SCIENTIFIC REPORTS
natureresearch

OPEN

Comparative analysis of corrected tiger genome provides clues to its neuronal evolution

Parul Mittal[1], Shubham K. Jaiswal[1], Nagarjun Vijay[2], Rituja Saxena[1] & Vineet K. Sharma[1*]

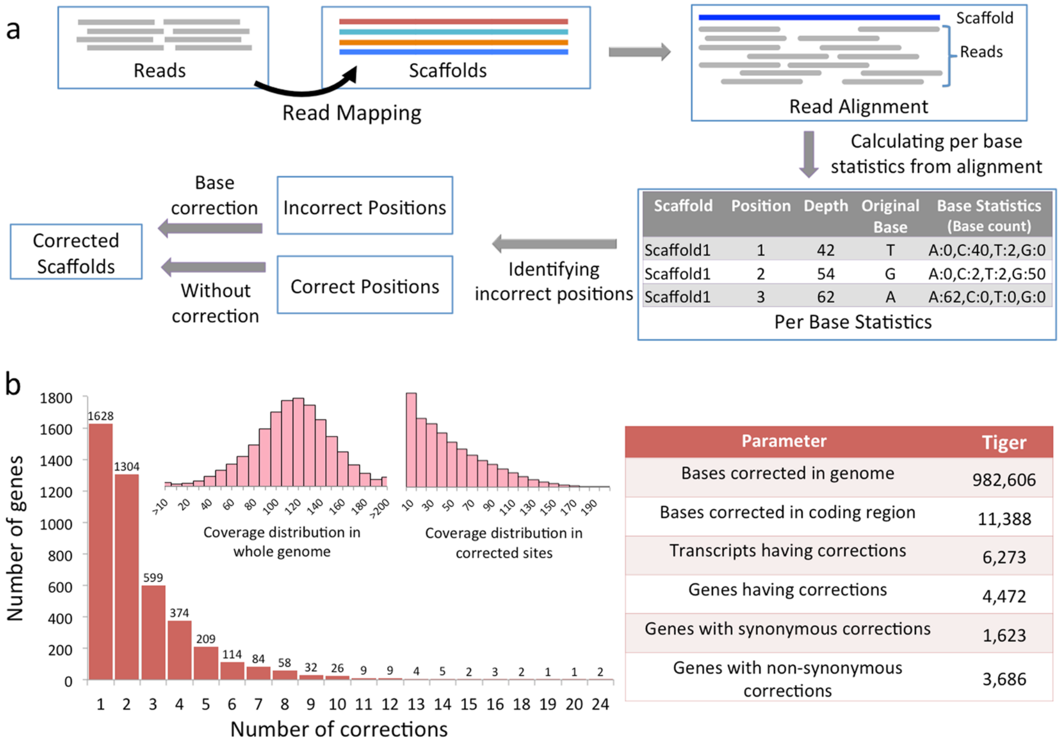
The availability of completed and draft genome assemblies of tiger, leopard, and other felids provides an opportunity to gain comparative insights on their unique evolutionary adaptations. However, genome-wide comparative analyses are susceptible to errors in genome sequences and thus require accurate genome assemblies for reliable evolutionary insights. In this study, while analyzing the tiger genome, we found almost one million erroneous substitutions in the coding and non-coding region of the genome affecting 4,472 genes, hence, biasing the current understanding of tiger evolution. Moreover, these errors produced several misleading observations in previous studies. Thus, to gain insights into the tiger evolution, we corrected the erroneous bases in the genome assembly and gene set of tiger using 'SeqBug' approach developed in this study. We sequenced the first Bengal tiger genome and transcriptome from India to validate these corrections. A comprehensive evolutionary analysis was performed using 10,920 orthologs from nine mammalian species including the corrected gene sets of tiger and leopard and using five different methods at three hierarchical levels, i.e. felids, *Panthera*, and tiger. The unique genetic changes in tiger revealed that the genes showing signatures of adaptation in tiger were enriched in development and neuronal functioning. Specifically, the genes belonging to the Notch signalling pathway, which is among the most conserved pathways involved in embryonic and neuronal development, were found to have significantly diverged in tiger in comparison to the other mammals. Our findings suggest the role of adaptive evolution in neuronal functions and development processes, which correlates well with the presence of exceptional traits such as sensory perception, strong neuro-muscular coordination, and hypercarnivorous behaviour in tiger.

The advancement in genomic sequencing technologies has provided a tremendous impetus for studying the molecular and genetic basis of adaptive evolution. A recent accomplishment is the genome sequencing of tiger, the largest felid and a model species to identify the molecular adaptations to hypercarnivory[1–3]. Tiger is a prominent member of the big cats, which are the topmost predators in the food chain, and play a key role in the ecological niche[4]. It is a solitary animal with extraordinary muscle strength and predatory capabilities[2,3]. The tiger genome sequencing revealed several molecular signatures of selection, particularly the rapid evolution in genes related to muscle strength, energy metabolism, and sensory nerves[1]. Similar studies in felids, including the tiger, had also shown strong positive selection in genes related to sensory perception and neurotransmitters[5].

While genome sequences are indispensable for comparative genome-wide evolutionary studies, quality of a genome is crucial for such analyses and in deriving reliable inferences[6–11]. The quality of a genome assembly is commonly assessed based on the N50 values of contigs and scaffolds and does not account for single nucleotide errors, which are mainly introduced by the read error correction tools or *de novo* assemblers[12–17]. Such sequence errors in genomes can produce drastically misleading results in comparative genomics and evolutionary studies[7,8,11]. We found a similar case in the tiger genome assembly reported by Cho *et al*. in 2013[1] and available at Ensembl release 94 (PanTig1.0)[18] and NCBI. The presence of several erroneous single nucleotide substitutions in the assembly bias the current understanding of the tiger evolution.

Therefore, to perform a comprehensive genome-wide analysis of tiger, we corrected the erroneous bases in the earlier-reported tiger genome assembly (PanTig1.0) using 'SeqBug' pipeline. We further validated the corrections by resequencing the genome and transcriptome of a male Bengal tiger. Using the corrected genome assembly and

[1]Metaomics and Systems Biology Group, Department of Biological Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal, India. [2]Computational Evolutionary Genomics Lab, Department of Biological Sciences, Indian Institute of Science Education and Research Bhopal, Bhopal, India. *email: vineetks@iiserb.ac.in



Parameter	Tiger
Bases corrected in genome	982,606
Bases corrected in coding region	11,388
Transcripts having corrections	6,273
Genes having corrections	4,472
Genes with synonymous corrections	1,623
Genes with non-synonymous corrections	3,686

Figure 1. Correction in the tiger genome. (**a**) Workflow of methods used to identify erroneous sites in the tiger genome assembly and their corrections. (**b**) The main bar plot represents the number of genes and the number of sites corrected in each gene. The left inset bar plot represents the distribution of coverage of each position in the tiger genome, and the right inset bar plot represents the distribution of coverage of only the corrected sites in the genome assembly. The table in the right represents the correction statistics for the tiger genome assembly.

gene set of tiger, we carried out a comparative genomic analysis of tiger with several other mammalian species, which provided novel insights into the adaptive evolution of the lineage leading to tiger.

