Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Pangolin genomes and the evolution of mammalian scales and immunity

Siew Woh Choo, Mike Rayko, Tze King Tan, et al.

Genome Res. published online August 10, 2016
Access the most recent version at doi:10.1101/gr.203521.115

P<P Published online August 10, 2016 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Creative Commons License This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Pangolin genomes and the evolution of mammalian scales and immunity

The International Pangolin Research Consortium (IPaRC)

Siew Woh Choo[1,10,14*ϕ], Mike Rayko[2*], Tze King Tan[1,10], Ranjeev Hari[1,10], Aleksey Komissarov[2], Wei Yee Wee[1,10], Andrey Yurchenko[2], Sergey Kliver[2], Gaik Tamazian[2], Agostinho Antunes[16,17], Richard K. Wilson[3], Wesley C. Warren[3], Klaus-Peter Koepfli[15], Patrick Minx[3], Ksenia Krasheninnikova[2], Antoinette Kotze[20,21], Desire L. Dalton[20,21], Elaine Vermaak[20], Ian C. Paterson[10,18], Pavel Dobrynin[2], Frankie Thomas Sitam[11], Jeffrine Rovie Ryan Japning[11], Warren E. Johnson[15], Aini Mohamed Yusoff[1,10], Shu-Jin Luo[9], Kayal Vizi Karuppannan[11], Gang Fang[19], Deyou Zheng[8], Mark B. Gerstein[5,6,7], Leonard Lipovich[4,12], Stephen J. O'Brien [2,13ϕ] and Guat Jah Wong[1]

[1]Genome Informatics Research Laboratory, High Impact Research (HIR) Building, University of Malaya, 50603 Kuala Lumpur, Malaysia.
[2]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, Russia.
[3]McDonnell Genome Institute, Washington University, St Louis, MO 63108, USA.
[4]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA.
[5]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA
[6]Department of Molecular Biophysics and Biochemistry, P.O. Box 208114, Yale University, New Haven, CT 06520, USA.
[7]Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA.
[8]Department of Neurology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA.
[9]Peking-Tsinghua Center for Life Sciences, College of Life Sciences, Peking University, Beijing 100871, China.
[10]Department of Oral and Craniofacial Sciences, Faculty of Dentistry, University of Malaya, 50603 Kuala Lumpur, Malaysia
[11]Ex-Situ Conservation Division, Department of Wildlife and National Parks, 10 Jalan Cheras, 56100 Kuala Lumpur, Malaysia.
[12]Department of Neurology, School of Medicine, Wayne State University, Detroit, MI 48201, USA.
[13]Oceanographic Center, Nova Southeastern University, Ft Lauderdale, Florida 33004, USA.
[14]Genome Solutions Sdn Bhd, Suite 8, Innovation Incubator UM, Level 5, Research Management & Innovation Complex, University of Malaya, 50603 Kuala Lumpur, Malaysia.
[15]National Zoological Park, Smithsonian Conservation Biology Institute, Washington, DC 20008, USA
[16]CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Rua dos Bragas, 177, 4050-123 Porto, Portugal.
[17]Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal.
[18]Oral Cancer Research and Coordinating Centre, Faculty of Dentistry, University of Malaya, 50603 Kuala Lumpur, Malaysia.
[19]NYU Shanghai, 1555 Century Ave, Pudong, Shanghai, China 200122.
[20]National Zoological Gardens of South Africa, P.O. Box 754, Pretoria 0001, South Africa.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

[21]Department of Genetics, University of the Free State, P.O. Box 339, Bloemfontein, 9300, South Africa

*=These authors contributed equally to this work.
ϕ=Corresponding authors:
Siew Woh Choo
l.choo@genomesolutions.com.my

Stephen J. O'Brien
lgdchief@gmail.com

Keyword

Malayan pangolin, Chinese pangolin, *IFNE*, *Manis javanica*, *Manis pentadactyla*, and genome

Abstract

Pangolins, unique mammals with scales over most of their body, no teeth, poor vision, and an acute olfactory system, comprise the only placental order (Pholidota) without a whole-genome map. To investigate pangolin biology and evolution, we developed genome assemblies of the Malayan (*Manis javanica*) and Chinese (*M. pentadactyla*) pangolins. Strikingly, we found that interferon epsilon (*IFNE*), exclusively expressed in epithelial cells and important in skin and mucosal immunity, is pseudogenised in all African and Asian pangolin species that we examined, perhaps impacting resistance to infection. We propose that scale development was an innovation that provided protection against injuries or stress and reduced pangolin vulnerability to infection. Further evidence of specialized adaptations was evident from positively-selected genes involving immunity-related pathways, inflammation, energy storage and metabolism, muscular and nervous systems, and scale/hair development. Olfactory receptor gene families are significantly expanded in pangolins, reflecting their well-developed olfaction system. This study provides insights into mammalian adaptation and functional diversification, new research tools and

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

questions, and perhaps a new natural *IFNE*-deficient animal model for studying mammalian immunity.

Introduction

Pangolins (also known as scaly anteaters) are mammals of the order Pholidota. Their name is derived from the Malay word "pengguling", meaning "something that rolls up". Eight pangolin species are recognised, four from Asia (*Manis javanica, M. pentadactyla, M. crassicaudata* and *M. culionensis)* and four from Africa (*M. tricuspis, M. tetradactyla, M. gigantea* and *M. temminckii*) (Timothy J. Gaudin 2009). Although pangolin morphology shares some analogies with South-American anteaters and armadillos (superorder Xenarthra), they are phylogenetically distinct, with the pangolin order Pholidota and the order Carnivora being a sister group (Ferae) nested within the superorder Lauraisatheria (Murphy et al. 2001).

Unlike other placental mammals, pangolin skin is covered by large and overlapping keratinized scales (Meyer et al. 2013). The selective forces underlining the origin of this unique mammalian trait remains a mystery, although observations of the eight modern species suggest a defensive armour function against predators. Furthermore, pangolins are edentulous or toothless, eating mostly ants and termites captured using long and muscular tongues. They also have a well-developed muscular system for fossoriality or arboreality and a remarkable olfactory system. Pangolins are the most poached and trafficked mammal in the world due to the huge demand of their meat as a delicacy and their scales for use in traditional medicines.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

The Malayan pangolin and the Chinese pangolin are classified as critically endangered by the IUCN Red List of Threatened Species (Challender 2014). Malayan pangolin is mainly found in Southeast Asia, whereas Chinese pangolins inhabit China, Taiwan and in some Northern of Southeast Asia (**Supplemental Fig. S1.1**). Here we present the first whole-genome sequencing and comparative analyses of these two unique species of the Pholidota, filling an important gap in our understanding of mammalian genome evolution and providing fundamental knowledge for further research in pangolin biology and conservation.

Results

Pangolin genome, divergence and heterozygosity

A female Malayan pangolin derived from a wild specimen from Malaysia and a female Chinese pangolin from Taiwan were sequenced using whole-genome shotgun sequencing strategies to a coverage ~145X and ~59X based on the estimated genome size of 2.5 gigabases (Gbp) and 2.7Gbp, respectively (**Supplemental Information 1.0**). We identified 23,446 and 20,298 protein-coding genes in the Malayan and Chinese pangolin genomes, respectively, using the MAKER annotation pipeline based on several sources of evidence from *ab initio* gene prediction, transcriptomic data and protein evidence from the *Canis familiaris* reference genome (Broad CanFam3.1/canFam3) and transcriptome (Lindblad-Toh et al. 2005; Cantarel et al. 2008). The assemblies of both Malayan and Chinese pangolins capture at least 99.7% and 94.8% of the expressed genes assembled from RNA-seq data generated from eight different organs, respectively (**Supplementaly Information 2.1**). The lower overall mapping rate of the Chinese pangolin is partially explained by the divergence of Malayan and Chinese pangolin, although it could also be impacted by the different sequencing depths. Furthermore, Core Eukaryotic Genes

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Mapping Approaches (CEGMA) and genome comparison analyses indicated that these genomes are good candidates for genome annotation (**Supplemental Information 2.2&2.3&3.0**).

