1, have become 'unusually
106 altered' in amino acid sequence during the evolution of two gerbils: sand rat (*Psammomys*
107 *obesus*) and Mongolian jird (*Meriones unguiculatus*). We approach this through a comparative
108 study, searching for genes that have changed in gerbils but are conserved in other mammals.
109 These may include genes that have accumulated potentially deleterious changes. Second, we
110 ask if the genes encoding these unusually altered proteins in gerbils are associated with GC
111 skew: an excess of AT to GC base pair changes. This question tests whether unusual protein
112 sequence change is being driven by the process of gBGC, rather than natural selection. Third,
113 we ask if any of the deviant proteins could feasibly be connected to propensity to type 2 diabetes
114 or related physiological disorders. Together, these three lines of investigation combine to test
115 whether 'blind' fixation of maladaptive mutations though biased gene conversion had
116 deleterious consequences for the biology of these desert-living rodents.

117 **Results**

118 *Many gerbil proteins are extremely divergent compared to their homologues*

119 To detect protein-coding genes that have become unusually divergent in gerbils, we use
120 measures of 'relative divergence' not 'absolute divergence' because different types of protein
121 evolve at different rates. High absolute sequence divergence could simply reflect low levels of
122 selective constraint on amino acid sequence, which would be manifest as higher sequence
123 divergence in all lineages. In contrast, high relative divergence reflects amino acid sequence

changes in one species (or lineage) that is greater than the expectation for that protein. Comparing predicted proteomes from the genomes of 12 rodents, plus human as an outgroup, gave a set of 10544 genes with one-to-one orthology between all species (Fig. 1; Additional file 2). To quantify relative divergence in amino acid sequence we used two different indices for measuring the biological and chemical relevance of amino acid changes: the Sneath Index and Epstein's Coefficient (32,33). We limited analysis to positions in multiple sequence alignments that had the same residue in three nested outgroup species (usually human, guinea pig and squirrel; asterisks in Fig. 1A); this conservation suggests that these residues represent the ancestral state for rodents and they are likely functionally conserved. Some proteins had few residues meeting this criterion, reducing the dataset to 9771 sand rat genes and 10069 Mongolian jird genes. For each sequence in two gerbil species and two murine species (mouse and rat), we scored the Sneath Index and Epstein's Coefficient for each residue relative to that of the outgroup species. The sum across all residues gives a 'Sneath value' and 'Epstein's Coefficient' for each protein; these are divided by the protein alignment length to give an 'adjusted Sneath value' or 'adjusted Epstein's Coefficient'.

For most genes, the adjusted Sneath value is similar between species as expected (most dots lie along the x=y axis in Fig. 1B-D). For example, there is very high correlation between values for two gerbil species (Pearson correlation coefficient $r = 0.973$, $p < 2.2e\text{-}16$; Fig. 1B) and also between two murid rodent species ($r = 0.943$, $p < 2.2e\text{-}16$; Fig. 1C). The comparison between gerbil and murid proteins includes several outlier sequences (dots away from the x=y line), although the overall correlation is still high ($r = 0.909$, $p < 2.2e\text{-}16$; Fig. 1D). Similar results were obtained using Epstein's coefficient: very high correlation two gerbil species ($r = 0.965$,

146 $p < 2.2e-16$), between two murid species ($r = 0.924$, $p < 2.2e-16$), and lower correlation between
147 murids and sand rat ($r = 0.884$, $p < 2.2e-16$) (Additional file 1: Fig. S1).

148 A small fraction of proteins lies away from x=y axis indicating differences in Sneath values. To
149 identify those proteins that have 'unusually high' sequence divergence in gerbils compared to
150 mice and rats, we first calculated the difference between adjusted Sneath values for sand rat
151 and their corresponding murid protein average values and then ranked these 'dissimilarity
152 differences' from largest to smallest. High ranking proteins are abnormally divergent in the sand
153 rat compared to their mouse and rat homologues, while low ranking proteins are divergent in
154 mouse and rat compared to the sand rat. Second, by comparing rank order with dissimilarity
155 difference, we looked for discontinuities in the distribution that could be used to identify a set
156 of 'unusually' divergent proteins (Fig. 2). Most proteins have a dissimilarity difference close to
157 0, while a small number of proteins differ drastically between the sand rat and murid lineage.

158 Discontinuities in the graph suggest there are 50 'aberrantly divergent' protein coding genes in
159 sand rat with difference values above 1.80, of which 16 are exceptionally unusual with Sneath
160 dissimilarity differences above 3. In contrast, murids have 18 aberrantly divergent protein
161 coding genes with difference values above 1.80, with just 2 highly unusual genes showing a
162 difference value above 3 (Fig. 2; Additional file 1: Fig. S2).

163 To further test if these results are driven by unusual evolution in gerbils, and not slow evolution
164 in mouse and rat, we also compared Sneath values for sand rat proteins to the average Sneath
165 values of their homologues across a wider clade encompassing kangaroo rat (*Dipodomys ordii*),
166 jerboa (*Jaculus jaculus*), blind mole rat (*Nannospalax galili)*, mouse and rat. This analysis did

not change the overall picture, identifying 67 sand rat proteins (dissimilarity difference > 1.80) including 17 exceptionally unusual (dissimilarity difference > 3) (Additional file 1: Fig.S3).

Aberrantly divergent gerbil proteins are associated with islands of high GC

It has been shown previously that gerbil genomes have islands of GC skew, and for the *Pdx1* gene (located in the island of most extreme GC skew) this is associated with radical protein sequence divergence (3,5). To test whether excess mutational change and/or GC-skew has driven unusual protein sequence divergence in other gerbil genes, we focused on the 50 genes with aberrantly high relative sequence divergence in sand rats (highest adjusted Sneath value).

First, we asked how many of the aberrantly divergent proteins are encoded by genes located in the most extreme GC-rich region of the sand rat genome (the set of genes including *Pdx1*) (3,5). Our dataset of 9771 orthologous genes includes 14 genes from the extreme GC-rich region in the sand rat genome, out of which 10 are present in the set of 50 most aberrantly divergent protein sequences (Fig. 3). Extreme GC-rich genes are significantly enriched in the set of aberrantly divergent proteins (Fisher's Exact Test; p-value = 4.7e-21). These 10 genes are *Pdx1*, *Medag*, *Pex11g*, *Tex45*, *Insr*, *Trappc5*, *Pan3*, *B3glct*, *Pdap1*, and *Cdx2*.

We extended our analysis beyond the extreme GC-rich genomic region and tested if other aberrantly divergent genes are associated with other less extreme high GC regions identified previously in the sand rat genome (3,5) (Additional file 1: Table S1). In total, we found 19 of the 50 most aberrantly divergent proteins are encoded by genes in known high GC regions (including the 10 noted above), compared to 277 out of 9721 non-divergent genes in these regions. This is a significant enrichment of aberrantly divergent proteins encoded within the

known GC-rich islands of the sand rat genome (Fisher's Exact Test; p-value = 1.0e-16). Even when we removed the genes located in the extreme high GC region from this analysis, the enrichment is significant (9/40 of divergent proteins in high GC regions, compared to 273/9717 non-divergent proteins; Fisher's Exact Test p-value = 1.5e-06). Using high GC region coordinates previously reported for the Mongolian jird (3,5), we also found a similar enrichment of aberrantly divergent proteins in GC regions (Fisher's Exact Test; p-value = 0.00096) (Additional file 1: Table S2 and Fig. S4).

The above tests focus on the chromosomal location of each gene and the nucleotide compositional effects occurring in genomic regions. We also analyzed genes individually to enable a higher resolution analysis of the patterns of nucleotide substitution underpinning protein divergence. Amino acid substitutions, which occurred to an extreme extent in the aberrantly divergent proteins, are caused by non-synonymous substitutions. However, here we focus on synonymous substitution rates (dS) to enable us to detect the presence of underlying mutational processes, independent of selection on amino acid changes. Specifically, we tested whether aberrantly divergent gerbil proteins are encoded by genes that have experienced GC-skew (an excess of weak to strong mutations) and/or an excess of mutational change in all categories. We calculated dS values for all categories of nucleotide changes: weak-to-strong (dS_{ws} = rate of A or T being altered to C or G), strong-to-weak (dS_{sw}: rate of G/C to A/T), weak-to-weak (dS_{ww}) and strong-to-strong (dS_{ss}). We defined mutational 'outliers' as genes with dS values more than 2.5 times the average dS for that category (5).

Out of the 50 genes encoding aberrantly divergent proteins, 52% are dS_{ws} outliers (26/50)

compared to only 4% of the remaining genes (358/9721), showing that dS_{ws} outlier genes are significantly enriched for aberrantly divergent proteins (Fisher's Exact Test; p-value = 6.3e-24). We also find that dS_{sw}, dS_{ww}, or dS_{ss} outliers are significantly enriched for aberrantly divergent proteins (dS_{sw} p-value = 4.4e-30; dS_{ww} p-value = 4.4e-19; dS_{ss} p-value = 8.8e-25) (Additional file 1: Table S3 – 5). These results indicate that aberrantly divergent gerbil proteins are largely encoded by genes that have experienced unusually high levels of mutational change, even at non-synonymous sites, and especially high levels of weak-to-strong mutations.

To formally test if these elevated mutational effects are associated with GC-skew (excess of change towards G or C nucleotides), we measured the ratio of dS_{ws} to dS_{sw} (high dS_{ws} causes GC content to increase, high dS_{sw} causes AT content to increase). Plotting dS_{ws} for each gene against dS_{sw} shows a large number of genes in gerbils with GC-skew, as previously reported (sand rat dS_{ws}/dS_{sw} = mean 1.81; Mongolian jird dS_{ws}/dS_{sw} = mean 1.99; Fig. 4A and 4B) (5). The same effect is not seen in murid rodents (mouse dS_{ws}/dS_{sw} = mean 1.20; rat dS_{ws}/dS_{sw} = mean 1.24; Fig. 4C and 4D). Interestingly, most of the aberrantly divergent proteins in gerbils lie above the x=y line, indicating they are encoded by genes that have strong GC-skew (sand rat aberrantly divergent genes dS_{ws}/dS_{sw} = mean 2.19; Mongolian jird aberrantly divergent genes dS_{ws}/dS_{sw} = mean 1.97; Fig. 4A and 4B). These results argue that high mutation accumulation and GC-skew are major factors contributing to the evolution of divergent gerbil proteins.

Phenotypic implications of aberrantly divergent gerbil genes

We asked if the evolution of aberrantly divergent proteins, driven by elevated mutational and gBGC processes, could have had phenotypic consequences for gerbils. Specifically, we wished to test in the evolution of divergent gerbil genes could be associated with the type 2 diabetes-prone phenotype observed in sand rats and Mongolian jirds.

First, we performed Gene Ontology (GO) analyses to test for over-representation of biological functions or molecular pathways in the set of aberrantly divergent genes, and as a control we also analyzed the initial set of 9771 genes. We found significant enrichment of a number of biological processes, molecular functions and cellular components in the dataset of 9771 sand rat genes when compared to the mouse genome (Fisher's Exact Test, q-value < 0.05), but overrepresentation was modest with maximum fold-change 2.01. Other functions showed underrepresentation including sensory perception of smell (GO:0007608; 20-fold underrepresentation, q-value = 5.44e-137), likely a result of only using 1-to-1 orthologues causing bias against dynamic gene families with extensive gene duplication and loss. We did not find significant overrepresentation or underrepresentation for biological processes, functions or cellular components in the aberrantly divergent proteins when compared against the dataset of 9771 sand rat genes (q-value < 0.05). We interpret these results to mean that the aberrantly divergent proteins that evolved in gerbils are associated with a diversity of biological functions, not only dietary metabolism.

To search for possible associations between aberrant protein divergence and propensity to type 2 diabetes, we took a candidate gene approach. Specifically, we asked if any of the proteins

showing extreme relative divergence in gerbils have been linked to dietary metabolism in other species. We note that amongst the top 10 most aberrantly divergent sand rat proteins, four are clearly associated with lipid or carbohydrate metabolism. The genes encoding these are all located in the extreme GC-rich region of sand rat genome (Table 1). These four proteins are the previously discussed PDX1 transcription factor, plus MEDAG (Mesenteric Estrogen Dependent Adipogenesis), INSR (Insulin Receptor) and SPP1 (Secreted Phosphoprotein 1 or osteopontin). In Fig. 5, we show protein sequence alignments of PDX1 and INSR to highlight some of the highly unusual amino acid changes observed in sand rat and related gerbil species, compared to other mammals (Fig. 5). In the PDX1 protein, there is a high degree of conservation across vertebrates of the hexapeptide domain (a cofactor binding domain) and the homeodomain (DNA-binding and sequence recognition domain), but extreme divergence in three gerbil species (Fig. 5A) (3). We note, however, that three amino acids in the PDX1 homeodomain that have been previously associated with T2D in humans are conserved in the three gerbils (Fig. 5A). For INSR, a strong candidate for association with metabolic function, we find many amino acid changes unique to the gerbil lineage throughout the protein, although again nearly all key amino acid sites previously associated with T2D in humans remained conserved (Fig. 5B). An exception is K1095Q located in the INSR tyrosine kinase domain; an identical point mutation is associated with insulin resistance and T2D in humans, although the biochemical effect of this missense mutation remains unknown (34).

