1   **Title:** Six new reference-quality bat genomes illuminate the molecular basis and evolution of

2   bat adaptations

3

4   **Authors:** David Jebb^[1,2,3], Zixia Huang ^[4] Martin Pippel^[1,3], Graham M. Hughes[4], Ksenia

5   Lavrichenko[5], Paolo Devanna[5], Sylke Winkler[1], Lars S. Jermiin[4,6,7], Emilia C. Skirmuntt[8], Aris

6   Katzourakis[8], Lucy Burkitt-Gray[9], David A. Ray[10], Kevin A. M. Sullivan[10], Juliana G.

7   Roscito[1,2,3], Bogdan M. Kirilenko[1,2,3], Liliana M. Dávalos[11,12], Angelique P. Corthals[13], Megan

8   L. Power[4], Gareth Jones[14], Roger D. Ransome[14], Dina Dechmann[15,16,17], Andrea G.

9   Locatelli[4], Sebastien J. Puechmaille[18,19], Olivier Fedrigo[20], Erich D. Jarvis[21,22], Mark S.

10   Springer[23], Michael Hiller*[1,2,3], Sonja C. Vernes*[5,24], Eugene W. Myers*[1,3,25], Emma C.

11   Teeling*[4]

12

13   [1]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany

14   [2]Max Planck Institute for the Physics of Complex Systems, Dresden, Germany

15   [3]Center for Systems Biology Dresden, Dresden, Germany

16   [4]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland

17   [5]Neurogenetics of Vocal Communication Group, Max Planck Institute for Psycholinguistics,

18   Nijmegen, The Netherlands

19   [6]Research School of Biology, Australian National University, Canberra, ACT, Australia

20   [7]Earth Institute, University College Dublin, Dublin, Ireland

21   [8]Peter Medawar Building for Pathogen Research, Department of Zoology, University of Oxford,

22   Oxford, United Kingdom

23   [9]Conway Institute of Biomolecular and Biomedical Science, University College Dublin, Dublin,

24   Ireland

25   [10]Department of Biological Sciences, Texas Tech University, Lubbock, USA

26   [11]Department of Ecology and Evolution, Stony Brook University, Stony Brook, Stony Brook, USA

27   [12]Consortium for Inter Disciplinary Environmental Research, Stony Brook University, Stony Brook,

28   USA

29   [13]Department of Sciences, John Jay College of Criminal Justice, New York, USA

30   [14]School of Biological Sciences, University of Bristol, Bristol, United Kingdom

31   [15]Department of Migration and Immuno-Ecology, Max Planck Institute of Animal Behavior,

32   Radolfzell, Germany

33   [16]Department of Biology, University of Konstanz, Konstanz, Germany

34   [17]Smithsonian Tropical Research Institute; Panama City, Panama

35   [18]ISEM, University of Montpellier, Montpellier, France

36   [19]Zoological Institute and Museum, University of Greifswald, Greifswald, Germany

37   [20]Vertebrate Genomes Laboratory, The Rockefeller University, New York, NY, USA

38   [21]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA

39   [22]Howard Hughes Medical Institute, Chevy Chase, MD, USA

40   [23]Department of Biology, University of California, Riverside, CA, USA

41   [24]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

42   [25]Faculty of Computer Science, Technical University Dresden, Dresden, Germany

43

44   ^Joint first authors

45   *joint senior/corresponding authors emails

46   Michael Hiller: hiller@mpi-cbg.de

47   Sonja Vernes: sonja.vernes@mpi.nl

48   Eugene Myers: gene@mpi-cbg.de

49   Emma C. Teeling: emma.teeling@ucd.ie

50   **Abstract:** Bats account for ~20% of all extant mammal species and are considered exceptional

51   given their extraordinary adaptations, including biosonar, true flight, extreme longevity, and

52   unparalleled immune systems. To understand these adaptations, we generated reference-quality

53   genomes of six species representing the key divergent lineages. We assembled these genomes

54   with a novel pipeline incorporating state-of-the-art long-read and long-range sequencing and

55   assembly techniques. The genomes were annotated using a maximal evidence approach, *de*

56   *novo* predictions, protein/mRNA alignments, Iso-seq long read and RNA-seq short read