Results

Comparative genomic analysis was performed to gain insights into the evolution of tiger with several other mammalian species. Tiger is a prominent member of the *Panthera* genus, which is a fast-evolving group that has undergone recent radiation with rapid functional diversification[19–22]. Thus, the comparative genome-wide study of tiger with respect to the closely related *Panthera* species and other mammals is likely to provide novel evolutionary insights into their adaptive evolution. The tiger genome assembly, reported by Cho *et al.*, 2013 (available at http://tigergenome.org), was retrieved and used for the comparative analysis. During the analysis, we observed that the assembly comprised of several erroneous single base substitutions, which were perhaps introduced by the *de novo* assembler or by the read correction tools (Supplementary Text S1 and Supplementary Table S1)[12–17]. Similar errors were also present in the genome assembly of tiger available at NCBI (GCA_000464555.1) and Ensembl (PanTig1.0). As observed from the analysis performed using the publicly available tiger genome assembly presented in the Supplementary Text S1, the above errors produced several misleading results (Supplementary Fig. S1,S2). For example, BEX3, a gene that plays an important role in the neuronal apoptosis[23,24], was found to be positively selected and showed nine unique amino acid substitutions in tiger. Our analysis revealed that eight of the nine substitutions in this gene were due to the single nucleotide errors in the publicly available genome assembly of tiger, which produced the incorrect result shown in Supplementary Fig. S2. We found several such cases in previous studies where the erroneous single nucleotide substitutions have led to incorrect evolutionary interpretations (Supplementary Text S2, Fig. S3,S4).

Therefore, to understand the adaptive evolution of tiger lineage, we first corrected the available reference assembly of the tiger genome, which was further validated by sequencing the genome and transcriptome of a Bengal tiger from India. A total of 175,680,850 paired-end reads (67 GB) and 32,252,904 reads (6 GB) were generated for the genome and transcriptome, respectively (Supplementary Table S2 and S3). Among the five extant species in the *Panthera* genus, the genome assemblies are publicly available for only two species, tiger and leopard[1,25]. Thus, in addition to tiger, we also generated the corrected genome assembly of leopard using the strategy shown in Fig. 1a, and briefly mentioned below.

Correcting the genome assembly and gene set of tiger and leopard.

The genomic reads of Amur tiger were mapped to the tiger genome assembly obtained from Ensembl (PanTig1.0), and the incorrect positions in the assembly were identified using the read alignments. The genome sequence data of the same Amur tiger individual, which was used to generate the tiger assembly (Cho *et al*., 2014), was used for mapping and subsequent correction in this study (Supplementary Table S4). We developed a pipeline, named 'SeqBug', for genome assembly corrections for single nucleotide errors introduced primarily by the read error correction or *de novo* assembler. The method is based upon the mapping of genomic reads to the genome assembly followed by identification of an incorrect base-pair and its correction. A given base-pair in the reference assembly (ref_base) was considered as an 'incorrect' base if its representation was less than or equal to one-fifth of the most frequent base (max_base), and subsequently, the ref_base was replaced by the max_base at that position. The criteria for base correction were optimized after several iterations (Supplementary Table S5) and provided in Methods. A total of 982,606 bases (0.04% of the genome) were corrected in tiger genome assembly (Supplementary Data Sheet 1). The distribution of per base coverage of these corrected positions showed a Poisson distribution (with a peak on the low coverage side) in comparison to the normal distribution for the whole genome (Fig. 1b), suggesting that a significant number of corrections were made in the low coverage regions.

The corrected genome assembly was used to construct the corrected gene set for tiger as per the gene structure information available at Ensembl release 94. A total of 14,145 codons corresponding to 6,273 transcripts and 4,472 genes were corrected. Of these, 3,686 genes (21%) had non-synonymous, and 1,623 genes (9.3%) had synonymous corrections (Supplementary Data Sheet 2). Not surprisingly, most of the corrected bases in the coding region of genes in tiger were found identical to the corresponding base present in the cat gene orthologs, which validates the correction methodology.

We also corrected the leopard genome assembly using the same approach in which a total of 58,566 bases (0.002% of the genome) were corrected. The corrections in the coding regions mapped to 194 codons in 165 transcripts corresponding to 125 genes, of which 40 genes had synonymous changes and 43 genes had non-synonymous changes. It is apparent that much fewer corrections were made in the leopard genome assembly in comparison to the tiger assembly. This was expected because the leopard genome was sequenced at a very high (~300 ×) coverage compared to tiger (~100 ×), and the mapping-based correction was previously performed by the authors[9,25]. Fewer corrections in the leopard genome assembly also indicate that the error correction method used in this study was specific enough to identify and correct only the erroneous positions. The corrected assemblies and gene sets of tiger and leopard were utilized to identify the molecular signatures underlying the adaptive evolution of tiger lineage.

Adaptive evolution analysis. We performed the adaptive evolution analysis for the lineage leading to tiger using five methods: A) Higher dN/dS using branch model: to identify genes with higher rate of evolution, B) High nucleotide substitution: to identify genes with a higher rate of mutation by comparing root-to-tip branch lengths, C) Positive selection using the branch-site model: to identify genes with positive selection in the selected clade, D) Unique substitution with functional impact: to identify genes with unique substitution in the selected clade that has significant impact on the protein function, and E) Positively selected amino acid sites: to identify the positively selected sites in a gene. The analysis was performed using nine mammalian species, including the corrected gene set of leopard and tiger, and the high-quality annotated gene sets of seven mammalian species (human, mouse, cow, horse, cat, ferret, dog) retrieved from Ensembl (release 94) (Supplementary Table S6)[18]. A total of 10,920 one-to-one orthologs for these nine species were identified using Ensembl BioMart[26]. The phylogenetic tree for these species was derived using the tree published by Nyakatura *et al.*[27], by employing the tree subset methodology from the "ape" package of R statistical software[28] (Fig. 2a). The adaptive evolution analysis for the lineage leading to tiger provided several new insights into the felid, *Panthera* and tiger evolution.

Evolutionary insights into lineage leading to tiger. Recent studies in felids have identified evolutionary signatures that are important for their unique sensory perception and hunting characteristics[1,5,19,25,29]. However, these studies were performed using the previous gene set (from Cho *et al*., 2013) of tiger, which consisted of erroneous base substitutions that can potentially bias the findings. Thus, the usage of corrected tiger gene set in this study is expected to identify the signatures of adaptive evolution in the lineage leading to tiger, i.e., felid, *Panthera* and tiger. A total of 766 genes showed faster evolution and 906 genes showed positive selection in felid in comparison to the other mammals. These genes showed enrichment for biological functions such as sensory perception, neuronal functioning, cell signalling, development, and stress response (Fig. 2b and Supplementary Table S7). Several genes that previously showed adaptive evolution in felids could not be identified in this study, whereas many additional genes were found to have evolved in felid (Supplementary Text S2). However, previous studies on the evolution of felids have also reported positive selection and adaptive evolution in the genes involved in the sensory perception and neuronal functioning[5]. This indicates that in terms of the broader biological processes, the results from the evolutionary analysis using the corrected genome assemblies corroborate with the previous study on felids[5].