Coalescent dating analysis suggest that pangolins diverged from their closest relatives, the Carnivora, ~56.8-67.1 millions years ago (MYA) (**Supplemental Information 4.0**) and that Malayan and Chinese pangolin species diverged from each other ~4-17,3MYA. Malayan pangolin has a more than three times higher heterozygosity rate than Chinese pangolin (1.55×10-3 and 0.4× 10-3, respectively), which likely reflects the demographic impacts of a founder population and smaller effective population in the latter species (**Supplemental Information 5.0)**. Interestingly, analysis of the trajectory of historical effective population size of the two pangolins using the Pairwise Sequential Markovian Coalescent model showed an inverted pattern during the Middle Pleistocene, reflecting the influence of temperature and sea level on both pangolin species **(Supplemental Fig. S6.1).**

Pangolin-specific phenotypes

To gain insight into possible genomic patterns linked with the unique traits of pangolins, we first searched for pseudogenized (presumably loss-of-function) genes. Gene loss through pseudogenization is increasingly recognized as an important factor in lineage and adaptive phenotypic diversification (Wang et al. 2006; Albalat and Canestro 2016).Given that pangolins are edentulous mammals, we screened 107 tooth development-related genes and found three pseudogenised candidate genes in pangolins. Most notable of these was *ENAM*, the largest protein of the enamel matrix and essential for normal tooth development (Meredith et al. 2009). *ENAM* mutations can lead to enamel defects (Hart et al. 2003). We identified several frameshift

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

indels and premature stop codons in the *ENAM* genes of both the Malayan and Chinese pangolins (confirmed by Sanger sequencing; sample size=8) **(Figure 1a & Supplemental Information 7.0)** (Meredith et al. 2009). *ENAM* pseudogenization has also been reported in the African tree pangolin (*M. tricuspis*) (Meredith et al. 2009). We also identified two other pseudogenised genes that have be linked with normal tooth development, amelogenin (*AMELX*) and ameloblastin (*AMBN*). These genes shared common mutations in both Asian species and all four African species (*M. tricuspis*, *M. tetradactyla*, *M. gigantea* and *M. temminckii*) that we examined, suggesting that pseudogenization occurred early in the evolutionary history of the Pholidota lineage (Supplemental Figs. S7.1&S7.2). Loss of function of these genes through pseudogenization has also been reported in other edentulous vertebrates such as toothless baleen whales, birds and turtles (Meredith et al. 2014).

Pangolins are mostly nocturnal (the only exception being the long-tailed pangolin) and are thought to have poor vision (Soewu and Ayodele 2009). We screened 217 vision-related genes and identified several candidate genes that were pseudogenized **(Figure 1b)**. One of these, *BFSP2*, encodes a lens-specific intermediate filament-like protein that is a unique cytoskeletal element in the ocular lens of vertebrates (Sandilands et al. 2003). Interestingly, one frameshift insertion and three premature stop codons were consistently identified in the *BFSP2* genes of both pangolin species (confirmed by Sanger sequencing; sample size=8), suggesting possible loss of optical clarity, which may lead to progressive cataracts (Alizadeh et al. 2004). Another candidate gene, guanylate cyclase activator 1C (*GUCA1C*), expresses proteins that stimulate photoreceptors GC1 and GC2 at low concentrations of free calcium, but inhibit GCs when the concentrations of free calcium ions are elevated, which is important for regulating the recovery of the dark state of rod photoreceptors following light exposure (Haeseleer et al. 1999). We

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

found identical frameshift mutations and premature stop codons in *GUCA1C* in both pangolin species (confirmed by Sanger sequencing; sample size=8), suggesting reduced stimulation of GC1 and GC2 and reduced rates of phototransduction in pangolins (Stephen et al. 2006). The *GUCA1C* gene was also pseudogenised in all African species with identical mutations (Supplemental Fig. S7.3). Interestingly, the pseudogenisation of *BFSP2* and *GUCA1C* have been reported in mice with reduced vision, supporting the hypothesis that the two genes are likely associated with the poor vision of pangolins (Imanishi et al. 2002; Sandilands et al. 2004; Song et al. 2009).

As an evolutionary consequence of being covered by scales, it is plausible that the pangolin immune system evolved differently than in other mammals. For example, genes such as *IFNE*, a unique interferon exclusively expressed in skin epithelial cells and inner mucosa-protected tisssues (e.g. lung, intestines and reproductive tissues), establish a first line of defense against pathogens in other placental mammals (Day et al. 2008; Ponten et al. 2008; Xi et al. 2012; Fung et al. 2013; Demers et al. 2014; Uhlen et al. 2015). Interferons (IFNs) are a cluster of highly conserved gene families that encode for cytokines expressed by host cells for communication between cells, leading to the activation of the immune system in the presence of pathogens (De Andrea et al. 2002; Fensterl and Sen 2009). Strikingly, the single copy intronless *IFNE* gene is pseudogenised in both pangolin species (confirmed by Sanger sequencing; sample size=8), but is intact in 21 other mammalian species **(Figure 2a)**. We found an insertion from positions 195 to position 219, and a point mutation at position c235C>T causing a premature stop codon. Other frameshift deletions at positions 264 and 540-546 would cause the loss of function. Although an alternate splicing pattern might avoid the premature stop codon and preserve the protein, high-throughput RNA sequencing of lung, cerebellum, cerebrum and skin transcriptomes failed to

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

detect *IFNE* expression (Fragments Per Kilobase Million (FPKM)=0.0), although it is expressed in these organs in other mammals (Uhlen et al. 2005; Demers et al. 2014). We were also unable to identify any sequencing reads that mapped to the remainder of the gene after the premature stop codon, which is consistent with the hypothesis of a loss of function even if the stop codon were rescued. Interestingly, many putative functional domains or key signatures are deleted in the pangolin *IFNE* protein due to the premature stop codon **(Figure 2b)**. This suggests that the *IFNE* protein is unlikely to function properly if expressed, indicating that resistance to infection may be impacted in pangolins. We also examined other IFN families (*IFNB*, *IFNK*, *IFNA*, *IFNG* and *IFNL*) and found that all families have intact gene copies in both pangolin species. It should be noted that the *IFNA* family consists of a cluster of 13 functional genes in human. Although we found that several intact genes from the IFN family in both pangolin species were retained, the number of *IFNA* genes is relatively low compared to human, suggesting diminished functions of this gene family in pangolins.

To test whether the *IFNE* was also pseudogenized in the African pangolins, we sequenced the gene in all African species using the primers that we designed based on the genome sequence of Malayan pangolin. Our data showed that the *IFNE* gene was also pseudogenised in all African species and shared ancestral mutations with the Asian pangolin species (Supplemental Fig. S7.4). Therefore, we suggest that the pseudogenisation of *IFNE* in pangolins most likely occured before the divergence of the African and Asian pangolins, or at least 19-26.9 MYA, based on prior estimates (Bininda-Emonds et al. 2007; Fritz et al. 2009; Meredith et al. 2011).

Comparative analyses among pangolins and mammals

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

To further explore genetic differences between pangolins and their closest relatives, we compared the overlap of gene families among the two pangolin species and dog, cat and giant panda. Our analysis identified 8,325 ancestral gene families common to all five species **(Figure 3a)**. Interestingly, many gene families were unique to either Malayan pangolin (4,958) or Chinese pangolin (3,465), confirming the high level of divergence within pangolins, likely resulting from evolutionary adaptations to different ecological environments. We also found 1,152 pangolin-specific gene families, of which 61% are expressed in at least one of the eight organs assessed, further suggesting that at least these genes are not assembly or annotation artefacts (**Figure 3b**). Functional enrichment analyses suggest that the pangolin-specific genes are significantly over-represented in signal transduction (496 genes; GO:0007165; FDR p-value=2.53 $x10^{-9}$), neurological system processes (262 genes; GO:0050877; FDR p-value=2.51 $x10^{-20}$), cytoskeleton organisation (99 genes; GO:0007010; FDR p-value=1.53 $x10^{-3}$) and cell junction organisation (28 genes; GO:0034330; FDR p-value= 1.32 $x10^{-2}$) (**Figure 3c**). The large number of pangolin-specific genes significantly enriched in signal transduction and neurological system processes suggest that pangolins evolved complex and unique signal transduction and neurological system networks. The large number of cytoskeleton organization associated genes may be linked with the assembly, formation and maintenance of the cytoskeletal elements involved in cell shape, structural integrity and motility of the pangolin scales.

Gene family analysis identified 147 families that are significantly expanded and 18 that are contracted in the pangolin lineage, indicating that gene expansion and contraction events played major role in the functional diversification of pangolins **(Figure 3d & Supplemental Information 8.0)**. Notably, pangolins have a very reduced number of interferon genes, which

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

have a major role in responding to infections, inflammation and healing of skin(Xi et al. 2012; Fung et al. 2013). Of the 10 genes in the interferon family, only three were found in Malayan pangolin and two in Chinese pangolin **(Figure 3e)**. The heat shock gene family also contracted significantly, possibly contributing to the sensitivity of pangolins to stress (Hua et al. 2015). However, we also document gene expansion in several important families, including ribosomal genes (17 families), olfactory receptor (OR) genes (6 families), cathepsin genes (1 family), and septin genes (1 family), which were recently shown to deter microbial pathogens and preventing them from invading other cells (**Supplemental Information 8.0)**(A 2011). The significantly expanded OR gene families suggest that pangolins have an enhanced sense of smell, possibly helping locate prey and counterbalancing poor vision **(Supplemental Fig. S8.2).** Strikingly, functional enrichment analysis of the 147 expanded gene families reveal these genes are significant over-represented in neurological system processes (including neuromuscular process and sensory perception) (207 genes; GO:0050877; FDR p-value=6.43 $x10^{-14}$), which is consistent with our finding in the functional analysis of pangolin-specific genes (**Supplemental Fig. S8.3)**. The gene families associated with symbiosis and host-parasite interactions were also over-represented (82 genes; GO:0044403; FDR p-value=2.67 $x10^{-13}$).