Discussion

When unusual or divergent proteins are observed in some species and not others, this is

generally thought to result from natural selection acting to adjust or optimize a protein for a new or modified function. The contributions of mutation rate variation and recombination frequency differences between species are easily overlooked. However, there is growing evidence that these genomic and chromosomal level processes can have dramatic effects on the way that genes and proteins evolve. In this study, we focus on two gerbil species that have the unusual GC-rich regions distributed through their genome, where recombination-related processes are affecting the evolution of genes, including genes that are highly conserved in other mammals (3,5). We ask if these gerbils have an abnormal number of aberrantly divergent genes compared to related rodent species, and test if these genes have become divergent because of association with GC-rich genomic regions and GC-skew, rather than because of positive selection. In addition, we ask whether these aberrant genes could be linked to physiological abnormalities observed in some gerbils.

Aberrantly divergent gerbil proteins have an extremely high number of amino acid changes at sites that are conserved in other rodent lineages, or indeed conserved across the vertebrates. The conservation of amino acid sequence is indicative of purifying (negative) selection acting to prevent change, and thus it is striking to see such major changes in conserved proteins. One possible hypothesis to explain the aberrantly divergent proteins is that strong positive selection and adaptive changes occurred in the gerbil lineage. However, we detect signals of positive selection in only 3 out of 50 aberrantly divergent proteins (Additional file 2). These 3 proteins are TEX37 (Testis Expressed 37), SYCP2L (Synaptonemal Complex Protein 2 Like) and IL15 (Interleukin 15); they are not associated with GC-skew nor located in GC-rich regions. TEX37 is predominantly expressed in the testis (35) while SYCP2L is mainly expressed in the ovary

(36). In addition, gene association studies report a correlation between *SYCP2L* variants and lipid metabolism, although the mechanism remains unknown (37,38). IL15 is a proinflammatory cytokine that activates T-cells and aberrant activity has been linked to destruction of pancreatic beta-cells in type 1 diabetes (T1D) (39,40). However, autoimmune destruction of β-cells has not been postulated as a cause of metabolic dysfunction in gerbils. In general, reproductive proteins and immunity-related proteins are under positive selection in many species (41,42). For the 47 other aberrantly divergent gerbil proteins, the extent of change exceeds other lineages and there is no clear signature of positive selection, making an adaptive hypothesis unlikely.

A second possibility is that these aberrant proteins have lost function and are randomly accumulating amino acid changes. This is again unlikely given that we do not observe early stop codons or frameshifts in the coding sequences of these genes. A third hypothesis is that genomic processes have inadvertently caused unusual sequence changes to become fixed in populations, even if maladaptive at the time of origination. For example, GC-biased gene conversion (gBGC) can convert heterozygous sites to homozygous sites, and overall gBGC promotes the fixation of neutral and slightly deleterious weak-to-strong substitutions (8,43). This hypothesis predicts that aberrantly divergent proteins would be encoded by genes located in GC rich genomic regions, as these are localized regions where gBGC has occurred at high rates. Consistent with this prediction, we find a significant enrichment of aberrantly divergent genes mapping to previously described high GC regions in the gerbil genome (3,5). Also consistent with this hypothesis, aberrant gerbil genes are enriched for dS_{ws} outliers and show strong GC-skew in gerbils (sand rat average dS_{ws}/dS_{sw} = 2.19) compared to murids (mouse

314 average dS_{ws}/dS_{sw} = 1.20).

315 We argue, therefore, that aberrantly divergent gerbil genes are likely to have arisen through
316 unusually high levels of nucleotide substitution, including especially high levels of weak-to-
317 strong substitutions resulting in GC-skew. Given the degree of amino acid change observed in
318 these 50 gerbil proteins, the high rates of nucleotide substitution have likely resulted in fixation
319 of slightly deleterious mutations in these proteins, with natural selection only successfully
320 removing individuals with the most severely deleterious amino acid changes. We suggest that
321 gerbil evolution has witnessed a tug-of-war between localized GC-skew in certain genomic
322 regions causing fixation of deleterious mutations and natural selection preserving essential
323 genes trapped in these 'quagmires' of GC-skew.

324 The two gerbil species focused on in this study have also been reported to become obese and
325 exhibit metabolic abnormalities when maintained on a standard laboratory diet or a high fat diet
326 (23,25,44,45). For example, sand rat islet cells show more pancreatic β-cell damage compared
327 to rat islets when exposed to high glucose concentrations, and even healthy sand rats on plant-
328 based diets do not have a strong response to human insulin (46,47). Could these metabolic
329 disorders be related to the accumulation of deleterious mutations in gerbils, driven by excessive
330 biased gene conversion? We do not have direct evidence linking specific genetic changes with
331 phenotype, but we argue there is a plausible connection. Here, we highlight four aberrantly
332 divergent gerbil proteins for which the human and mouse orthologues have been associated
333 with dietary metabolism: these proteins are PDX1, INSR, MEDAG and SPP1. In two cases, the
334 human orthologues have been associated directly with type 2 diabetes or other metabolic

335 diseases, and these associations can be tracked down to single amino acid substitutions or other
336 small mutations. In gerbils we see dramatic amino acid change at multiple conserved sites. In
337 the other two cases, functional studies in mouse or human indicate key roles in adipose tissue.

338 PDX1 is a homeodomain transcription factor essential for vertebrate pancreatic development
339 and normal β-cell function (15–17). Given its essential roles in embryonic development and
340 adult physiology, it is not surprising that PDX1 is highly conserved across vertebrates. Indeed,
341 *PDX1* mutations have been linked metabolic disease in humans spanning various degrees of
342 severity. The homeodomain is the mostly highly conserved protein domain in PDX1 and point
343 mutations in this region severely compromise PDX1 function in humans. For example, an
344 individual born with pancreas agenesis was reported to carry a E164D substitution on one allele
345 and a E178K substitution on the other, with both mutations shown to significantly shorten
346 PDX1 protein half-life (19). In addition, the mutation R197H has >50% decreased binding
347 affinity with the human insulin promoter sequence and is associated with type 2 diabetes (20).
348 These conditions in humans are associated with single amino acid changes to the homeodomain,
349 yet by contrast the sand rat PDX1 homeodomain has a total of 15 amino acid differences
350 compared to the conserved sequence (Fig. 5A). It is the most aberrantly divergent protein out
351 of 9771 analyzed proteins in the present study. Although the 15 substitutions do not include
352 changes at the three best-characterized disease-associated sites in humans, it is likely that the
353 combination of large number of substitutions would radically compromise protein function.

354 INSR, the insulin receptor, is the cell surface receptor for insulin and IGF peptides and initiates
355 the insulin signaling pathway in vertebrates (48). In sand rat, the INSR protein is the 6th most

aberrantly divergent protein, carrying many amino acid substitutions at otherwise conserved sites. Other vertebrates carry small changes to the INSR protein, and these highlight the possible phenotypic consequences of amino acid substitutions. In particular, recent work has uncovered adaptive change in the INSR protein of Mexican cavefish, notably a P211L substitution mutation (equivalent to P220L in human INSR) that has been fixed by positive selection (49). This mutation reduces binding capability to the insulin peptide and results in insulin resistance in the cavefish, which paradoxically may be a trait beneficial to survival in a cave environment lacking a stable food supply (49). In humans, deleterious mutations affecting INSR function have also been reported, with over 30 different point mutations described in clinical reports (31). In gerbils, such as sand rat, we find that although there are a large number of unusual changes, these generally do not match the individual mutations associated with human metabolic conditions, apart from point mutation K1095Q, nor the P220L mutation reported in Mexican cavefish populations. Concerning point mutation K1095Q, this was reported in a small cohort study of diabetes patients with insulin resistance, but the exact biochemical effects of this point mutation are not yet known (34). In summary, these observations suggest that severely detrimental amino acid changes occurring in PDX1 and INSR of gerbils are likely to have been removed by natural selection, while changes in residues with lower levels of functional importance have been allowed, even though these may have been deleterious at the time of origin. Nonetheless, it is difficult to predict the consequences of all amino acid changes in the gerbil based on these comparisons, especially as there are no reports approaching this dramatic extent of amino acid change in other organisms.

For the two other aberrant proteins putatively associated with dietary metabolism, MEDAG and

SPP1, we cannot make comparisons to variants associated with human phenotypes. However, the known functions of these proteins suggest plausible association with dietary metabolism. MEDAG is a well-conserved pro-inflammatory protein (91% sequence identity between humans and mice) that promotes adipocyte differentiation and regulates adipocyte glucose uptake (50). A study comparing islet transcriptomes of diabetic and healthy individuals reported significantly different expression of *MEDAG* between these two groups (51), and a clinical study has reported association between copy number variation in the *MEDAG* loci and obesity (52). SPP1, also known as osteopontin, is a multifunctional protein best known for its role in biomineralization and bone remodeling. However, SPP1 also has roles in in adipose tissue and contributes to adipose tissue inflammation and insulin resistance (53). Mice fed on a high fat diet, or with genetic predisposition to obesity, show enormous elevation of SPP1 expression levels (54), while blocking SPP1 function by antibodies or genetic mutation reduces adipose inflammation and improves insulin sensitivity (55,56). In humans, SPP1 expression is also reported to be elevated in adipose tissue of obese individuals (54,57). We speculate, therefore, that the amino acid changes in these gerbil proteins likely cause changes in protein function and could contribute to the unusual dietary physiology of gerbils.

Conclusions

In this paper, we show there are more than twice the number of aberrantly divergent proteins in gerbils compared to mice and rats, and that to a large extent the genes that encode these aberrant proteins are associated with GC-skew and previously reported GC-rich genomic regions (3,5). We propose that a tug-of-war between GC-skew and natural selection has been taking place

399 during gerbil evolution, leaving behind an excess of deleterious mutations in several aberrant
400 gerbil genes, several of which may be partly responsible for the abnormal metabolic phenotypes
401 observed in some gerbil species.

402 **Methods**

403 *Curation and alignment of 1-to-1 orthologues*

404 We identified and aligned groups of orthologous sequences following a previously described
405 approach (5). Briefly, we obtained predicted gene sequences for 12 rodents and human from
406 Ensembl (release 95, accession IDs of genome annotations in Additional file 1: Table S6 and
407 cds and protein IDs in Additional file 2), retrieved the longest transcript for each gene and
408 identified groups of orthologous transcripts between these species using Orthofinder version
409 2.2.7 (58,59), applying default parameters, the diamond aligner version 0.9.21 (60) and the
410 species tree topology shown in Fig. 1A. From these, we identified groups of orthologous genes
411 that have a single sequence in (a) both murids (mouse and rat) and the outgroup species (human),
412 (b) at least one gerbil (sand rat and/or Mongolian jird) but not more than one sequence in either
413 species, and (c) all but one or two of the remaining species, in which case we removed those
414 species from the analysis. This approach was chosen to maximize the number of genes analyzed
415 despite assembly errors and occasional gene duplications or deletions in rodent genomes.

416 Protein and nucleotide sequences of each orthologous gene set were aligned using MACSE
417 v2.03 with default parameters and recoded with the MACSE exportAlignment function (options
418 '-codonForInternalStop NNN -codonForExternalFS --- -codonForInternalFS ---' (61). We
419 removed low similarity regions using the default parameters of HmmCleaner.pl version

0.180750 (62). Additional filtering was performed with MACSE reportMaskAA2NT function, with parameters '-min_NT_to_keep_seq 30 -mask_AA $ -min_seq_to_keep_site 4 -min_percent_NT_at_ends 0.3 -dist_isolate_AA 3 -min_homology_to_keep_seq 0.3 -min_internal_homology_to_keep_seq 0.5'.We also filtered each alignment by removing any sequence for which gaps represented more than 70% of the non-gap length of the longest gene in the alignment. Since this removes species from alignments, we again applied the species filter used above (requiring both murids and human to be single copy, at least one gerbil to be single copy, and all but one or two other species to be single copy. DNA alignments for each orthologue dataset can be found in Additional file 3.

dN and dS calculation

For each orthologue dataset, the tree in Fig. 1A was trimmed to only include the species in that alignment. Branch lengths for each alignment were optimized using the BppML subprogram version 2.3.1 of BppSuite (63), the YN98 (F3X4) model (64) and parameters available online (65). Synonymous and nonsynonymous substitution rates (dS and dN, respectively) were estimated for each mutational category (weak-to-strong, strong-to-weak, weak-to-weak, and strong-to-strong) using the parameter 'map.type = Combination(reg1 = dNdS, reg2 = SW)' of the BppML subprogram MapNH version 1.1.1 (66). Branch lengths were summed from the Muridae node to the tree tip to retrieve rate measurements for mouse, rat, sand rat and Mongolian jird.