We observed several felid-specific amino acid substitutions in the AgRP gene (Supplementary Figure S5) expressed in AGRP neurons, which is involved in regulating the feeding behaviour in animals[30–32]. The injection of AgRP peptides into the brain in rats was found to induce voracious eating behaviour even in well-fed mice. AgRP polymorphisms have been associated with diet, leanness, obesity, type-2 diabetes and anorexia nervosa[32–34]. The felid-specific unique substitutions in the AgRP gene were also found to have significant functional impact predicted using SIFT, and thus, could be associated with the voracious feeding behaviour shown by felids[35,36].

A total of 1,450 genes showing positive selection were identified in *Panthera*, which were functionally enriched in sensory perception, regulation of protein serine/threonine kinase activity, gene expression regulation, stress response, and development (Supplementary Table S8). A total of 917 genes showed a faster rate of evolution (branch model) and were enriched in cell-cell signalling and early development functions. Further, 797 genes showed amino acid substitutions unique to *Panthera* with significant functional impact. These genes were enriched in the biological functions related to sperm motility, development, fatty acid metabolism, and

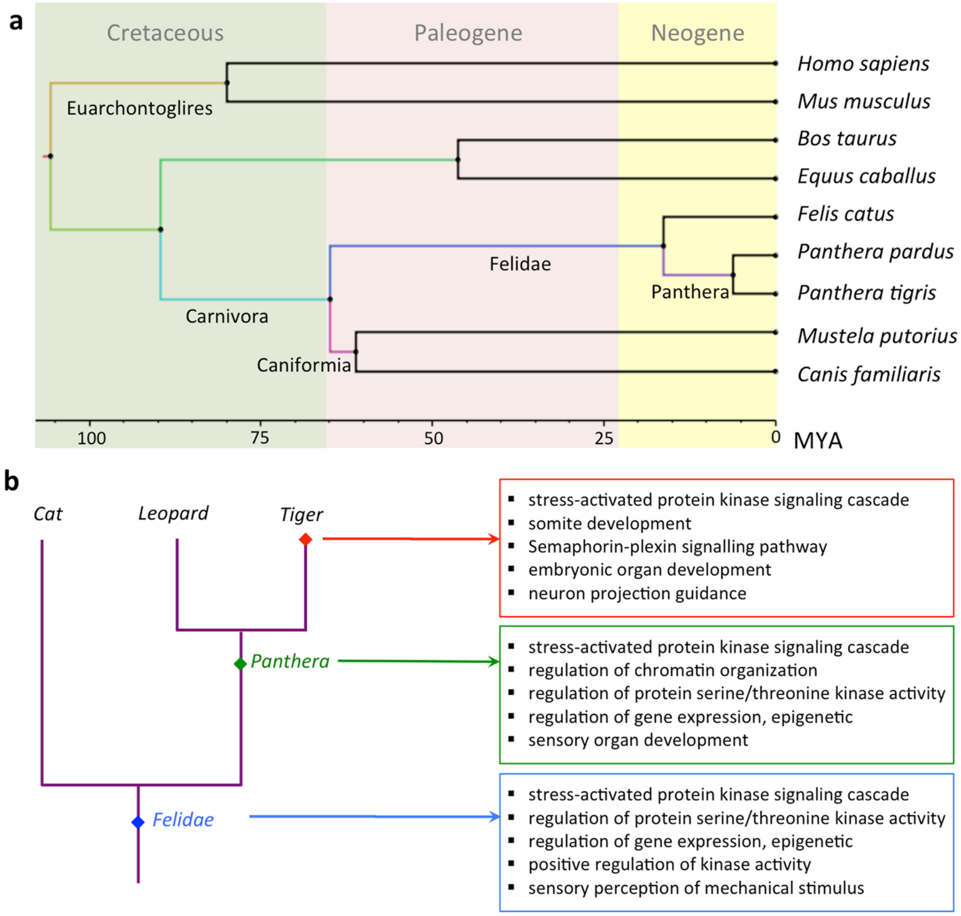

a
Cretaceous
Paleogene
Neogene
Homo sapiens
Mus musculus
Bos taurus
Equus caballus
Felis catus
Panthera pardus
Panthera tigris
Mustela putorius
Canis familiaris
Euarchontoglires
Felidae
Panthera
Carnivora
Caniformia
100
75
50
25
0
MYA
b
Cat
Leopard
Tiger
Panthera
Felidae
stress-activated protein kinase signaling cascade
somite development
Semaphorin-plexin signalling pathway
embryonic organ development
neuron projection guidance
stress-activated protein kinase signaling cascade
regulation of chromatin organization
regulation of protein serine/threonine kinase activity
regulation of gene expression, epigenetic
sensory organ development
stress-activated protein kinase signaling cascade
regulation of protein serine/threonine kinase activity
regulation of gene expression, epigenetic
positive regulation of kinase activity
sensory perception of mechanical stimulus

Figure 2. Phylogeny and positive selection in the lineage leading to tiger. (**a**) The phylogenetic tree of the nine mammalian species used in the study. (**b**) The top five enriched GO categories of positively selected genes identified in felid, *Panthera* and tiger lineages.

DNA repair. The genes showing positive selection in tiger (branch-site model), faster evolution in tiger (branch model), high nucleotide divergence rate and unique substitutions with functional impact were enriched for functional categories such as early development, fatty acid metabolism neuronal functioning and sensory perception (Supplementary Table S9-S12; Supplementary Text S3 for details).

Insights into the evolution of tiger using genes with multiple signs of adaptation. The genes with multiple signs of adaptation (MSA) were identified as the genes that showed three or more signs of adaptive evolution out of the five methods used for the adaptive evolution analysis (A. Higher dN/dS analysis using the branch model, B. Higher nucleotide substitution, C. Positive selection using the branch-site model, D. Unique substitution with functional impact, and E. Positively selected amino acid sites). A total of 955 genes showed MSA in tiger in comparison to all the other species, including the closely related leopard genome. A total of 83 genes showed all the five signs of adaptive evolution, and a maximum of 348 genes showed four signs of adaptive evolution including higher branch dN/dS, positive selection, unique substitution with functional impact, and positively selected sites.