Positive selection analysis

To identify signatures of natural selection, we used a set of 8,250 protein-coding orthologs shared among the dog, cat, panda, cow, horse, mouse, human and megabat genomes (**Supplemental Information 9.0**). Branch site tests identified evidence of positive selection along the pangolin lineage in 427 genes that had significant signals ($p<0.05$ adjusted; **Supplemental Table S9.1**).

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

We hypothesize that the reduced IFN-mediated immunity from the loss of *IFNE* and the contraction of interferon gene family in pangolins imposed strong selective pressure on immunity-related genes. We identified a large proportion of genes under selection in the pangolin lineage, that involve a wide range of immunity-related pathways including hematopoietic cell lineage, cytosolic DNA-sensing pathway, complement and coagulation cascades, cytokine-cytokine receptor interaction and the phagosome pathway **(Supplemental Table S9.2)**. For instance, in the hematopoietic cell lineage, the colony stimulating factor 3 receptor (granulocyte) (*CSF3R*) gene encodes a transmembrane receptor for a cytokine (granulocyte colony stimulating factor 3) (**Figure 4a**). CSF3R proteins present on precursor cells in the bone marrow, playing important roles in the proliferation and differentiation into mature neutrophilic granulocytes and macrophages which are essential for combating infection(Liongue and Ward 2014). As predicted by Protein Variation Effect Analyzer (PROVEAN), CSF3R protein contains a critical pangolin-specific mutation at position 247(G->H) in the functionally relevant Fibronection Type III domain that likely affecting its protein function in pangolins (**Figure 4c**). Another gene, integrin alpha M (*ITGAM*) mediates leukocyte activation and migration, phagocytosis and neutrophil apoptosis. It has been reported that *ITGAM* can migrate neutrophils across intestinal epithelium to maintain intestinal homeostasis and eliminate pathogens that have translocated across the single layer of mucosal epithelial cells(Parkos et al. 1991). We identified two pangolin-specific amino acid changes in the functional Integrin alpha-2 domain of *ITGAM* which likely affect protein function (**Figure 4c**). In the cytosolic DNA-sensing pathway, transmembrane protein 173 (*TMEM173*), also known as Stimulator of interferon genes, plays an important role in innate immunity by sensing cytosolic foreign DNA

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

and inducing type I interferon production when cells are infected with pathogens (**Figure 4b**)(Ran et al. 2014). We identified two critical amino acid changes in pangolins at positions 216(D->G) and 217(P->L) in the functionally relevant region, suggesting that they may have functional impact on *TMEM173* (**Figure 4c**).

We further postulate that genetic changes during the evolution of hair-derived pangolin scales likely involved genes related to hair formation in general, and particularly keratins, which are essential component of scales and hairs (Meyer et al. 2013). Among the positively-selected candidate genes associated with skin formation are *KRT36, KRT75, KRT82* and *KRTAP3-1*. For instance, Type II keratin 75 (*KRT75*), a hair follicle-specific keratin, has an essential role in hair and nail integrity (Chen et al. 2008). Type I keratin 36 (*KRT36*), a hair or "hard" sulfur-rich keratin, is mainly responsible for the extraordinary high degree of filamentous cross-linking by specialized keratin-associated proteins (Moll et al. 2008). Many critical pangolin-specific amino acid changes were identified in keratin genes in functionally relevant domains **(Figure 4d)**. These critical amino changes are also present in both *KRT36* and *KRT75* genes in the distantly related African pangolin species that we examined, suggesting that they potentially contributed to the development of pangolin scales (**Figure 4d & Supplemental Fig. S7.4&S7.5**). We also identified genes related to energy storage and metabolism (12 genes) and mitochondrial metabolism (11 genes), perhaps an adaptive response to reduce their metabolic rate given their large body size, but low energy diet of ants and termites (da Fonseca et al. 2008). Other candidate genes are associated with the development of the robust muscular system of pangolins, which assists its fossorial lifestyle (10 genes), and the nervous system (10 genes) (**Supplemental Table S9.2**).

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Furthermore, we explored whether there was evidence that genes under selection were associated with any response to diseases, particularly those related with infection- or skin/mucosal organs. We identified genes associated with inflammation (18), cancer or viral infection (25), bacterial infection (10), pneumonia and gastrointestinal disease (21) and skin diseases (15) (**Supplemental Table S9.3**). Interestingly, one gene, lactotransferrin *(LTF*) encodes a multifunctional immune protein found at mucosal surfaces, providing the first line of defense to the host against inflammation and infection (Ward et al. 2002). It has been shown that *LTF* has antimicrobial activities and induces both systematic and mucosal immune responses, for example, against lung and gut-related systemic infections (Debbabi et al. 1998). Two pangolin-specific amino acid changes were detected in the LTF protein sequence, potentially having impact on LTF function (**Supplemental Fig. S9.1)**. Together, evidence of selection on these genes provides possible links with the unique phenotypes and physiological requirements of pangolins. Importantly, the immunity- or inflammation/infection- related genes provide avenues to further explore the evolution of mammalian immunity.

Pangolins have noncanonical repeat layout

Transposable elements accounted for approximately 28.9-30.0% of the pangolin genomes, which are lower than in other Carnivora such as the giant panda (35.3%), tiger (36.7%), cat (37.2) and dog (36.2%), suggesting divergence in genome architecture between pangolins and other placental mammals **(Supplemental Fig. S10.1a)**. Pangolins also have relatively low proportions of SINE repeats (2.6-2.8% vs. 8.2-11%), mainly due to a relatively smaller number of tRNA SINEs **(Supplemental Fig. S10.1a & b)**. A search using a *de novo* constructed repeat library supports the comparatively low proportion of this repeat family (Supplemental Table S10.2).

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Furthermore, both pangolin species have relatively low proportion of intronic repeats (SINEs, LINEs and LTRs) **(Supplemental Fig. S10.1c)**.

Discussion and conclusion

Our analyses have revealed that, similar to observations in other lineages (Wang et al. 2006; Meredith et al. 2014), gene loss through pseudogenization has had a significant impact on the evolution and diversification of pangolin genomes and their biology. We detected and validated pseudogenes in pathways related to multiple morphological or physiological functions including dentition, vision, and immunity. The pseudogenization of the *IFNE* gene in pangolins, otherwise functional in all other mammal genomes surveyed so far, is especially noteworthy. In mice, a deficiency of *IFNE* in epithelial cells of the female reproductive tract (FRT) can increase susceptibility to microbial infection (Fung et al. 2013). Therefore, the pseudogenisation of this interferon gene in pangolins suggests that innate immunity may be compromised, resulting in an increased susceptibility to infection, particularly in the skin and mucosa-protected organs. This is further supported by evidence that the Malayan pangolin used here had been significantly colonised by *Burkhoderia* sp. in tissues that express *IFNE* (cerebrum, cerebellum and lung) in other species, but not in tissues (liver, kidney, thymus, spleen and heart) where the gene is usually not expressed in other mammals **(**data not shown) (Ponten et al. 2008; Demers et al. 2014; Uhlen et al. 2015). Since pangolins may have intrinsically weak mucosal immunity, *Burkholderia* sp. might easily infect the lung by penetrating lung epithelial cells and the brain by penetrating nasal mucosa and migrating to the brain through olfactory nerves has previously been shown with other *Burkholderia* (Owen et al. 2009; Sim et al. 2009; Taylor et al. 2010; Dando et al. 2014; St John et al. 2014). Furthermore, captive pangolins are prone to frequently

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

fatal gastrointestinal disease, pneumonia, and skin maladies (Clark L 2008; Hua et al. 2015). Pangolins are notoriously difficult to maintain in captivity, and the stress of captivity and/poor husbandry might render pangolins even more vulnerable to infections or diseases by suppressing immune responses. Whether these afflict free ranging pangolins where pathogen exposure is rampant remains to be seen.

Our study raises the fundamental evolutionary question of how pangolin adaptations occurred given their seemingly increased risk to infection if their skin and mucosal surfaces are constantly being exposed to pathogens? Our analyses suggest several evolutionary changes in genes that may respond or interact with pathogens, including immunity-related genes/pathways, the significant expansion of the septin gene family and the enrichment of the genes in symbiosis. These changes may enhance the immunity system of pangolins, although we do not know the relative efficiency of these adaptations compared to interferon-mediated immunity. We propose that pangolin scales may be an important morphological innovation to compensate for decreased immunity normally provided by the skin. These hard and overlapping scales may act as defensive armor to protect pangolins against injuries (or stress) which would make pangolins even more vulnerable to infection or the invasion of pathogens. Moreover, pangolins curl into near impregnable balls covering their scaleless abdomen using their well-developed neuromuscular system during sleep or when threatened, which would also support the hypothesis that these adaptations serve to protect pangolins from skin injuries. It is noteworthy that besides the anti-microbial effects, plasmacytoid dendritic cell (pDC)-produced type 1 IFNA/B/ help heal wounds through reepithelization of injured skin (Gregorio et al. 2010). It is unknown whether keratinocyte-produced type-I *IFNE* is also involved in skin wound healing (e.g. recruiting pDCs

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

to the wound). Pangolins have a marked contraction of interferon genes that likely facilitates wound healing. Future studies are needed to test the relationship between *IFNE,* skin healing and scale development.