57   transcripts, and gene projections from our new TOGA pipeline, retrieving virtually all (>99%)

58   mammalian BUSCO genes. Phylogenetic analyses of 12,931 protein coding-genes and 10,857

59   conserved non-coding elements identified across 48 mammalian genomes helped to resolve

60   bats' closest extant relatives within Laurasiatheria, supporting a basal position for bats within

61   Scrotifera. Genome-wide screens along the bat ancestral branch revealed (a) selection on

62   hearing-involved genes (e.g *LRP2, SERPINB6, TJP2)*, which suggest that laryngeal

63   echolocation is a shared ancestral trait of bats; (b) selection (e.g *INAVA, CXCL13, NPSR1*) and

64   loss of immunity related proteins (e.g. *LRRC70, IL36G*), including pro-inflammatory NF-kB

65   signalling; and (c) expansion of the *APOBEC* family, associated with restricting viral infection,

66   transposon activity and interferon signalling. We also identified unique integrated viruses,

67   indicating that bats have a history of tolerating viral pathogens, lethal to other mammal species.

68   Non-coding RNA analyses identified variant and novel microRNAs, revealing regulatory

69   relationships that may contribute to phenotypic diversity in bats. Together, our reference-

70   quality genomes, high-quality annotations, genome-wide screens and *in-vitro* tests revealed

71   previously unknown genomic adaptations in bats that may explain their extraordinary traits.

72

73   **Keywords:** Bats, genomes, immunity, flight, ageing, miRNA, viruses, longevity, echolocation

2

74 **Introduction.**

75 With more than ~1400 species identified to date[1], bats (Chiroptera) account for ~20% of all

76 currently recognised, extant, mammal species, are found around the globe, and successfully

77 occupy diverse ecological niches[1,2]. Their global success is attributed to their extraordinary

78 suite of adaptations including: powered flight, laryngeal echolocation for orientation and

79 hunting in complete darkness, exceptional longevity, and a unique immune system that enables

80 bats to tolerate viruses that are typically lethal in other mammals (e.g., rabies, SARS, MERS)[2].

81 It has been proposed that the evolution of extended longevity and immunity in bats was driven

82 by the acquisition of flight, which has a high metabolic cost[3-5], but the mechanisms underlying

83 these adaptations are unknown and their potential connection to flight is still debated[6,7]. Given

84 bats' distinctive adaptations, they represent important model systems to uncover the molecular

85 basis and evolution of extended healthspan[7,8], enhanced disease tolerance[9] and sensory

86 perception[10,11]. To understand the evolution of such traits, one needs to understand bats'

87 evolutionary history. However, key aspects of that evolutionary history such as monophyly of

88 echolocating bats and the single origin of laryngeal echolocation[10,12] remain debated, partially

89 stemming from a poor fossil record[13], incongruent phylogenetic analyses[14], and importantly the

90 limited quality of available genome assemblies.

91

92 Here, we generated the first reference-quality genomes of six bats as part of the Bat1K global

93 genome consortium[2] (http://bat1k.com) in coordination with the Vertebrate Genome Project

94 (https://vertebrategenomesproject.org/). Species were chosen to enable capture of the major

95 ecological trait space and life histories observed in bats while representing deep phylogenetic

96 divergences. These six bat species belong to five families that represent key evolutionary

97 clades, unique adaptations and span both major lineages in Chiroptera estimated to have

98 diverged ~64 MYA[15]. In the suborder Yinpterochiroptera we sequenced *Rhinolophus*

99 *ferrumequinum* (Greater horseshoe bat; family Rhinolophidae) and *Rousettus aegyptiacus*

100 (Egyptian fruit bat; Pteropodidae), and in Yangochiroptera we sequenced *Phyllostomus*

101 *discolor* (Pale spear-nose bat; Phyllostomidae), *Myotis myotis* (Greater mouse-eared bat;

102 Vespertilionidae), *Pipistrellus kuhlii* (Kuhl's pipistrelle; Vespertilionidae) and *Molossus*

103 *molossus* (Velvety free-tailed bat; Molossidae) (Table S1). These bat genera represent the

104 extremes in known bat longevity[16]. They also represent major adaptations in bat sensory

105 perception and ecological diversity[2], and include species considered key viral reservoirs and

106 asymptomatic hosts[9,17].