Among the five signatures of adaptive evolution used in this study, the higher branch dN/dS, positive selection, and higher nucleotide divergence can identify the 'gene-wide' signals of evolution, suggesting that the complete gene is evolving. On the other hand, unique substitution with functional impact and positively selected sites indicate the evolution of only specific sites in the gene, thus can identify the 'site-specific' signals of evolution. Among the MSA, seven genes did not show positive selection (gene-wide signal) in tiger, though they had statistically significant positively selected amino acid sites (site-specific signal), suggesting that the sites evolving under purifying selection or neutrality masked the effect of these positively selected sites. Thus, the usage of five different evolutionary analyses helped to identify both site-specific and gene-wide signals of evolution in genes. Using these methods, among the MSA genes, a total of 111 genes showed all the three gene-wide signals of evolution, and a total of 580 genes showed the two site-specific signals of evolution in tiger (Fig. 3a).

Evolution of developmental and neuronal processes in tiger. The GO enrichment for the MSA genes was performed to identify the underlying biological processes of genes showing adaptive evolution, and the enriched categories (p-value < 0.01) were visualized as a network using Cytoscape v3.2.1[37]. The nodes in the network

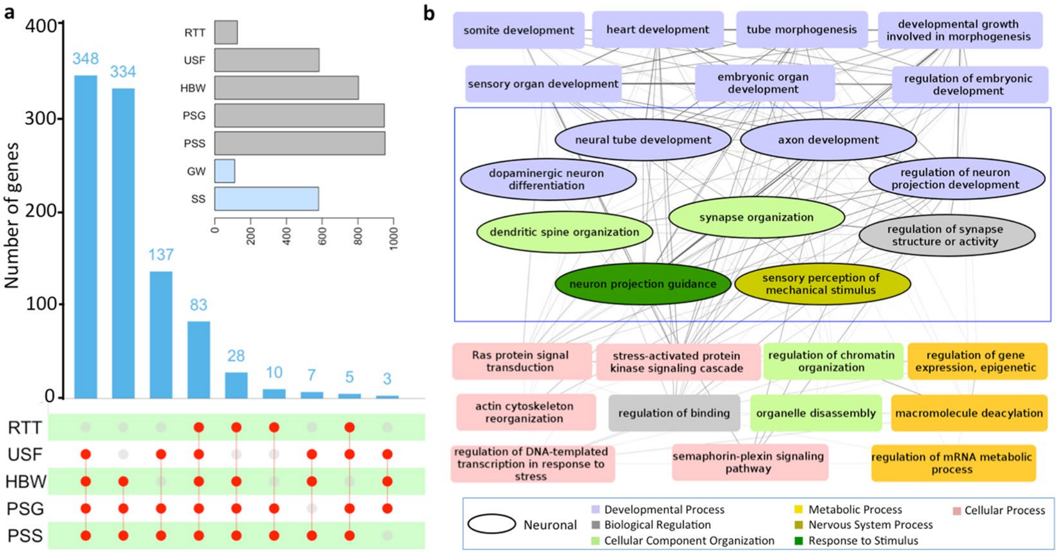


Figure 3. The genes showing multiple signatures of adaptation in tiger. (**a**) The upSet plot of the number of genes shared by the combination of the five methods used to test for adaptive evolution. The matrix layout was constructed using the upSET package in R[71]. The connection between the red circles shows the intersection of different methods with the intersection value depicted as a bar plot. RTT: higher root-to-tip branch length, USF: Unique substitution with functional impact, HBW: Higher branch dN/dS (ω), PSG: Positively selected genes, PSS: Positively selected sites, GW: Gene-wide, SS: Site-specific. Gene-wide (GW) represents the genes showing all three signs of adaptation among the MSA (HBW, RTT, PSG), which takes into account the evolution of the complete gene. Site-specific (SS) represents the genes showing the two signs of adaptation among the MSA (USF and PSS), which takes into account the evolution of specific sites in a gene. (**b**) Network diagram of GO biological processes enriched (p-value < 0.01) in the MSA genes. The nodes represent the GO biological processes, and the edges represent the number of MSA genes shared among the enriched categories.

represent the individual GO categories, and the width of edges represents the number of shared genes among the GO categories (Fig. 3b). It is interesting to note that one-third of the enriched categories were involved in neuronal functioning and development. The genes in these categories perform diverse functions such as regulation, cellular component organization, developmental process, nervous system process, and response to stimulus (Fig. 3b). It is also apparent from the network that several GO categories, including neuronal-related functions, belonged to a broader GO term "Developmental process". These GO categories showed dense connections with each other, which indicates that a large number of genes are common among these functional categories. This suggests that these developmental genes with adaptive divergence in tiger have pleiotropic functions, where one gene can regulate multiple developmental processes. Further, the eggNOG classification of the MSA genes revealed 'signal transduction mechanisms' as the most enriched category (Supplementary Table S13). Taken together, it points towards the differential evolution of neuronal functioning and developmental processes genes in tiger.

The highly evolved Notch signalling pathway in tigers. The pathway enrichment analysis performed using Fisher's exact test and network enrichment method revealed the Notch signalling pathway to be the most significantly enriched pathway in tiger (Supplementary Table S14). The regression of XD-score, which is a measure of network interconnectivity, and Fisher's test (with Benjamini-Hochberg adjusted q-value) also revealed that after applying these tests, only the Notch signalling pathway was above the significance threshold (Fig. 4a). This kind of framework for the pathway enrichment is more accurate than the classical overrepresentation-based method, as it also includes the protein interaction network information[38]. In the Notch signalling pathway, 11 genes showed adaptive evolution in tiger among which, CTBP1 gene showed all the five signs of adaptation. The 11 genes include the notch receptor (NOTCH3), ligand (DLL3), intracellular and extracellular regulators (DVL3, NUMB, LFNG, ADAM17), transcription factor (RBPJL), and its regulators (CREBBP, NCOR2, CTBP1). Protein-protein interaction data was obtained for the Notch signalling pathway genes using the STRING database[39], and a network diagram was constructed using Cytoscape v3.2.1[37] where the edges represent the physical interaction. From the network diagram, it was apparent that these 11 genes interact with all the genes and regulators of the Notch pathway (Fig. 4b,c). Thus, it is apparent that every crucial step of the Notch signalling pathway has evolved in tiger in comparison to the other mammalian species, including the close relative leopard. The genes of this pathway are evolutionarily conserved in multi-cellular organisms and regulate the cell-fate determination and tissue homeostasis, thus play an important role in embryonic development[40,41]. Using the juxtacrine signalling method, it regulates the development and functioning of cardiac, neuronal, immune, and endocrine system[41–43]. The tissue expression data from GNF Atlas[44] revealed that these 11 adaptively evolved Notch pathway genes also show high expression in the temporal lobe, whole brain, cerebellum peduncles, and prostate (Supplementary Table S15).