Protein dissimilarity calculation and ranking

From the protein alignments, we removed any sites containing a stop codon, an unknown

441 residue, or a missing residue in one or more species. We then identified all sites which had the
442 same amino acid in three outgroup species. Outgroups were usually human (*Homo sapiens*),
443 guinea pig (*Cavia porcellus*) and squirrel (*Ictidomys tridecemlineatus*), but for seven
444 alignments these were human, degu (*Octodon degus*) and chinchilla (*Chinchilla lanigera*) due
445 to lack of guinea pig and squirrel sequence, for 376 alignments these were human, guinea pig
446 and jaculus (*Jaculus jaculus*) due to lack of squirrel sequence, and for 270 alignments we used
447 the human, squirrel and degu sequence due to lack of guinea pig sequence. Sites with the same
448 amino acid in all three outgroups were defined as 'conserved positions'; only these sites were
449 used to calculate protein dissimilarity. Proteins with <50 residues meeting this criterion were
450 excluded, reducing the dataset to 9771 genes for sand rat and 10069 for Mongolian jird.

451 To calculate a Sneath score (32), each site which differed from a 'conserved position' was
452 assigned a value dictated by the Sneath Index, and values summed to produce the Sneath value
453 for the protein in that species. Epstein's Coefficient assigns different values to each amino acid
454 difference; these were summed in the same way (33). Scores were divided by the length of the
455 parsed alignment to produce adjusted protein dissimilarity values.

456 *GC-skew analysis and correlation with high GC peaks*

457 To measure GC-skew, synonymous substitution rates for weak to strong changes (dS_{ws}) were
458 divided by dS for strong to weak changes (dS_{sw}), estimated using the species tree and calculated
459 from the murid-gerbil node to the terminal branch tips. Genes with $dS_{ws}/dS_{sw} >> 1$ are under
460 strong GC-skew, and genes with $dS_{ws}/dS_{sw} << 1$ are under strong AT-skew. To test for correlation
461 with position, 10000 random samples of 50 sand rat genes from the 9771 proteins dataset or 50

462 Mongolian jird genes from the 10069 orthologues dataset were selected and chromosomal
463 location in mouse recorded; samplings were compared to locations for the 50 genes with highest
464 adjusted Sneath value. To test for correlation with high GC peaks or islands in the sand rat and
465 Mongolian jird genomes, we used peaks defined previously using sliding window analysis of
466 dS_{ws} (5) (coordinates in Additional file 1: Table S1).

467 *Analysis of protein functions*

468 To test if 'aberrantly divergent' sand rat proteins are enriched for biological functions, we tested
469 for statistical overrepresentation using PANTHER and Fisher's Exact test with calculation of
470 False Discovery Rate as the correction method (67,68). PANTHER identified mouse orthologue
471 IDs for 9671 of the 9771 proteins of interest (including 49 out of 50 aberrantly divergent
472 proteins).

473 **List of abbreviations**

474 GC: guanine and cytosine

475 AT: adenine and thymine

476 gBGC: GC-biased gene conversion

477 T2D: type 2 diabetes

478 GO: Gene Ontology

479 RTK: receptor-tyrosine kinase

480 **Declarations**

481 *Ethics approval and consent to participate*

482 Not applicable

483 *Consent for publication*

484 Not applicable

485 *Availability of data and materials*

486 All data generated or analyzed during this study are included in this published article and its
487 additional information files. Scripts used for analyses have been deposited on GitHub (69).

488 *Competing interests*

489 The authors declare that they have no competing interests.

490 *Funding*

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

493 *Authors' contributions*

494 YD and PWHH conceived this study. YD and RP performed the analysis. YD and PWHH
495 drafted a first version of this manuscript. All authors contributed to revision and editing of the
496 manuscript and read and approved the final version.

the sand rat. Am J Physiol Content. 1965 Feb 1;208(2):297–300.

2. Leibowitz G, Ferber S, Apelqvist A, Edlund H, Gross DJ, Cerasi E, et al. IPF1/PDX1 deficiency and beta-cell dysfunction in Psammomys obesus, an animal With type 2 diabetes. Diabetes. 2001 Aug 1;50(8):1799–806.

3. Hargreaves AD, Zhou L, Christensen J, Marlétaz F, Liu S, Li F, et al. Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. Proc Natl Acad Sci. 2017;114(29):7677–82.

4. Dai Y, Holland PWH. The Interaction of Natural Selection and GC Skew May Drive the Fast Evolution of a Sand Rat Homeobox Gene. Su B, editor. Mol Biol Evol. 2019 Jul 1;36(7):1473–80.

5. Pracana R, Hargreaves AD, Mulley JF, Holland PWH. Runaway GC Evolution in Gerbil Genomes. Wilke C, editor. Mol Biol Evol. 2020 Apr 24;

6. Oliver JL, Carpena P, Román-Roldán R, Mata-Balaguer T, Mejías-Romero A, Hackenberg M, et al. Isochore chromosome maps of the human genome. Gene. 2002;300(1–2):117–27.

7. Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. Genome Res. 2010;20(8):1001–9.

8. Duret L, Galtier N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annu Rev Genomics Hum Genet. 2009;10(1):285–311.

9. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature. 2008;454(7203):479–85.

10. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics. 2001 Jan;159:907–11.

11. Boĺivar P, Mugal CF, Nater A, Ellegren H. Recombination rate variation modulates gene sequence evolution mainly via GC-Biased gene conversion, not Hill-Robertson interference, in an avian system. Mol Biol Evol. 2016;33(1):216–27.

12. Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. PLoS Genet. 2013;9(8).

13. Figuet E, Ballenghien M, Romiguier J, Galtier N. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol Evol. 2015;7(1):240–50.

14. Montoya-Burgos JI, Boursot P, Galtier N. Recombination explains isochores in mammalian genomes. Trends Genet. 2003 Mar;19(3):128–30.

15. Jonsson J, Carlsson L, Edlund T, Edlund H. Insulin-promoter-factor 1 is required for pancreas development in mice. Nature. 1994;371(6498):606–9.

16. Offield MF, Jetton TL, Labosky PA, Ray M, Stein RW, Magnuson MA, et al. PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. Development. 1996;122:983–95.

554 17. Ahlgren U, Jonsson J, Jonsson L, Simu K, Edlund H. beta -Cell-specific inactivation of the
555 mouse Ipf1/Pdx1 gene results in loss of the beta -cell phenotype and maturity onset diabetes.
556 Genes Dev. 1998 Jun 15;12(12):1763–8.

557 18. Glaspole IN, Watson AL, Allan H, Chapman S, Cooper WA, Corte TJ, et al. Determinants and
558 outcomes of prolonged anxiety and depression in idiopathic pulmonary fibrosis. Eur Respir J.
559 2017 Aug 17;50(2):1700168.

560 19. Schwitzgebel VM, Mamin A, Brun T, Ritz-Laser B, Zaiko M, Maret A, et al. Agenesis of
561 human pancreas due to decreased half-life of insulin promoter factor 1. J Clin Endocrinol
562 Metab. 2003;88(9):4398–406.

563 20. Macfarlane WM, Frayling TM, Ellard S, Evans JC, Allen LIS, Bulman MP, et al. Missense
564 mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. J Clin Invest.
565 1999 Nov 1;104(9):R33–9.

566 21. Leibowitz G, Yuli M, Donath MY, Nesher R, Melloul D, Cerasi E, et al. Beta-cell
567 glucotoxicity in the Psammomys obesus model of type 2 diabetes. Diabetes. 2001 Feb
568 1;50(Supplement 1):S113–7.

569 22. Gadot M, Leibowitz G, Shafrir E, Cerasi E, Gross D., Kaiser N. Hyperproinsulinemia and
570 Insulin Deficiency in the Diabetic Psammomys obesus. Endocrinology. 1994;135(2):610–6.

571 23. Schmidt-Nielsen K, Haines HB, Hackel DB. Diabetes Mellitus in the Sand Rat Induced by
572 Standard Laboratory Diets. Science (80-). 1964 Feb 14;143(3607):689–90.

29

573 24. Shafrir E, Ziv E, Mosthaf L. Nutritionally induced insulin resistance and receptor defect
574 leading to β-cell failure in animal models. Ann N Y Acad Sci. 1999;892:223–46.

575 25. Boquist L. Obesity and pancreatic islet hyperplasia in the Mongolian gerbil. Diabetologia. 1972
576 Aug;8(4):274–82.

577 26. Li X, Lu J, Wang Y, Huo X, Li Z, Zhang S, et al. Establishment and Characterization of a
578 Newly Established Diabetic Gerbil Line. Bader M, editor. PLoS One. 2016 Jul
579 18;11(7):e0159420.

580 27. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The
581 genetic architecture of type 2 diabetes. Nature. 2016;536(7614):41–7.

582 28. Gaulton KJ. Mechanisms of Type 2 Diabetes Risk Loci. Curr Diab Rep. 2017;17(9).

583 29. Gloyn AL, Ellard S. Defining the genetic aetiology of monogenic diabetes can improve
584 treatment. Expert Opin Pharmacother. 2006;7(13):1759–67.

585 30. Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM. Positional cloning of the
586 mouse obese gene and its human homologue. Nature. 1994 Dec;372(6505):425–32.

587 31. Taylor SI. Lilly Lecture: Molecular Mechanisms of Insulin Resistance: Lessons From Patients
588 With Mutations in the Insulin-Receptor Gene. Diabetes. 1992 Nov 1;41(11):1473–90.

589 32. Sneath PHA. Relations between chemical structure and biological activity in peptides. J Theor
590 Biol. 1966 Nov;12(2):157–95.

591 33. Epstein CJ. Non-randomness of Ammo-acid Changes in the Evolution of Homologous

592 Proteins. Nature. 1967 Jul;215(5099):355–9.

593 34. O'Rahilly S, Choi WH, Patel P, Turner RC, Flier JS, Moller DE. Detection of Mutations in
594 Insulin-Receptor Gene in NIDDM Patients by Analysis of Single-Stranded Conformation
595 Polymorphisms. Diabetes. 1991 Jun 1;40(6):777–82.

596 35. Khan M, Jabeen N, Khan T, Hussain HMJ, Ali A, Khan R, et al. The evolutionarily conserved
597 genes: Tex37, Ccdc73, Prss55 and Nxt2 are dispensable for fertility in mice. Sci Rep. 2018 Dec
598 21;8(1):4975.

599 36. Zhou J, Stein P, Leu NA, Chmátal L, Xue J, Ma J, et al. Accelerated reproductive aging in
600 females lacking a novel centromere protein SYCP2L. Hum Mol Genet. 2015 Nov
601 15;24(22):6505–14.

602 37. Jha P, McDevitt MT, Halilbasic E, Williams EG, Quiros PM, Gariani K, et al. Genetic
603 Regulation of Plasma Lipid Species and Their Association with Metabolic Phenotypes. Cell
604 Syst. 2018 Jun;6(6):709-721.e6.

605 38. Tintle NL, Pottala JV, Lacey S, Ramachandran V, Westra J, Rogers A, et al. A genome-wide
606 association study of saturated, mono- and polyunsaturated red blood cell fatty acids in the
607 Framingham Heart Offspring Study. Prostaglandins, Leukot Essent Fat Acids. 2015
608 Mar;94:65–72.

609 39. Anderson DM, Johnson L, Glaccum MB, Copeland NG, Gilbert DJ, Jenkins NA, et al.
610 Chromosomal assignment and genomic structure of Il15. Genomics. 1995 Feb;25(3):701–6.

611 40. Chen J, Feigenbaum L, Awasthi P, Butcher DO, Anver MR, Golubeva YG, et al. Insulin-
612 dependent diabetes induced by pancreatic beta cell expression of IL-15 and IL-15R. Proc Natl
613 Acad Sci. 2013 Aug 13;110(33):13534–9.

614 41. Vallender EJ. Positive selection on the human genome. Hum Mol Genet. 2004 Oct
615 1;13(suppl_2):R245–54.

616 42. Torgerson DG, Kulathinal RJ, Singh RS. Mammalian Sperm Proteins Are Rapidly Evolving:
617 Evidence of Positive Selection in Functionally Diverse Genes. Mol Biol Evol. 2002 Nov
618 1;19(11):1973–80.

619 43. Rousselle M, Laverré A, Figuet E, Nabholz B, Galtier N. Influence of Recombination and GC-
620 biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals
621 versus Birds. Singh N, editor. Mol Biol Evol. 2019 Mar 1;36(3):458–71.

