Long-read sequencing reveals rapid evolution of immunity- and cancer-related genes in bats

Armin Scheben[1], Olivia Mendivil Ramos[2], Melissa Kramer[2], Sara Goodwin[2], Sara Oppenheim[3], Daniel J Becker[4], Michael C Schatz[1,5], Nancy B Simmons[6], Adam Siepel[1*], W Richard McCombie[2*]

[1] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
[2] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
[3] American Museum of Natural History, Sackler Institute for Comparative Genomics, New York, NY
[4] Department of Biology, University of Oklahoma, Norman, OK
[5] Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD
[6] Department of Mammalogy, Division of Vertebrate Zoology, American Museum of Natural History, New York, NY

*To whom correspondence should be addressed: asiepel@cshl.edu, mccombie@cshl.edu

Author ORCID iD

Armin Scheben: https://orcid.org/0000-0002-2230-2013
Olivia Mendivil Ramos: https://orcid.org/0000-0003-3740-1943
Melissa Kramer: N/A
Sara Goodwin: N/A
Sara Oppenheim: https://orcid.org/0000-0003-1109-3639
Daniel J Becker: https://orcid.org/0000-0003-4315-8628
Michael C Schatz: https://orcid.org/0000-0002-4118-4446
Nancy B Simmons: http://orcid.org/0000-0001-8807-7499
Adam Siepel: https://orcid.org/0000-0002-3557-7219
W Richard McCombie: https://orcid.org/0000-0003-1899-0682

Abstract

Bats are exceptional among mammals for their powered flight, extended lifespans, and robust immune systems. To investigate the genomic underpinnings of unique bat adaptations, we sequenced the genomes of the Jamaican fruit bat (*Artibeus jamaicensis*) and the Mesoamerican mustached bat (*Pteronotus mesoamericanus*) and compared them to a diverse collection of 13 additional bat species together with other mammals. We used the Oxford Nanopore Technologies long-read platform to generate highly complete assemblies (N50: 28-29Mb) and facilitate analysis of complex genomic regions containing duplicated genes. Using gene family size analysis, we found that the type I interferon locus was contracted by eight genes in the most recent common ancestor (MRCA) of bats, shifting the proportion of interferon-ω to interferon-α and making interferon-ω the most common type I interferon in bats. Antiviral genes stimulated by type I interferons were also rapidly evolving, with interferon-induced transmembrane genes experiencing a lineage-specific duplication and strong positive selection in the gene *IFIT2*. Moreover, the lineage of phyllostomid bats showed an unprecedented expansion of *PRDM9*, a recombination-related gene also involved in infection responses, raising the possibility that this gene contributes to bat antiviral defenses. These modifications in the bat innate immune system may be important adaptations allowing them to harbor viruses asymptomatically. We additionally found evidence of positive selection on the branch leading to the MRCA of bats acting on 33 tumor suppressors and six DNA repair genes, which may contribute to the low cancer rates and longevity observed across bats. These new genomic resources enable insights into the extraordinary adaptations of bats, with implications for mammalian evolutionary studies and public health.

Introduction

Bats (order Chiroptera) form the second largest order of mammals and are known for a wide variety of remarkable adaptations including powered flight[1], laryngeal echolocation[2,3], unusual longevity[4], and low rates of cancer[5]. Bats are also hosts of diverse viruses[6,7] and have played roles in outbreaks of emerging zoonotic viruses including Marburg virus[8], Nipah virus[9], and both severe acute respiratory syndrome coronavirus 1 (SARS-CoV-1)[10] and SARS-CoV-2[11], either through direct human contact or via bridge hosts. It has been suggested that their tolerance of many viral infections stems from unusual features of their innate immune response[12]. Together, these adaptations make bats a powerful system for investigating a wide variety of genotype-to-phenotype relationships, including several with implications for human health. For example, by better understanding the mechanisms of the bat immune system that allow them to tolerate viral infections[5], we may be better able to prevent zoonotic outbreaks. In addition, comparative genomic analyses of bats and cancer-susceptible mammals may shed new light on the causes of cancer and links between cancer and immunity[13]. Importantly, such studies of bats and other non-model organisms are highly complementary to studies based on mouse models, which are far more amenable to experimental manipulation but exhibit fewer natural adaptations relevant to human disease.

With these goals in mind, investigators have sequenced and assembled the genomes of at least 43 bat species over the past decade (**Table S1**). Recently, sequencing efforts in bats have been accelerated by the Bat1K global genome consortium[14] (http://bat1k.com), DNA Zoo[15] (https://www.dnazoo.org/) and Vertebrate Genome Project (https://vertebrategenomesproject.org). These new genome sequences have revealed numerous intriguing features of the immune systems of bats[12,16–22]. In particular, several genes with key roles in the innate immune system appear to have adaptively evolved in bats, including primary lines of inducible host defences such as pathogen sensors[16,23], type I interferons (IFNs)[12,24] and antiviral genes[25]. Specifically, bats have lost the mammalian PYHIN DNA-sensing gene family[17,26], they show evidence of positive selection in pathogen-sensing Toll-like receptors (TLRs) [23], and they display copy-number variation in type I IFN cytokines[12,24], which are induced by TLRs. Bat-specific modifications in tumor suppressors, DNA damage checkpoint-DNA repair pathway genes[17], and growth hormone[27] may be associated with cancer resistance. It is thought that these adaptations in innate immunity and cancer resistance may have arisen as a result of coevolution of bats with viruses[18,28], and that a need for enhanced DNA repair in the face of elevated reactive oxygen species (ROS) may have been a consequence of powered flight[17].

In this study, we augment previously existing genome sequences with new Oxford Nanopore Technologies (ONT)-based long-read assemblies for the Jamaican fruit bat (*Artibeus jamaicensis*) and the Mesoamerican mustached bat (*Pteronotus mesoamericanus*) (**Figure S1**). We present a comprehensive analysis of these genome sequences together with 13 previously assembled bats and other mammalian genomes. Importantly, long-read assemblies enable dramatic improvements in the characterization of gene duplications and losses, and of genomic repeats[29–31]. These benefits are of particular value in studies of mammalian immunity-related genes, many of which fall in highly repetitive genomic regions including large arrays of duplicated

genes[32]. Our comparative genomic analysis of these genome sequences provides several new insights into unique features of innate immune response and cancer resistance in bats.

Results

Genomic structure of *A. jamaicensis* and *P. mesoamericanus*

Except for two recently published studies[12,18], most previous genomic investigations in bats have relied on short-read assemblies based on Illumina DNA sequencing technology, which has limited the study of complex regions. In this case, we were able to leverage the ONT long-read sequencing platform and an optimized flye[33]-PEPPER[34]-POLCA[35] assembly and polishing strategy (**Methods**) to generate reference-quality assemblies for *A. jamaicensis* and *P. mesoamericanus* with contig N50 values of 28-29Mb (**Figure S2**, **Table S2**) and POLCA consensus accuracy >99.99%. Using EVidenceModeler, we annotated 21,621 genes in *A. jamaicensis* and 21,269 genes in *P. mesoamericanus*. Based on a BUSCO protein assessment of our annotations, the gene sets in both bats are highly complete at 98.3% and 98.2% respectively, comparable to the values of 97.4-98.3% reported for six recent PacBio-based bat assemblies (**Figure S3**). Notably, all of these long-read bat assemblies have BUSCO scores approaching those of the human (99.9%) and mouse (99.9%) annotations. Orthofinder orthology detection produced 19,935 orthogroups for 15 bats and 5 outgroup mammals, of which 12,517 single-copy orthogroups were set aside for our positive selection analysis (below). Total fractions of 39.2% and 37.9% of the *A. jamaicensis* and *P. mesoamericanus* genomes consisted of repeats, respectively, with 0.4% in each genome attributed to recently active transposons including hAT, TcMariner, and piggyBac elements (**Figure S4**, **Table S3**). We also detected non-retroviral endogenous viral elements, predominantly derived from Bornaviridae and Parvoviridae (**Table S4**). We provide our annotations, aligned evidence, and multiple genome alignments as a public UCSC genome browser instance (http://compgen.cshl.edu/bat/).

Gene family expansion and contraction analysis

Changes in gene family size have played an important role in shaping the immune systems of bats[17,26]. To facilitate further analysis of gene family expansions and contractions, we focused on our new ONT-based assemblies and the previously published long-read bat genome sequences. By comparing these bat genomes with mammalian outgroups (human, mouse, dog, pig, and horse), we identified four expanded gene families and 47 contracted gene families in the most recent common ancestor of bats (hereafter, the “bat MRCA”; **Table S5**). Twenty-five of the gene families changing size on the bat ancestral branch were related to immune system processes, including the previously identified PYHIN gene family (PTHR12200)[26], which we confirmed to be entirely absent in all bats including the two newly sequenced species.

The type I interferon locus is contracted in bats

The type I IFN immune response is a critical component of the mammalian innate immune system, and is responsible for activating the expression of a battery of antiviral genes following induction by pathogen-sensing components such as PYHINs, TLRs, and *cGAS-STING*[36]. Previous comparative analyses of the type I IFN locus have shown that it is highly structurally variable in bats and other mammals[24], with some bats, such as *Pteropus alecto,* showing a contraction[24], whereas others such as *Rousettus aegyptiacus*[12], *Pteropus vampyrus,* and *Myotis lucifugus*[37] show evidence of expansions. However, this locus is generally large and highly duplicated across mammals (e.g., in humans, it spans ~400kb and contains 16 IFN genes, including 13 IFN-α genes and one IFN-ω gene), making it challenging to assemble and analyze.

Using our expanded set of long-read-based bat genomes, we found clear evidence of a major contraction (-8 genes; Viterbi method, see ref.[38], p=6.4e-11) of the type I IFN gene family in the bat MRCA (**Figure 1**, **Figure S5**). Interestingly, we found that this contraction was driven specifically by loss of IFN-α genes, with gene counts of 0-4 in bats compared to 10-18 in the outgroup mammals (**Figure 1**). In contrast, IFN-ω gene counts in the bat ancestor were largely unchanged (Viterbi method, p=0.51), ranging from 3-7 in bats and 0-8 in other mammals. As a consequence, IFN-ω was 11-fold enriched relative to IFN-α in bats compared to the other mammals (Fisher's exact test, p=4.5e-11, odds ratio=10.80) and this enrichment was observed in every bat species (**Table S6**). Considering the relative ligand-binding and antiproliferative properties of IFN-ω and IFN-α[39,40], these changes in gene number could potentially be responsible for more potent responses to viral infections in bats relative to other mammalian orders (see **Discussion**).

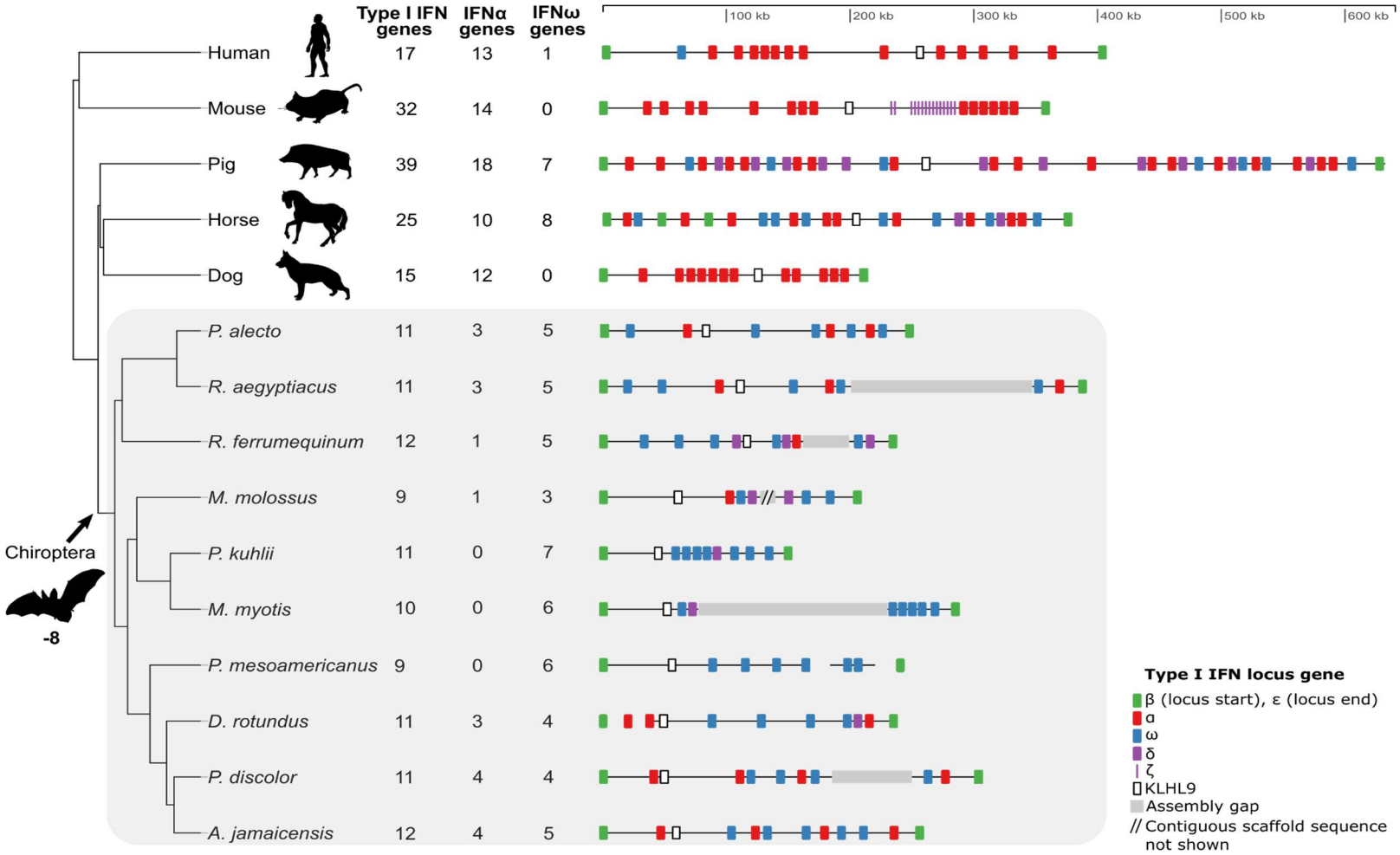


Figure 1. Contraction of the type I interferon (IFN) locus in bats compared to other mammals. A loss of eight genes in the bat MRCA was estimated by CAFE[38] (Viterbi method, *p*=6.4e-11). The reduction in locus size occurred together with a significant loss of IFN-α genes but not IFN-ω genes in the bat MRCA (Fisher's exact test, *p*=4.5e-11, odds ratio=10.80). The type I IFN loci in bats (grey background) are shown for long-read based assemblies as well as a BAC-based locus assembly[24] (*P. alecto*) and an Illumina-based assembly[19] (*Desmodus rotundus*). Abbreviated species names: *Rhinolophus ferrumequinum*, *Molossus molossus*, *Pipistrellus kuhlii*, *Myotis myotis*, and *Phyllostomus discolor*.