Our analyses showed that pangolin-specific genes were significantly over-represented in the GO categories governing cytoskeleton organisation, neurological system process and signal transduction. Further, certain genes involved in muscular and neuron system showed evidence of positive selection. These could be linked to the unique adaptations and behaviours of pangolins, particularly those related with their highly sophisticated musculoskeletal system **(Supplemental Information 12).**

While the Chinese pangolin had a significantly larger predicted genome size but significantly fewer genes than the Malayan pangolin, it does not represent the fact in the final assembly. Despite this limitation, we found segmental duplications were identified in a higher proportion in Chinese pangolin compared to the Malayan pangolin, suggesting that the segmental duplication might partly contribute to the large genome size of Chinese pangolin (**Supplemental Information 3.4**). It should be noted that a previous first comparative chromosome map between the Malayan (2n=38) and Chinese (2n=36 to 42) pangolins revealed many chromosomal rearrangements between the two species (Nie et al. 2009). Besides that, we did not observe significant recent repeat family activity unique to Chinese pangolins, suggesting that the genome expansion is unlikely directly accounted by repeats (Supplemental Fig. S10.2). However, we cannot rule out the possibility that the recent segmental duplications and repeats were underestimated (e.g. some might collapse together due to high sequence similarity during

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

assembly since we sequenced the pangolin genomes using the short read Illumina technology), which would complicate the analyses.

Our data also showed that the number of gene families unique to each of the pangolin species is 3-4 times as great as the number of shared gene families. In this analysis, we faced the common problem of the almost all draft genomes - difference between the real and available gene set. We can expect a number of misannotated genes because of fragmented assembly, and this can partly inflate the number of genes for each species. But it should not affect subsequent analyses because we had refined pangolin-specific gene set by orthology clustering as described in the Methods section. It is also worth noting here that the discrepancy in the unique gene family numbers indeed can be explained by the natural reasons such as high evolution rate. It was previously shown that the chromosome complements of Malayan pangolin and Chinese pangolin differ by seven Robertsonian rearrangements (3 centric fissions and 4 centric fusions), so it also could affect the number of the unique gene families for each species after their divergence from the common ancestor. (Nie et al. 2009).

In conclusion, this study provides new insights into the biology, evolution and diversification of pangolins. The annotated reference genomes will be invaluable for future studies addressing issues of species conservation and gene function and may provide a new natural animal model for understanding mammalian immunity.

Methods

Genome sequencing, assembly and annotation

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Genomic DNA from female pangolins was extracted and sequenced using Illumina sequencing platform. Different insert size of short read and mate-pair libraries ranging from ~180bp-8kbp were used. The genome sequence of Malayan pangolin was assembled using CLC Assembly Cell 4.10, SGA 0.10.10 and SOAPdenovo2 and the best assembly was chosen for downstream analyses (Supplemental Information 1.3.1) (Luo et al. 2012; Simpson and Durbin 2012). The genome of Chinese pangolin was assembled using SOAPdenovo v1.0.5 (Supplemental Information 1.3.2). The Malayan pangolin assembly has a genomic length of ~2.5Gbp with an assembled N50 scaffold size of 204,525bp, whereas the Chinese pangolin scaffold size was ~2.2Gbp with an assembled N50 scaffold size of 157,892bp (**Supplemental Information 1.3**).

Protein-coding genes in the pangolin genomes were predicted using MAKER annotation pipeline by combining *de novo* gene prediction, RNA-seq data and homology-based methods **(Supplemental Information 3.0) (Cantarel et al. 2008).** Identified protein-coding genes in the Malayan and Chinese pangolin genomes are supported by functional assignments from at least one biological database by InterProScan 5 pipeline (**Supplemental Information 3.0**) (Jones et al. 2014).

Genome size was estimated by *k*-mer analysis using short-insert sequencing reads (**Supplemental Information 1.4**). The genome size of Malayan pangolin and Chinese pangolin was estimated to be 2,492,544,425 bp and 2,696,930,760 bp, respectively.

Gene family clustering

OrthoMCL (Li et al. 2003) was used to identify homologous protein sequences among Malayan pangolin, Chinese pangolin, cat, dog and giant panda. BLASTALL (Pearson 2014) was first

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

performed among the protein sequences. For each pair of gene members of a gene family/cluster, both local and global matched regions must be at least 40% of the longer gene protein sequence. The minimum score value of 50 and e-value of 1 $x10^{-8}$ in BLAST was used. After clustering the protein sequences into gene families, pangolin-specific genes were retrieved for downstream analysis using in-house Perl scripts.

Functional enrichment analysis

Enrichment analysis of pangolin-specific and the significantly expanded gene families were performed using Blast2GO with all pangolin genes used as background for comparisons (Conesa et al. 2005). To identify significantly over-represented gene ontology categories (e.g. molecular function and biological processes), Fisher's exact test was used with a cut-off of a False Discovery Rate of 0.05 after multiple test correction.

Repetitive element analysis

Repetitive elements in the pangolin genomes and other Carnivora genomes (giant panda, tiger, cat and dog) were identified using RepeatMasker open-4.0.5 (Smit 2013-2015) against the Repbase TE database (version 2014-01-31) (**Supplemental Information 6**) (Jurka 2000).

RNA-seq expression analysis

The RNA-seq raw data were from our Malayan pangolin transcriptome project. Briefly, we sequenced the transcriptomes of cerebellum, cerebrum, lung, heart, kidney, liver, spleen and thymus using Illumina HiSeq technology platform (100bp Paired End (PE) strategy). The data from each sample ranged from 41-53 million of PE reads or ~8.2-10.6Gbp. For each organ, the

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

PE reads were mapped to the genome assembly using TopHat 2.0.11 (Trapnell et al. 2009). Properly mapped PE reads with the best match were sorted and indexed with SAMtools (Li et al. 2009). The expression of each genes were calculated and represented in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) based on the coordinates of the Maker-generated gene models. Heat map showing expression of pangolin-specific genes across the eight different pangolin organs were generated using in-house R scripts. Using the same approach, we also calculated the expression level of pseudogenised *IFNE* gene in the transcriptomes of three organs (lung, cerebrum and cerebellum) of Malayan pangolin, as well as a transcriptome of skin tissue (250bp PE; # of PE reads=~17 millions) which we generated in a separate project using Illumina MiSeq technology.

Functional impact of substitutions in positive selected genes

We assessed potential functional impact of pangolin-specific amino acid changes in the positively selected genes of interest in the pangolin ancestor using PROVEAN (Choi and Chan 2015). Amino acid substitutions were considered "deleterious" if the PROVEAN score was $\leqslant -2.5$ and "neutral replacements" if score was >-2.5.

Data access

The Malayan pangolin and Chinese pangolin BioProjects are accessible at the NCBI BioProject (BioProject: http://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA256023 and PRJNA20331, respectively. The assembled whole-genome sequences have been deposited at GenBank under the accessions JSZB00000000.1 (Malayan pangolin) and JPTV00000000.1 (Chinese pangolin). The raw sequencing reads of Malayan pangolin and Chinese pangolin have

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

been deposited in the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) with the accession numbers SRP078903 and SRX247981, respectively. All RNA-seq reads can be accessed at the NCBI SRA under the accession number SRP064341. The assembled transcriptome sequences of Malayan pangolinhave been deposited at Transcriptome Shotgun Assembly (TSA; http://www.ncbi.nlm.nih.gov/genbank/tsa) with the accession number GEUV00000000. All PCR trace data have been deposited to NCBI Trace (Trace; http://www.ncbi.nlm.nih.gov/Traces) with the trace IDs (SID 271513, SID 271514, SID 271515, SID 271516 and SID 271517). In addition, the pangolin genome sequences, raw sequencing reads, RNA-seq data of Malayan pangolin are also accessible at our pangolin genome hub (http://pangolin-genome.um.edu.my/).

Acknowledgements

We would like to thank Kerstin Lindblad-Toh and Jessica Alfoldi for generously providing guidance and sharing their advice. We also thank all members of Genome Informatics Research Laboratory, University of Malaya in contributing to this research. Special thank to Tan Shi Yang for providing IT and bioinformatics support in this project. This project was mainly supported by University of Malaya and Ministry of Education, Malaysia under the High Impact Research (HIR) grant UM.C/HIR/MOHE/08 and UMRG grant (grant number: RG541-13HTM) from University of Malaya and Ministry of Education, Malaysia. This research was also supported in part by Russian Ministry of Science (Mega-grant no.11.G34.31.0068; SJ O'Brien Principal Investigator) and the funding to Richard K Wilson was provided by NIH-NHGRI grant 5U54HG00307907.