622 44. Sahraoui A, Dewachter C, Vegh G, Mc Entee K, Naeije R, Bouguerra SA, et al. High fat diet
623 altered cardiac metabolic gene profile in Psammomys obesus gerbils. Lipids Health Dis. 2020
624 Dec 3;19(1):123.

625 45. Sahraoui A, Dewachter C, de Medina G, Naeije R, Aouichat Bouguerra S, Dewachter L.
626 Myocardial Structural and Biological Anomalies Induced by High Fat Diet in Psammomys
627 obesus Gerbils. Chen X, editor. PLoS One. 2016 Feb 3;11(2):e0148117.

628 46. Gross DJ, Leibowitz G, Cerasi E, Kaiser N. Increased susceptibility of islets from diabetes-
629 prone Psammomys obesus to the deleterious effects of chronic glucose exposure.
630 Endocrinology. 1996 Dec 1;137(12):5610–5.

631 47. Shafrir E, Ziv E, Saha AK, Ruderman NB. Regulation of muscle malonyl-CoA levels in the
632 nutritionally insulin-resistant desert gerbil,Psammomys obesus. Diabetes Metab Res Rev. 2002
633 May;18(3):217–23.

634 48. Petersen MC, Shulman GI. Mechanisms of Insulin Action and Insulin Resistance. Physiol Rev.
635 2018 Oct 1;98(4):2133–223.

636 49. Riddle MR, Aspiras AC, Gaudenz K, Peuß R, Sung JY, Martineau B, et al. Insulin resistance in
637 cavefish as an adaptation to a nutrient-limited environment. Nature. 2018 Mar
638 21;555(7698):647–51.

639 50. Zhang H, Chen X, Sairam MR. Novel Genes of Visceral Adiposity: Identification of Mouse
640 and Human Mesenteric Estrogen-Dependent Adipose (MEDA)-4 Gene and Its Adipogenic
641 Function. Endocrinology. 2012 Jun;153(6):2665–76.

642 51. Li L, Pan Z, Yang S, Shan W, Yang Y. Identification of key gene pathways and coexpression
643 networks of islets in human type 2 diabetes. Diabetes Metab Syndr Obes. 2018 Sep;11:553–63.

644 52. D'Angelo CS, Varela MC, de Castro CIE, Otto PA, Perez ABA, Lourenço CM, et al.
645 Chromosomal microarray analysis in the genetic evaluation of 279 patients with syndromic
646 obesity. Mol Cytogenet. 2018 Dec 5;11(1):14.

647 53. Kahles F, Findeisen HM, Bruemmer D. Osteopontin: A novel regulator at the cross roads of
648 inflammation, obesity and diabetes. Mol Metab. 2014 Jul;3(4):384–93.

649 54. Kiefer FW, Zeyda M, Todoric J, Huber J, Geyeregger R, Weichhart T, et al. Osteopontin

650 Expression in Human and Murine Obesity: Extensive Local Up-Regulation in Adipose Tissue
651 but Minimal Systemic Alterations. Endocrinology. 2008 Mar;149(3):1350–7.

652 55. Nomiyama T, Perez-Tilve D, Ogawa D, Gizard F, Zhao Y, Heywood EB, et al. Osteopontin
653 mediates obesity-induced adipose tissue macrophage infiltration and insulin resistance in mice.
654 J Clin Invest. 2007 Oct 1;117(10):2877–88.

655 56. Chapman J, Miles PD, Ofrecio JM, Neels JG, Yu JG, Resnik JL, et al. Osteopontin Is Required
656 for the Early Onset of High Fat Diet-Induced Insulin Resistance in Mice. Polidori C, editor.
657 PLoS One. 2010 Nov 12;5(11):e13959.

658 57. Gómez-Ambrosi J, Catalán V, Ramírez B, Rodríguez A, Colina I, Silva C, et al. Plasma
659 osteopontin levels and expression in adipose tissue are increased in obesity. J Clin Endocrinol
660 Metab. 2007 Sep;92(9):3719–27.

661 58. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons
662 dramatically improves orthogroup inference accuracy. Genome Biol. 2015 Dec 6;16(1):157.

663 59. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics.
664 Genome Biol. 2019 Dec 14;20(1):238.

665 60. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat
666 Methods. 2015 Jan 17;12(1):59–60.

667 61. Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. MACSE v2: Toolkit for the
668 Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Wilke C, editor.

669 Mol Biol Evol. 2018 Oct 1;35(10):2582–4.

670 62. Di Franco A. HmmCleaner.pl - Removing low similarity segments from your MSA [Internet].
671 https://metacpan.org/pod/HmmCleaner.pl. 2018 [cited 2020 Mar 19]. Available from:
672 https://metacpan.org/pod/HmmCleaner.pl

673 63. Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, Rochette NC, et al. Bio++: Efficient
674 Extensible Libraries and Tools for Computational Molecular Evolution. Mol Biol Evol. 2013
675 Aug;30(8):1745–50.

676 64. Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of
677 mammals. J Mol Evol. 1998 Apr;46(4):409–18.

678 65. Pracana R, Hargreaves A. Analyses of GC evolution in gerbil genomes [Internet]. Oxford
679 University Research Archive. 2019 [cited 2020 Jun 15]. Available from:
680 https://ora.ox.ac.uk/objects/uuid:54a81d9b-c493-48ba-bda1-23e14ec3ae7a

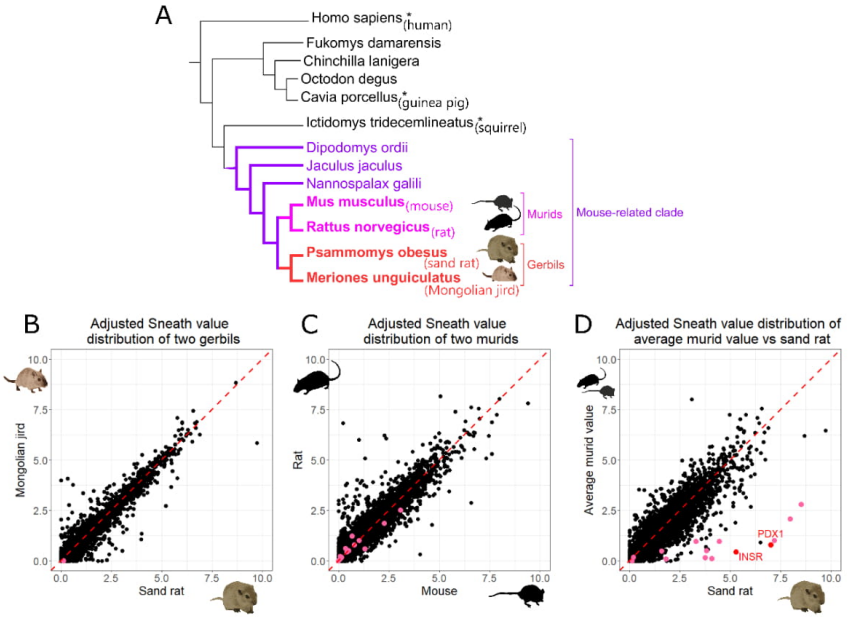
681 66. Romiguier J, Figuet E, Galtier N, Douzery EJP, Boussau B, Dutheil JY, et al. Fast and Robust
682 Characterization of Time-Heterogeneous Sequence Evolutionary Processes Using Substitution
683 Mapping. Liberles D, editor. PLoS One. 2012 Mar 27;7(3):e33852.

684 67. Thomas PD. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function.
685 Genome Res. 2003 Sep 1;13(9):2129–41.

686 68. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: modeling the evolution of gene
687 function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res.

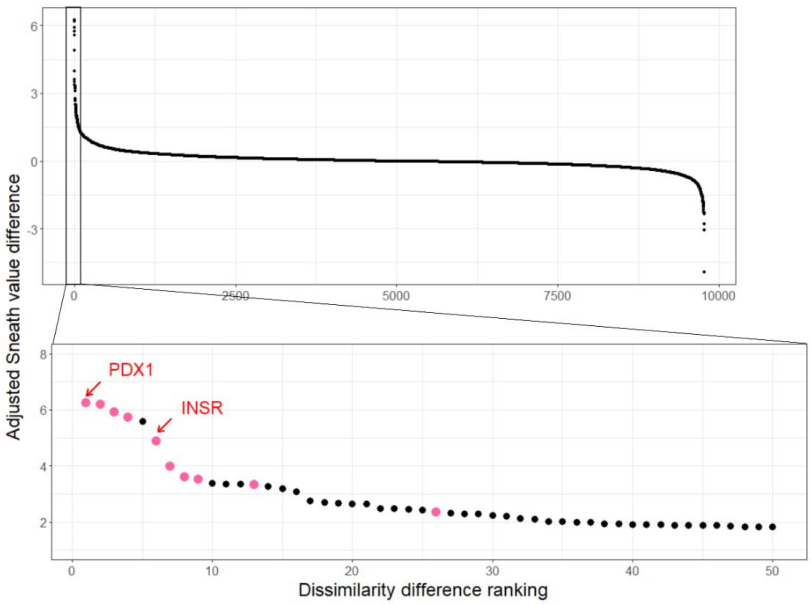
2012 Nov 26;41(D1):D377–86.

69. Pracana R. Divergent gerbil genes [Internet]. GitHub. 2020 [cited 2020 Jun 15]. Available from: https://github.com/roddypr/divergent_gerbil_genes



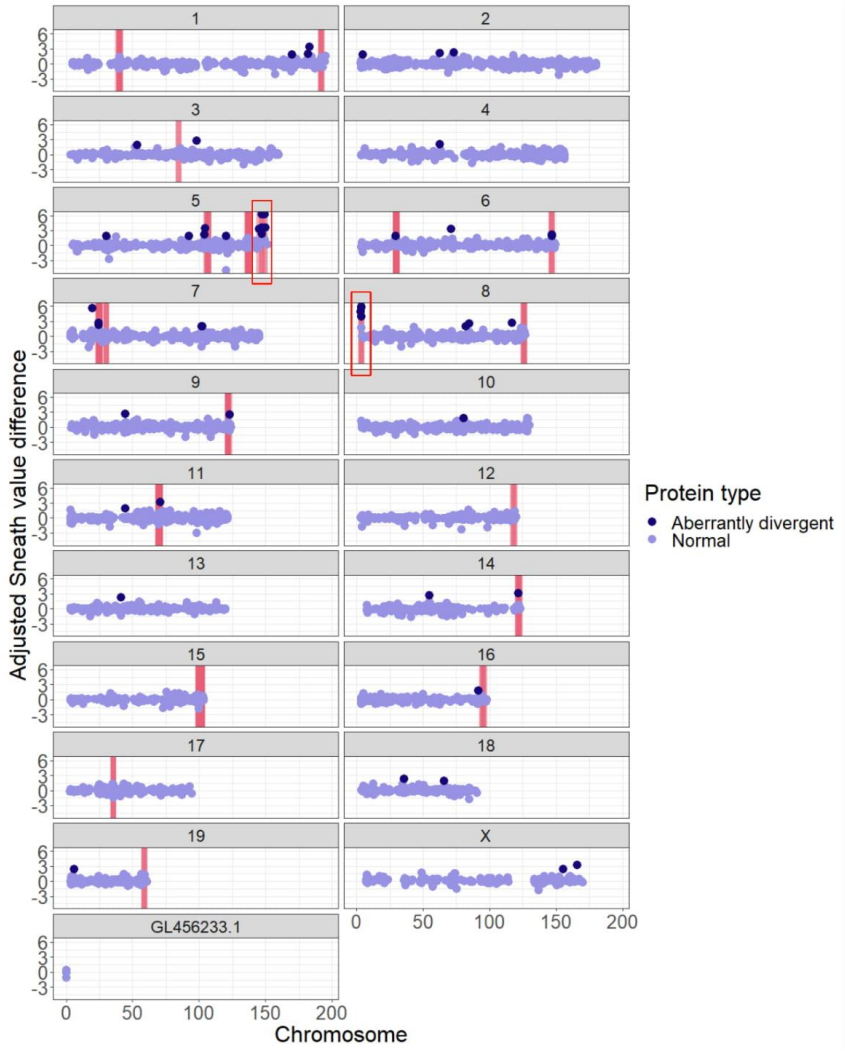
693

694 Fig. 1 Phylogenetic relationships between species analyzed and patterns of protein sequence
695 divergence. (A) Phylogenetic tree showing evolutionary relationships between the species
696 analyzed. Outgroup species used for most proteins are marked with an asterisk, murid species
697 in pink, gerbil species in red. (B–D) Adjusted Sneath values orthologous proteins compared
698 between (B) two gerbil species, (C) two murid species, (D) the average of two murid species
699 and the sand rat. Each point represents one protein; proteins encoded by genes in the extreme
700 GC-rich region of sand rat are shown in pink or (for PDX1 and INSR) in red. Only one protein
701 encoded within the extreme GC-rich region is present in (B) due to missing genes in the
702 Mongolian jird genome assembly.