Antiviral interferon-induced transmembrane genes are expanded in yangochiropteran bats

An important downstream consequence of the activation of type I IFNs is the expression of various antiviral IFN-stimulated genes (ISGs)[41]. In bats, several of these genes, such as *tetherin* and *APOBEC3,* have also been shown to be under positive selection[42] or are duplicated[18]. Among antiviral ISGs, we observed an expansion of the immune-related IFN-induced transmembrane (IR-IFITM) gene family (+2 genes; Viterbi method, p=1.6e-3) on the branch leading to the Yangochiroptera, the suborder that includes most microbats (**Figure 2**). The IR-IFITMs—which have previously been reported to be under positive selection in bats[43]—are potent broad-spectrum antiviral factors that help to prevent infection before a virus passes the lipid bilayer of the cell[44,45]. By applying a dN/dS-based branch-site likelihood ratio test (**Methods**), we confirmed using our data that the IR-IFITMs show evidence of positive selection in the bat MRCA (p=6.4e-3). Furthermore, we identified seven particular codon sites that show signs of episodic diversifying selection (**Table S7**), including three sites in the CD225 domain (codons 68, 70, and 117). Notably, codons 68 and 70 in the CD255 domain are among several sites previously shown to be critical for blocking viral infection[43]. Furthermore, the methionine-to-phenylalanine substitution at codon 68, occurs in an amphipathic helix previously shown to be essential for blocking viral infection[46]. Together these observations of gene duplications and positive selection in functional domains in IR-IFITMs suggests that this gene family may have played an important role in the evolution of antiviral responses in bats.

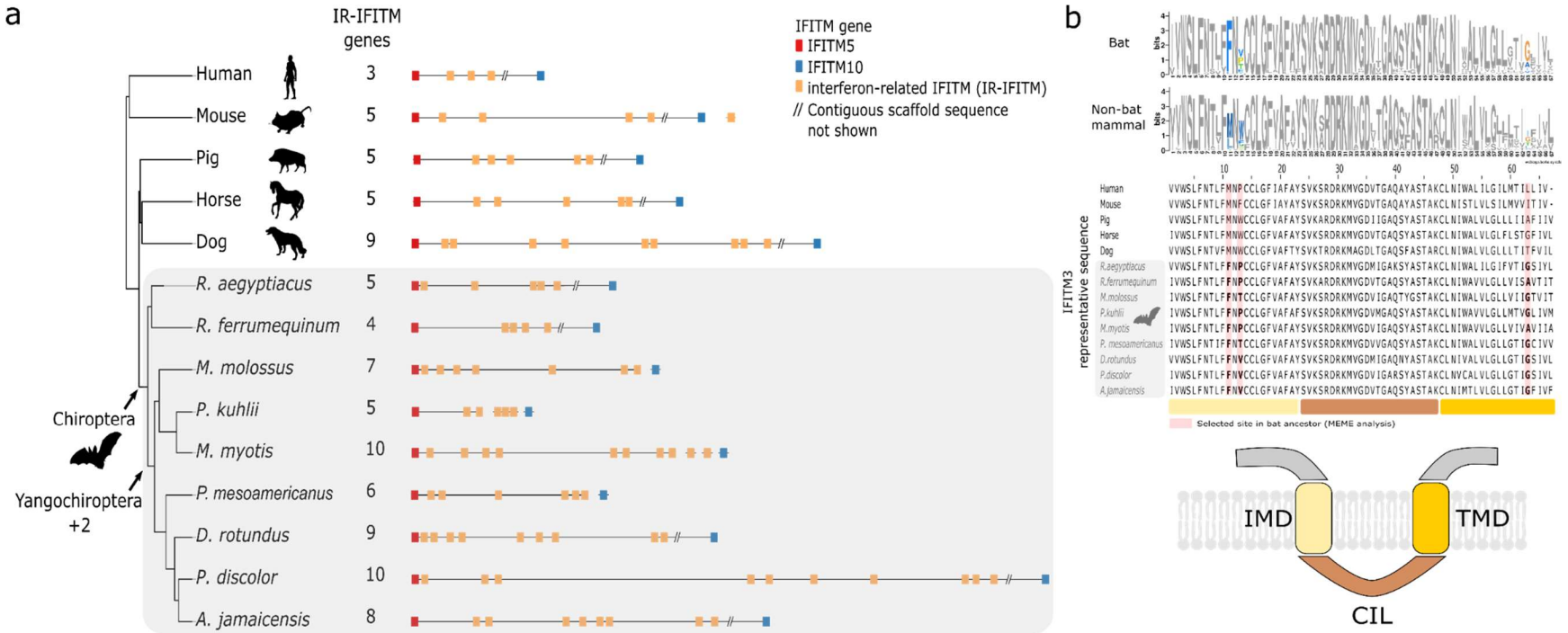


Figure 2. **IFITM gene family expansion and positive selection associated with the bat antiviral immune response.** a) Phylogeny of bats and other mammals, showing a significant increase in gene copy number at the IFITM locus in yangochiropteran bats based on CAFE[38] analysis. Immune-related interferon-induced transmembrane (IR-IFITM) genes are shown in yellow. b) Three codon sites in the IFITM transmembrane domains IMD and TMD in bats show evidence of positive selection (see also **Table S7**). The sequence logo (top) compares a 67-amino-acid region spanning these domains with orthologous regions from other mammals (selected sites in color). The sequence alignment (middle) compares human *IFITM3* with the most similar representative IR-IFITM ortholog from the other species (selected sites highlighted with red background and shown in bold). The protein model (bottom) is based on a published transmembrane topology[47].

Expansion of *PRDM9* in phyllostomid bats and expansion of heat shock proteins in *P. mesoamericanus*

A third gene family to emerge from our survey of gene expansions and contractions was *PRDM9* (**Table S8**), which specifies the location of meiotic recombination sites[48] and is known to evolve rapidly in vertebrates[49]. *PRDM9* may play a role in speciation[50] and is also upregulated upon viral infection[51]. We found that *PRDM9* experienced a striking expansion in phyllostomid bats (+5 genes; Viterbi method, *p*=1.2e-09; **Figure S6**), far beyond anything observed in other mammals. Intriguingly, many of the *PRDM9* copies in phyllostomid bats have lost the KRAB domain (**Table S9**), which is thought to play an important role in recruiting the recombination machinery[49], suggesting they may have alternate functions. Finally, we observed a major expansion in *P. mesoamericanus* of heat-shock proteins across multiple gene families, including heat-shock protein 70kDa (PTHR19375, +11 genes; Viterbi method, *p*=8.2E-14). Interestingly, this expansion was largely restricted to *P. mesoamericanus* and was not observed across other bats. Overexpression of heat-shock proteins can modulate immune responses[52], therefore this duplication may have implications for immunity in *P. mesoamericanus*.

Positive selection analysis

Having identified signatures of positive selection at the amino-acid level in several gene families of interest, we applied similar branch-site likelihood ratio tests[53–55] genome-wide, focusing on 12,517 single-copy orthologs present in bats and outgroup mammals. Because we were most interested in adaptations shared by most bats, we mainly focused on a test for positive selection on the branch leading to the bat MRCA, where we expected to have reasonably good statistical power. However, we also tested for positive selection on the lineages leading to each of the two newly sequenced bats. Observing a highly skewed distribution of nominal *p*-values (*p*>0.98 in 86% of tests), we opted to follow a recent study[56] and omit a correction for multiple comparisons across orthologs, instead adjusting only for testing on three different branches (but see **Table S10** and **Data S1** for more conservative adjusted *p*-values). Based on this strategy, we identified 468 positively selected genes (PSGs) with adjusted *p*<0.05 in the bat MRCA (**Figure 3, Table S10**). This number is roughly comparable to the 298 PSGs recently identified for the Phyllostomidae bat lineage using a similar filtering approach[56] but somewhat larger than the numbers found on the branch leading to the bat MRCA in recent studies that used more stringent criteria (181 PSGs[20] and 23 PSGs[18]). These PSGs were strongly enriched for immune-related functions (**Table S11**), including the major GO biological processes “regulation of inflammatory response” (GO:0050727, *p*=6.1e-4) and “innate immune response” (GO:0045087, *p*=6.2e-4). In total, 125 PSGs (27% of the 468) were annotated with the parent term “immune system process” (GO:0002376). The bat MRCA branch was also enriched for PSGs involved in “positive regulation of reactive oxygen species (ROS)” (GO:2000379, *p*=3.5e-4), possibly suggesting adaptations associated with heightened metabolic rates owing to flight (see **Discussion**). Moreover, we detected six autophagy-related PSGs (**Table S10**), including autophagy regulator *ATG9B* which was previously implicated in bat longevity[57]. Below, we further discuss some specific PSGs falling in two major classes: (1) pathogen sensors, cytokines, and antiviral genes; and (2) DNA repair genes and tumor suppressors.

Figure 3. **Positive selection scan on the bat ancestral branch.** a) Maximum likelihood phylogeny based on codon-site partitioned analysis of 3,632 gene alignments, with the bat ancestral branch indicated. b) TopGO hierarchical gene ontology enrichment analysis of the positively selected genes against the background set of tested genes, suggesting strong enrichment of innate immunity genes. The ten most significant GO terms are shown, eight of which are related to innate immunity including genes involved in interleukin regulation and interferon pathways.

Pathogen sensors, interleukins, and antiviral genes in bats are rapidly evolving

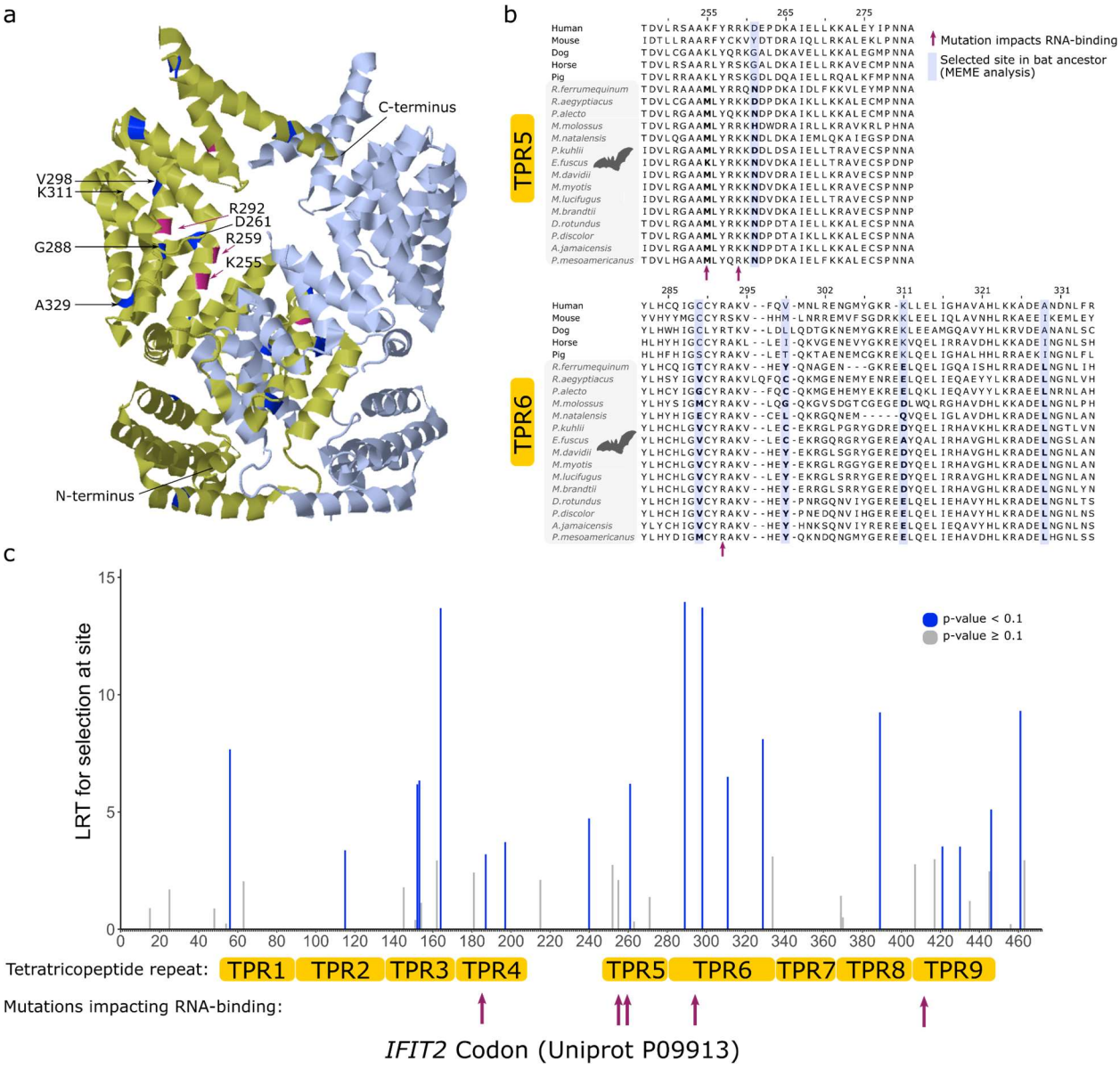
The 468 PSGs on the bat ancestral branch included several genes that encode proteins with pathogen-sensing roles. For example, the TLR-encoding genes *TLR7* and *TLR8*, are included in our set (**Table S10**), and the related *TLR9* (which shows reduced activation in bats[58]) was on the threshold of statistical significance (p=0.05). All of these genes have been identified in previous scans for positive selection[16,23]. Another previously identified PSG in our set is the IFN stimulator *STING*[59]. The TLRs and *STING*, as well as the PSG *NLRP3*[17,60], all play important roles in inflammation and are considered therapeutic targets for inflammatory disease[61–64]. Positive selection in these genes may play a role in dampening downstream responses to pathogens. Interestingly, the location of a bat-specific substitution in codon 358 of *STING* that was linked to dampened IFN activation[59] was identified in our scan as one of ten positively selected sites (**Table S7**). In addition to these previously identified PSGs, we identified *TLR2* to be under positive selection. Unlike the nucleic-acid sensing *TLRs 7*,*8,* and *9*, *TLR2* recognizes lipoproteins of pathogens such as bacteria and enveloped viruses[65]. *TLR2* signalling also induces the inflammatory cytokine *TNF-α* in infections with viruses like SARS-CoV-2, and blocking *TLR2* protects against pathogenesis caused by the 'cytokine storm'[66]. We additionally found evidence of positive selection in the key TLR regulator *UNC93B1*[67], which is regulated by type I IFNs[68]. *UNC93B1* is thought to regulate nucleic-acid sensing TLRs such as *TLR7* and *TLR9* upstream of the process of TLR trafficking from the endoplasmic reticulum to endolysosomes[69,70].

Another group prominently represented in our PSGs are genes encoding the interleukins, a collection of cytokines with diverse functions in immunity and inflammation[71,72]. In bats, earlier work identified reduced expression of *interleukin-8*[58], positive selection of *interleukin-1β,* and loss of *interleukin-36γ*[18], suggesting that rapid evolution of interleukins may have contributed to unique adaptations in bat immunity. We found that five interleukin-related GO categories were significantly enriched for PSGs in bats (**Table S11**), including "interleukin-1β production" (GO:0032611, p=0.00075) and "positive regulation of interleukin-6 production" (GO:0032755, p=0.00028). In addition, we identified the pleiotropic cytokine-encoding genes *interleukin-6* and *interleukin-15* as PSGs as well as the genes encoding several interleukin-associated receptors (**Table S10**). *Interleukin-6*, which encodes one of the most important cytokines during infection[73], showed six sites predicted to be positively selected in bats and a high ratio of nonsynonymous to synonymous mutations on the bat MRCA branch (dN/dS=2.72). Inflammatory cytokine-encoding genes such as *interleukin-6* and *interleukin-15* that were under positive selection in the bat MRCA could be additional contributors to dampened inflammation in bats.