Disclosure Declaration

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

All the authors declare they have no competing interest.

Figure Legends

Figure 1. Case studies of pseudogenised genes in pangolins. (a) Three tooth development related genes were pseudogenised and may be related to the lack of teeth in pangolins. There was a frameshift deletion in positions c1311-c1324, a single base pair deletion in c1476 and an insertion of AGAT at position 1621, resulting another premature stop codon in the *AMBN* gene. The *AMBN* gene contains a large frameshift deletion at position c396-c455. (b) Two genes were pseudogenised and may be related to poor vision of pangolins. Blue=insertion; green=deletion; and Pink=stop codon.

Figure 2. Multiple sequence alignment of all mammalian *IFNE* genes. 73 mammalian species have available *IFNE* sequences and used for alignment. (a) Nucleotide sequence alignment of *IFNE* genes across different mammalian species. Blue=insertion, green=deletion and pink=stop codon. Protein sequence alignment of *IFNE* genes across different mammalian species is also shown. We identified an insertion (blue) in *IFNE* starting from nucleotide position of 195th but not in other mammalian species, indicating that this insertion is specific to pangolins and possibly a marker to differentiate pangolins and other mammalian species. A premature stop codon or frameshift (orange) in the *IFNE* gene was consistently detected in both pangolin genomes. These mutations were validated by Sanger sequencing in eight Malayan pangolins. Protein sequences highlighted in red are the frameshift mutations and premature stop codon. (b) Comparison between pangolin and human (reference) *IFNE* protein sequences. Pangolins have a short putative protein sequence because of a premature stop codon. Predicted functional domains and signatures in the human *IFNE* gene are represented by color boxes. Yellow and pink boxes represent the predicted binding residues to IFNAR2 and IFNAR1, respectively, and which are bounded by interferons, which we obtained from a previosuly published paper(Fung et al. 2013). IFabd (SM00076) is a conserved functional domain in known interferons. The main conserved structural feature of interferons is a disulphide bond. INTERFERONAB (PR00266) is a 3-element fingerprint that provides a signature

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

for alpha, beta and omega interferons. The elements 1 and 3 contain Cys residues involved in disulphide bond formation.

Figure 3. Comparative analysis between pangolins and mammals. (a) Venn diagram showing the unique and shared gene families among pangolins and their closest relatives (cat, dog and giant panda). (b) Heatmap showing the expression level of pangolin-specific genes across different pangolin organs, which represented by FPKM values. Any FPKM values >5 were set to 5 in the heatmap for visualisation purpose. Only genes expressed (FPKM>=0.3) in at least one organ were shown. (c) GO enrichment analysis of 1,152 pangolin-specific genes. Significantly enriched GO terms are shown for the categories of Cellular Compartment (blue), Molecular Function (yellow) and Biological Process (purple). (d) Phylogenetic tree and gene family expansion and contraction. Expanded gene families are indicated in blue, whereas contracted gene families are indicated in red. The proportion of expanded and contracted gene families is also shown in pie charts. (e) Phylogenetic tree showing significant contraction of the interferon gene family in pangolins.

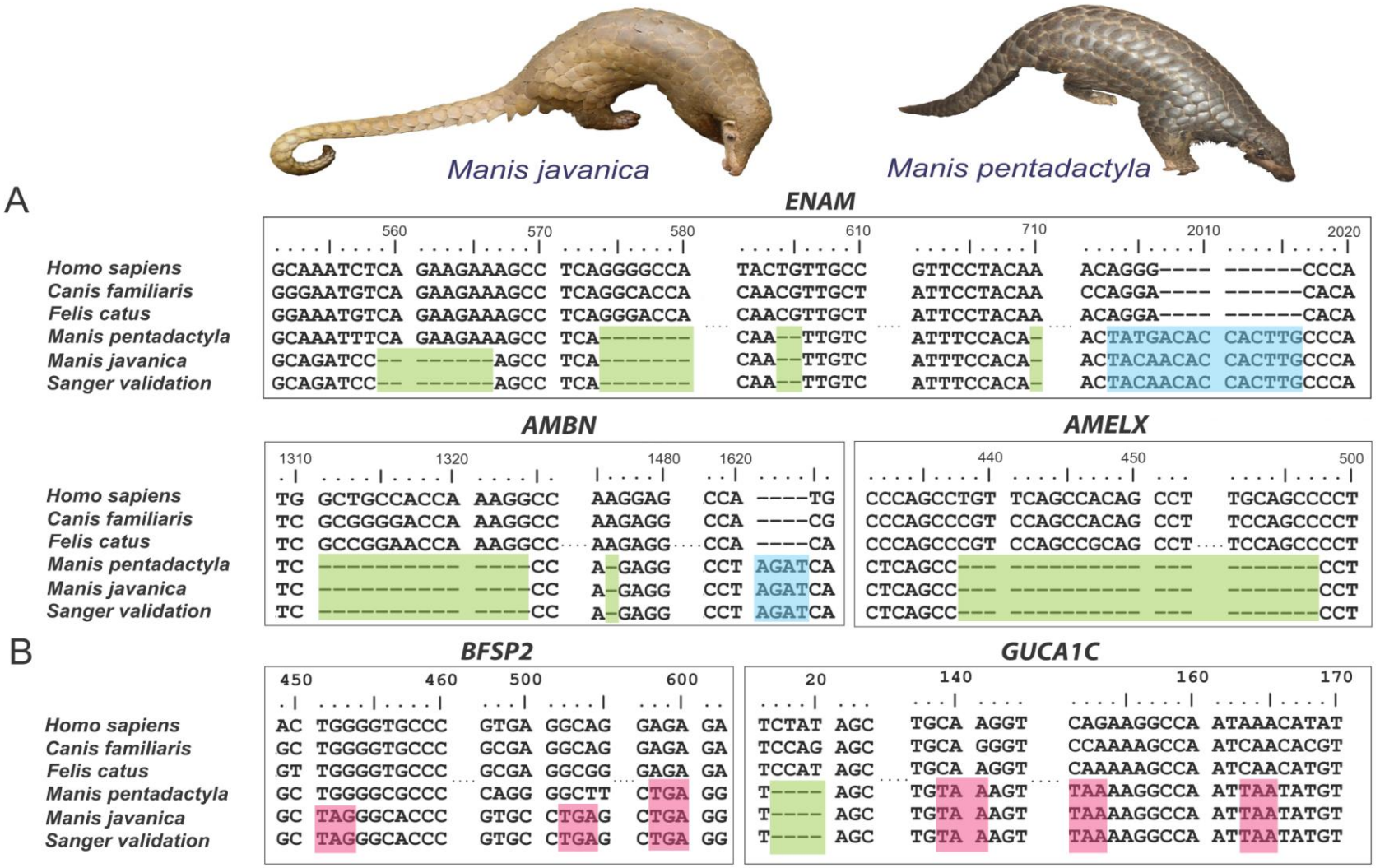
Figure 4. Evolution in the immunity-related pathways and scale-related genes in the pangolin ancestor. Genes under positive selection at the pangolin lineage in hematopoietic cell lineage (a) and cytosolic DNA-sensing pathway (b) are highlighted in red colour. (c) Several critical pangolin-specific mutations were identified in the functionally relevant Fibronection Type III signatures (INTERPRO ID=IPR003961) of CSF3R, the Integrin alpha-2 signatures (IPR013649) of ITGAM and the Stimulator of Interferon Genes Protein region (INTERPRO ID=IPR029158) of TMEM173. For the CSF3R, ITGAM and TMEM173, p-values of the detection of positive selection were 1.58×10^{-2}, 3.75×10^{-2} and 2.92×10^{-5}, respectively. (d) Several critical pangolin-specific amino acid changes were detected in hair/scale-related keratin proteins, KRT36 and KRT75, which located in functionally relevant regions that may affect protein functions. For the KRT36 and KRT75, p-values of the detection of positive selection were 2.24 $\times 10^{-2}$ and 1.33×10^{-5}, respectively. We also examined whether African species (*M. tricuspis, M.*

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

tetradactyla and *M. temminckii*) have these critical pangolin-specific amino acid changes that we observed in the Asian pangolins. Circles at the bottom indicated our preliminary results: Red circle= all African species that we examined have the identical amino acid change; Yellow circle= Not all African species that we examined have identical amino acid change. But they all have amino changes/deletion that likely affect protein function as predicted by PROVEAN; Brown circle=All African species that we examined have the same amino acids like the human reference sequence; and White circle=data not available. The alignment results are shown in Supplemental Fig. 7.5.

Figures

Figure 1. Case studies of pseudogenised genes in pangolins.



Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 2. Multiple sequence alignment of all mammalian *IFNE* genes.