107

3

108   To obtain genome assemblies of high contiguity and completeness, we developed novel
109   pipelines incorporating state-of-the-art sequencing and assembly. To ascertain the position of
110   Chiroptera within Laurasiatheria and thus resolve a long-standing phylogenetic debate[14], we
111   mined these near complete genomes to produce a comprehensive orthologous gene data set
112   (12,931), including data from 42 other representative mammalian genomes (TableS1), and
113   applied a suite of diverse phylogenetic approaches. To identify molecular changes in regions
114   of the genome - both coding and non-coding - that underlie bat adaptations, we carried out
115   selection tests, analysed gains and losses of genes, and experimentally validated novel bat
116   microRNAs. We focussed on assessing the shared commonalities between the bat species
117   enabling us to infer the ancestral selection driving key bat adaptations. We elucidated the
118   diversity of endogenous viruses contained within the bat genomes, exploring bats' putative
119   history with these viruses. Herein, we present the first six reference-quality bat genomes, which
120   we make available in the open access Bat1K browser (also available on NCBI and GenomeArk)
121   and demonstrate both the value of highly contiguous and highly complete genomes and the
122   utility of bats as model organisms to address fundamental questions in biology.
123

Genome Sequencing and Assembly

125   For each of the six bats, we generated: (i) PacBio long reads (52-70X in reads $\geq$4 kb; N50 read
126   length 14.9-24.5 kb), (ii) 10x Genomics Illumina read clouds (43-104X), (iii) Bionano optical
127   maps (coverage of molecules $\geq$150 kb 89-288X), and (iv) Hi-C Illumina read pairs (15-95X).
128   PacBio reads were assembled into contigs with a customized assembler called DAmar, a hybrid
129   of our earlier Marvel[18], Dazzler (https://dazzlerblog.wordpress.com/), and Daccord[19,20] systems
130   (Fig. 1a). Next, we used PacBio reads and 10x read cloud Illumina data to remove base errors,
131   which was followed by identifying and phasing all regions of the contigs that had a sufficient
132   rate of haplotype heterogeneity. We retained one haplotype for each region, yielding primary
133   contigs. These primary contigs were then scaffolded using a Bionano optical map and the Hi-C
134   data (see supplementary methods section 2).
135

136   For all six bats, this sequencing and assembly strategy produced assemblies with contig N50
137   values ranging from 10.6 to 22.2 Mb (Fig. 1b, Table S2). Thus, our contigs are $\geq$355 times
138   more contiguous than the recent *Miniopterus* assembly generated from short read data[21], and
139    $\geq$7 times more contiguous than a previous *Rousettus* assembly generated from a hybrid of short
140   and long read data[22] (Fig. 1b). Our scaffold N50 values ranged from 80.2 to 171.1 Mb and were
141   often limited by the size of chromosomes (Fig. 1b, Table S2). We estimated that 87 to 99% of

4

142   each assembly is in chromosome-level scaffolds (Table S3). Consensus base accuracies across

143   the entire assembly range from QV 40.8 to 46.2 (Table S2) for the six bats (where QV 40

144   represents 1 error in 10,000 bp). Since the algorithms for assembling, scaffolding, and

145   haplotyping are an active area of research[23], we expect that in the future even more complete

146   genome reconstructions can be produced with the data we collected. Even so, our current

147   strategy and algorithms generated chromosome-level assemblies of the six bats with

148   unprecedented contiguity, which are comparable to the best reference-quality genomes

149   currently generated for any eukaryotic species with a complex, multi-gigabyte genome[24].

150   Importantly, they meet the Vertebrate Genome Project (VGP) minimum standard of 3.4.2QV40

151   and have been added to the VGP collection.

152

153   To assess genome completeness, we first evaluated the presence of 4,104 genes that are highly

154   conserved among mammals (BUSCO, Benchmarking Universal Single-Copy Orthologs[25]).

155   Between 92.9 and 95.8% of these genes were completely present in our assemblies, which is

156   comparable to the assemblies of human, mouse, and other Laurasiatheria (Fig. 1c, Table S4).