Our PSGs also included several antiviral IFN-stimulated genes, which, as noted above, are activated by type I IFNs[41] and in several cases (including *APOBEC3* and *Mx*) have been shown to be duplicated[18] and/or under positive selection[25,74] in bats. For example, we detected positive selection in *PARP9*, whose gene product interacts with those of *DTX3L* and *STAT1* to enhance IFN responsiveness[75,76]. We also found strong evidence of positive selection, affecting at least 18 amino-acid sites, in *IFIT2*, whose product inhibits replication of a broad range of RNA and DNA viruses[77,78] (**Figure 4**). Ten of the positively selected sites overlap or are physically close to sites known to impact RNA-binding in the TPR4, TPR5, TPR6 and TPR9 motifs[79], and a bat-specific lysine-to-methionine substitution occurs at codon 255 which is involved in RNA-binding[79] (**Figure**

231 **4**). Positive selection of *IFIT2* in bats may therefore alter, and possibly enhance, expression of
232 numerous antiviral response genes. Taken together, our results provide novel evidence for rapid
233 evolution in the innate immune system of bats (**Figure 5**).

234



235

236 **Figure 4**. **The antiviral gene *IFIT2* is positively selected in RNA-binding regions in bats.** a)
237 3D-structure of the IFIT2 protein (PDB:4G1T) showing sites selected in the bat ancestor (blue)
238 and sites known to be involved in RNA-binding function of the protein (purple). b) Amino acid
239 alignment of the tetratricopeptide repeat (TPR) regions TPR5 and TPR6 (based on UniProt
240 annotation for P09913) showing sites selected in the bat ancestor (blue highlighting and bold) and
241 a fixed bat lysine-to-methionine substitution in codon 255 which is located in a region involved in

RNA-binding (indicated with purple arrow). c) MEME[105] analysis of *IFIT2* codon sites showing 18 sites selected in bats based on a likelihood ratio test.

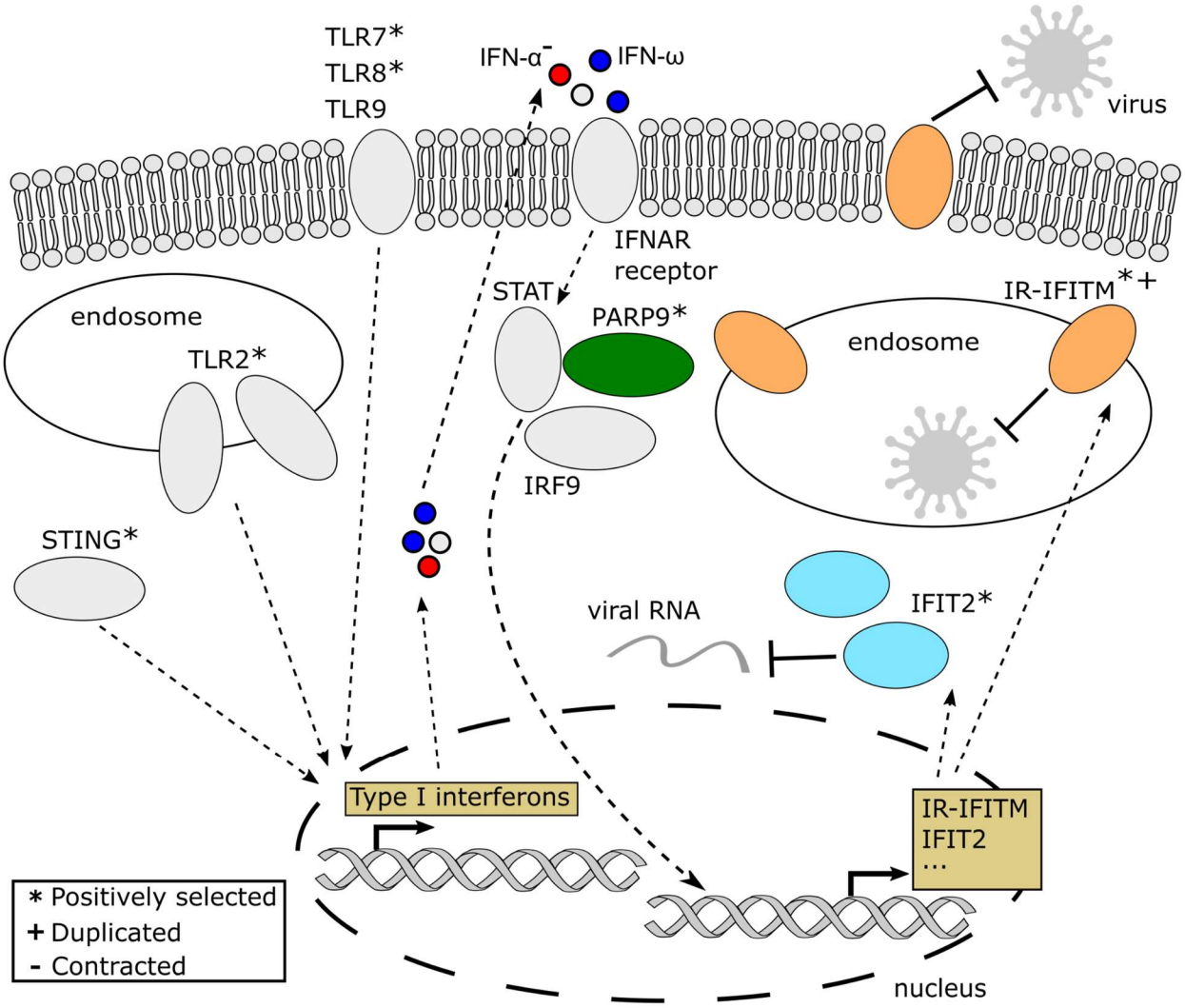


Figure 5. **Schema of cellular innate immunity processes associated with genes rapidly evolving in bats.** Proteins shown in colour are the most significant innate immunity proteins highlighted in this study. Pathogen sensing pathways involving toll-like receptors (TLRs) and *cGAS-STING* can induce expression of type I interferons (IFNs) including IFN-α and IFN-ω. IFN-α genes were lost in the bat ancestor, potentially giving a greater role to IFN-ω. The type I IFNs trigger the induction of IFN-stimulated genes via pathways including *STAT*, which interacts with the positively selected *PARP9*. The IFN-stimulated genes include the positively selected *IFIT2* gene and the immune-related (IR) IFITM genes duplicated in yangochiropteran bats, with both *IFIT2* and IR-IFITM genes playing prominent roles in antiviral defences. The overall schema is based on reviews of type I interferons, PARP9, IFIT and IFITM proteins[78,80] and studies on IFITM interactions with RNA viruses[81,82].

DNA repair genes and tumor suppressors are positively selected in bats

Enhanced DNA repair has been proposed as a mechanism for longevity and cancer resistance in various mammals including bats[27,83]. We identified six DNA repair-related PSGs and 46 PSGs that were "cancer-related" (**Table S10**), meaning they were included in either the Tumor Suppressor Database[84] or the Catalogue Of Somatic Mutations In Cancer[85]. Notably, cancer-related genes were enriched more than two-fold among PSGs on the bat ancestral branch relative to a set of mammalian branches (Fisher's exact test: p=9e-3, odds ratio=2.2). Among DNA repair genes, we detected evidence of positive selection in the tumor suppressor-encoding *PALB2* and in four DNA polymerase-encoding genes (*POLA1*, *POLD1*, *POLK,* and *POLM*). *PALB2* is a crucial component of the BRCA complex and is required for homologous recombination repair[86–88]. In bats, it shows three sites under selection as well as seven bat-specific coding indels, including a 21-nucleotide deletion in a *RAD51*/*BRCA1*-interacting region (Uniprot annotation: Q86YC2). Despite previous evidence in a long-lived bat[17], we did not find a signal for selection in *BRCA2*, although it does contain 14 bat-specific indels (**Data S2**). Similarly, we did not identify the DNA repair genes *RAD50* and *KU80* (cf. ref.[17]) or the key tumor suppressor-encoding *TP53* as PSGs, but we did find four sites (codons 35,38,54,97) in *TP53* that are potentially selected in bats as well as a previously described bat-specific indel in the nuclear localization signal domain[17] (**Figure S7**).

While *TP53* did not appear among our PSGs, we did identify genes encoding two other tumor suppressors that interact with it (**Table 1**): *BIK* and *LATS2*. Both genes showed highly significant signals of selection in bats in our data set but have not been identified in tests of other mammalian branches[89,90] or in earlier studies in bats[18,20]. *LATS2* is a kinase that modulates the functions of tumor suppressors such as *TP53* as well as canonical growth-related Hippo signaling effectors *YAP*/*TAZ*[91]. We found that *LATS2* is predominantly under negative selection (bat MRCA dN/dS =0.23, outgroup mammals dN/dS=0.11) but nevertheless contains 13 selected nonsynonymous substitutions with evidence of selection in the bat MRCA as well as three bat-specific deletions in its coding region (**Figure S8, Data S2**). Four of the substitutions fall within known functional domains of the protein, with a codon 134 glutamic-acid-to-aspartic-acid substitution in the LATS conserved domain 1 (LCD1) predicted to have an effect on the protein based on SNPeffect[92]. Previous experimental work in mice has shown that the LCD1 domain is critical for tumor suppressor activity of *LATS2*[93]. The second tumor suppressor, *BIK*, belongs to the pro-apoptotic BH3-only family of proteins that are upregulated in response to various stress signals and act as antagonists of pro-survival proteins[94,95]. *BIK* is regulated by *TP53*[96,97] and contributes to the apoptotic response induced by chemotherapy treatments[94,98]. The rapid molecular evolution of *LATS2* and *BIK* suggests that these genes may contribute to bat adaptations in tumor suppression.

The two cancer-related genes displaying the strongest signals of positive selection, *CDH1* and *CAT* (**Table 1**), also play roles in tumor development. *CDH1* belongs to a family of transmembrane glycoproteins that mediate cell-cell adhesion and regulate cell growth, making them therapeutic targets for preventing tumor progression[99]. *CAT* is an antioxidant enzyme that can protect from ROS-induced stress[100] and plays a role in *TP53*-mediated ROS regulation in response to DNA damage[101]. Although these functions make *CAT* and *CDH1* intriguing examples of genes that

have potentially evolved to enhance bat cancer resistance, previous studies have also found a signal for selection in *CAT*[102,103] and *CDH1*[102–104] in other mammals. Experimental validation of the bat-specific mutations detected here will thus be important to demonstrate their potential adaptive value.

Table 1. Positively selected genes involved in cancer (*p*<1.0e-3 with at least one site selected in the bat MRCA) that showed the strongest statistical significance. The *p*-values shown (aBSREL *p*-value) are derived from the branch-site likelihood ratio test and adjusted for multiple comparisons across branches but not across genes (see text). For contrast, we also show *p*-values from a similar test (M2a vs. M1a implemented in PAML) applied to a set of mammalian branches in studies by ref.[89] and ref.[104]. For ref.[89], nominal *p*-values are shown, with nominal *p*<1.1e−3 corresponding to FDR<0.05; for ref.[104], adjusted *p*-values are shown, with adjusted *p*<0.05. Missing values are indicated by dashes.

Symbol	Name	aBSREL *p*-value	Sites under selection*	Bat branch dN/dS	Outgroup branch dN/dS	Ref.[34] *p*-value	Ref.[35] *p*-value
CDH1	*cadherin-1*	2.0e-6	3	0.17	0.29	1	2.7e-2
CAT	*catalase*	7.0e-6	1	0.38	-	1	-
BIK	*BCL2 interacting killer*	3.7e-5	1	1.92	0.89	1	-
PALB2	*partner and localizer of BRCA2*	1.8e-4	3	0.67	0.43	1	6.0e-13
LATS2	*large tumor suppressor kinase 2*	2.5e-4	10	0.23	0.11	-	-
SLC39A4	*solute carrier family 39 member 4*	3.8e-4	3	0.55	0.42	0.11	-
SPARCL1	*SPARC like 1*	4.0e-4	4	0.88	0.95	1	-
PLA2G7	*phospholipase A2 group VII*	8.5e-4	5	-	1.70	1	-
GALR1	*galanin receptor 1*	8.7e-4	1	0.10	0.08	1	-
CD79A	*CD79a molecule*	9.0e-4	1	-	0.37	5.4e-9	9.5e-15
MYO1A	*myosin IA*	9.8e-4	7	0.94	0.14	1	-

*Sites with evidence of positive selection in the bat MRCA were identified using site-wise MEME[105] analysis with a significance threshold of *p*<0.1

Discussion

High-quality and complete genome sequences are indispensable for revealing patterns of genomic variation within and between species. In this study, we used long-read sequencing to assemble the genomes of the bats *A. jamaicensis* and *P. mesoamericanus* and analysed them together with 13 additional bat genomes to provide insights into unique bat adaptations. We found evidence of genomic adaptation in three key components of the bat innate immune system, including pathogen sensing, type I IFN cytokine signalling, and IFN-stimulated antiviral genes

(see schema in **Figure 5**). In particular, the loss of IFN-α genes and the potentially increased reliance of bats on IFN-ω may play a role in their tolerance of viral infections. The expansion of antiviral IR-IFITM genes and *PRDM9* in major bat clades may have further contributed to lineage-specific adaptations. We also found evidence of positive selection in the bat MRCA in 46 cancer-related genes, suggesting a possible link to the unusually low incidence of cancer in bats.

Perhaps our most striking finding, building on earlier comparative genomic studies of the complex type I IFN locus[12,24], is that eight type I IFNs were lost in the bat MRCA. Notably, we found that bats have lost most, or—in the case of *Pipistrellus kuhlii*, *M. myotis* and *P. mesoamericanus* —all of their IFN-α genes, making their type I IFN locus particularly distinct among mammals. The overall contraction of the type I IFN locus in bats has been hypothesized to allow a smaller number of constitutively expressed IFN-α genes to perform the functions of the 13 IFN-α genes in humans[24]. The constitutive expression and rapid evolution of bat IFNs may also partly result from dampened inflammation caused by evolutionary changes to inflammosome genes such as *AIM2*, *caspase-1* and *IL-1β*[106], because inflammosomes may negatively regulate type I IFN sensors[107]. Indeed, the regulation of type I IFNs in bats also appears to have evolved adaptively, as positive selection of the key transcription factor *IRF3* in bats was previously shown to enhance antiviral responses via type I IFN activation[108]. However, our results suggest that constitutive expression of IFN-α is not common to all bats, in line with expression analyses in *R. aegyptiacus*[12]. We hypothesize that by relying on the potentially more potent IFN-ω rather than IFN-α, bats may further enhance their antiviral responses. Although further work will be needed to demonstrate a functional shift to IFN-ω, the lack of any IFN-α genes in three bat species strongly suggests a shift has occurred at least in these cases. It is possible that an enhanced antiviral response owing to IFN-ω helps to compensate for an overall dampened inflammatory response in bats. If these properties of IFN-ω can be established, they may open the door to new therapeutic uses of IFN-ω[109,110].