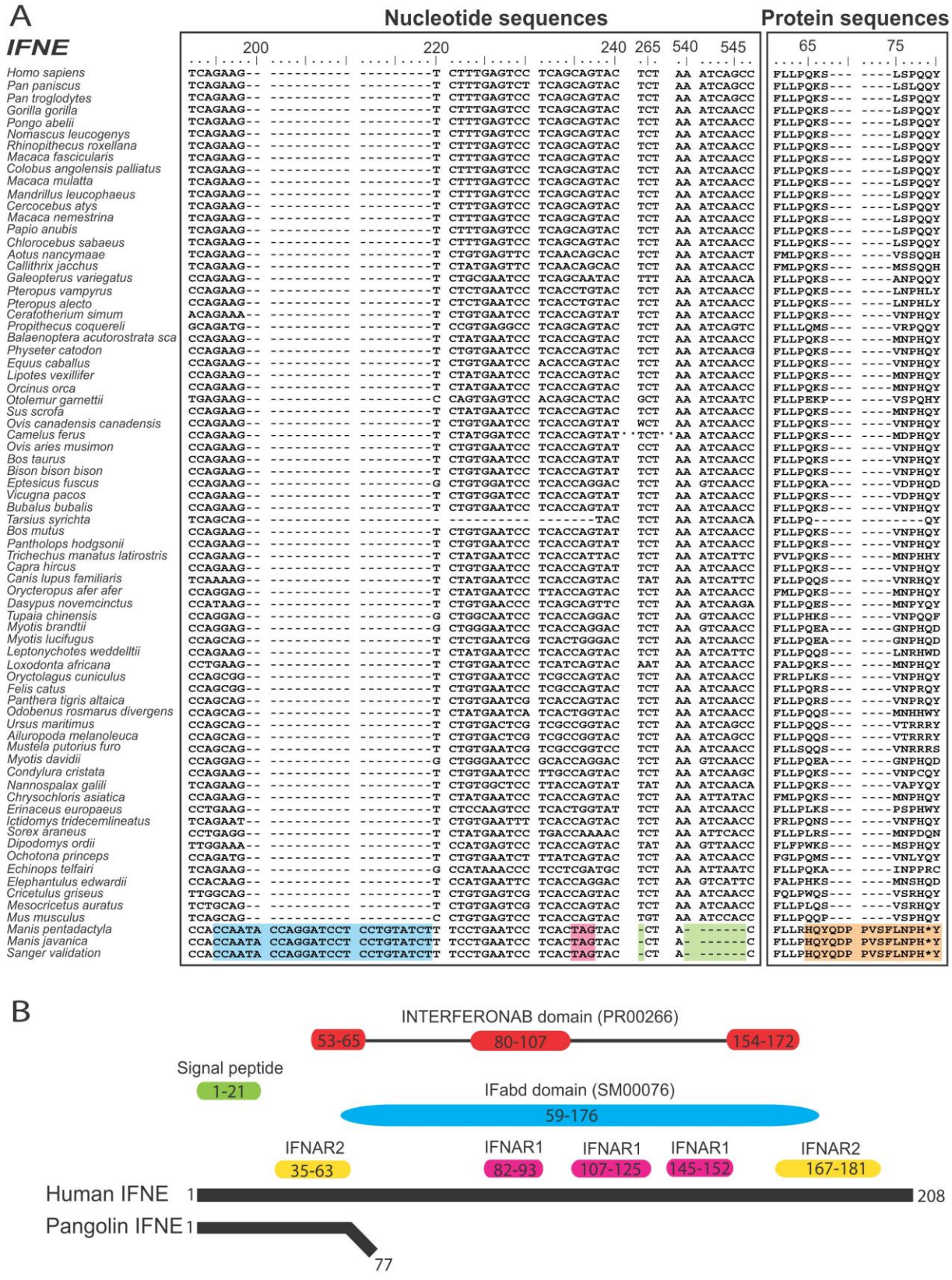
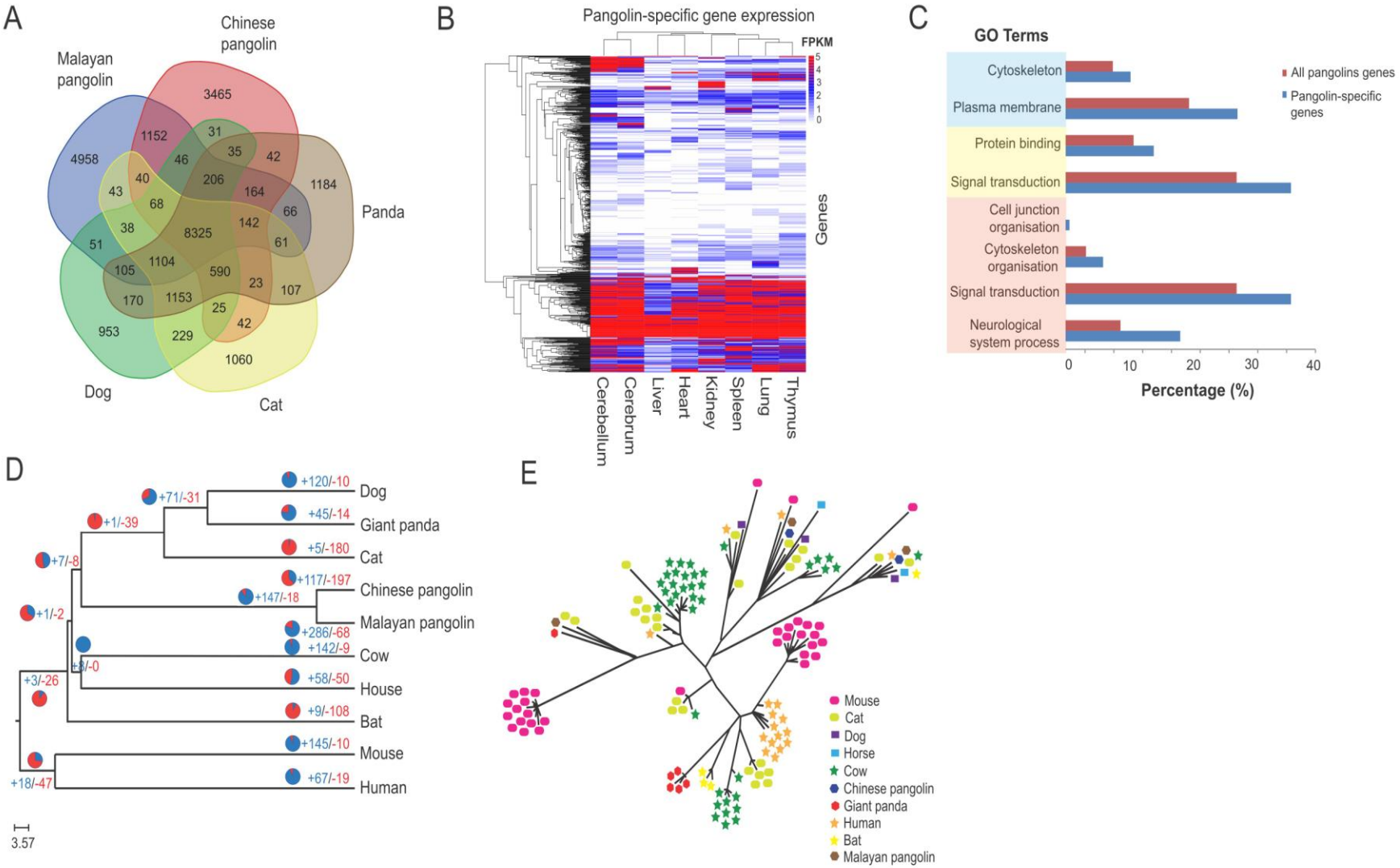
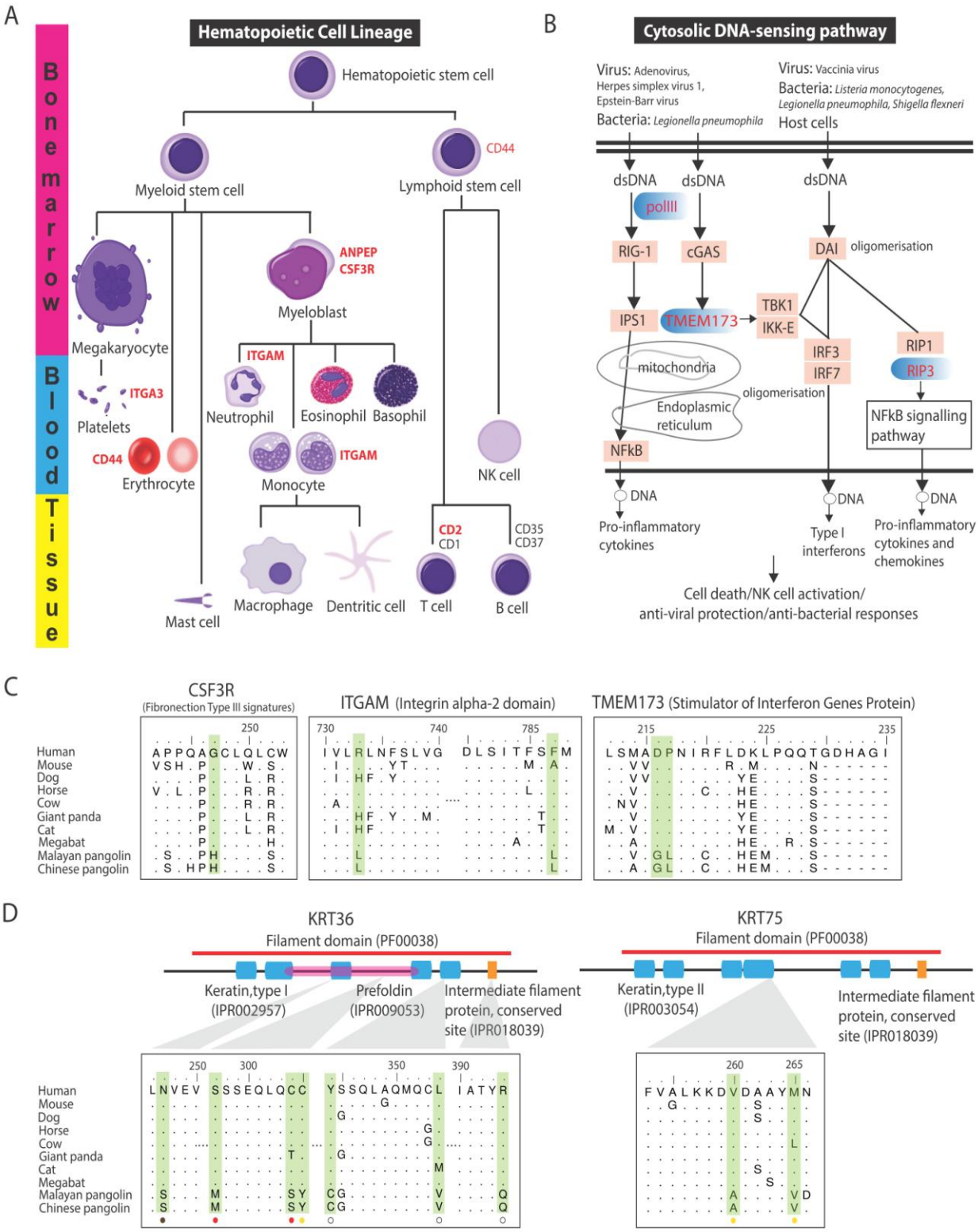