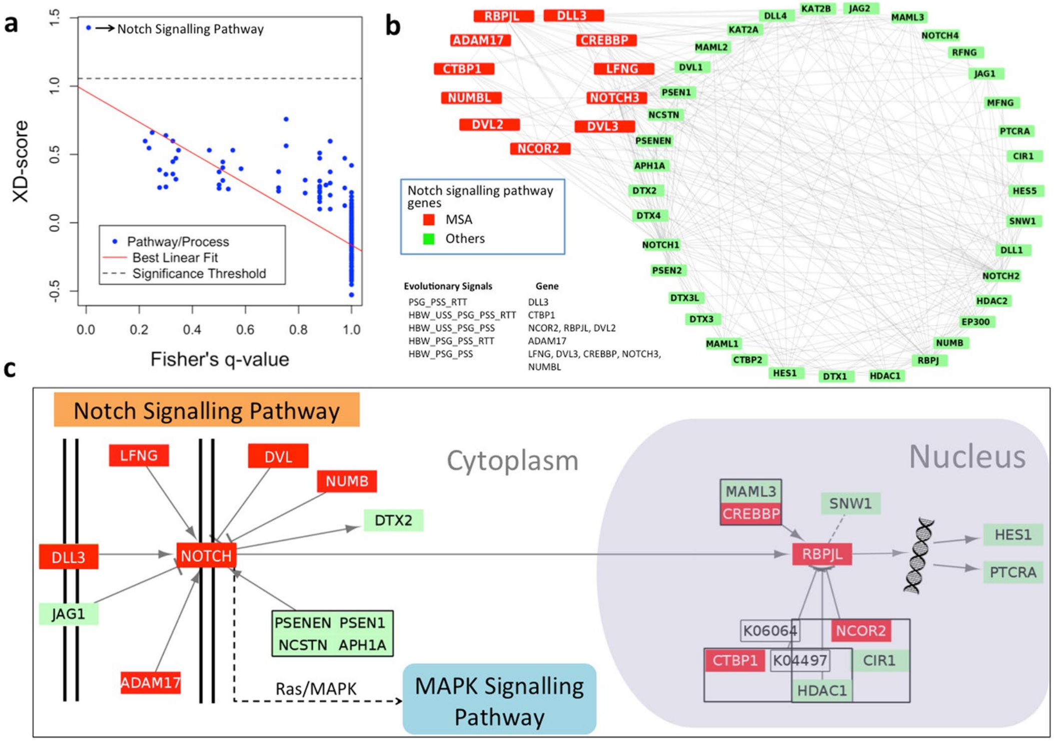


Figure 4. The adaptive divergence of Notch signalling pathway in tiger. (**a**) The regression of XD-score and Fisher's q-value of KEGG pathways. (**b**) Network diagram of genes showing the interaction of MSA genes with the rest of the genes of the Notch signalling pathway. The nodes represent individual genes of the NOTCH pathway in which MSA genes are shown in red and other genes of the pathway are shown in green. The edges in the network represent the protein-protein interactions among the genes obtained from STRING database. (**c**) Schematic representation of the Notch signalling pathway from KEGG with the genes showing multiple signatures of adaptation highlighted in Red. The pathway diagram was constructed using KEGGscape plug-in[72] in Cytoscape.

Thus, the evolution of Notch signalling relates well with the differential neural morphology observed in tiger in comparison to the other mammals[45,46].

Discussion

Genome sequencing followed by genome-wide comparative analysis has become a powerful tool to study the patterns of evolution in different lineages. The genome sequencing of tiger has provided novel insights into their unique adaptations and divergence from other species, and among its subspecies. The genome sequencing of tiger is significant since it is a part of charismatic megafauna that has captivated human interest, is the largest felid and is among one of the most endangered species with less than 4,000 individuals remaining in the wild[47].

In this study, while using the publicly available genome sequence assembly of tiger, we found that the assembly consisted of several errors, which also led to several incorrect interpretations in recent other studies. For example, Figueiró *et al.*, 2017 identified that the ESRP1 gene, important for craniofacial robustness, and has a positively selected I298Y substitution in jaguar[19]. This substitution was found positively selected in jaguar due to the presence of "I" in the respective ortholog in tiger, which was a result of single nucleotide error in the tiger genome assembly (Supplementary Fig. S3 and S4). From the above example and the other cases described in the Supplementary Text, it is apparent that evolutionary studies are susceptible to nucleotide errors present in the genome assemblies and gene sets, where even single nucleotide errors can produce drastically misleading results in the analyses.

Thus, to understand the adaptive evolution of the lineage leading to tiger, the publicly available tiger genome assembly was corrected for such single nucleotide errors using the 'SeqBug' pipeline developed in this study. The corrections were validated using the resequencing data of a new male Bengal tiger genome and transcriptome. Using this approach, 95% of the corrected sites could be verified using the Bengal tiger sequence data and thus, validating the 'SeqBug' approach. Since 5% of the corrected bases could not be validated, it represents the upper limit of over-correction (or Type-1 error) associated with the correction methodology. The 'SeqBug' method is prone to Type-2 error primarily at the sites with low coverage depth in the genome assembly since the approach uses a minimum coverage cut-off of 10 ×, which however can be minimized with the usage of high depth sequencing data.

The identification of errors in 4,472 genes and 982,606 bases (0.04% genome) in the tiger genome put forth the need for correction. Further, the incorrect positions were mostly present in regions of low coverage (<30) and were much fewer in the leopard genome that was sequenced at three times higher coverage than tiger[9,25]. These observations underscore the need for higher genome coverage along with mapping-based correction to produce a more accurate assembly. The estimated genome size of tiger is 2.44 Gb (Ensembl browser 94), and the reported corrected assembly of tiger is 2.33 Gb, which suggests that the assembly is 95% complete (N50 value of 8.8 Mb). The genome sequence of another tiger individual sequenced in this study and the construction of corrected genome sequence and gene set of tiger are likely to be beneficial for further comparative studies.

After correction, most of the bases corrected (89.3%) in the coding genome of tiger were identical to the corresponding bases in the cat genome, thus, validating our correction methodology. This further indicates that the divergence time of tiger calculated using the genetic differences in the previous studies could suffer from an over-estimation because of these erroneous substitutions[8]. Considering errors of 0.9 million bases and a mutation rate of 1.1e-09 per base per year for tiger[1], the estimated divergence time of tiger can be affected by 0.37 million years.

The usage of corrected tiger and leopard coding genome, a large number of orthologs, and five different evolutionary analyses in this study made the evolutionary assessments more reliable and was also successful in revealing the signatures of adaptive divergence in felids, *Panthera* and tiger lineages. Previous reports identified evolutionary adaptations in genes related to muscle strength, hypercarnivorous diet, sensory perception, and craniofacial and limb development in the *Panthera*/Felidae lineage[1,5,19,25]. Similar categories were also found as adaptively evolved in the *Panthera*/Felidae lineage in this study. However, large differences in the gene sets were observed, which further highlights the impact of incorrect genomes on the evolutionary analysis.

One of the unique findings was the enrichment of neuronal functioning and developmental processes in genes showing multiple signs of adaptive evolution in tiger. Notably, the Notch signalling pathway emerged as the most diverged pathway in tiger, which was not found as adaptively evolved in the previous studies. The observation is significant since the Notch pathway plays key roles in diverse developmental processes, including neurogenesis, neural differentiation, and cell fate determination[41,48]. Also, the observed divergence at almost every step of the Notch signalling pathway, which is a highly conserved pathway throughout the animal kingdom, further, indicates the adaptive evolution in neuronal functioning and development genes in tiger.