In addition, our findings suggest that, compared to other parts of the mammalian phylogeny, the lineage leading to bats is enriched for positively selected cancer-related genes. Rapid evolution of DNA repair genes as well as tumor suppressors has been proposed as a mechanism for cancer resistance in other long-lived mammals such as whales[111]. Intriguingly, it was previously hypothesized that selection of cancer resistance components such DNA repair genes in bats resulted from a need to reduce the negative effects of ROS generated as a consequence of flight[17]. This suggestion appears consistent with our findings of positively selected ROS regulators and DNA repair genes. The evolution of cancer resistance in bats may also be associated with adaptations in the bat immune system. There is substantial overlap of genes related to cancer and the immune system[112], with immune-related genes being known to play a role in cancer surveillance[113] and tumor suppression[114–116]. Specific examples include immune-related genes that are rapidly evolving in bats. For example, ROS plays an important role in *NLRP3* inflammasome activation[117,118], *IFIT2* plays a role in apoptosis and cancer progression[119] and IFNs contribute to anti-tumor activity[120,121]. Positive selection in such genes may be driven by fitness trade-offs between roles in immunity and cancer[112]. For example, inhibition of inflammation can promote longevity[122,123] and suppress tumor growth[124,125]**.** Comparative analyses of gene expression across mammals and experimental validation may help resolve the different roles of ROS regulation, DNA repair, inflammation and immunity in cancer resistance. We anticipate that

366 our initial evolutionary findings and the novel genomic resources we have made available
367 (including a genome browser: http://compgen.cshl.edu/bat/) will encourage and facilitate further
368 genomic research in bats, particularly as models that can lead to new strategies for addressing
369 major challenges to human health such as COVID-19 and cancer.

370 Acknowledgements

371 This research was supported, in part, by US National Institutes of Health (NIH) grants P30-
372 CA045508 (to D. Tuveson and colleagues) and R35-GM127070 (to A. Siepel), and by the Simons
373 Center for Quantitative Biology. Computational work was performed with assistance from NIH
374 Grant S10OD028632-01, and the genome browser was set up with help from Ritika Ramani. We
375 thank the CSHL Cancer Center and Director David Tuveson for support and encouragement. We
376 would further like to acknowledge funding support from the CSHL/Northwell Health Affiliation for
377 purchase of the ONT PromethION sequencer used in this study. We also thank Semir Beyaz for
378 helpful comments on the manuscript. Finally, we thank Brock Fenton, Neil Duncan, the staff at
379 the Lamanai Field Research Center, and many other colleagues who assisted with the fieldwork
380 necessary for this study. The content is solely the responsibility of the authors and does not
381 necessarily represent the official views of the NIH or other funding sources.
382

383 Data availability

384

385 All raw sequencing data and final genome assemblies have been made available via NCBI
386 (PRJNA751559). Gene annotations can be downloaded or viewed as UCSC genome browser
387 tracks from our laboratory website (http://compgen.cshl.edu/bat/). Supplementary tables and data
388 are available via FigShare (https://doi.org/10.6084/m9.figshare.15223014.v1).

Materials and methods

Sample background and acquisition

Fresh liver samples from a single individual of *A. jamaicensis* (AMNH 279493, male) and one *P. mesoamericanus* (AMNH 279536, male) were collected by NBS in April 2017 at the Lamanai Archaeological Reserve in Orange Walk District, Belize (17.75117°N, 88.65446°W). Sampling followed best practices for humane capture, handling, and euthanasia of live mammals outlined by the American Society of Mammalogists[126]. All work was conducted with permission of the Belize Forest Department under permit number WL/2/1/17(19) with IACUC approval from the American Museum of Natural History (AMNHIACUC-20170403) and University of Georgia (A2014 04-016-Y3-A5). Bats were captured in ground-level mist nets and placed in individual cloth bags for transport to the Lamanai Field Research Center. After identification, the bats were euthanized using isoflurane, and the liver was removed immediately after death. Samples were placed in multiple individual 2ml cryotubes and flash-frozen by placement in a liquid nitrogen dry shipper. The cold chain was maintained through shipment to the AMNH, storage in the AMNH Ambrose Monell Cryo Collection, and subsequent sample processing and transfers.

Genome sequencing and assembly

Approximately 40 mg of liver tissue from each bat was received at CSHL and stored at -80C. The liver samples were crushed with a micro pestle and mixed with 10 ml of TLB buffer and 50µl of RNase A immediately before use. After a one-hour incubation at 37C, 50 µll of Proteinase K was added, followed by incubation at 50C for 3 hours with hourly inversion mixing. After addition of 10ml of a phenol chloroform / isoamyl alcohol mixture, each sample was rocked for 10 minutes and then centrifuged at 4500RPM for 10 minutes. The top aqueous layer was retained and an equal volume of chloroform / isoamyl alcohol was added, and rocking and centrifugation was repeated as above. The top layer was transferred to a fresh tube with 4ml of 5M ammonium acetate. Following the addition of 30ml of ice cold 100% ethanol, the sample was rocked for 10 minutes. The visible DNA was then extracted with a glass pipet and placed in a 1.5ml tube. The sample was washed once with 100% ethanol and centrifuged for 5 minutes at 10,000 RPM. The ethanol was removed and any remaining ethanol was evaporated on a 37C heat block for 10 minutes. The final DNA was resuspended in 10mM Tris HCl pH 8.5 and stored overnight at 4C.

DNA was then sheared to ~50-75kb using a Diagnode Megarupter following manufacturer's recommendations. DNA was further enriched for long fragments via the Circulomics small read eliminator XL kit, which iteratively degrades short fragments. DNA was prepared for Nanopore sequencing using the ONT 1D sequencing by ligation kit (SQK-LSK109). Briefly, 1-1.5ug of fragmented DNA was repaired with the NEB FFPE repair kit, followed by end repair and A-tailing with the NEB Ultra II end-prep kit. After an Ampure clean-up step, prepared fragments were ligated to ONT specific adapters via the NEB blunt/TA master mix kit. The library underwent a final clean-up and was loaded onto a PromethION PRO0002 flowcell per manufacturer's instructions. The flowcells were sequenced with standard parameters for 3 days. Basecalling was

performed in real time with Guppy 3.2. Nanopore reads were filtered for minimum length of 10 and minimum 85% accuracy using filtLong (http//github.com/rrwick/Filtlong). Illumina short read libraries were prepared from the same tissue as above with the Illumina TruSeq DNA kit, targeting a 550bp insert size with PCR enrichment.

Libraries were sequenced at the New York Genome Center, on a NovaSeq S4 flowcell in paired end 150bp format to ~30x genome coverage. Reads were assembled using flye 2.8.3[33], after evaluation of additional assemblies generated with wtbg2[127], NextDenovo (https://github.com/Nextomics/NextDenovo) and Shasta[128] (**Table S12**). Assembly was followed by long-read polishing using minimap 2.17[129] for alignment and PEPPER 1.0[34] for polishing. Next, bwa-mem 0.7.17-r1188[130] was used to align the Illumina short-read data to the long-read polished assembly, and short-read polishing was carried out with POLCA[35]. To compare assembly contiguity, error rates and completeness, assemblies were then assessed using Merqury 1.0[131], as well as with Benchmarking Sets of Universal Single-Copy Orthologs 4.0.5 (BUSCO)[132] for mammals (odb9). Additionally, we used python scripts to compute the cumulative sum of contigs length (N(X) length) versus the cumulative sum of N(X)% of the total genome. Further assembly statistics were calculated using BBTools 38.86 (http//sourceforge.net/projects/bbmap/). Duplicated haplotypes in the assemblies were purged using purge_dups 1.2.3[133].

The assemblies were aligned to human, mouse, and pig assemblies as well as to four bat assemblies (*M. myotis*, *P. discolor*, *D. rotundus,* and *R. ferrumequinum*) using Cactus[134]. Bat-specific small indels (<1kb) were called relative to the human reference from the multiple alignment using a custom python script. Indels were called in alignment blocks with at least 10 bases and seven species aligned including at least 2 non-bat mammals. Only fixed indels present in all bats or all bats but one were retained. Bat-specific insertions and deletions relative to the human reference were distinguished based on the sequence of the pig outgroup.

Gene annotation

Public RNA-seq data from SRA were downloaded for *A. jamaicensis* and *Pteronotus parnelli* (**Table S13**) and proteins for human (GCF_000001405.39), mouse (GCF_000001635.26), and seven bat species (*Myotis brandtii*, GCF_000412655.1; *M. davidii*,GCF_000327345.1; *M. lucifugus*, GCF_000147115.1; *Phyllostomus discolor*, GCF_004126475.1; *Pteropus alecto*, GCF_000325575.1; *Rhinolophus ferumequinum*, GCF_004115265.1; *R. aegyptiacus*, GCF_001466805.2) were downloaded from RefSeq. RNA-seq reads were aligned to the new reference genomes using HISAT 2.2.0[135] with the parameters "--no-mixed --no-discordant --downstream-transcriptome-assembly" and transcripts were assembled using StringTie 2.1.1[136]. To reduce potential loss of transcripts due to limitations of short-read alignment, transcriptomes were also *de novo* assembled with Trinity 2.9.1[137]. PASA 2.4.1[138] was used to generate the final set of transcripts based on alignment with GMAP[139] and BLAT[140] using minimum thresholds of 90% of transcript length aligned at 90% identity. GeMoMa was used to project gene annotations from six bat assemblies[18] to the genomes of *A. jamaicensis* and *P. mesoamericanus*. Transdecoder was used to predict coding sequences within the assembled transcripts[141]. We used the transcripts and proteins sequences as evidence for the MAKER3 annotation pipeline[142] with the *ab initio* gene predictors SNAP 2006.07.28[143] and Augustus 3.3.3[144]. GlimmerHMM[145]

and GeneMark 4.68[146] were used to generate further *ab initio* gene predictions. Final annotations were generated with Evidence Modeler[147]. To reduce the number of potentially missing genes caused by lack of protein and RNA-seq evidence, we used intact TOGA (https://github.com/hillerlab/TOGA) gene projections from the recently assembled *A. jamaicensis* (GCF_014825515.1) to add gene annotations that did not intersect the Evidence Modeler annotations. The completeness of the final predicted protein set was assessed using BUSCO.

Repeat analysis

Repeat masking was carried out using an iterative masking and *de novo* repeat detection approach. After masking repeats with Repeatmasker 4.0.9[148] and the combined RepBase-20181026 and Dfam-3.1 repeat databases, novel repeats were detected in the masked genomes with RepeatModeler 2.0.1. The consensus de novo repeats longer than 100 bp were then concatenated with a vertebrate repeat library including novel bat repeats[18] and clustered with CD-HIT 4.81[149]. All clustered novel sequences with >80% sequence similarity across >80% of the length of the clustered sequences were excluded[150]. Novel repeats were then aligned to the nt database using BLAST and all repeats matching annotated transcripts were removed. A further BLAST analysis was used to exclude repeats with fewer than 10 alignments to the reference genome from which they were derived. Transposable elements were classified and assigned families using TEclass[151] and DeepTE[152]. The consensus *de novo* repeats were then concatenated with the vertebrate repeat library and a final masking of the genome was carried out with RepeatMasker using the 'sensitive' setting. The recently diverged repeat landscape was analysed using the RepeatMasker script calcDivergenceFromAlign.pl with correction of substitution rates based on the Kimura 2-Parameter model. Transposons were considered recently diverged at 7% divergence from the consensus[18], which approximates an insertion <30 mya, assuming a mammalian substitution rate of 2.2×10^{-9} [153]. To allow comparison of the repeat landscape within the Noctilionoidea bat clade, the same repeat masking approach was carried out for several closely related bats (*D. rotundus*, *P. discolor*, S. *hondurensis* and *M. myotis*).

Mining of endogenous viral elements

To scan for endogenous viral elements that might provide evidence of past infections, we aligned the complete RVDB-prot viral protein database[154] to our genomes using BLAST with an e-value threshold of 1e-5. Alignments intersecting coding sequence and those shorter than 100aa were excluded. Non-retroviral matches were aligned to the NR protein database using blastx with an e-value threshold of 1e-5 and TaxonKit 0.6.0[155] was used to retain sequences with a best match to a viral lineage.

Identification of orthologous gene groups

We analysed bat genes for lineage-specific signals of positive selection and gene duplications based on clustered ortholog groups. Coding sequences for 13 bats and five outgroup mammals (human, mouse, dog, pig, and horse) were downloaded from RefSeq and the Hiller lab (https://bds.mpi-cbg.de/hillerlab/Bat1KPilotProject/). Annotated open reading frames that did not show a nucleotide number that was a multiple of three or that contained internal stop codons were

discarded. The longest isoform for each gene was retained. All proteins were clustered with the proteins of our bats using OrthoFinder 2.3.11[156]. A set of single copy orthologs was extracted from the OrthoFinder results, retaining orthologs with at least 12 bat species and three mammalian outgroups. Genes were then aligned using PRANK v.170427[157] with the species tree provided to guide the alignment. Cancer-associated ortholog clusters were identified based on the databases TSG[84] and COSMIC[5].

Gene family expansion analysis

Alignments of 3,632 single copy genes that were present in all species were concatenated into a single alignment, which was divided into three partitions corresponding to codon positions. A maximum-likelihood phylogeny was inferred using RAxML 8.2.12[158] rapid bootstrapping under the GTR+G+I model with 100 bootstraps. A dated phylogeny was then generated using MCMCtree in PAML. Node ages were calibrated based on TimeTree[159] ages for Euarchontoglires, bats, Yangochiroptera and Yinpterochiroptera. Convergence was assessed based on analysis of two replicate runs with tracer[160]. The dated phylogeny was used to calculate gene family expansions and contractions with CAFE 4.21[38] based on a *p*-value threshold of 0.05. Orthofinder orthogroups were collapsed into 7,405 PANTHER gene families based on BiomaRt[161] data and 1,448 unassigned Orthogroups. For each of 19,935 orthogroups, a representative gene was selected to obtain a PANTHER assignment. For 2,466 orthogroups, where no human or mouse gene was represented, we selected a gene from a different species for sequence-based annotation using interproscan and eggnog. As CAFE assumes all genes were present in the common ancestor of all analysed species, orthogroups represented in less than two bats and two outgroup mammals were excluded. Final copy numbers and genomic coordinates for IFITM (**Table S14**) and type I IFN (**Table S15**) genes were visualised using genoPlotR[162]. For the analysis of the *PRDM9* expansion in phyllostomids, recently generated short-read assemblies of Sturnira hondurensis (GCA_014824575.2) and A. jamaicensis (GCA_014825515.1) were included in the phylogenetic analysis (**Figure S6**) but not the CAFE analysis. Sequences and RAxML phylogenies are provided for IFITM, IFN, and *PRDM9* orthogroups (**Data S3**).