Figure 3. Comparative analysis between pangolins and mammals.



Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 4. Evolution in the immunity-related pathways and scale-related genes in the pangolin ancestor.



Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Author information

Affiliations

Genome Informatics Research Laboratory, High Impact Research (HIR) Building, University of Malaya, Malaysia
Siew Woh Choo, Tze King Tan, Ranjeev Hari, Wei Yee Wee, Aini Mohamed Yusoff & Guat Jah Wong

Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, Russia
Stephen J. O'Brien, Mike Rayko, Andrey Yurchenko, Aleksey Komissarov, Sergey Kliver, Pasha Dobrynin, Gaik Tamazian & Ksenia Krasheninnikova

McDonnell Genome Institute, Washington University, St Louis, MO 63108, USA.
Wesley C. Warren, Richard K. Wilson & Patrick Minx

Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA.
Mark B. Gerstein

Department of Molecular Biophysics and Biochemistry, P Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA.
Mark B. Gerstein

Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA.
Mark B. Gerstein

Department of Neurology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA.
Deyou Zheng

Peking-Tsinghua Center for Life Sciences, College of Life Sciences, Peking University, Beijing 100871, China.
Shu-Jin Luo

Department of Oral Biology and Biomedical Sciences, Faculty of Dentistry, University of Malaya, 50603 Kuala Lumpur, Malaysia
Siew Woh Choo & Ian C Paterson

Ex-Situ Conservation Division, Department of Wildlife and National Parks, 10 Jalan Cheras, 56100 Kuala Lumpur, Malaysia.
Kayal Vizi Karuppannan, Frankie Thomas Sitam & Jeffrine Rovie Ryan Japning

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Department of Neurology, School of Medicine, Wayne State University, Detroit, MI 48201, USA.
Leonard Lipovich

Oceanographic Center, Nova Southeastern University, Ft Lauderdale, Florida 33004
Stephen J. Obrien

Genome Solutions Sdn Bhd, Suite 8, Innovation Incubator UM, Level 5, Research Management & Innovation Complex, University of Malaya, 50603 Kuala Lumpur, Malaysia.
Siew Woh Choo

National Zoological Park, Smithsonian Conservation Biology Institute, Washington, DC 20008, USA
Warren E. Johnson & Klaus Koepfli

CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Rua dos Bragas, 177, 4050-123 Porto, Portugal
Agostinho Antunes

Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal.
Agostinho Antunes

Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA.
Leonard Lipovich

Oral Cancer Research and Coordinating Centre, Faculty of Dentistry, University of Malaya, 50603 Kuala Lumpur, Malaysia.
Ian C. Paterson

NYU Shanghai, 1555 Century Ave, Pudong, Shanghai, China 200122.
Gang Fang

National Zoological Gardens of South Africa, .O. Box 754, Pretoria 0001, South Africa.
Antoinette Kotze, Desire L. Dalton & Elaine Vermaak

Department of Genetics, University of the Free State, P.O. Box 339, Bloemfontein, 9300, South Africa.
Antoinette Kotze & Desire L. Dalton

Contributions

The project is a part of the International Pangolin Research Consortium (IPaRC). The analyses were mainly conducted by members from the Genome Informatics Research Laboratory,

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

University of Malaya and the Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, Russia. For full details of author contributions please see below:

S.W.C., S.J.O, M.R., G.J.W, W.E.J., and A.M.Y. conceived and coordinated this project. S.W.C., K.V.K, A.M.Y., W.G.J., J.R.R.J and F.A.T.S performed animal handling, sampling and tagging. S.W.C., R.H. and I.C.P coordinated and performed DNA extraction. S.W.C., R.K.W. and P.M. led and designed library construction and NGS experiments. S.W.C. and T.K.T designed primers, PCR and Sanger sequencing validation experiments. A.K, D.L.D, E.V. performed Sanger sequencing validation experiments for African pangolins. S.W.C., M.R., T.K.T., R.H., W.Y.W., A.A., K.P.K., A.K., Y.A., M.R., S.K., P.D., R.K.W., P.M., D.Z., M.B.G. and G.F. performed data analyses, data interpretation and oversaw various analyses. S.W.C., T.K.T., M.R., A.K., R.H., & W.G.J. wrote manuscript. S.J.O., A.A., K.P.K., W.C.W, W.E.J., I.C.P. and L.L. revised manuscript. S.W.C. and S.J.O. were the principal investigators of this project.

Corresponding authors

Siew Woh Choo

Email: lchoo@um.edu.my or l.choo@genomesolutions.com.my

Stephen J. O'Brien

Email: lgdchief@gmail.com

References

A M. 2011. Septin proteins take bacterial prisoners: A cellular defence against microbial pathogens holds therapeutic potential. *Nature*.

Albalat R, Canestro C. 2016. Evolution by gene loss. *Nat Rev Genet*.

Alizadeh A, Clark J, Seeberger T, Hess J, Blankenship T, FitzGerald PG. 2004. Characterization of a mutation in the lens-specific CP49 in the 129 strain of mouse. *Invest Ophthalmol Vis Sci* **45**(3): 884-891.

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* **446**(7135): 507-512.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**(1): 188-196.

Challender Dea. 2014. Manis javanica. The IUCN Red List of Threatened Species. Version 2014.3.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Chen J, Jaeger K, Den Z, Koch PJ, Sundberg JP, Roop DR. 2008. Mice expressing a mutant Krt75 (K6hf) allele develop hair and nail defects resembling pachyonychia congenita. *The Journal of investigative dermatology* **128**(2): 270-279.

Choi Y, Chan AP. 2015. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**(16): 2745-2747.

Clark L VTN, Phuong TQ. 2008. A long way from home: the health status of Asian pangolins confiscated from the illegal wildlife trade in Viet Nam. In: Pantel S, Chin SY. (Eds). In *Proceedings of the Workshop on Trade and Conservation of Pangolins Native to South and Southeast Asia TRAFFIC Southeast Asia, Singapore Zoo*, pp. 111–118, Singapore.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18): 3674-3676.

da Fonseca RR, Johnson WE, O'Brien SJ, Ramos MJ, Antunes A. 2008. The adaptive evolution of the mammalian mitochondrial genome. *BMC Genomics* **9**: 119.

Dando SJ, Mackay-Sim A, Norton R, Currie BJ, St John JA, Ekberg JA, Batzloff M, Ulett GC, Beacham IR. 2014. Pathogens penetrating the central nervous system: infection pathways and the cellular and molecular mechanisms of invasion. *Clinical microbiology reviews* **27**(4): 691-726.

Day SL, Ramshaw IA, Ramsay AJ, Ranasinghe C. 2008. Differential effects of the type I interferons alpha4, beta, and epsilon on antiviral activity and vaccine efficacy. *Journal of immunology* **180**(11): 7158-7166.