The evolution of the neuronal related genes in the tiger lineage is informative but not very surprising given their unique physiological and behavioural characteristics[2,3,36,49–51]. Several studies show that the feeding, drinking, aggression, predation and sexual behaviour, and the energy homeostasis of an organism are primarily governed by neuronal circuitry[30,31,52–54]. The felids, particularly the big cats being the large hypercarnivores, show a very distinct aggressive and predatory behaviour. They require strong neuro-muscular coordination, sensory perception and timed actions for successful hunting[36,55]. Thus, it is tempting to speculate the role of evolution in the neural development and processes for attaining unique phenotypes, behaviour, and dietary patterns. This notion also gets support from the previous studies that showed differences in neuronal morphology in tiger in comparison to other mammals, including its closest relative leopard[45,46]. The dendrites of typical pyramidal neurons in tiger are very complex, and the dendritic measures of these neurons are disproportionally large relative to body/brain size[46]. To summarize, the identification of adaptive evolution in the neuronal functioning genes in tiger indicates the plausible role of evolution in neural processes in achieving exceptional sensory perception, neuro-muscular coordination, faster reflex actions, predatory capabilities and hypercarnivorous behaviour in tiger.

Methods

Sample collection and sequencing of the Bengal tiger genome and transcriptome.

Blood sample was obtained from a four years old male tiger at Van Vihar National Park, Bhopal, India. This study was performed in accordance with the approval of Institute Ethics committee of Indian Institute of Science Education and Research (IISER) Bhopal, and the permission to collect blood sample of Bengal tiger was obtained from Principal Chief Conservator of Forests (PCCF) Bhopal. The genomic DNA was extracted using DNeasy Blood and Tissue Kit (Qiagen, USA) following the manufacturer's protocol. Multiple shotgun genomic libraries were prepared using Illumina TruSeq DNA PCR-free library preparation kit and Nextera XT sample preparation kit (Illumina Inc., USA). The insert size for the TruSeq libraries was 350 and 550 bp, and the average insert size for Nextera XT libraries was ~650 bp. The normalised TruSeq 550 bp and Nextera XT libraries were loaded on Illumina NextSeq. 500 platform using NextSeq 500/550 v2 sequencing reagent kit (Illumina Inc., USA) and 150 bp paired-end sequencing was performed. The TruSeq libraries of 350 bp were sequenced on Illumina HiSeq platform to generate 250 bp paired-end reads. Total RNA extraction was carried out from the blood sample for transcriptomic analysis. The transcriptomic libraries were prepared from the total RNA using the SMARTer universal low input RNA kit and TruSeq RNA sample prep kit v2 using the manufacturer's instructions, and 100 bp paired end sequencing was performed on the Illumina HiSeq platform (Supplementary Text S4 for details).

Data download and preparation.

For assembly correction, the latest assemblies of tiger (*Panthera tigris altaica*) and leopard (*Panthera pardus*) genome were retrieved from Ensembl release 94 (PanTig1.0 and PanPar1.0)[18]. The reads data was retrieved from NCBI SRA with Accession SRX272981, SRX272988, SRX272991, SRX272997, SRX273000, SRX273020 and SRX273023 for Amur tiger. For leopard, the reads data with SRA Accession SRX1495683, SRX1495735, and SRX1495737 were retrieved from NCBI SRA. To construct the corrected CDS, the reference gtf was downloaded from Ensembl release 94 (Panthera pardus 1.0.93 and Panthera tigris altaica 1.0.93)[18]. The retrieved raw reads of Amur tiger were mapped to the tiger reference assembly and raw reads of leopard were mapped to the leopard genome assembly using bwa mem (v0.7.12)[56] using default parameters. The genomic reads of the same individual, which was used for assembly, were aligned with the reference assembly to reduce the type-I error arising due to SNPs. The alignment file was sorted and split scaffold-wise using Samtools (v1.4)[57].

Genomic and CDS correction. The per-nucleotide metrics for each scaffold was calculated using bam-readcount tool (github/genome/bam-readcount) using minimum mapping quality 25, minimum base quality 25, and maximum depth 400. The base positions with less than 10x coverage or more than 200x coverage, or percent indel $>10\%$ were filtered out to remove the low coverage, potentially repetitive and indel regions, respectively, and the remaining bases were analyzed further. For each position, the base with highest frequency (max_base) was identified. At a given position, if the representation of the base in the reference assembly (ref_base) was less than or equal to one-fifth of the most frequent base (max_base), the ref_base was considered as an error and subsequently the ref_base was replaced by the max_base at that position. The adoption of this stringent criteria ensured that only those positions were corrected where a sufficient coverage of the max_base relative to the ref_base was available to justify the replacement of the ref_base. The above criteria were optimized after several iterations and visualization of randomly selected regions with the mapped reads in the IGV software (Supplementary Table S6)[58]. The corrected genome assembly was used to construct the corrected gene set using the gene structure information available at Ensembl.

To validate, the genomic data of newly sequenced Bengal tiger was aligned with the corrected tiger genome assembly using bwa mem. Similarly, the transcriptome reads of Bengal tiger were aligned against the corrected gene sets of tiger using bwa. Using bam-readcount tool, per base statistics were obtained for each position of the aligned reads. For each corrected site, the per base statistics at that position were extracted. The site was validated if the most frequent base was identical to the corrected base.

Orthologous gene set construction. An orthologous gene set was constructed using nine species – *Homo sapiens* (human), *Mus musculus* (mouse), *Bos taurus* (cow), *Equus caballus* (horse), *Canis familiaris* (dog), *Mustela putorius* (ferret), *Felis catus* (cat), *Panthera tigris altaica* (tiger) and *Panthera pardus* (leopard). The gene sets were retrieved from Ensembl release 94 (Cat: Felis_catus_9.0, Cow: UMD3.1, Dog: CanFam3.1, Horse: EquCab2, Human: GRCh38, Ferret: MusPutFur1.0, Mouse: Mus_musculus.GRCm38)[18]. The corrected gene sets of tiger and leopard were used in the analysis. Information on one-to-one orthologs for the above species was retrieved from BioMart (Ensembl browser 94)[26].

Protein and nucleotide alignment. The one-to-one orthologs were filtered for the presence of premature stop codons (non-sense mutations). The gene phylogeny of each ortholog was inferred from the species phylogeny and was subjected to protein alignment using SATé-II[59], which implemented PRANK for alignment, Muscle for merging the alignment, and RAxML for tree estimation. The protein-based nucleotide alignment was carried out using 'tranalign' tool in EMBOSS package[60].