Positive selection analysis

We aimed to determine whether each gene was positively selected in one or more of three groups: the bat MRCA, *P. mesoamericanus* and *A. jamaicensis*. We also detected positive selection using the adaptive branch-site random effects model (aBSREL) method implemented in HyPhy 2.5.12[55]. A likelihood ratio test was used to determine whether a lineage-specific group of codons in the alignment is experiencing significant positive selection. Unlike the branch-site tests implemented in PAML[163], aBSREL allows rate variation among the background branches, which can reduce false positive errors[55]. In addition, we also used PAML 4.9 to apply Test 2, a branch-site test of positive selection. This test compares the alternative model where some branches are under positive selection and thus exhibit sites with $\omega>1$ with the corresponding null model where omega is fixed as 1. We computed *p*-values according to a χ^2 distribution with one degree of freedom. The *p*-values calculated by PAML and HyPhy were adjusted for multiple testing of three branches per gene using the Benjamini-Hochberg method for controlling false discovery rate (FDR) implemented in base R. We also corrected for multiple testing of all genes using FDR (**Table**

S10). We used the maximum-likelihood species tree inferred using 3,048 orthologs (see **Methods** section 'Gene family expansion analysis') to provide the topology for the positive selection scans. In this tree, bats are the sister group to Fereuungulata (Cetartiodactyla, Perissodactyla, Carnivora, and Pholidota)[18]. Relationships within Fereungulata remain challenging to resolve[18,164] and here we followed the TimeTree[159] topology that places Perissodactyla+Carnivora as sister to Cetartiodactyla. For comparison with a previous scan for positive selection on all branches of a mammalian phylogeny (including human, chimpanzee, macaque, mouse, rat, and dog), we used the set of 8,594 genes that were tested for positive selection using PAML in both this study and in ref[89]. MEME[105] analysis was used to identify sites potentially under selection in the bat MRCA with a significance threshold of $p<0.1$. Alignment of genes positively selected in the bat MRCA are provided in **Data S4**.

Gene ontology enrichment

GO, Reactome and KEGG annotations for genes were obtained via BiomaRt. A total of 18,452 orthogroups were annotated with at least one feature and 17,840 were assigned GO features. We carried out enrichment analysis on groups of genes that were positively selected or that showed gene family expansions. Enrichment analysis was performed using topGO 2.44[165], with the elim algorithm and Fisher's exact test ($p<0.01$). All genes tested for selection were used as the background. The elim algorithm is a conservative approach that processes the GO graph from the bottom up, to eliminate higher order terms that would otherwise appear enriched due to correlation with more specific terms. GO terms with fewer than 10 genes annotated were excluded.

Supplementary tables

Supplementary tables are available online (https://doi.org/10.6084/m9.figshare.15223014.v1).

Table S1. GenBank assemblies for the taxonomic order Chiroptera (bats). List generated on 4th April 2021.

Table S2. Comparison of assembly statistics of the unscaffolded genomes of *Artibeus jamaicensis* and *Pteronotus mesoamericanus* with five scaffolded bat genomes assembled using a range of technologies.

Table S3. Recently diverged repetitive elements in the genomes of five noctilionoid bats and the related bat *Myotis myotis*. Repeats were detected and Kimura divergence from the consensus was calculated using RepeatMasker with a custom bat repeat database (available from http://compgen.cshl.edu/bat/).

Table S4. Non-retroviral endogenous viral elements (EVEs) detected in the genomes of Artibeus jamaicensis and Pteronotus mesoamericanus. EVEs were identified based on tblastn alignments of the reference viral database (https://rvdb-prot.pasteur.fr/) to the genomes, followed by filtering of sequences <100aa and retroviral sequences. Bat sequences with a best hit to a viral protein or EVE in the NCBI nr protein database using blastp were retained.

Table S5. Significantly expanded or contracted gene families in the bat lineage detected using CAFE. Ortholog clusters were assigned to PANTHER gene families based on human protein annotations from biomaRt and gene family descriptions were obtained from pantherdb.org. CAFE results for the Yangochiroptera clade and for *Artibeus jamaicensis* and *Pteronotus mesoamericanus* are also shown.

Table S6. Fisher's exact test using a 2-by-2 contingency table for interferon-α and interferon-ω copy number in a set of outgroup mammals and in individual bat species.

Table S7. Significantly selected sites in the genes IFITM3, IFIT2 and STING in the bat lineage based on MEME analysis. Codon positions are based on the human reference sequence. Protein mutagenesis and domain information was extracted from uniprot.org.

Table S8. Significantly expanded or contracted gene families in the bat subclade Yangochiroptera and in the species *Artibeus jamaicensis* and *Pteronotus mesoamericanus* detected using CAFE. Ortholog clusters were assigned to PANTHER gene families based on human protein annotations from biomaRt and gene family descriptions were obtained from pantherdb.org.

Table S9. Domain analysis of all protein sequences in the *PRDM9* ortholog cluster for seventeen bats and four other mammals. Protein sequences were analysed using PfamScan (https://www.ebi.ac.uk/Tools/pfa/pfamscan/). There is no *PRDM9* ortholog present in the dog. Protein IDs for genomes not generated in this study are based on the GenBank annotations and annotations from https://bds.mpi-cbg.de/hillerlab/Bat1KPilotProject/.

Table S10. Positively selected single copy orthologs in bats (order Chiroptera). Gene annotations are based on biomaRt data for the human protein, or, when not available, the protein of mouse or another mammal. Published p-values for the same genes were extracted based on publicly available data from eight studies. Branch-site likelihood ratio test *p*-values are shown for aBSREL and PAML M2a tests for the bat lineage as well as for *Artibeus jamaicensis* (Ajamaicensis) and

Pteronotus mesoamericanus (Pmesoericanus). Uncorrected *p*-values are shown as well as *p*-values adjusted with FDR for multiple tested branches and for multiple tested branches and genes.

Table S11. TopGO gene ontology enrichment analysis of positively selected genes in the bat ancestor. The 'elim' algorithm was used and GO enrichments with $p<0.01$ and >=10 annotated genes were retained for further analysis.

Table S12. Comparison of assembly and polishing statistics using different approaches. Polished assemblies were assessed using BUSCO v4 (Mammalia odb9) in genome mode and Merqury v1.3. RNAseq reads were aligned using hisat2.

Table S13. SRA accession numbers for RNAseq data used in the annotation of *Artibeus jamaicensis* and *Pteronotus mesoamericanus*.

Table S14. Genomic coordinates of the genes at the IFITM locus in nine bats and five other mammals. Coordinates for genomes not generated in this study are based on GenBank annotations.

Table S15. Genomic coordinates of the genes at the type I interferon locus in 10 bats and five other mammals. Coordinates for genomes not generated in this study are based on GenBank annotations.

Supplementary data

Supplementary data are available online (https://doi.org/10.6084/m9.figshare.15223014.v1).

Data S1. Positive selection analysis of orthogroups in bats (order Chiroptera). All orthogroup clusters generated by Orthofinder are shown, including multicopy orthogroups and lineage-specific orthogroups, but only single-copy orthogroups were analysed for positive selection. Gene annotations are based on biomaRt data for the human protein, or, when not available, the protein of mouse or another mammal. Published *p*-values for the same genes were extracted based on publicly available data from eight studies. Branch-site likelihood ratio test p-values are shown for aBSREL and PAML M2a tests for the bat lineage as well as for *Artibeus jamaicensis* (Ajamaicensis) and *Pteronotus mesoamericanus* (Pmesoericanus). Uncorrected p-values are shown as well as *p*-values adjusted with FDR for multiple tested branches and for multiple tested branches and genes.

Data S2. Gene-associated microindels detected in five or more bats from a Cactus multiple alignment of six bats with human, mouse and pig. The bats included in the alignment were *D. rotundus*, *A. jamaicensis*, *P. mesoamericanus*, *P. discolor*, *M. myotis* and *R. ferrumequinum*. Only indels occurring within coding sequences, within 1 kilobase of coding sequence or within the core promoter (based on human core promotors downloaded from https://epd.epfl.ch//index.php) were included. Coordinates are based on the human hg38 assembly and coordinates of aligned sequence from *A. jamaicensis* and *P. mesoamericanus* are also provided.

Data S3. Coding sequences and RAxML phylogenies for the type I interferon, interferon-induced transmembrane and PRDM9 orthogroups.

Data S4. Unfiltered PRANK alignments of all single-copy orthologs where positive selection was detected on the branch leading to the bat most recent common ancestor.

Supplementary figures

Figure S1. **Photographs of the bats *A. jamaicensis* and *P. mesoamericanus* sequenced in this study.** Photographs were provided by Brock & Sherri Fenton.

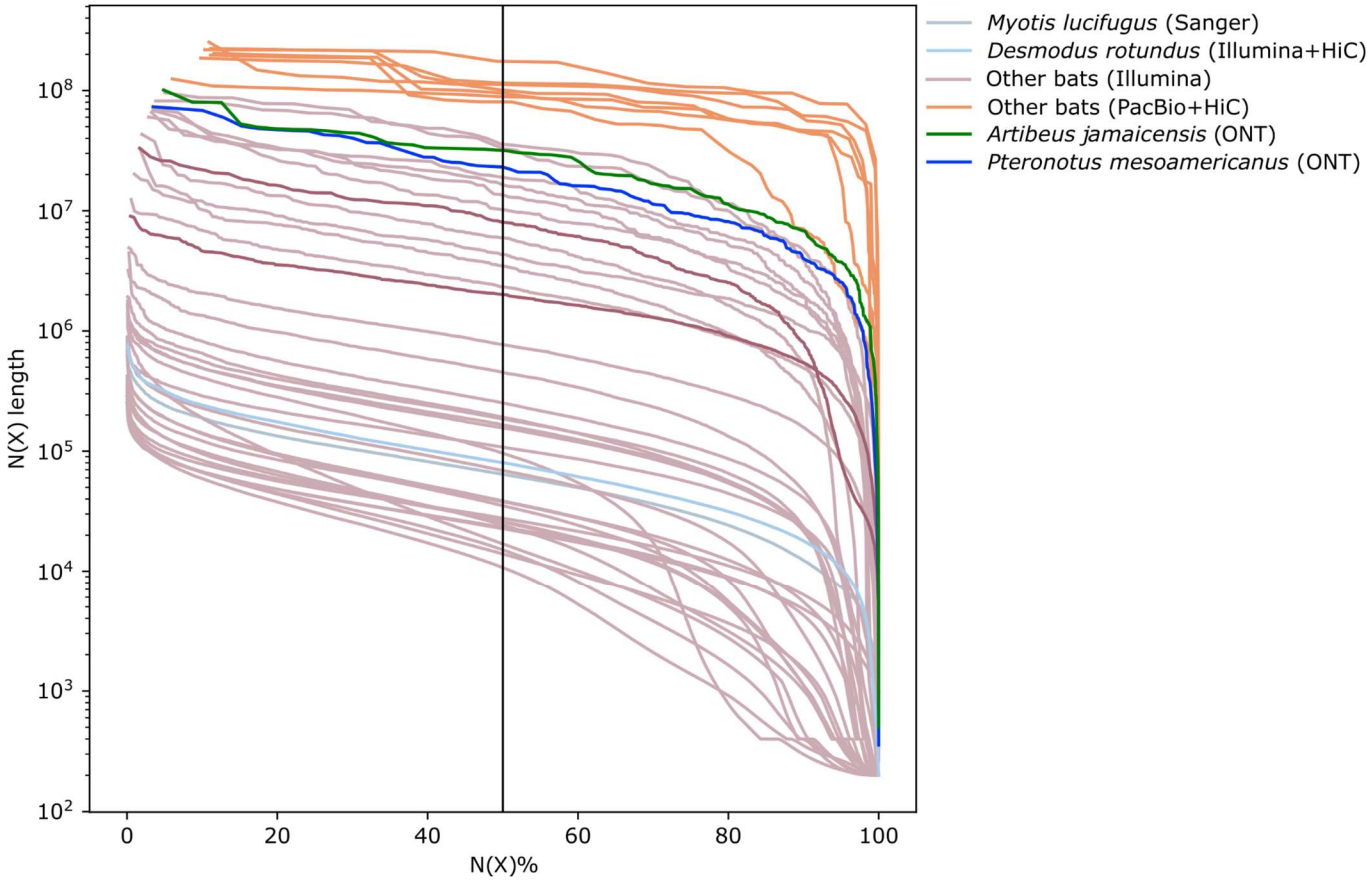
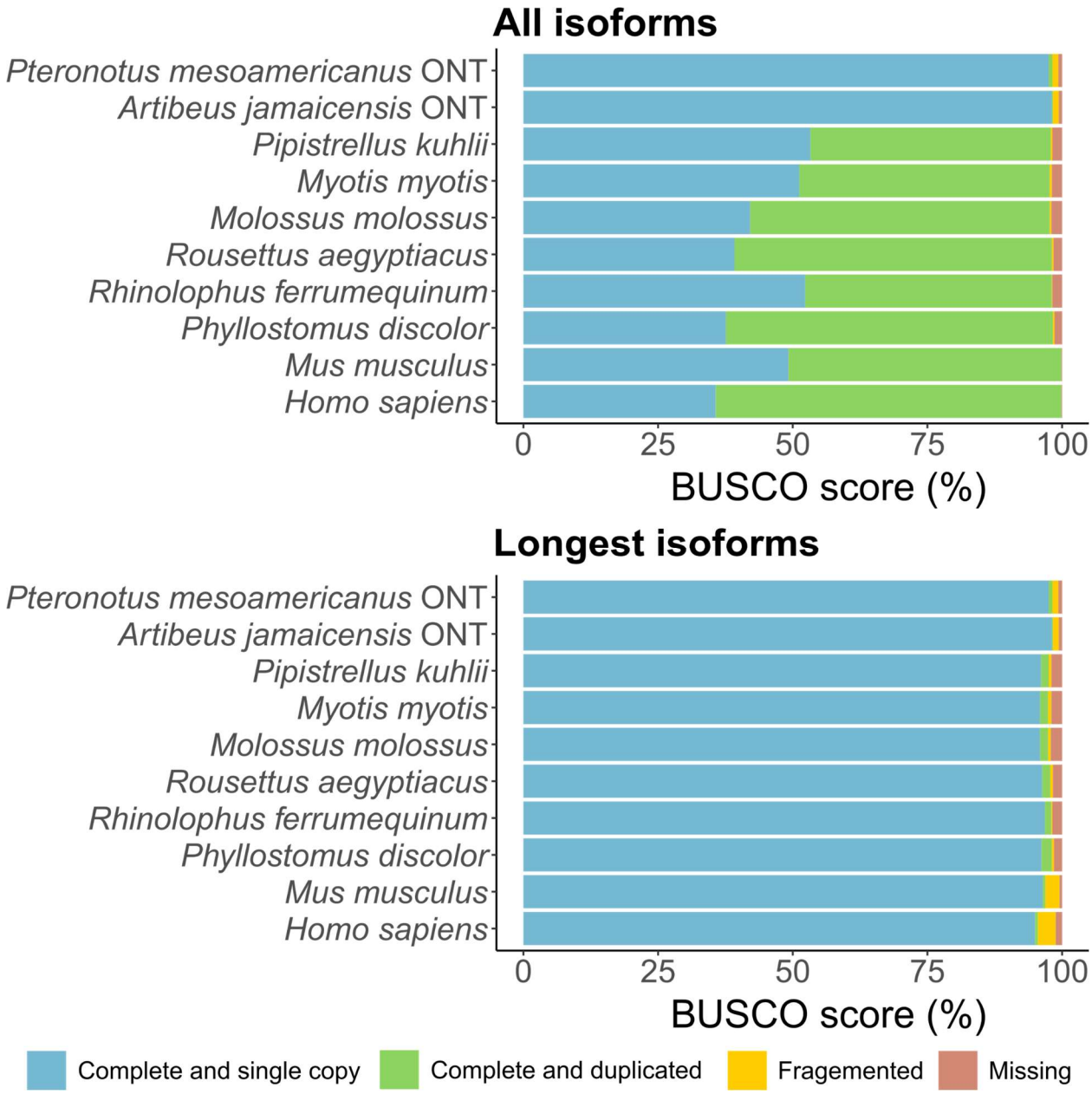
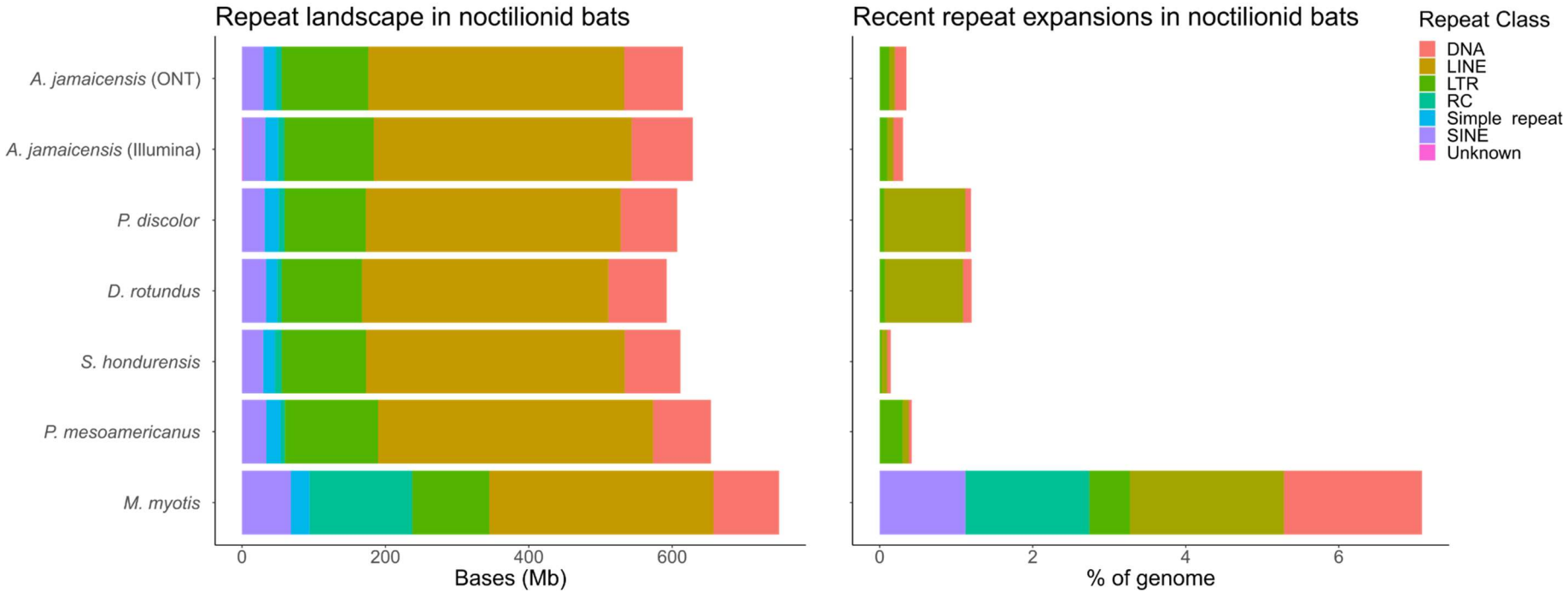