703

704 Fig. 2 Dissimilarity difference ranking for 9,771 sand rat proteins against the difference in
705 adjusted Sneath value compared to the murid homologue. The top 61 ranked proteins are
706 enlarged in the bottom plot with proteins PDX1 (rank 1) and INSR (rank 6) marked with arrows.
707 Amongst these proteins, those encoded by genes in the extreme GC-rich region are shown in
708 pink.



709

710 Fig. 3 Aberrantly divergent sand rat proteins are frequently encoded by genes in GC-rich islands.

711 Each panel shows one mouse chromosome (scale in Mb) to which the locations of sand rat

712 orthologues are mapped. All analyzed sand rat genes are displayed as dots plotted according to

713 the midpoint position of their corresponding mouse orthologue. The position of each dot on the

714 y-axis shows the difference in adjusted Sneath value between the sand rat and mouse

39

orthologues. Pink lines indicate locations of GC-rich regions identified previously (3,5). 'Clusters' of aberrantly divergent proteins with more than two proteins mapped to regions less than 1 Mb apart are marked with open red boxes.

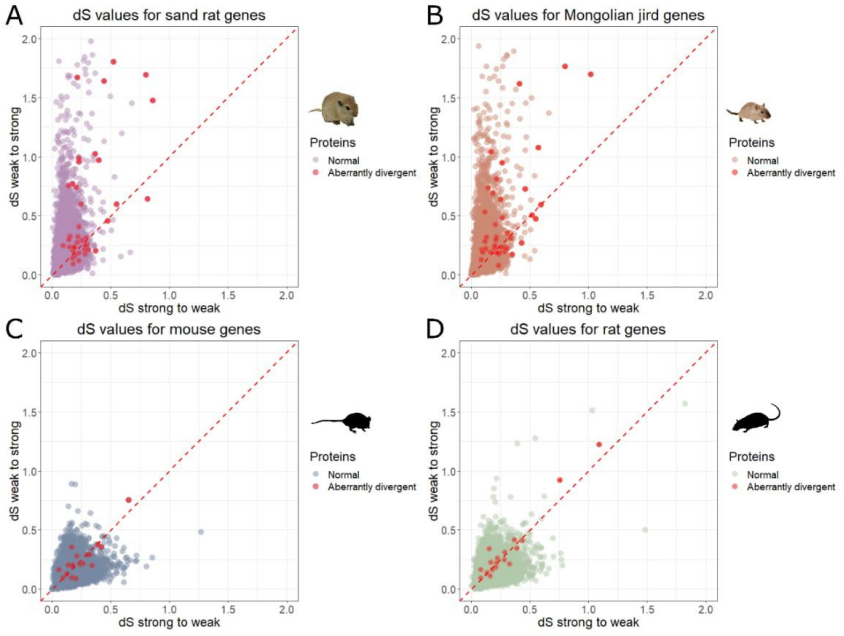


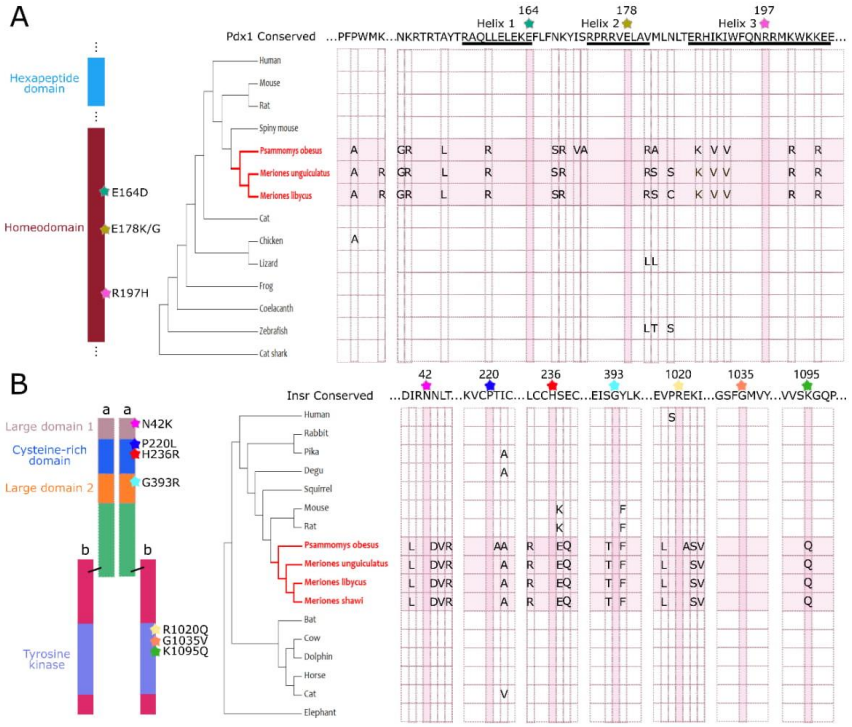
Fig. 4 Mutational patterns at synonymous sites for orthologous genes from four rodent species. Genes encoding aberrantly divergent proteins are highlighted with red circles. (A) Graph comparing weak-to-strong (dS_{ws}) and strong-to-weak (dS_{sw}) synonymous mutation rates for sand rat genes. The 'chimney' shape indicates that many sand rat genes have undergone GC-skew; most of the aberrantly divergent proteins are in this category. Eighteen sand rat genes with dS_{ws} values larger than 2 are omitted from the graph, among which 12 encode aberrantly divergent sand rat proteins. (B) Graph comparing weak-to-strong (dS_{ws}) and strong-to-weak (dS_{sw}) synonymous mutation rates for Mongolian jird genes. The 'chimney' shape indicates

727 GC-skew; most aberrantly divergent proteins are in this category. Ten Mongolian jird genes
728 with dS_{ws} values larger than 2 are omitted, among which two encode aberrantly divergent
729 Mongolian jird proteins. (C) Graph comparing dS_{ws} and dS_{sw} for mouse genes. (D) Graph
730 comparing dS_{ws} and dS_{sw} for rat genes.

731

Rank	Gene	GC-rich region	Sand rat dS_{ws} value	Encoded protein
1	*Pdx1*	Yes	20.09	Pancreatic duodenal homeobox. Transcription factor involved in pancreas development and β-cell function.
2	*Medag*	Yes	3.09	Mesenteric estrogen dependent adipogenesis. Involved in adipocyte differentiation and glucose uptake.
3	*Pex11g*	Yes	4.58	Peroxisomal membrane protein. Involved in growth and division of peroxisomes (organelles partaking in lipid biosynthesis).
4	*Tex45*	Yes	1.48	Testis-expressed protein. Understudied gene expressed in testis.
5	*Cblc*	No	0.60	E3 ubiquitin-protein ligase. Involved in protein degradation; down-regulates receptor-tyrosine kinase (RTK) signaling.
6	*Insr*	Yes	7.08	Insulin receptor. RTK receptor involved in insulin signaling and stimulation of glucose uptake.
7	*Trappc5*	Yes	34.40	Trafficking protein particle complex subunit. Implicated in vesicular transport from ER to Golgi.
8	*Pan3*	Yes	10.66	Subunit of PAN2-PAN3 deadenylase complex. Trim poly(A) tail prior to degradation.
9	*B3glct*	Yes	2.29	Beta 3-glucosyltransferase. Transfers glucose onto fucose on TSR repeat proteins.
10	*Spp1*	No	0.22	Secreted phosphoprotein 1 or osteopontin. Implicated in bone mineralization, immune system, and diet-induced insulin resistance.

732 Table 1 A list of the sand rat genes that encode the most aberrantly divergent sand rat proteins ordered by ranking, whether these genes are in GC-rich regions,
733 dS_{ws} value and a short description of the function of their encoded protein.



734

735 Fig. 5 Alignment of key functional domains in PDX1 and INSR proteins. (A) Alignment of the
736 conserved PDX1 hexapeptide and homeodomain sequence from representative vertebrate
737 species. The three gerbil species are shown in red and sites where amino acid substitutions are
738 associated with T2D in humans are marked with a star. (B) Alignment of representative regions
739 in the INSR protein, including both large domains, the cysteine-rich domain and the tyrosine
740 kinase domain. The four gerbil species are shown in red and sites where amino acid
741 substitutions are associated with T2D in humans are marked with a star.

742

References

1. Lynch, V. J., Leclerc, R. D., May, G. & Wagner, G. P. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat. Genet.* **43**, 1154–1159 (2011).

2. Stone, J. R. & Wray, G. A. Rapid Evolution of cis-Regulatory Sequences via Local Point Mutations. *Mol. Biol. Evol.* **18**, 1764–1770 (2001).

3. Kappen, C., Schughart, K. & Ruddle, F. H. Two steps in the evolution of Antennapedia-class vertebrate homeobox genes. *Proc. Natl. Acad. Sci.* **86**, 5459–5463 (1989).

4. Eichler, E. E. & Sankoff, D. Structural Dynamics of Eukaryotic Chromosome Evolution. *Science.* **301**, 793–797 (2003).

5. Dehal, P. & Boore, J. L. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol.* **3**, e314 (2005).

6. Macaya, G., Thiery, J.-P. & Bernardi, G. An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.* **108**, 237–254 (1976).

7. Cortadas, J., Macaya, G. & Bernardi, G. An Analysis of the Bovine Genome by Density Gradient Centrifugation: Fractionation in Cs2SO4/3,6-Bis(acetatomercurimethyl)dioxane Density Gradient. *Eur. J. Biochem.* **76**, 13–19 (1977).

8. Cuny, G., Soriano, P., Macaya, G. & Bernardi, G. The Major Components of the Mouse and Human Genomes. *Eur. J. Biochem.* **115**, 227–233 (1981).

9. Bernardi, G. *et al.* The Mosaic Genome of Warm-Blooded Vertebrates. *Science.* **228**, 953–958 (1985).

10. Bernardi, G. The isochore organization of the human genome and its evolutionary history — a review. *Gene* **135**, 57–66 (1993).

11. Cammarano, R., Costantini, M. & Bernardi, G. The isochore patterns of invertebrate genomes. *BMC Genomics* **10**, 538 (2009).

12. The Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honey bee Apis mellifera. *Nature* **443**, 931–949 (2006).

13. Li, J. & Heinz, K. M. Genome complexity and organization in the red imported fire ant Solenopsis invicta Buren. *Genet. Res.* **75**, 129–135 (2000).

14. Chapman, J. A. *et al.* The dynamic genome of Hydra. *Nature* **464**, 592–596 (2010).

15. Wurm, Y. *et al.* The genome of the fire ant Solenopsis invicta. *Proc. Natl. Acad. Sci.* **108**, 5679–5684 (2011).

16. Touchon, M. *et al.* Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLoS Genet.* **5**, e1000344 (2009).

17. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241**, 3–17 (2000).

18. Tatarinova, T. V., Alexandrov, N. N., Bouck, J. B. & Feldmann, K. A. GC3 biology in corn, rice, sorghum and other grasses. *BMC Genomics* **11**, 308 (2010).

19. Lander, S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

20. St John, J. A. *et al.* Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biol.* **13**, 415 (2012).

21. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).

22. Dehal, P. The Draft Genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins. *Science.* **298**, 2157–2167 (2002).

23. Cohen, N., Dagan, T., Stone, L. & Graur, D. GC composition of the human genome: In search of isochores. *Mol. Biol. Evol.* **22**, 1260–1272 (2005).

24. Clay, O. & Bernardi, G. How not to search for isochores: A reply to Cohen et al. *Mol. Biol. Evol.* **22**, 2315–2317 (2005).

25. Costantini, M., Clay, O., Auletta, F. & Bernardi, G. An isochore map of human chromosomes. *Genome Res.* **16**, 536–541 (2006).

26. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).

27. Arbeithuber, B., Betancourt, A. J., Ebner, T. & Tiemann-Boege, I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc. Natl. Acad. Sci.* **112**, 2109–2114 (2015).

28. Castellano, D., Eyre-Walker, A. & Munch, K. Impact of mutation rate and selection at linked sites on DNA variation across the genomes of humans and other homininae. *Genome Biol. Evol.* **12**, 3550–3561 (2019).

29. Ehrlich, M. & Wang, R. Y.-H. 5-Methylcytosine in Eukaryotic DNA. *Science.* **212**, 1350 (1981).

30. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombinations through a sequence level genetic map. *Science.* **363**, eaau1043 (2019).

31. Tyekucheva, S. *et al.* Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol.* **9**, R76 (2008).

32. Smith, T. C. A., Arndt, P. F. & Eyre-Walker, A. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLOS Genet.* **14**, e1007254 (2018).

33. Foerstner, K. U., von Mering, C., Hooper, S. D. & Bork, P. Environments shape the nucleotide composition of genomes. *EMBO Rep.* **6**, 1208–1213 (2005).

34. Kawashima, T. *et al.* Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium. *Proc. Natl. Acad. Sci.* **97**, 14257–14262 (2000).

35. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat. Genet.* **31**, 180–183 (2002).

36. Carels, N. & Bernardi, G. Two classes of genes in plants. *Genetics* **154**, 1819–1825 (2000).

37. D'Onofrio, G., Ghosh, T. C. & Saccone, S. Different functional classes of genes are characterized by different compositional properties. *FEBS Lett.* **581**, 5819–5824 (2007).

38. Duret, L., Mouchiroud, D. & Gautier, C. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **40**, 308–317 (1995).

39. Dekker, J. GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. *Genome Biol.* **8**, (2007).

40. Mouchiroud, D. *et al.* The distribution of genes in the human genome. *Gene* **100**, 181–187 (1991).

41. dos Reis, M. & Wernisch, L. Estimating Translational Selection in Eukaryotic Genomes. *Mol. Biol. Evol.* **26**, 451–461 (2009).

42. Sémon, M., Mouchiroud, D. & Duret, L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. *Hum. Mol. Genet.* **14**, 421–427 (2005).

43. Pouyet, F., Mouchiroud, D., Duret, L. & Sémon, M. Recombination, meiotic expression and human codon usage. *Elife* **6**, 1–19 (2017).

44. Vaillant, I. & Paszkowski, J. Role of histone and DNA methylation in gene regulation. *Curr. Opin. Plant Biol.* **10**, 528–533 (2007).

45. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16**, 990–995 (2009).

46. Duret, L. & Galtier, N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).

47. Backström, N., Ceplitis, H., Berlin, S. & Ellegren, H. Gene Conversion Drives the Evolution of HINTW, an Ampliconic Gene on the Female-Specific Avian W Chromosome. *Mol. Biol. Evol.* **22**, 1992–1999 (2005).

48. Coop, G. & Przeworski, M. An evolutionary view of human recombination. *Nat. Rev. Genet.* **8**, 23–34 (2007).

49. Whitby, M. C. Making crossovers during meiosis. *Biochem. Soc. Trans.* **33**, 1451 (2005).

50. Storlazzi, A., Xu, L., Cao, L. & Kleckner, N. Crossover and noncrossover recombination during meiosis: timing and pathway relationships. *Proc. Natl. Acad. Sci.* **92**, 8512–8516 (1995).

51. Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C. & Patrinos, G. P. Gene conversion: mechanisms, evolution and human disease. *Nat. Rev. Genet.* **8**, 762–775 (2007).

52. Mancera, E., Bourgon, R., Brozzi, A., Huber, W. & Steinmetz, L. M. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454**, 479–485 (2008).

53. Odenthal-Hesse, L., Berg, I. L., Veselis, A., Jeffreys, A. J. & May, C. A. Transmission Distortion Affecting Human Noncrossover but Not Crossover Recombination: A Hidden Source of Meiotic Drive. *PLoS Genet.* **10**, e1004106 (2014).

54. Williams, A. L. *et al.* Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife* **4**, 1–21 (2015).

55. Halldorsson, B. V. *et al.* The rate of meiotic gene conversion varies by sex and age. *Nat. Genet.* **48**, 1377–1384 (2016).

56. Li, R. *et al.* A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat. Commun.* **10**, 24–29 (2019).

57. Pessia, E. *et al.* Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.* **4**, 675–682 (2012).

58. Figuet, E., Ballenghien, M., Romiguier, J. & Galtier, N. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. *Genome Biol. Evol.* **7**, 240–250 (2015).

59. Bolívar, P., Guéguen, L., Duret, L., Ellegren, H. & Mugal, C. F. GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biol.* **20**, 5 (2019).

60. Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics* **159**, 907–911 (2001).

61. Duret, L., Eyre-Walker, A. & Galtier, N. A new perspective on isochore evolution. *Gene* **385**, 71–74 (2006).

62. Lan, Y., Liu, M. & Cao, Y. The complete mitochondrial genome of Psammomys obesus (Rodentia: Muridae). *Mitochondrial DNA Part B Resour.* **3**, 97–98 (2018).

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

64. American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. *Diabetes Care* **41**, S13–S27 (2018).

65. Fu, Z., R. Gilbert, E. & Liu, D. Regulation of Insulin Synthesis and Secretion and Pancreatic Beta-Cell Dysfunction in Diabetes. *Curr. Diabetes Rev.* **9**, 25–53 (2012).

66. Mosley, A. L. & Özcan, S. Glucose Regulates Insulin Gene Transcription by Hyperacetylation of Histone H4. *J. Biol. Chem.* **278**, 19660–19666 (2003).

67. Huang, X. F. & Arvan, P. Intracellular Transport of Proinsulin in Pancreatic β-Cells. *J. Biol. Chem.* **270**, 20417–20423 (1995).

68. Dean, P. M. Ultrastructural morphometry of the pancreatic beta-cell. *Diabetologia* **9**, 115–119 (1973).

69. Jensen, M. V. *et al.* Metabolic cycling in control of glucose-stimulated insulin secretion. *Am. J. Physiol. - Endocrinol. Metab.* **295**, 1287–1297 (2008).

70. Gauthier, B. R. *et al.* Synaptotagmin VII splice variants α, β, and δ are expressed in pancreatic β-cells and regulate insulin exocytosis. *FASEB J.* **22**, 194–206 (2008).

71. Hiriart, M. & Aguilar-Bryan, L. Channel regulation of glucose sensing in the pancreatic β-cell. *Am. J. Physiol. - Endocrinol. Metab.* **295**, 1298–1306 (2008).

72. Wiederkehr, A. *et al.* Matrix alkalinization: A novel mitochondrial signal for sustained pancreatic β-cell activation. *EMBO J.* **28**, 417–428 (2009).

73. Haeusler, R. A., McGraw, T. E. & Accili, D. Biochemical and cellular properties of insulin receptor signalling. *Nat. Rev. Mol. Cell Biol.* **19**, 31–44 (2018).

74. Petersen, M. C. & Shulman, G. I. Mechanisms of Insulin Action and Insulin Resistance. *Physiol. Rev.* **98**, 2133–2223 (2018).

75. Hubbard, S. R. The Insulin Receptor: Both a Prototypical and Atypical Receptor Tyrosine Kinase. *Cold Spring Harb. Perspect. Biol.* **5**, a008946–a008946 (2013).

76. Petersen, M. C. & Shulman, G. I. Mechanisms of Insulin Action and Insulin Resistance. *Physiol. Rev.* **98**, 2133–2223 (2018).

77. Haeusler, R. A., McGraw, T. E. & Accili, D. Metabolic Signalling: Biochemical and cellular properties of insulin receptor signalling. *Nat. Rev. Mol. Cell Biol.* **19**, 31–44 (2018).

78. Zhang, J., Hiken, J., Davis, A. E. & Lawrence, J. C. Insulin stimulates dephosphorylation of phosphorylase in rat epitrochlearis muscles. *J. Biol. Chem.* **264**, 17513–17523 (1989).

79. Hancock, M. L. *et al.* Insulin Receptor Associates with Promoters Genome-wide and Regulates Gene Expression. *Cell* **177**, 722-736.e22 (2019).

80. Schwartz, S. S. *et al.* A Unified Pathophysiological Construct of Diabetes and its Complications. *Trends Endocrinol. Metab.* **28**, 645–655 (2017).

81. Macfarlane, W. M. *et al.* Missense mutations in the insulin promoter factor-1 gene predispose to type 2 diabetes. *J. Clin. Invest.* **104**, R33–R39 (1999).

82. Fletcher, B., Gulanick, M. & Lamendola, C. Risk Factors for Type 2 Diabetes Mellitus. *J. Cardiovasc. Nurs.* **16**, 17–23 (2002).

83. Kahn, S. E., Hull, R. L. & Utzschneider, K. M. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* **444**, 840–846 (2006).

84. Gunasekaran, U. & Gannon, M. Type 2 diabetes and the aging pancreatic beta cell. *Aging (Albany. NY).* **3**, 565–575 (2011).

85. Ahlgren, U., Jonsson, J., Jonsson, L., Simu, K. & Edlund, H. beta-Cell-specific inactivation of the mouse Ipf1/Pdx1 gene results in loss of the beta-cell phenotype and maturity onset diabetes. *Genes Dev.* **12**, 1763–1768 (1998).

86. Kitamura, T. *et al.* The forkhead transcription factor Foxo1 links insulin signaling to Pdx1 regulation of pancreatic β cell growth. *J. Clin. Invest.* **110**, 1839–1847 (2002).

87. Taylor, S. I. Lilly Lecture: Molecular Mechanisms of Insulin Resistance: Lessons From Patients With Mutations in the Insulin-Receptor Gene. *Diabetes* **41**, 1473–1490 (1992).

88. George, S. *et al.* A family with severe insulin resistance and diabetes due to a mutation in AKT2. *Science.* **304**, 1325–1328 (2004).

89. Cho, H. *et al.* Insulin Resistance and a Diabetes Mellitus-Like Syndrome in Mice Lacking the Protein Linked references are available on JSTOR for this article : Insulin Resistance and a Diabetes Mellitus-Like Syndrome in Mice Lacking the Protein Kinase Akt2 (PKBI). *Science.* **292**, 1728–1731 (2001).

90. Defronzo, R. A. & Lang, R. Hypophosphatemia and Glucose Intolerance: Evidence for Tissue Insensitivity to Insulin. *N. Engl. J. Med.* **303**, 1259–1263 (1980).

91. Moran, A. *et al.* Insulin resistance during puberty: Results from clamp studies in 357 children. *Diabetes* **48**, 2039–2044 (1999).

92. Schmidt-Nielsen, K., Haines, H. B. & Hackel, D. B. Diabetes Mellitus in the Sand Rat Induced by Standard Laboratory Diets. *Science.* **143**, 689–690 (1964).

93. Miki, E., Like, A. A., Stuart Soeldner, J., Steinke, J. & Cahill, G. F. Acute ketotic-type diabetic syndrome in sand rats (Psammomys obesus) with special reference to the pancreas. *Metabolism* **15**, 749–760 (1966).

94. Brodoff, B. N., Penhos, J. C. & Levine, R. The effects of feeding and various hormones on the glucose tolerance of the sand rat (psammomys obesus). *Diabetologia* **3**, 167–170 (1967).

95. Kaiser, N. *et al.* Psammomys obesus, a model for environment-gene interactions in type 2 diabetes. *Diabetes* **54**, (2005).

96. Kalderon, B., Gutman, A., Levy, E., Shafrir, E. & Adler, J. H. Characterization of Stages in Development of Obesity-Diabetes Syndrome in Sand Rat (Psammomys obesus). *Diabetes* **35**, 717–724 (1986).

97. 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).

98. Shafrir, E., Ziv, E., Saha, A. K. & Ruderman, N. B. Regulation of muscle malonyl-CoA levels in the nutritionally insulin-resistant desert gerbil,Psammomys obesus. *Diabetes. Metab. Res. Rev.* **18**, 217–223 (2002).

99. Kaiser, N. *et al.* Dynamic Changes in -Cell Mass and Pancreatic Insulin During the Evolution of Nutrition-Dependent Diabetes in Psammomys obesus: Impact of Glycemic Control. *Diabetes* **54**, 138–145 (2005).

100. Boquist, L. Obesity and pancreatic islet hyperplasia in the Mongolian gerbil. *Diabetologia* **8**, 274–282 (1972).

101. Li, X. *et al.* Establishment and Characterization of a Newly Established Diabetic Gerbil Line. *PLoS One* **11**, e0159420 (2016).

102. Hargreaves, A. D. *et al.* Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. *Proc. Natl. Acad. Sci.* **114**, 7677–7682 (2017).

103. Gehring, W. J. *et al.* The structure of the homeodomain and its functional implications. *Trends Genet.* **6**, 323–329 (1990).

104. Slack, J. M., Holland, P. W. & Graham, C. F. The zootype and the phylotypic stage. *Nature* **361**, 490–492 (1993).

105. Glaspole, I. N. *et al.* Determinants and outcomes of prolonged anxiety and depression in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **50**, 1700168 (2017).

106. Pollard, S. L. & Holland, P. W. H. Evidence for fourteen homeobox gene clusters in human genome ancestry. *Curr. Biol.* **10**, 1059–1062 (2000).

107. Li, H., Zeitler, P. S., Valerius, M. T., Small, K. & Potter, S. S. Gsh-1, an orphan Hox gene, is required for normal pituitary development. *EMBO J.* **15**, 714–724 (1996).

108. Szucsik, J. C. *et al.* Altered forebrain and hindbrain development in mice mutant for the Gsh- 2 homeobox gene. *Dev. Biol.* **191**, 230–242 (1997).

109. Wright, C. V. E., Schnegelsberg, P. & Robertis, E. M. DE. XlHbox 8: a novel Xenopus homeo protein restricted to a narrow band of endoderm. *Development* **105**, 787–794 (1989).