Evolutionary analysis. *Higher branch dN/dS.* The variation in ω ratio between lineages on individual genes was calculated using the branch model in CodeML from the PAML software package (v4.9a)[61]. The codons with any ambiguity site were removed from the analyses. The genes that qualified likelihood ratio test using a conservative 5% false-discovery-rate criterion against the null model (One ratio) were considered for further analysis. Also, the genes with dN/dS values >3 were not used for further analysis[62,63]. The genes having a higher branch dN/dS values for foreground lineage compared to the background lineage were considered to show divergence (HBW: higher branch omega).

Positively selected genes and sites. To identify positively selected genes, a branch-site model was used in PAML software package (v4.9a)[61]. The codons with any ambiguity site among the nine species were removed from the analyses. The genes that qualified the likelihood ratio test against the null model (fixed omega) with 5% false-discovery-rate were considered as positively selected genes (PSG). The sites with greater than 0.95 probability value for foreground lineages in Bayes Empirical Bayes analysis were considered as positively selected sites (PSS).

Unique substitutions and functional impact. Unique substitutions in amino acid in tiger were identified using the aligned protein sequences. The positions identical in all species but different in tiger were considered as a unique substitution in tiger. Any gap or unknown position was ignored. Five sites around any gap in the protein alignment were also ignored from the analysis. Unique substitutions in *Panthera* and felids were identified using the same approach. Functional impact of the substitutions were identified using Sorting Intolerant From Tolerant (SIFT)[64] tool and the UniProt database[65] was used for reference.

Higher nucleotide divergence. The maximum likelihood phylogenetic tree for each gene using its CDS alignments was constructed using PhyML package v3.1[66]. The root-to-tip branch length distances were calculated for each species using the 'adephylo' package in R[67,68]. The genes with a significantly higher root-to-tip branch length for lineage leading to tiger compared to all other lineages were considered to show higher nucleotide divergence in tiger.

Identification of genes with multiple signs of adaptation. Genes showing more than two signs of adaptive divergence among the five signs (Unique substitution, higher dN/dS, positive selection, positively selected sites, and higher nucleotide divergence) used in the study were considered to be the genes with multiple signs of adaptation (MSA). Enrichment of MSA genes was carried out using WebGeStalt web server[69]. The GO enrichment with p-value < 0.05 in over-representation enrichment analysis were considered to be enriched. The eggNOG analysis of the MSA genes was performed using the eggNOG v4.5.1[70]. The network-based pathway enrichment analysis was carried out based on the methodology implemented in EnrichNet[38] to identify the network interconnectivity score (XD-score) and classical overlap-based enrichment score (Fisher's exact test adj. using

Benjamini-Hochberg) using KEGG as the reference database. The significance threshold was calculated by performing a linear regression of network interconnectivity score (XD-score) and enrichment score (Fisher's q-value). The pathways above the significance threshold were considered as enriched.

Data availability

The sequence data of the Bengal tiger genome and transcriptome is available on NCBI SRA with Accessions-SRR8617820, SRR8617821, SRR8617819, SRR8617818, and SRR8617817. The corrected genome assemblies, and gene sets are available at http://metagenomics.iiserb.ac.in/tiger/data.php, and correction pipeline 'SeqBug' developed in this study are available at https://github.com/MetabiosysLab/seqbug.

Received: 14 March 2019; Accepted: 14 November 2019;
Published online: 05 December 2019

References

1. Cho, Y. S. *et al*. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat Commun* **4**, 2433 (2013).
2. Mazak, V. Panthera tigris. *Mammalian species*, 1–8 (1981).
3. Seidensticker, J. & McDougal, C. Tiger predatory behaviour, ecology and conservation. (1993).
4. Terborgh, J. *et al*. The role of top carnivores in regulating terrestrial ecosystems. (1999).
5. Montague, M. J. *et al*. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc Natl Acad Sci USA* **111**, 17230–17235 (2014).
6. Denton, J. F. *et al*. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol* **10**, e1003998 (2014).
7. Prosdocimi, F., Linard, B., Pontarotti, P., Poch, O. & Thompson, J. D. Controversies in modern evolutionary biology: the imperative for error detection and quality control. *BMC Genomics* **13**, 5 (2012).
8. Clark, A. G. & Whittam, T. S. Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution* **9**, 744–752 (1992).
9. Hubisz, M. J., Lin, M. F., Kellis, M. & Siepel, A. Error and error mitigation in low-coverage genome assemblies. *PloS one* **6**, e17034 (2011).
10. Hoeppner, M. P. *et al*. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One* **9**, e91172 (2014).
11. Hughes, A. L., Friedman, R. & Murray, M. Genomewide pattern of synonymous nucleotide substitution in two complete genomes of Mycobacterium tuberculosis. *Emerg Infect Dis* **8**, 1342–1346 (2002).
12. Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y. & Fostier, J. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. *BMC Bioinformatics* **18**, 374 (2017).
13. Yang, X., Chockalingam, S. P. & Aluru, S. A survey of error-correction methods for next-generation sequencing. *Brief Bioinform* **14**, 56–66 (2013).
14. Salzberg, S. L. *et al*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* **22**, 557–567 (2012).
15. Del Angel, V. D. *et al*. Ten steps to get started in Genome Assembly and Annotation. *F1000Research* **7** (2018).
16. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one* **9**, e112963 (2014).
17. Ronen, R., Boucher, C., Chitsaz, H. & Pevzner, P. SEQuel: improving the accuracy of genome assemblies. *Bioinformatics* **28**, i188–i196 (2012).
18. Cunningham, F. *et al*. Ensembl 2019. *Nucleic acids research* **47**, D745–D751 (2018).
19. Figueiro, H. V. *et al*. Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci Adv* **3**, e1700299 (2017).
20. Johnson, W. E. *et al*. The late Miocene radiation of modern Felidae: a genetic assessment. *Science* **311**, 73–77 (2006).
21. Li, G., Davis, B. W., Eizirik, E. & Murphy, W. J. Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome research* **26**, 1–11 (2016).
22. Davis, B. W., Li, G. & Murphy, W. J. Supermatrix and species tree methods resolve phylogenetic relationships within the big cats, Panthera (Carnivora: Felidae). *Molecular Phylogenetics and Evolution* **56**, 64–76 (2010).
23. Calvo, L. *et al*. Bex3 Dimerization Regulates NGF-Dependent Neuronal Survival and Differentiation by Enhancing trkA Gene Transcription. *J Neurosci* **35**, 7190–7202 (2015).
24. Yi, J. S., Lee, S. K., Sato, T. A. & Koh, J. Y. Co-induction of p75(NTR) and the associated death executor NADE in degenerating hippocampal neurons after kainate-induced seizures in the rat. *Neurosci Lett* **347**, 126–130 (2003).
25. Kim, S. *et al*. Comparison of carnivore, omnivore, and herbivore mammalian genomes with a new leopard assembly. *Genome Biol* **17**, 211 (2016).
26. Kasprzyk, A. BioMart: driving a paradigm change in biological data management. *Database (Oxford)* **2011**, bar049 (2011).
27. Nyakatura, K. & Bininda-Emonds, O. R. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. *BMC Biol* **10**, 12 (2012).
28. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, (289–290 (2004).
29. Dobrynin, P. *et al*. Genomic legacy of the African cheetah, Acinonyx jubatus. *Genome Biol* **16**, 277 (2015).
30. Sternson, S. M. Hypothalamic survival circuits: blueprints for purposive behaviors. *Neuron* **77**, 810–824 (2013).
31. Sternson, S. M. & Eiselt, A.-K. Three pillars for the neural control of appetite. *Annual review of physiology* **79**, 401–423 (2017).
32. Ilnytska, O. & Argyropoulos, G. The role of the Agouti-Related Protein in energy balance regulation. *Cell Mol Life Sci* **65**, 2721–2731 (2008).
33. Vink, T. *et al*. Association between an agouti-related protein gene polymorphism and anorexia nervosa. *Mol Psychiatry* **6**, 325–328 (2001).
34. Marks, D. L. *et al*. Ala67Thr polymorphism in the Agouti-related peptide gene is associated with inherited leanness in humans. *Am J Med Genet A* **126A**, 267–271 (2004).
35. Van Valkenburgh, B. & Wayne, R. K. Carnivores. *Curr Biol* **20**, R915–919 (2010).
36. Sunquist, M. & Sunquist, F. Wild cats of the world. (University of chicago press, 2017).
37. Su, G., Morris, J. H., Demchak, B. & Bader, G. D. Biological network exploration with Cytoscape 3. *Curr Protoc Bioinformatics* **47**(8), 13 11–24 (2014).
38. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* **28**, i451–i457 (2012).
39. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res* **28**, 3442–3444 (2000).