Figure S2. **Cumulative sum of the length of the ordered contigs versus the cumulative sum of the total proportion of the genome.** The vertical black line indicates the N50 metric. The set of 'Other bats (PacBio+HiC)' is based on ref[18] and 'Other bats (Illumina)' are a different set of publicly available bat genomes downloaded from GenBank (see **Table S1** for all GenBank bat assemblies).



673
674
675 **Figure S3**. **Protein-based BUSCO v4 analysis of eight bats as well as human and mouse.**
676 The mammalian BUSCO set (odb9) was used. Proteins for bat species not sequenced in this
677 study are based on annotations downloaded from https://bds.mpi-
678 cbg.de/hillerlab/Bat1KPilotProject/.



679

680 **Figure S4. RepeatMasker analysis of repeat classes in noctilionoid genomes.** Noctilionoids exhibit a homogeneous landscape of
681 repeat classes that almost completely lacks the rolling circle (RC) repeats prevalent in *M. myotis*. Long interspersed nuclear elements
682 (LINEs) make up the largest repeat class in all bats. Recent repeat expansion make up less than 2% of the genome in all noctilionoids.
683 A list of the recently expanded repeats is provided in **Table S3**.

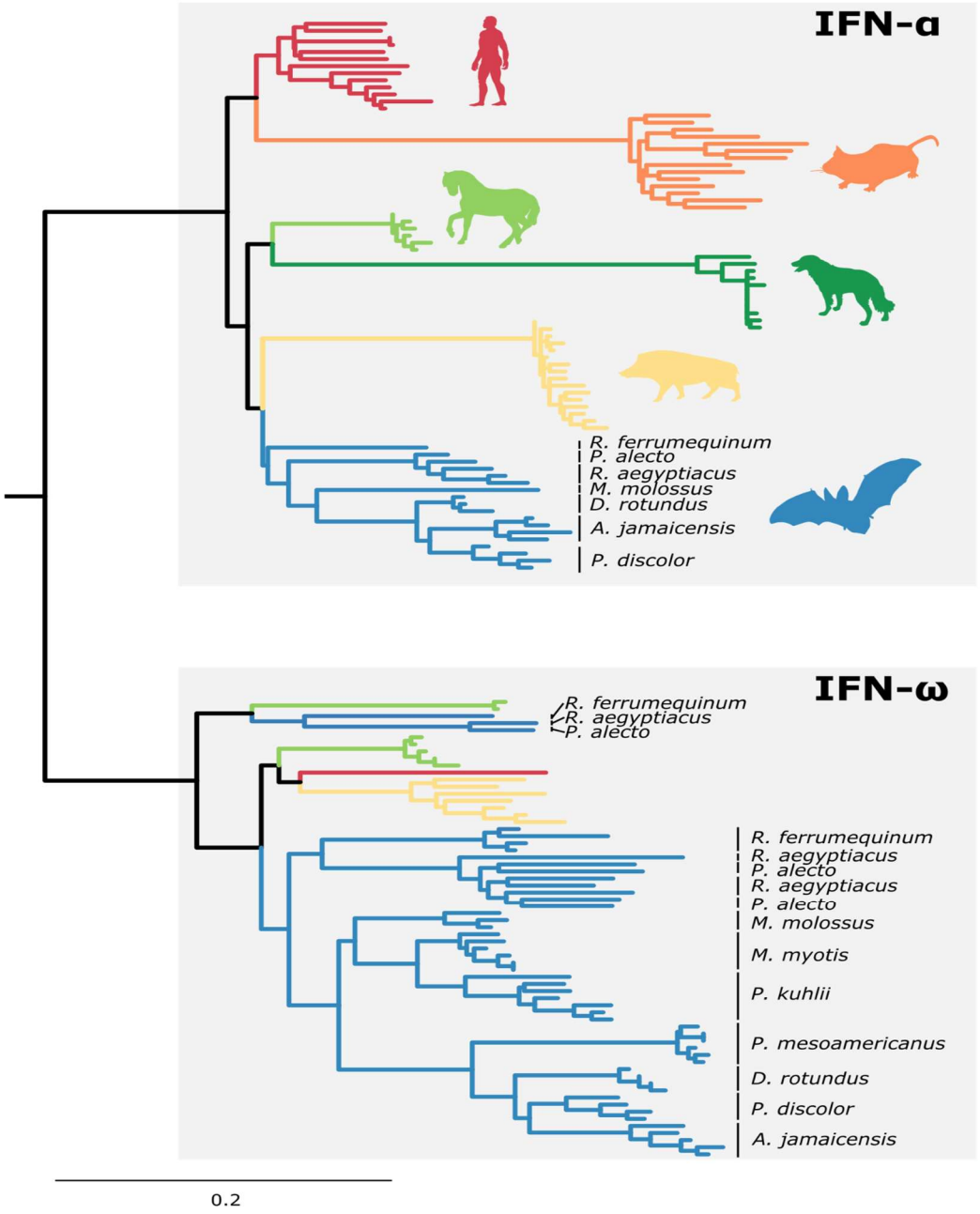


Figure S5. Maximum-likelihood phylogeny of mammalian IFN-α and IFN-ω genes showing the shift in the ratio of IFN-α to IFN-ω copy number in bats (shown in blue) compared to other mammals. The phylogeny was inferred with RAxML under a GTRGAMMA model using alignments partitioned by codon sites. IFN-ω genes are not present in the dog or mouse. Although the topology mostly reflects the expected phylogenetic relationships between species, the tree is intended to show copy number variation in IFN-α and IFN-ω, and it should be noted that internal branches within each ortholog cluster are not robustly supported. The sequence and topology are provided in **Data S3**.

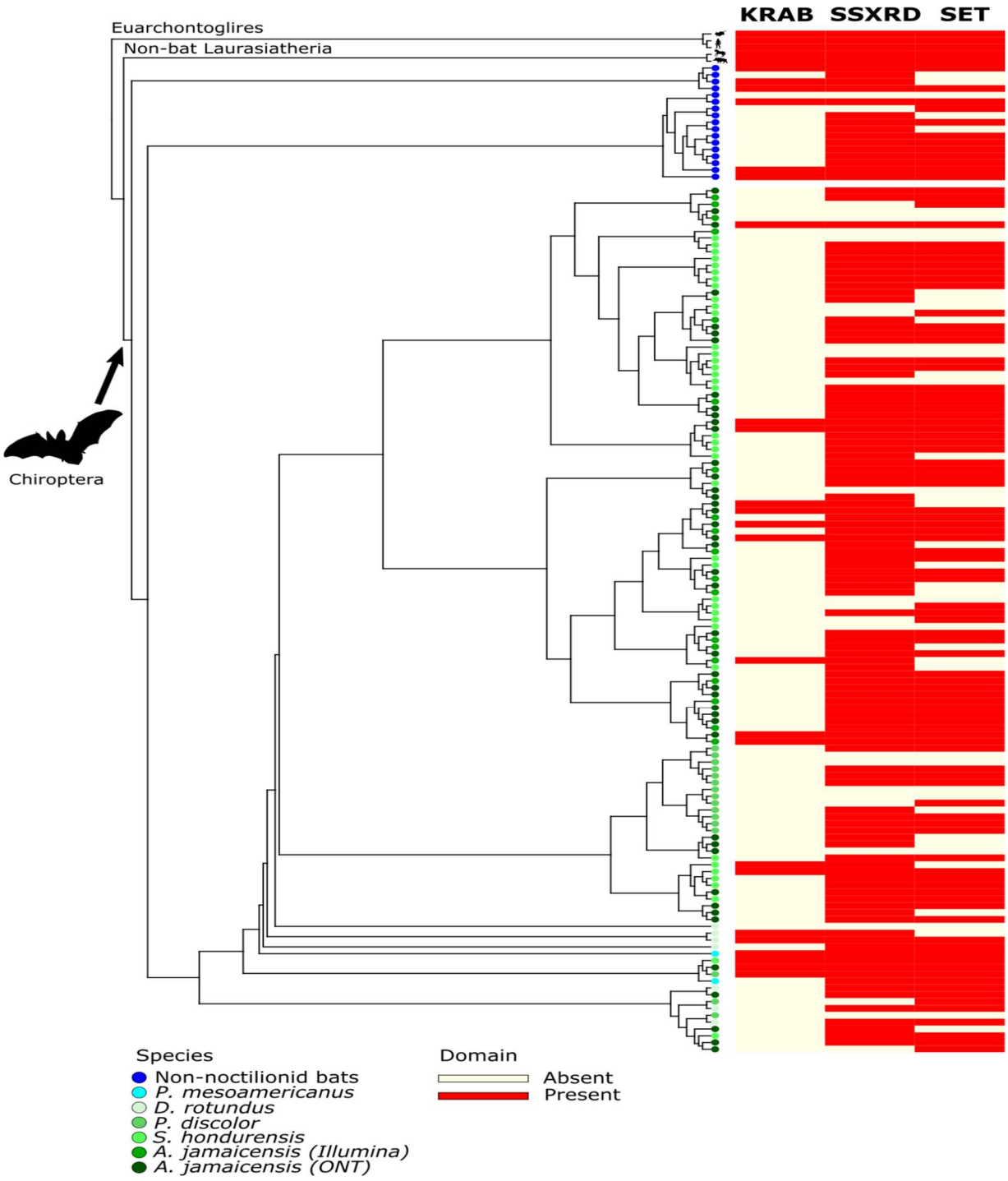


Figure S6. Maximum likelihood phylogeny of the *PRDM9* orthogroup generated using RAxML under the GTRGAMMA model. Presence and absence of the key domains KRAB, SSXRD and SET was determined using pfamscan (https://www.ebi.ac.uk/Tools/pfa/pfamscan/) with default e-value thresholds. The *PRDM9* orthologs underwent an expansion in phyllostomid bats (shown in shades of green). The sequence and topology are provided in **Data S3**.

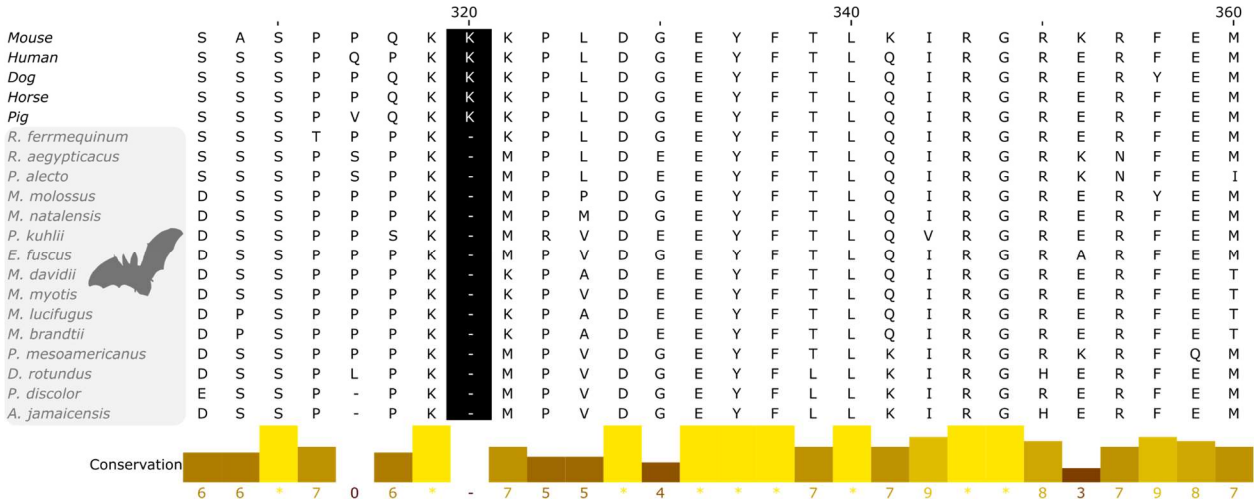


Figure S7. **Multiple alignment of *TP53* showing a bat-specific deletion (codon 320) in the nuclear localization signal domain.** Amino acid positions are based on the human protein (NP_000537.3). Image was exported from JalView 2.11.1.4.