De Andrea M, Ravera R, Gioia D, Gariglio M, Landolfo S. 2002. The interferon system: an overview. *European journal of paediatric neurology : EJPN : official journal of the European Paediatric Neurology Society* **6 Suppl A**: A41-46; discussion A55-48.

Debbabi H, Dubarry M, Rautureau M, Tome D. 1998. Bovine lactoferrin induces both mucosal and systemic immune response in mice. *The Journal of dairy research* **65**(2): 283-293.

Demers A, Kang G, Ma F, Lu W, Yuan Z, Li Y, Lewis M, Kraiselburd EN, Montaner L, Li Q. 2014. The mucosal expression pattern of interferon-epsilon in rhesus macaques. *J Leukoc Biol* **96**(6): 1101-1107.

Fensterl V, Sen GC. 2009. Interferons and viral infections. *BioFactors* **35**(1): 14-20.

Fritz SA, Bininda-Emonds OR, Purvis A. 2009. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecology letters* **12**(6): 538-549.

Fung KY, Mangan NE, Cumming H, Horvat JC, Mayall JR, Stifter SA, De Weerd N, Roisman LC, Rossjohn J, Robertson SA et al. 2013. Interferon-epsilon protects the female reproductive tract from viral and bacterial infection. *Science* **339**(6123): 1088-1092.

Gregorio J, Meller S, Conrad C, Di Nardo A, Homey B, Lauerma A, Arai N, Gallo RL, Digiovanni J, Gilliet M. 2010. Plasmacytoid dendritic cells sense skin injury and promote wound healing through type I interferons. *The Journal of experimental medicine* **207**(13): 2921-2930.

Haeseleer F, Sokal I, Li N, Pettenati M, Rao N, Bronson D, Wechter R, Baehr W, Palczewski K. 1999. Molecular characterization of a third member of the guanylyl cyclase-activating protein subfamily. *The Journal of biological chemistry* **274**(10): 6526-6535.

Hart TC, Hart PS, Gorry MC, Michalec MD, Ryu OH, Uygur C, Ozdemir D, Firatli S, Aren G, Firatli E. 2003. Novel ENAM mutation responsible for autosomal recessive amelogenesis imperfecta and localised enamel defects. *J Med Genet* **40**(12): 900-906.

Hua L, Gong S, Wang F, Li W, Ge Y, Li X, Hou F. 2015. Captive breeding of pangolins: current status, problems and future prospects. *ZooKeys*(507): 99-114.

Imanishi Y, Li N, Sokal I, Sowa ME, Lichtarge O, Wensel TG, Saperstein DA, Baehr W, Palczewski K. 2002. Characterization of retinal guanylate cyclase-activating protein 3 (GCAP3) from zebrafish to man. *The European journal of neuroscience* **15**(1): 63-78.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**(9): 1236-1240.

Jurka J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**(9): 418-420.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**(9): 2178-2189.

Lindblad-Toh K Wade CM Mikkelsen TS Karlsson EK Jaffe DB Kamal M Clamp M Chang JL Kulbokas EJ, 3rd Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**(7069): 803-819.

Liongue C, Ward AC. 2014. Granulocyte colony-stimulating factor receptor mutations in myeloid malignancy. *Frontiers in oncology* **4**: 93.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**(1): 18.

Meredith RW, Gatesy J, Murphy WJ, Ryder OA, Springer MS. 2009. Molecular decay of the tooth gene Enamelin (ENAM) mirrors the loss of enamel in the fossil record of placental mammals. *PLoS genetics* **5**(9): e1000634.

Meredith RW, Janecka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simao TL, Stadler T et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science* **334**(6055): 521-524.

Meredith RW, Zhang G, Gilbert MT, Jarvis ED, Springer MS. 2014. Evidence for a single loss of mineralized teeth in the common avian ancestor. *Science* **346**(6215): 1254390.

Meyer W, Liumsiricharoen M, Suprasert A, Fleischer LG, Hewicker-Trautwein M. 2013. Immunohistochemical demonstration of keratins in the epidermal layers of the Malayan pangolin (Manis javanica), with remarks on the evolution of the integumental scale armour. *European journal of histochemistry : EJH* **57**(3): e27.

Moll R, Divo M, Langbein L. 2008. The human keratins: biology and pathology. *Histochemistry and cell biology* **129**(6): 705-733.

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**(6820): 614-618.

Nie W, Wang J, Su W, Wang Y, Yang F. 2009. Chromosomal rearrangements underlying karyotype differences between Chinese pangolin (Manis pentadactyla) and Malayan pangolin (Manis javanica) revealed by chromosome painting. *Chromosome Res* **17**(3): 321-329.

Owen SJ, Batzloff M, Chehrehasa F, Meedeniya A, Casart Y, Logue CA, Hirst RG, Peak IR, Mackay-Sim A, Beacham IR. 2009. Nasal-associated lymphoid tissue and olfactory epithelium as portals of entry for Burkholderia pseudomallei in murine melioidosis. *The Journal of infectious diseases* **199**(12): 1761-1770.

Parkos CA, Delp C, Arnaout MA, Madara JL. 1991. Neutrophil migration across a cultured intestinal epithelium. Dependence on a CD11b/CD18-mediated event and enhanced efficiency in physiological direction. *The Journal of clinical investigation* **88**(5): 1605-1612.

Pearson WR. 2014. BLAST and FASTA similarity searching for multiple sequence alignment. *Methods in molecular biology* **1079**: 75-101.

Ponten F, Jirstrom K, Uhlen M. 2008. The Human Protein Atlas--a tool for pathology. *The Journal of pathology* **216**(4): 387-393.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Ran Y, Shu HB, Wang YY. 2014. MITA/STING: a central and multifaceted mediator in innate immune response. *Cytokine & growth factor reviews* **25**(6): 631-639.

Sandilands A, Prescott AR, Wegener A, Zoltoski RK, Hutcheson AM, Masaki S, Kuszak JR, Quinlan RA. 2003. Knockout of the intermediate filament protein CP49 destabilises the lens fibre cell cytoskeleton and decreases lens optical quality, but does not induce cataract. *Experimental eye research* **76**(3): 385-391.

Sandilands A, Wang X, Hutcheson AM, James J, Prescott AR, Wegener A, Pekny M, Gong X, Quinlan RA. 2004. Bfsp2 mutation found in mouse 129 strains causes the loss of CP49 and induces vimentin-dependent changes in the lens fibre cell cytoskeleton. *Experimental eye research* **78**(1): 109-123.

Sim SH, Liu Y, Wang D, Novem V, Sivalingam SP, Thong TW, Ooi EE, Tan G. 2009. Innate immune responses of pulmonary epithelial cells to Burkholderia pseudomallei infection. *PloS one* **4**(10): e7308.

Simpson JT, Durbin R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22**(3): 549-556.

Smit A, Hubley, R & Green, P. 2013-2015. RepeatMasker Open-4.0.

Soewu DA, Ayodele IA. 2009. Utilisation of Pangolin (Manis sps) in traditional Yorubic medicine in Ijebu province, Ogun State, Nigeria. *Journal of ethnobiology and ethnomedicine* **5**: 39.

Song S, Landsbury A, Dahm R, Liu Y, Zhang Q, Quinlan RA. 2009. Functions of the intermediate filament cytoskeleton in the eye lens. *The Journal of clinical investigation* **119**(7): 1837-1848.

St John JA, Ekberg JA, Dando SJ, Meedeniya AC, Horton RE, Batzloff M, Owen SJ, Holt S, Peak IR, Ulett GC et al. 2014. Burkholderia pseudomallei penetrates the brain via destruction of the olfactory and trigeminal nerves: implications for the pathogenesis of neurological melioidosis. *mBio* **5**(2): e00025.

Stephen R, Palczewski K, Sousa MC. 2006. The crystal structure of GCAP3 suggests molecular mechanism of GCAP-linked cone dystrophies. *J Mol Biol* **359**(2): 266-275.

Taylor JB, Hogue LA, LiPuma JJ, Walter MJ, Brody SL, Cannon CL. 2010. Entry of Burkholderia organisms into respiratory epithelium: CFTR, microfilament and microtubule dependence. *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society* **9**(1): 36-43.

Timothy J. Gaudin RJRJE. 2009. The Phylogeny of Living and Extinct Pangolins (Mammalia, Pholidota) and Associated Taxa: A Morphology Based Analysis. *J Mammal Evol* **16**: 235–305.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.

Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, Andersen E, Andersson AC, Angelidou P, Asplund A, Asplund C et al. 2005. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & cellular proteomics : MCP* **4**(12): 1920-1932.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* **347**(6220): 1260419.

Wang X, Grus WE, Zhang J. 2006. Gene losses during human origins. *PLoS biology* **4**(3): e52.

Ward PP, Uribe-Luna S, Conneely OM. 2002. Lactoferrin and host defense. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **80**(1): 95-102.

Xi Y, Day SL, Jackson RJ, Ranasinghe C. 2012. Role of novel type I interferon epsilon in viral infection and mucosal immunity. *Mucosal Immunol* **5**(6): 610-622.