110. Jonsson, J., Carlsson, L., Edlund, T. & Edlund, H. Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* **371**, 606–609 (1994).

111. Duprey, P. *et al.* A mouse gene homologous to the Drosophila gene caudal is expressed in epithelial cells from the embryonic intestine. *Genes Dev.* **2**, 1647–1654 (1988).

112. Beck, F., Erler, T., Russell, A. & James, R. Expression of Cdx-2 in the mouse embryo and placenta: Possible role in patterning of the extra-embryonic membranes. *Dev. Dyn.* **204**, 219–227 (1995).

113. Gamer, L. W. & Wright, C. V. E. Murine Cdx-4 bears striking similarities to the Drosophila caudal gene in its homeodomain sequence and early expression pattern. *Mech. Dev.* **43**, 71–81 (1993).

114. Boam, D. S. W. & Docherty, K. A tissue-specific nuclear factor binds to multiple sites in the human insulin-gene enhancer. *Biochem. J.* **264**, 233–239 (1989).

115. Scott, V., Clark, A. R., Hutton, J. C. & Docherty, K. Two proteins act as the IUF1 insulin gene enhancer binding factor. *FEBS Lett.* **290**, 27–30 (1991).

116. Melloul, D., Ben-Neriah, Y. & Cerasi, E. Glucose modulates the binding of an islet-specific factor to a conserved sequence within the rat I and the human insulin promoters. *Proc. Natl. Acad. Sci.* **90**, 3865–3869 (1993).

117. MacFarlane, W. M. *et al.* Impaired expression of transcription factor IUF1 in a pancreatic β-cell line derived from a patient with persistent hyperinsulinaemic hypoglycaemia of infancy (nesidioblastosis). *FEBS Lett.* **413**, 304–308 (1997).

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

119. 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).

120. Leonard, J. *et al.* Characterization of somatostatin transactivating factor-1, a novel homeobox factor that stimulates somatostatin expression in pancreatic islet cells. *Mol. Endocrinol.* **7**, 1275–1283 (1993).

121. Offield, M. F. *et al.* PDX-1 is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* **122**, 983–995 (1996).

122. Ramond, C. *et al.* Understanding human fetal pancreas development using subpopulation sorting, RNA sequencing and single-cell profiling. *Development* **3**, dev.165480 (2018).

123. Hebrok, M., Kim, S. K. & Melton, D. A. Notochord repression of endodermal Sonic hedgehog permits pancreas development. *Genes Dev.* **12**, 1705–1713 (1998).

124. Wells, J. M. & Melton, D. A. Early mouse endoderm is patterned by soluble factors from adjacent germ layers. *Development* **127**, 1563–1572 (2000).

125. Zhu, Z. *et al.* Genome Editing of Lineage Determinants in Human Pluripotent Stem Cells Reveals Mechanisms of Pancreatic Development and Diabetes. *Cell Stem Cell* **18**, 755–768 (2016).

126. Khoo, C. *et al.* Research Resource: The Pdx1 Cistrome of Pancreatic Islets. *Mol. Endocrinol.* **26**, 521–533 (2012).

127. Johnson, J. D. *et al.* Increased islet apoptosis in Pdx1+ /– mice. *J. Clin. Invest.* **111**, 1147–1160 (2003).

128. Gao, T. *et al.* Pdx1 maintains β cell identity and function by repressing an α cell program. *Cell Metab.* **19**, 259–271 (2014).

129. Brissova, M. *et al.* Reduction in Pancreatic Transcription Factor PDX-1 Impairs Glucose-stimulated Insulin Secretion. *J. Biol. Chem.* **277**, 11225–11232 (2002).

130. Yang, Y., Thorel, F., Boyer, D. F., Herrera, P. L. & Wright, C. V. E. Context-specific a-to-b-cell reprogramming by forced Pdx1 expression. *Genes Dev.* **25**, 1680–1685 (2011).

131. Bernardo, A. S., Hay, C. W. & Docherty, K. Pancreatic transcription factors and their role in the birth, life and survival of the pancreatic β cell. *Mol. Cell. Endocrinol.* **294**, 1–9 (2008).

132. Hay, C. W. & Docherty, K. Comparative analysis of insulin gene promoters: Implications for diabetes research. *Diabetes* **55**, 3201–3213 (2006).

133. German, M. *et al.* The insulin gene promoter: A simplified nomenclature. *Diabetes* vol. 44 1002–1004 (1995).

134. Marshak, S., Totary, H., Cerasi, E. & Melloul, D. Purification of the beta-cell glucose-sensitive factor that transactivates the insulin gene differentially in normal and transformed islet cells. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 15057–15062 (1996).

135. Melloul, D. Transcription factors in islet development and physiology: Role of PDX-1 in beta-cell function. *Ann. N. Y. Acad. Sci.* **1014**, 28–37 (2004).

136. MacFarlane, W. M., Read, M. L., Gilligan, M., Bujalska, I. & Docherty, K. Glucose modulates the binding activity of the beta-cell transcription factor IUF1 in a phosphorylation-dependent manner. *Biochem. J.* **303**, 625–31 (1994).

137. Mosley, A. L. & Özcan, S. 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 (2004).

138. Mosley, A. L., Corbett, J. A. & Özcan, S. Glucose Regulation of Insulin Gene Expression Requires the Recruitment of p300 by the β-Cell-Specific Transcription Factor Pdx-1. *Mol. Endocrinol.* **18**, 2279–2290 (2004).

139. Gauthier, B. R. *et al.* PDX1 Deficiency Causes Mitochondrial Dysfunction and Defective Insulin Secretion through TFAM Suppression. *Cell Metab.* **10**, 110–118 (2009).

140. Laybutt, D. R. *et al.* Endoplasmic reticulum stress contributes to beta cell apoptosis in type 2 diabetes. *Diabetologia* **50**, 752–763 (2007).

141. Sachdeva, M. M. *et al.* Pdx1 (MODY4) regulates pancreatic beta cell susceptibility to ER stress. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19090–19095 (2009).

142. Khoo, C. *et al.* Research Resource: The Pdx1 Cistrome of Pancreatic Islets. *Mol. Endocrinol.* **26**, 521–533 (2012).

143. Hattersley, A. T. & Patel, K. A. Precision diabetes: learning from monogenic diabetes. *Diabetologia* **60**, 769–777 (2017).

144. Wang, X. *et al.* Genome-wide analysis of PDX1 target genes in human pancreatic progenitors. *Mol. Metab.* **9**, 57–68 (2018).

145. Heuvel-Borsboom, H., de Valk, H. W., Losekoot, M. & Westerink, J. Maturity onset diabetes of the young: Seek and you will find. *Neth. J. Med.* **74**, 193–200 (2016).

146. Stoffers, D. A., Zinkin, N. T., Stanojevic, V., Clarke, W. L. & Habener, J. F. Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nat. Genet.* **15**, 106–110 (1997).

147. Schwitzgebel, V. M. *et al.* Agenesis of human pancreas due to decreased half-life of insulin promoter factor 1. *J. Clin. Endocrinol. Metab.* **88**, 4398–4406 (2003).

148. Gragnoli, C. *et al.* IPF-1/MODY4 gene missense mutation in an Italian family with type 2 and gestational diabetes. *Metabolism.* **54**, 983–988 (2005).

149. Wang, X. *et al.* Point mutations in the PDX1 transactivation domain impair human β-cell development and function. *Mol. Metab.* 1–18 (2019) doi:10.1016/J.MOLMET.2019.03.006.

150. Hartono, S. R., Korf, I. F. & Chédin, F. GC skew is a conserved property of unmethylated CpG island promoters across vertebrates. *Nucleic Acids Res.* **43**, gkv811 (2015).

151. Blanga-Kanfi, S. *et al.* Rodent phylogeny revised: Analysis of six nuclear genes from all major rodent clades. *BMC Evol. Biol.* **9**, 1–12 (2009).

152. Montgelard, C., Forty, E., Arnal, V. & Matthee, C. A. Suprafamilial relationships among Rodentia and the phylogenetic effect of removing fast-evolving nucleotides in mitochondrial, exon and intron fragments. *BMC Evol. Biol.* **8**, 1–16 (2008).

153. She, J. X., Bonhomme, F., Boursot, P., Thaler, L. & Catzeflis, F. M. Molecular phylogenies in the genus Mus: Comparative analysis of electrophoretic, scnDNA hybridization, and mtDNA RFLP data. *Biol. J. Linn. Soc.* **41**, 83–103 (1990).

154. Catzeflis, F. M., Aguilar, J. P. & Jaeger, J. J. Muroid rodents: Phylogeny and evolution. *Trends Ecol. Evol.* **7**, 122–126 (1992).

155. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci.* **99**, 803–808 (2002).

156. Medarde, N., López-Fuster, M. J., Muñoz-Muñoz, F. & Ventura, J. Spatio-temporal variation in the structure of a chromosomal polymorphism zone in the house mouse. *Heredity (Edinb).* **109**, 78–89 (2012).

157. Aniskin, V. M. *et al.* Unusually extensive karyotype reorganization in four congeneric *Gerbillus* species (Muridae: Gerbillinae). *Cytogenet. Genome Res.* **112**, 131–140 (2006).

158. Capanna, E. & Merani, M. S. Karyotypes of Somalian rodent populations. *Monit. Zool. Ital. Suppl.* **14**, 227–240 (1981).

159. Lay, D. M. & Nadler, C. F. Hybridization in the rodent genus *Meriones*. *Cytogenet. Genome Res.* **8**, 35–50 (1969).

160. Weiss, L., Mayeda, K. & Dully, M. The Karyotype of the Mongolian Gerbil, Meriones unguiculatus. *Cytologia (Tokyo).* **35**, 102–106 (1970).

161. Mostafa, B. H., Abderrazak Souha, B., Sabeh, F., Noureddine, C. & Riadh, B. I. Evidence for the existence of two distinct species: Psammomys obesus and Psammomys vexillaris within the sand rats (Rodentia, Gerbillinae), reservoirs of cutaneous leishmaniasis in Tunisia. *Infect. Genet. Evol.* **6**, 301–308 (2006).

162. Andrews, S. *et al.* FastQC. (2012).

163. Joshi, N. A. & Fass, J. N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. (2011).

164. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–52 (2011).

165. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. & Jones, S. J. M. ABySS : A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).

166. Chikhi, R. & Medvedev, P. Informed and Automated k-Mer Size Selection for Genome Assembly. *Bioinformatics* **30**, 31–37 (2013).

167. Sanger Institute UK. assembly-stats. (2014).

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

169. Harrison, R. G., Bogdanowicz, S. M., Hoffmann, R. S., Yensen, E. & Sherman, P. W. Phylogeny and evolutionary history of the ground squirrels (Rodentia: Marmotinae). *J. Mamm. Evol.* **10**, 249–276 (2003).

170. Brown, J. & Ojeda, R. Granivory: patterns, processes, and consequences of seed consumption on two continents. *Rev. Chil. Hist. Nat.* **60**, 337–349 (1987).

171. Reichman, O. J. Relation of Desert Rodent Diets to Available Resources. *J. Mammal.* **56**, 731–751 (1975).

172. Glémin, S. Surprising Fitness Consequences of GC-Biased Gene Conversion: I. Mutation Load and Inbreeding Depression. *Genetics* **185**, 939–959 (2010).

173. Capra, J. A., Hubisz, M. J., Kostka, D., Pollard, K. S. & Siepel, A. A Model-Based Analysis of GC-Biased Gene Conversion in the Human and Chimpanzee Genomes. *PLoS Genet.* **9**, (2013).

174. Knecht, E. *et al.* Intracellular protein degradation in mammalian cells: Recent developments. *Cell. Mol. Life Sci.* **66**, 2427–2443 (2009).

175. Boucher, M. J., Selander, L., Carlsson, L. & Edlund, H. Phosphorylation marks IPF1/PDX1 protein for degradation by glycogen synthase kinase 3-dependent mechanisms. *J. Biol. Chem.* **281**, 6395–6403 (2006).

176. Claiborn, K. C. *et al.* Pcif1 modulates Pdx1 protein stability and pancreatic β cell function and survival in mice. *J. Clin. Invest.* **120**, 3713–3721 (2010).

177. Ciechanover, A. & Ben-Saadon, R. N-terminal ubiquitination: More protein substrates join in. *Trends Cell Biol.* **14**, 103–106 (2004).

178. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).

179. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team. {nlme}: Linear and Nonlinear Mixed Effects Models. (2019).

180. R Development Core Team. R: A Language and Environment for Statistical Computing. (2008).

181. Longo, A., Guanga, G. P. & Rose, R. B. Structural basis for induced fit mechanisms in DNA recognition by the Pdx1 homeodomain. *Biochemistry* **46**, 2948–57 (2007).

182. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).

183. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

184. Zhou, M., Wu, X. & Ginsberg, H. N. Evidence That a Rapidly Turning Over Protein, Normally Degraded by Proteasomes, Regulates hsp72 Gene Transcription in HepG2 Cells. *J. Biol. Chem.* **271**, 24769–24775 (1996).

185. Pickart, C. M. Mechanisms Underlying Ubiquitination. *Annu. Rev. Biochem.* **70**, 503–533 (2001).

186. Kerscher, O., Felberbaum, R. & Hochstrasser, M. Modification of Proteins by Ubiquitin and Ubiquitin-Like Proteins. *Annu. Rev. Cell Dev. Biol.* **22**, 159–180 (2006).

187. Chen, Z., Zhou, Y., Song, J. & Zhang, Z. HCKSAAP-UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties. *Biochim. Biophys. Acta - Proteins Proteomics* **1834**, 1461–1467 (2013).

188. Gao, T. *et al.* UUCD: A family-based database of ubiquitin and ubiquitin-like conjugation. *Nucleic Acids Res.* **41**, 445–451 (2013).

189. Wang, J. R. *et al.* ESA-UbiSite: Accurate prediction of human ubiquitination sites by identifying a set of effective negatives. *Bioinformatics* **33**, 661–668 (2017).

190. Mallon, D. P. The mammals of the Mongolian People's Republic. *Mamm. Rev.* **15**, 71–102 (1985).

191. Shafrir, E. & Gutman, A. Psammomys Obesus of the Jerusalem Colony: A Model For Nutritionally Induced, Non-Insulin-Dependent Diabetes. *J. Basic Clin. Physiol. Pharmacol.* **4**, 83–100 (1993).

192. Leibowitz, G. *et al.* IPF1/PDX1 deficiency and beta-cell dysfunction in Psammomys obesus, an animal With type 2 diabetes. *Diabetes* **50**, 1799–806 (2001).

193. Jeffares, D. C., Tomiczek, B., Sojo, V. & dos Reis, M. A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. in *Parasite Genomics Protocols* (ed. Peacock, C.) 65–90 (Springer New York, 2015). doi:10.1007/978-1-4939-1438-8_4.

194. Ohneda, K., Mirmira, R. G., Wang, J., Johnson, J. D. & German, M. S. The Homeodomain of PDX-1 Mediates Multiple Protein-Protein Interactions in the Formation of a Transcriptional Activation Complex on the Insulin Promoter. *Mol. Cell. Biol.* **20**, 900–911 (2000).

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

196. Kishi, A., Nakamura, T., Nishio, Y., Maegawa, H. & Kashiwagi, A. Sumoylation of Pdx1 is associated with its nuclear localization and insulin gene activation. *Am. J. Physiol. Metab.* **284**, E830–E840 (2003).

197. Frogne, T., Sylvestersen, K. B., Kubicek, S., Nielsen, M. L. & Hecksher-Sørensen, J. Pdx1 is post-translationally modified in vivo and serine 61 is the principal site of phosphorylation. *PLoS One* **7**, (2012).

198. Melloul, D., Kaiser, N. & Cerasi, E. Chronic exposure to high glucose impairs multiple β-cell functions in cultured human pancreatic islets. *Curr. Opin. Endocrinol. Diabetes* **7**, 44–48 (2000).

199. Chakrabarti, S. K., James, J. C. & Mirmira, R. G. Quantitative assessment of gene targeting in vitro and in vivo by the pancreatic transcription factor, Pdx1. Importance of chromatin structure in directing promoter binding. *J. Biol. Chem.* **277**, 13286–13293 (2002).

200. Waeber, G., Thompson, N., Nicod, P. & Bonny, C. Transcriptional activation of the GLUT2 gene by the IPF-1/STF-1/IDX-1 homeobox factor. *Mol. Endocrinol.* **10**, 1327–34 (1996).

201. Raum, J. C. *et al.* FoxA2, Nkx2.2, and PDX-1 Regulate Islet beta-Cell-Specific mafA Expression through Conserved Sequences Located between Base Pairs -8118 and -

7750 Upstream from the Transcription Start Site. *Mol. Cell. Biol.* **26**, 5735–5743 (2006).

202. Stanojevic, V., Habener, J. F. & Thomas, M. K. Pancreas duodenum homeobox-1 transcriptional activation requires interactions with p300. *Endocrinology* **145**, 2918–2928 (2004).

203. Liberzon, A. Role of intrinsic DNA binding specificity in defining target genes of the mammalian transcription factor PDX1. *Nucleic Acids Res.* **32**, 54–64 (2004).

204. Farack, L. *et al.* Transcriptional Heterogeneity of Beta Cells in the Intact Pancreas. *Dev. Cell* **48**, 115-125.e4 (2019).

205. Asfari, M. *et al.* Establishment of 2-mercaptoethanol-dependent differentiated insulin-secreting cell lines. *Endocrinology* **130**, 167–178 (1992).

206. Montoya-Burgos, J. I., Boursot, P. & Galtier, N. Recombination explains isochores in mammalian genomes. *Trends Genet.* **19**, 128–130 (2003).

207. Pracana, R., Hargreaves, A. D., Mulley, J. F. & Holland, P. W. H. Runaway GC Evolution in Gerbil Genomes. *Mol. Biol. Evol.* **37**, 2197–2210 (2020).

208. Shafrir, E., Ziv, E. & Mosthaf, L. Nutritionally induced insulin resistance and receptor defect leading to β-cell failure in animal models. *Ann. N. Y. Acad. Sci.* **892**, 223–246 (1999).

209. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).

210. Gaulton, K. J. Mechanisms of Type 2 Diabetes Risk Loci. *Curr. Diab. Rep.* **17**, (2017).

211. Gloyn, A. L. & Ellard, S. Defining the genetic aetiology of monogenic diabetes can improve treatment. *Expert Opin. Pharmacother.* **7**, 1759–1767 (2006).

212. Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432 (1994).

213. Pracana, R. Divergent gerbil genes. *GitHub* https://github.com/roddypr/divergent_gerbil_genes (2020).

214. Sneath, P. H. A. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–195 (1966).

215. Epstein, C. J. Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. *Nature* **215**, 355–359 (1967).

216. Thomas, P. D. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res.* **13**, 2129–2141 (2003).

217. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377–D386 (2012).

218. Miyata, T., Miyazawa, S. & Yasunaga, T. Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* **12**, 219–236 (1979).

219. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science.* **185**, 862–864 (1974).

220. O'Rahilly, S. *et al.* Detection of Mutations in Insulin-Receptor Gene in NIDDM Patients by Analysis of Single-Stranded Conformation Polymorphisms. *Diabetes* **40**, 777–782 (1991).

221. Zorio, D. A. R. *et al.* De novo sequencing and initial annotation of the Mongolian gerbil (Meriones unguiculatus) genome. *Genomics* 0–1 (2018) doi:10.1016/j.ygeno.2018.03.001.

222. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).

223. Khan, M. *et al.* The evolutionarily conserved genes: Tex37, Ccdc73, Prss55 and Nxt2 are dispensable for fertility in mice. *Sci. Rep.* **8**, 4975 (2018).

224. Zhou, J. *et al.* Accelerated reproductive aging in females lacking a novel centromere protein SYCP2L. *Hum. Mol. Genet.* **24**, 6505–6514 (2015).

225. Jha, P. *et al.* Genetic Regulation of Plasma Lipid Species and Their Association with Metabolic Phenotypes. *Cell Syst.* **6**, 709-721.e6 (2018).

226. Tintle, N. L. *et al.* A genome-wide association study of saturated, mono- and polyunsaturated red blood cell fatty acids in the Framingham Heart Offspring Study. *Prostaglandins, Leukot. Essent. Fat. Acids* **94**, 65–72 (2015).

227. Anderson, D. M. *et al.* Chromosomal assignment and genomic structure of Il15. *Genomics* **25**, 701–706 (1995).

228. Chen, J. *et al.* Insulin-dependent diabetes induced by pancreatic beta cell expression of IL-15 and IL-15R. *Proc. Natl. Acad. Sci.* **110**, 13534–13539 (2013).

229. Vallender, E. J. Positive selection on the human genome. *Hum. Mol. Genet.* **13**, R245–R254 (2004).

230. Torgerson, D. G., Kulathinal, R. J. & Singh, R. S. Mammalian Sperm Proteins Are Rapidly Evolving: Evidence of Positive Selection in Functionally Diverse Genes. *Mol. Biol. Evol.* **19**, 1973–1980 (2002).

231. Rousselle, M., Laverré, A., Figuet, E., Nabholz, B. & Galtier, N. Influence of Recombination and GC-biased Gene Conversion on the Adaptive and Nonadaptive Substitution Rate in Mammals versus Birds. *Mol. Biol. Evol.* **36**, 458–471 (2019).

232. Sahraoui, A. *et al.* Myocardial Structural and Biological Anomalies Induced by High Fat Diet in Psammomys obesus Gerbils. *PLoS One* **11**, e0148117 (2016).

233. Sahraoui, A. *et al.* High fat diet altered cardiac metabolic gene profile in Psammomys obesus gerbils. *Lipids Health Dis.* **19**, 123 (2020).

234. Gross, D. J., Leibowitz, G., Cerasi, E. & Kaiser, N. Increased susceptibility of islets from diabetes-prone Psammomys obesus to the deleterious effects of chronic glucose exposure. *Endocrinology* **137**, 5610–5 (1996).

235. De Meyts, P. Insulin and its receptor: Structure, function and evolution. *BioEssays* **26**, 1351–1362 (2004).

236. Riddle, M. R. *et al.* Insulin resistance in cavefish as an adaptation to a nutrient-limited environment. *Nature* **555**, 647–651 (2018).

237. Zhang, H., Chen, X. & Sairam, M. R. Novel Genes of Visceral Adiposity: Identification of Mouse and Human Mesenteric Estrogen-Dependent Adipose (MEDA)-4 Gene and Its Adipogenic Function. *Endocrinology* **153**, 2665–2676 (2012).

238. Li, L., Pan, Z., Yang, S., Shan, W. & Yang, Y. Identification of key gene pathways and coexpression networks of islets in human type 2 diabetes. *Diabetes. Metab. Syndr. Obes.* **11**, 553–563 (2018).

239. D'Angelo, C. S. *et al.* Chromosomal microarray analysis in the genetic evaluation of 279 patients with syndromic obesity. *Mol. Cytogenet.* **11**, 14 (2018).

240. Kahles, F., Findeisen, H. M. & Bruemmer, D. Osteopontin: A novel regulator at the cross roads of inflammation, obesity and diabetes. *Mol. Metab.* **3**, 384–393 (2014).

241. Kiefer, F. W. *et al.* Osteopontin Expression in Human and Murine Obesity: Extensive Local Up-Regulation in Adipose Tissue but Minimal Systemic Alterations. *Endocrinology* **149**, 1350–1357 (2008).

242. Nomiyama, T. *et al.* Osteopontin mediates obesity-induced adipose tissue macrophage infiltration and insulin resistance in mice. *J. Clin. Invest.* **117**, 2877–2888 (2007).

243. Chapman, J. *et al.* Osteopontin Is Required for the Early Onset of High Fat Diet-Induced Insulin Resistance in Mice. *PLoS One* **5**, e13959 (2010).

244. Gómez-Ambrosi, J. *et al.* Plasma osteopontin levels and expression in adipose tissue are increased in obesity. *J. Clin. Endocrinol. Metab.* **92**, 3719–27 (2007).

245. Haenel, Q., Laurentino, T. G., Roesti, M. & Berner, D. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Mol. Ecol.* **27**, 2477–2497 (2018).

246. Milewski, W. M., Duguay, S. J., Chan, S. J. & Steiner, D. F. Conservation of PDX-1 Structure, Function, and Expression in Zebrafish. *Endocrinology* **139**, 1440–1449 (2014).

247. Gonet, A. E., Stauffacher, W., Pictet, R. & Renold, A. E. Obesity and diabetes mellitus with striking congenital hyperplasia of the islets of langerhans in spiny mice (Acomys Cahirinus). *Diabetologia* **1**, 162–171 (1966).

248. Gutzeit, A. *et al.* Evidence for decreased sensitivity to glucose of isolated islets from spiny mice (Acomys cahirinus). *Diabetologia* **10**, 661–665 (1974).

249. Shafrir, E., Ziv, E. & Kalman, R. Nutritionally Induced Diabetes in Desert Rodents as Models of Type 2 Diabetes: Acomys cahirinus (Spiny Mice) and Psammomys obesus (Desert Gerbil). *ILAR J.* **47**, 212–224 (2006).

250. Chattopadhyay, S. *et al.* High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. *Proc. Natl. Acad. Sci.* **106**, 12412–12417 (2009).

251. Xie, K. T. *et al.* DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science.* **363**, 81–84 (2019).

252. Webb, A. E. *et al.* Adaptive Evolution as a Predictor of Species-Specific Innate Immune Response. *Mol. Biol. Evol.* **32**, 1717–1729 (2015).