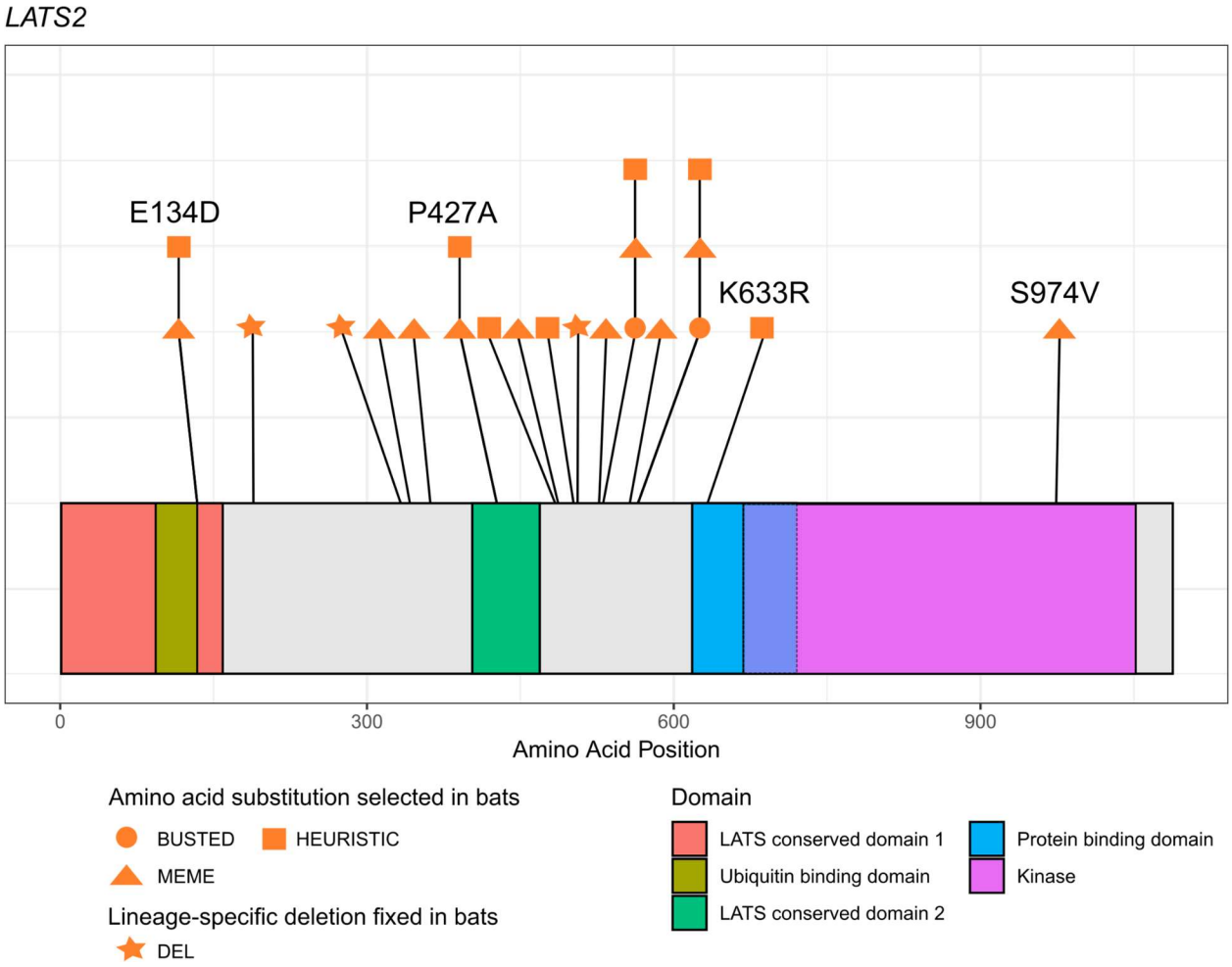


Figure S8. **Candidate sites under selection in bats in the *LATS2* gene.** Substitutions with evidence of selection were identified using BUSTED and MEME[55] as well as a simple heuristic (complete bat-specific fixation of a substitution that is not fixed in the outgroup mammals). Bat-specific deletions (DEL) were observed to begin in codons 189, 333, and 506 (coordinates based on human protein XP_016876030.1). Domain coordinates are based on UniProt and ref[93] (LCD1:1-159, LCD2:403-469, PBD:618-720, KINASE:668-1052, UBA:98-139).

References

1. Simmons, N. B., Seymour, K. L., Habersetzer, J. & Gunnell, G. F. Primitive Early Eocene bat from Wyoming and the evolution of flight and echolocation. *Nature* **451**, 818–821 (2008).
2. Simmons, N. B. & Geisler, J. H. Phylogenetic relationships of Icaronycteris, Archaeonycteris, Hassianycteris, and Palaeochiropteryx to extant bat lineages, with comments on the evolution of echolocation and foraging strategies in Microchiroptera. Bulletin of the AMNH ; no. 235. *Relationships of Eocene bats* (1998).
3. Moss, C. F. & Surlykke, A. Auditory scene analysis by echolocation in bats. *J. Acoust. Soc. Am.* **110**, 2207–2226 (2001).
4. Wilkinson, G. S. & Adams, D. M. Recurrent evolution of extreme longevity in bats. *Biol. Lett.* **15**, 20180860 (2019).
5. Wang, L.-F., Walker, P. J. & Poon, L. L. M. Mass extinctions, biodiversity and mitochondrial function: are bats 'special' as reservoirs for emerging viruses? *Curr. Opin. Virol.* **1**, 649–657 (2011).
6. Olival, K. J. *et al.* Host and viral traits predict zoonotic spillover from mammals. *Nature* **546**, 646–650 (2017).
7. Calisher, C. H., Childs, J. E., Field, H. E., Holmes, K. V. & Schountz, T. Bats: Important Reservoir Hosts of Emerging Viruses. *Clin. Microbiol. Rev.* **19**, 531–545 (2006).
8. Amman, B. R. *et al.* Seasonal pulses of Marburg virus circulation in juvenile *Rousettus aegyptiacus* bats coincide with periods of increased risk of human infection. *PLoS Pathog.* **8**, e1002877 (2012).
9. Pulliam, J. R. C. *et al.* Agricultural intensification, priming for persistence and the emergence of Nipah virus: a lethal bat-borne zoonosis. *J. R. Soc. Interface* **9**, 89–101 (2012).
10. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).

736 11. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal
737 origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
738 12. Pavlovich, S. S. *et al.* The Egyptian rousette genome reveals unexpected features of bat
739 antiviral immunity. *Cell* **173**, 1098-1110.e18 (2018).
740 13. Gonzalez, H., Hagerling, C. & Werb, Z. Roles of the immune system in cancer: from
741 tumor initiation to metastatic progression. *Genes Dev.* **32**, 1267–1284 (2018).
742 14. Teeling, E. C. *et al.* Bat Biology, Genomes, and the Bat1K Project: To Generate
743 Chromosome-Level Genomes for All Living Bat Species. *Annu. Rev. Anim. Biosci.* **6**, 23–46
744 (2018).
745 15. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields
746 chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
747 16. Escalera-Zamudio, M. *et al.* The evolution of bat nucleic acid-sensing Toll-like receptors.
748 *Mol. Ecol.* **24**, 5899–5909 (2015).
749 17. Zhang, G. *et al.* Comparative analysis of bat genomes provides insight into the evolution
750 of flight and immunity. *Science* **339**, 456–460 (2013).
751 18. Jebb, D. *et al.* Six reference-quality genomes reveal evolution of bat adaptations. *Nature*
752 **583**, 578–584 (2020).
753 19. Zepeda Mendoza, M. L. *et al.* Hologenomic adaptations underlying the evolution of
754 sanguivory in the common vampire bat. *Nat. Ecol. Evol.* **2**, 659–668 (2018).
755 20. Hawkins, J. A. *et al.* A metaanalysis of bat phylogenetics and positive selection based on
756 genomes and transcriptomes from 18 species. *Proc. Natl. Acad. Sci.* **116**, 11351–11360
757 (2019).
758 21. Irving, A. T., Ahn, M., Goh, G., Anderson, D. E. & Wang, L.-F. Lessons from the host
759 defences of bats, a unique viral reservoir. *Nature* **589**, 363–370 (2021).
760 22. Gorbunova, V., Seluanov, A. & Kennedy, B. K. The world goes bats: living longer and
761 tolerating viruses. *Cell Metab.* **32**, 31–43 (2020).

762 23. Jiang, H. *et al.* Selective evolution of Toll-like receptors 3, 7, 8, and 9 in bats.
763 *Immunogenetics* **69**, 271–285 (2017).
764 24. Zhou, P. *et al.* Contraction of the type I IFN locus and unusual constitutive expression of
765 IFN-α in bats. *Proc. Natl. Acad. Sci.* **113**, 2696–2701 (2016).
766 25. Fuchs, J. *et al.* Evolution and antiviral specificities of interferon-induced Mx proteins of
767 bats against ebola, influenza, and other RNA viruses. *J. Virol.* **91**, e00361-17 (2017).
768 26. Ahn, M., Cui, J., Irving, A. T. & Wang, L.-F. Unique loss of the PYHIN gene family in bats
769 amongst mammals: implications for inflammasome sensing. *Sci. Rep.* **6**, 21722 (2016).
770 27. Seim, I. *et al.* Genome analysis reveals insights into physiology and longevity of the
771 Brandt's bat *Myotis brandtii*. *Nat. Commun.* **4**, 2212 (2013).
772 28. Taylor, D. J., Dittmar, K., Ballinger, M. J. & Bruenn, J. A. Evolutionary maintenance of
773 filovirus-like genes in bat genomes. *BMC Evol. Biol.* **11**, 336 (2011).
774 29. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate
775 species. *Nature* **592**, 737–746 (2021).
776 30. Halo, J. V. *et al.* Long-read assembly of a Great Dane genome highlights the
777 contribution of GC-rich sequence and mobile elements to canine genomes. *Proc. Natl. Acad.*
778 *Sci.* **118**, (2021).
779 31. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat.*
780 *Methods* **16**, 88–94 (2019).
781 32. He, K., Minias, P. & Dunn, P. O. Long-read genome assemblies reveal extraordinary
782 variation in the number and structure of MHC loci in birds. *Genome Biol. Evol.* **13**, (2021).
783 33. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads
784 using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
785 34. Shafin, K. *et al. Haplotype-aware variant calling enables high accuracy in nanopore*
786 *long-reads using deep neural networks*. 2021.03.04.433952
787 https://www.biorxiv.org/content/10.1101/2021.03.04.433952v1 (2021)

788 doi:10.1101/2021.03.04.433952.

789 35. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and
790 accurate corrections in genome assemblies. *PLOS Comput. Biol.* **16**, e1007981 (2020).

791 36. McNab, F., Mayer-Barber, K., Sher, A., Wack, A. & O'Garra, A. Type I interferons in
792 infectious disease. *Nat. Rev. Immunol.* **15**, 87–103 (2015).

793 37. Kepler, T. B. *et al.* Chiropteran types I and II interferon genes inferred from genome
794 sequencing traces by a statistical gene-family assembler. *BMC Genomics* **11**, 444 (2010).

795 38. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for
796 the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).

797 39. Jaks, E., Gavutis, M., Uzé, G., Martal, J. & Piehler, J. Differential receptor subunit
798 affinities of type I interferons govern differential signal activation. *J. Mol. Biol.* **366**, 525–539
799 (2007).

800 40. Thomas, C. *et al.* Structural linkage between ligand discrimination and receptor
801 activation by type I interferons. *Cell* **146**, 621–632 (2011).

802 41. de Weerd, N. A., Samarajiwa, S. A. & Hertzog, P. J. Type I interferon receptors:
803 biochemistry and biological functions. *J. Biol. Chem.* **282**, 20053–20057 (2007).

804 42. Hayward, J. A. *et al.* Unique evolution of antiviral tetherin in bats. *bioRxiv*
805 2020.04.08.031203 (2020) doi:10.1101/2020.04.08.031203.

806 43. Benfield, C. T. *et al.* Bat IFITM3 restriction depends on S-palmitoylation and a
807 polymorphic site within the CD225 domain. *Life Sci. Alliance* **3**, e201900542 (2019).

808 44. Bailey, C. C., Zhong, G., Huang, I.-C. & Farzan, M. IFITM-family proteins: the cell's first
809 line of antiviral defense. *Annu. Rev. Virol.* **1**, 261–283 (2014).

810 45. Desai, T. M. *et al.* IFITM3 restricts influenza A virus entry by blocking the formation of
811 fusion pores following virus-endosome hemifusion. *PLoS Pathog.* **10**, e1004048 (2014).

812 46. Chesarino, N. M. *et al.* IFITM3 requires an amphipathic helix for antiviral activity. *EMBO*
813 *Rep.* **18**, 1740–1751 (2017).

814 47. Bailey, C. C., Kondur, H. R., Huang, I.-C. & Farzan, M. Interferon-induced
815 transmembrane protein 3 is a type II transmembrane protein *. *J. Biol. Chem.* **288**, 32184–
816 32193 (2013).
817 48. Paigen, K. & Petkov, P. M. PRDM9 and its role in genetic recombination. *Trends Genet.*
818 **34**, 291–300 (2018).
819 49. Baker, Z. *et al.* Repeated losses of PRDM9-directed recombination despite the
820 conservation of PRDM9 across vertebrates. *eLife* **6**, e24133 (2017).
821 50. Schwartz, J. J., Roach, D. J., Thomas, J. H. & Shendure, J. Primate evolution of the
822 recombination regulator PRDM9. *Nat. Commun.* **5**, 4370 (2014).
823 51. Xie, X., Liu, P.-S. & Percipalle, P. Analysis of global transcriptome change in mouse
824 embryonic fibroblasts after dsDNA and dsRNA viral mimic stimulation. *Front. Immunol.* **0**,
825 (2019).
826 52. Tsan, M.-F. & Gao, B. Heat shock proteins and immune system. *J. Leukoc. Biol.* **85**,
827 905–910 (2009).
828 53. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood
829 method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479
830 (2005).
831 54. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at
832 individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002).
833 55. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for
834 efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).
835 56. Potter, J. H. T. *et al.* Dietary diversification and specialization in Neotropical bats
836 facilitated by early molecular evolution. *Mol. Biol. Evol.* (2021) doi:10.1093/molbev/msab028.
837 57. Kacprzyk, J. *et al.* Evolution of mammalian longevity: age-related increase in autophagy
838 in bats compared to other mammals. *Aging* **13**, 7998–8025 (2021).
839 58. Banerjee, A., Rapin, N., Bollinger, T. & Misra, V. Lack of inflammatory gene expression

in bats: a unique role for a transcription repressor. *Sci. Rep.* **7**, (2017).

59. Xie, J. *et al.* Dampened STING-dependent interferon activation in bats. *Cell Host Microbe* **23**, 297-301.e4 (2018).

60. Ahn, M. *et al.* Dampened NLRP3-mediated inflammation in bats and implications for a special viral reservoir host. *Nat. Microbiol.* **4**, 789–799 (2019).

61. Decout, A., Katz, J. D., Venkatraman, S. & Ablasser, A. The cGAS–STING pathway as a therapeutic target in inflammatory diseases. *Nat. Rev. Immunol.* 1–22 (2021) doi:10.1038/s41577-021-00524-z.