40. Artavanis-Tsakonas, S., Rand, M. D. & Lake, R. J. Notch signaling: cell fate control and signal integration in development. *Science* **284**, 770–776 (1999).
41. Lowell, S., Benchoua, A., Heavey, B. & Smith, A. G. Notch promotes neural lineage entry by pluripotent embryonic stem cells. *PLoS Biol* **4**, e121 (2006).
42. Chau, M. D., Tuft, R., Fogarty, K. & Bao, Z. Z. Notch signaling plays a key role in cardiac cell differentiation. *Mech Dev* **123**, 626–640 (2006).
43. de la Pompa, J. L. *et al*. Conservation of the Notch signalling pathway in mammalian neurogenesis. *Development* **124**, 1139–1148 (1997).
44. Su, A. I. *et al*. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101**, 6062–6067 (2004).
45. Jacobs, B. *et al*. Comparative morphology of gigantopyramidal neurons in primary motor cortex across mammals. *Journal of Comparative Neurology* **526**, 496–536 (2018).
46. Johnson, C. B. *et al*. Neocortical neuronal morphology in the Siberian Tiger (Panthera tigris altaica) and the clouded leopard (Neofelis nebulosa). *J Comp Neurol* **524**, 3641–3665 (2016).
47. Seidensticker, J. Saving wild tigers: a case study in biodiversity loss and challenges to be met for recovery beyond 2010. *Integr Zool* **5**, 285–299 (2010).
48. Wakeham, A. *et al*. Conservation of the Notch signalling pathway in mammalian neurogenesis. *Development* **124**, 1139–1148 (1997).
49. Sunquist, M. Ecology, behaviour and resilience of the tiger and its conservation needs. *Riding the tiger: tiger conservation in human dominated landscapes*, 5–18 (1999).
50. Valkenburgh, B. V. & Ruff, C. Canine tooth strength and killing behaviour in large carnivores. *Journal of Zoology* **212**, 379–397 (1987).
51. Eisert, R. Hypercarnivory and the brain: protein requirements of cats reconsidered. *J Comp Physiol B* **181**, 1–17 (2011).
52. Panchin, Y. V. *et al*. Neuronal basis of hunting and feeding behaviour in the pteropod mollusc Clione limacina. *Netherlands journal of zoology* **44**, 170–183 (1993).
53. Han, W. *et al*. Integrated Control of Predatory Hunting by the Central Nucleus of the Amygdala. *Cell* **168**, 311–324 e318 (2017).
54. Jennings, J. H. *et al*. Interacting neural ensembles in orbitofrontal cortex for social and feeding behaviour. *Nature* **565**, 645 (2019).
55. Heffner, R. S. & Heffner, H. E. Hearing range of the domestic cat. *Hear Res* **19**, 85–88 (1985).
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
57. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
58. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192 (2013).
59. Liu, K. *et al*. SATe-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic biology* **61**, 90–106 (2011).
60. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends in genetics* **16**, 276–277 (2000).
61. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586–1591 (2007).
62. Uebbing, S. *et al*. Divergence in gene expression within and between two closely related flycatcher species. *Molecular ecology* **25**, 2015–2028 (2016).
63. Axelsson, E. *et al*. Natural selection in avian protein-coding genes expressed in brain. *Mol Ecol* **17**, 3008–3017 (2008).
64. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812–3814 (2003).
65. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699 (2018).
66. Guindon, S. *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321 (2010).
67. Jombart, T., Dray, S. & Dray, M. S. Package 'adephylo' (2017).
68. Jombart, T. & Dray, S. adephylo: exploratory analyses for the phylogenetic comparative method. *Bioinformatics* **26**, 1–21 (2010).
69. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research* **45**, W130–W137 (2017).
70. Huerta-Cepas, J. *et al*. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic acids research* **44**, D286–D293 (2015).
71. Lex, A. & Gehlenborg, N. Points of view: Sets and intersections. *Nature Methods* **11**, 779 (2014).
72. Nishida, K., Ono, K., Kanaya, S. & Takahashi, K. KEGGscape: a Cytoscape app for pathway data integration. *F1000Research* **3** (2014).

Acknowledgements

We thank Dr. Atul Gupta, Wildlife Veterinary Officer, Van Vihar National Park, Bhopal, and Director, Van Vihar National Park, Bhopal, India for providing the blood samples of tiger. We also acknowledge the help of Drs. Tista Joseph and Niraj Dahe, Wildlife Veterinary Officers (Wildlife SOS India) at Van Vihar National Park for carrying out the sample collection procedure. We thank the NGS facility at IISER Bhopal. SKJ and RS thank the Department of Science and Technology for the DST-INSPIRE fellowship. We also thank the intramural research funds provided by IISER Bhopal.

Author contributions

V.K.S. conceived and coordinated the project. R.S. prepared the DNA samples and performed genome sequencing. P.M. and S.K.J. generated the corrected genome assemblies and gene sets. P.M. performed the branch dN/dS, positive selection, unique substitution and SIFT analyses. S.K.J. performed the higher nucleotide divergence analysis. P.M., S.K.J., N.V. and V.K.S. analyzed the data and wrote the manuscript. P.M. created the manuscript figures. All the authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-54838-z.

Correspondence and requests for materials should be addressed to V.K.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

© The Author(s) 2019