62. Fitzgerald, K. A. & Kagan, J. C. Toll-like receptors and the control of immunity. *Cell* **180**, 1044–1066 (2020).

63. Hennessy, E. J., Parker, A. E. & O'Neill, L. A. J. Targeting Toll-like receptors: emerging therapeutics? *Nat. Rev. Drug Discov.* **9**, 293–307 (2010).

64. Gouravani, M. *et al.* The NLRP3 inflammasome: a therapeutic target for inflammation-associated cancers. *Expert Rev. Clin. Immunol.* **16**, 175–187 (2020).

65. de Oliviera Nascimento, L., Massari, P. & Wetzler, L. M. The Role of TLR2 in infection and immunity. *Front. Immunol.* **3**, (2012).

66. Zheng, M. *et al.* TLR2 senses the SARS-CoV-2 envelope protein to produce inflammatory cytokines. *Nat. Immunol.* **22**, 829–838 (2021).

67. Kim, Y.-M., Brinkmann, M. M., Paquet, M.-E. & Ploegh, H. L. UNC93B1 delivers nucleotide-sensing toll-like receptors to endolysosomes. *Nature* **452**, 234–238 (2008).

68. Panchanathan, R., Liu, H. & Choubey, D. Expression of murine Unc93b1 is up-regulated by interferon and estrogen signaling: implications for sex bias in the development of autoimmunity. *Int. Immunol.* **25**, 521–529 (2013).

69. Pelka, K. *et al.* The chaperone UNC93B1 regulates Toll-like receptor stability independently of endosomal TLR transport. *Immunity* **48**, 911-922.e7 (2018).

70. Lee, B. L. *et al.* UNC93B1 mediates differential trafficking of endosomal TLRs. *eLife* **2**,

866 e00291 (2013).

867 71. Garlanda, C., Dinarello, C. A. & Mantovani, A. The interleukin-1 family: back to the
868 future. *Immunity* **39**, 1003–1018 (2013).

869 72. Veldhoen, M. Interleukin 17 is a chief orchestrator of immunity. *Nat. Immunol.* **18**, 612–
870 621 (2017).

871 73. Velazquez-Salinas, L., Verdugo-Rodriguez, A., Rodriguez, L. L. & Borca, M. V. The role
872 of interleukin 6 during viral infections. *Front. Microbiol.* **10**, (2019).

873 74. Hayward, J. A. *et al.* Differential evolution of antiretroviral restriction factors in pteropid
874 bats as revealed by APOBEC3 gene complexity. *Mol. Biol. Evol.* **35**, 1626–1637 (2018).

875 75. Zhang, Y. *et al.* PARP9-DTX3L ubiquitin ligase targets host histone H2BJ and viral 3C
876 protease to enhance interferon signaling and control viral infection. *Nat. Immunol.* **16**, 1215–
877 1227 (2015).

878 76. Xing, J. *et al.* Identification of poly(ADP-ribose) polymerase 9 (PARP9) as a
879 noncanonical sensor for RNA virus in dendritic cells. *Nat. Commun.* **12**, 2681 (2021).

880 77. Fensterl, V. & Sen, G. C. Interferon-induced IFIT proteins: their role in viral
881 pathogenesis. *J. Virol.* **89**, 2462–2468 (2014).

882 78. Diamond, M. S. & Farzan, M. The broad-spectrum antiviral functions of IFIT and IFITM
883 proteins. *Nat. Rev. Immunol.* **13**, 46–57 (2013).

884 79. Yang, Z. *et al.* Crystal structure of ISG54 reveals a novel RNA binding structure and
885 potential functional mechanisms. *Cell Res.* **22**, 1328–1338 (2012).

886 80. Zhu, H. & Zheng, C. When PARPs meet antiviral innate immunity. *Trends Microbiol.* **29**,
887 776–778 (2021).

888 81. Foster, T. L. *et al.* Resistance of transmitted founder HIV-1 to IFITM-mediated restriction.
889 *Cell Host Microbe* **20**, 429–442 (2016).

890 82. Brass, A. L. *et al.* The IFITM proteins mediate cellular resistance to influenza A H1N1
891 virus, West Nile virus, and dengue virus. *Cell* **139**, 1243–1254 (2009).

892 83. Tollis, M., Schiffman, J. D. & Boddy, A. M. Evolution of cancer suppression as revealed
893 by mammalian comparative genomics. *Curr. Opin. Genet. Dev.* **42**, 40–47 (2017).
894 84. Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-
895 based knowledgebase for tumor suppressor genes. *Nucleic Acids Res.* **44**, D1023-1031
896 (2016).
897 85. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of
898 Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
899 86. Sy, S. M. H., Huen, M. S. Y. & Chen, J. PALB2 is an integral component of the BRCA
900 complex required for homologous recombination repair. *Proc. Natl. Acad. Sci.* **106**, 7155–
901 7160 (2009).
902 87. Deveryshetty, J. *et al.* Novel RNA and DNA strand exchange activity of the PALB2 DNA
903 binding domain and its critical role for DNA repair in cells. *eLife* **8**, e44063 (2019).
904 88. Belotserkovskaya, R. *et al.* PALB2 chromatin recruitment restores homologous
905 recombination in BRCA1-deficient cells depleted of 53BP1. *Nat. Commun.* **11**, 819 (2020).
906 89. Kosiol, C. *et al.* Patterns of positive selection in six mammalian genomes. *PLoS Genet.*
907 **4**, (2008).
908 90. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and
909 chimpanzees. *PLOS Biol.* **3**, e170 (2005).
910 91. Furth, N. & Aylon, Y. The LATS1 and LATS2 tumor suppressors: beyond the Hippo
911 pathway. *Cell Death Differ.* **24**, 1488–1501 (2017).
912 92. Reumers, J. *et al.* SNPeffect: a database mapping molecular phenotypic effects of
913 human non-synonymous coding SNPs. *Nucleic Acids Res.* **33**, D527–D532 (2005).
914 93. Yu, T., Bachman, J. & Lai, Z.-C. Mutation analysis of large tumor suppressor genes
915 LATS1 and LATS2 supports a tumor suppressor role in human cancer. *Protein Cell* **6**, 6–11
916 (2015).
917 94. Zou, Y. *et al.* Systemic tumor suppression by the proapoptotic gene bik. *Cancer Res.* **62**,

8–12 (2002).

95. Huang, D. C. & Strasser, A. BH3-Only proteins-essential initiators of apoptotic cell death. *Cell* **103**, 839–842 (2000).

96. López, I. *et al.* p53-mediated suppression of BiP triggers BIK-induced apoptosis during prolonged endoplasmic reticulum stress. *Cell Death Differ.* **24**, 1717–1729 (2017).

97. Hur, J. *et al.* Regulation of expression of BIK proapoptotic protein in human breast cancer cells: p53-dependent induction of BIK mRNA by fulvestrant and proteasomal degradation of BIK protein. *Cancer Res.* **66**, 10153–10161 (2006).

98. Real, P. J. *et al.* Transcriptional activation of the proapoptotic bik gene by E2F proteins in cancer cells. *FEBS Lett.* **580**, 5905–5909 (2006).

99. Yu, W., Yang, L., Li, T. & Zhang, Y. Cadherin signaling in cancer: its functions and role as a therapeutic target. *Front. Oncol.* **9**, (2019).

100. Glorieux, C. & Calderon, P. B. Catalase, a remarkable enzyme: targeting the oldest antioxidant enzyme to find a new cancer treatment approach. *Biol. Chem.* **398**, 1095–1108 (2017).

101. Kang, M. Y. *et al.* The critical role of catalase in prooxidant and antioxidant function of p53. *Cell Death Differ.* **20**, 117–129 (2013).

102. Slodkowicz, G. & Goldman, N. Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals. *Proc. Natl. Acad. Sci.* **117**, 5977–5986 (2020).

103. Enard, D., Cai, L., Gwennap, C. & Petrov, D. A. Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).

104. Vicens, A. & Posada, D. Selective Pressures on Human Cancer Genes along the Evolution of Mammals. *Genes* **9**, 582 (2018).

105. Murrell, B. *et al.* Detecting individual sites subject to episodic diversifying selection. *PLOS Genet.* **8**, e1002764 (2012).

944 106. Goh, G. *et al.* Complementary regulation of caspase-1 and IL-1β reveals additional
945 mechanisms of dampened inflammation in bats. *Proc. Natl. Acad. Sci.* **117**, 28939–28949
946 (2020).
947 107. Yu, X. *et al.* Inflammasome activation negatively regulates MyD88-IRF7 type I IFN
948 signaling and anti-malaria immunity. *Nat. Commun.* **9**, 4964 (2018).
949 108. Banerjee, A. *et al.* Positive selection of a serine residue in bat IRF3 confers enhanced
950 antiviral protection. *iScience* **23**, 100958 (2020).
951 109. Li, S. *et al.* Interferon-omega: Current status in clinical applications. *Int.*
952 *Immunopharmacol.* **52**, 253–260 (2017).
953 110. Zhao, X. *et al.* Characterization and virus-induced expression profiles of the porcine
954 interferon-omega multigene family. *J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon*
955 *Cytokine Res.* **29**, 687–693 (2009).
956 111. Tejada-Martinez, D., de Magalhães, J. P. & Opazo, J. C. Positive selection and gene
957 duplications in tumour suppressor genes reveal clues about how cetaceans resist cancer.
958 *Proc. R. Soc. B Biol. Sci.* **288**, 20202592 (2021).
959 112. Fonseca, R. R. da, Kosiol, C., Vinař, T., Siepel, A. & Nielsen, R. Positive selection on
960 apoptosis related genes. *FEBS Lett.* **584**, 469–476 (2010).
961 113. Ostrand-Rosenberg, S. Immune surveillance: a balance between protumor and
962 antitumor immunity. *Curr. Opin. Genet. Dev.* **18**, 11–18 (2008).
963 114. Li, D. *et al.* KLF4-mediated negative regulation of IFITM3 expression plays a critical role
964 in colon cancer pathogenesis. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **17**, 3558–
965 3568 (2011).
966 115. Siegrist, F., Ebeling, M. & Certa, U. The small interferon-induced transmembrane genes
967 and proteins. *J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon Cytokine Res.* **31**, 183–
968 197 (2011).
969 116. Alteber, Z. *et al.* The anti-inflammatory IFITM genes ameliorate colitis and partially

970 protect from tumorigenesis by changing immunity and microbiota. *Immunol. Cell Biol.* **96**,
971 284–297 (2018).
972 117. Yang, Y., Wang, H., Kouadir, M., Song, H. & Shi, F. Recent advances in the
973 mechanisms of NLRP3 inflammasome activation and its inhibitors. *Cell Death Dis.* **10**, 1–11
974 (2019).
975 118. Moossavi, M., Parsamanesh, N., Bahrami, A., Atkin, S. L. & Sahebkar, A. Role of the
976 NLRP3 inflammasome in cancer. *Mol. Cancer* **17**, 158 (2018).
977 119. Pidugu, V. K., Pidugu, H. B., Wu, M.-M., Liu, C.-J. & Lee, T.-C. Emerging Functions of
978 Human IFIT Proteins in Cancer. *Front. Mol. Biosci.* **0**, (2019).
979 120. Lu, C. *et al.* Type I interferon suppresses tumor growth through activating the STAT3-
980 granzyme B pathway in tumor-infiltrating cytotoxic T lymphocytes. *J. Immunother. Cancer* **7**,
981 157 (2019).
982 121. Aricò, E., Castiello, L., Capone, I., Gabriele, L. & Belardelli, F. Type I interferons and
983 cancer: an evolving story demanding novel clinical applications. *Cancers* **11**, 1943 (2019).
984 122. Marín-Aguilar, F. *et al.* NLRP3 inflammasome suppression improves longevity and
985 prevents cardiac aging in male mice. *Aging Cell* **19**, e13050 (2020).
986 123. Youm, Y.-H. *et al.* Canonical NLRP3 inflammasome links systemic low-grade
987 inflammation to functional decline in aging. *Cell Metab.* **18**, 519–532 (2013).
988 124. Lee, H. E. *et al.* Inhibition of NLRP3 inflammasome in tumor microenvironment leads to
989 suppression of metastatic potential of cancer cells. *Sci. Rep.* **9**, 12277 (2019).
990 125. Tengesdal, I. W. *et al.* Targeting tumor-derived NLRP3 reduces melanoma progression
991 by limiting MDSCs expansion. *Proc. Natl. Acad. Sci.* **118**, (2021).
992 126. Sikes, R. S. 2016 Guidelines of the American Society of Mammalogists for the use of
993 wild mammals in research and education. *J. Mammal.* **97**, 663–688 (2016).
994 127. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**,
995 155–158 (2020).

128. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).

129. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

130. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013).

131. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

132. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

133. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

134. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

135. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

136. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, (2019).

137. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

138. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

139. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for

1022 mRNA and EST sequences. *Bioinforma. Oxf. Engl.* **21**, 1859–1875 (2005).

1023 140. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

1024 141. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the

1025 Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).

1026 142. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging

1027 model organism genomes. *Genome Res.* **18**, 188–196 (2008).

1028 143. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

1029 144. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids*

1030 *Res.* **34**, W435–W439 (2006).

1031 145. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open

1032 source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

1033 146. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene

1034 identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**,

1035 6494–6506 (2005).

1036 147. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using

1037 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7

1038 (2008).

1039 148. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in

1040 genomic sequences. *Curr. Protoc. Bioinforma.* **Chapter 4**, Unit 4.10 (2009).

1041 149. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-

1042 generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

1043 150. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements.

1044 *Nat. Rev. Genet.* **8**, 973–982 (2007).

1045 151. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for

1046 automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**,

1047 1329–1330 (2009).

1048 152. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for de novo
1049 classification of transposons with convolutional neural network. *Bioinformatics*
1050 doi:10.1093/bioinformatics/btaa519.
1051 153. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl. Acad.*
1052 *Sci.* **99**, 803–808 (2002).
1053 154. Bigot, T., Temmam, S., Pérot, P. & Eloit, M. RVDB-prot, a reference viral protein
1054 database and its HMM profiles. *F1000Research* **8**, 530 (2019).
1055 155. Shen, W. & Xiong, J. TaxonKit: a cross-platform and efficient NCBI taxonomy toolkit.
1056 *bioRxiv* 513523 (2019) doi:10.1101/513523.
1057 156. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
1058 genomics. *Genome Biol.* **20**, 238 (2019).
1059 157. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol. Clifton NJ*
1060 **1079**, 155–170 (2014).
1061 158. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1062 large phylogenies. *Bioinforma. Oxf. Engl.* **30**, 1312–1313 (2014).
1063 159. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: a resource for timelines,
1064 timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
1065 160. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior
1066 summarization in Bayesian phylogenetics using tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
1067 161. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological
1068 databases and microarray data analysis. *Bioinforma. Oxf. Engl.* **21**, 3439–3440 (2005).
1069 162. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and
1070 genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
1071 163. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood
1072 method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479
1073 (2005).

1074 164. Doronina, L. *et al.* Speciation network in Laurasiatheria: retrophylogenomic signals.

1075 *Genome Res.* **27**, 997–1003 (2017).

1076 165. Alexa, A. & Rahnenfuhrer. J. topGO: Enrichment Analysis for Gene Ontology. R

1077 package version 2.44.0 (2021).