157   Second, to assess completeness in non-exonic genomic regions, we determined how many of

158   197 non-exonic ultraconserved elements (UCEs)[26] align at ≥85% identity to the human

159   sequence. As expected, the vast majority of UCEs were detected in all assemblies (Fig. 1d).

160   Two to four UCEs were not detected in *Miniopterus*, dog, cat, and cow due to assembly

161   incompleteness (i.e. assembly gaps; Table S5, Fig. S1). In the bat genomes reported herein, no

162   UCEs were missing due to assembly incompleteness. Instead, one to three UCEs were not

163   detected in our *Myotis* and *Pipistrellus* assemblies because the UCE sequences are more than

164   85% diverged (Table S5), a striking result given that UCE's are highly conserved across other

165   more divergent mammals (e.g. human-mouse-rat comparison). To determine if this sequence

166   divergence was caused by base errors in the assemblies, we aligned raw PacBio and Illumina

167   reads and sequencing data of related bats, which confirmed that these UCEs are truly diverged

168   (Figs. S1-S5). In summary, our six bat assemblies are highly complete and revealed the first

169   examples of highly diverged UCEs.

170

171   **Genome Annotation**

172   To comprehensively annotate genes, we integrated a variety of evidence (Fig. 1e). First, we

173   aligned protein and cDNA sequences of a related bat species to each of our six genomes (Table

174   S6). Second, we projected genes annotated in human, mouse[27] and two bat assemblies (*Myotis*

175   *lucifugus* (Ensembl) and *Myotis myotis* (Bat1K)) to our genomes via whole-genome

5

176   alignments[28]. Third, we generated *de novo* gene predictions by applying Augustus[29] with a

177   trained bat-specific gene model in single-genome mode to individual genomes, and in

178   comparative mode to a multiple genome alignment including our bat assemblies. Fourth, we

179   integrated transcriptomic data from both publicly available data sources and our own Illumina

180   short read RNA-seq data (Table S7). Additionally, we generated PacBio long read RNA

181   sequences (Iso-seq) from all six species to capture full-length isoforms and accurately annotate

182   untranslated regions (UTRs) (Table S8). Iso-seq data were processed using the TAMA

183   pipeline[30] which allowed capturing a substantially greater diversity of transcripts and isoforms

184   than the default pipeline (https://github.com/PacificBiosciences/IsoSeq3). All transcriptomic,

185   homology-based and *ab initio* evidence were integrated into a consensus gene annotation that

186   we further enriched for high-confidence transcript variants and filtered for strong coding

187   potential.

188

189   For the six bats, we annotated between 19,122 and 21,303 coding genes (Fig. 1f). These

190   annotations completely contain between 99.3 and 99.7% of the 4,104 highly conserved

191   mammalian BUSCO genes (Fig. 1f, Table S4), showing that our six bat assemblies are highly

192   complete in coding sequences. Since every annotated gene is by definition present in the

193   assembly, one would expect that BUSCO applied to the protein sequences of annotated genes

194   and BUSCO applied to the genome assembly should yield highly similar statistics. However,

195   the latter finds only 92.9 to 95.8% of the exact same gene set as completely present, showing

196   that BUSCO applied to an assembly only, underestimates the number of completely contained

197   genes. Importantly, this gene annotation completeness of our bats is higher than the Ensembl

198   gene annotations of dog, cat, horse, cow and pig, and is only surpassed by the gene annotations

199   of human and mouse, which have received extensive manual curation of gene models (Fig. 1f,

200   Table S4). This suggests reference-quality genome assemblies and the integration of various

201   gene evidence as detailed above, can be used to generate high-quality and near-complete gene

202   annotations of bats as well as other species too. All individual evidence and the final gene set

203   can be visualized and obtained from the Bat1K genome browser (https://genome-

204   public.pks.mpg.de).

205

206   **Genome Sizes and Transposable Elements**

207   At ~2 Gb, bat genomes are generally smaller than genomes of other placental mammals that

208   are typically between 2.5 and 3.5 Gb[2]. Nevertheless, our assemblies revealed noticeable

209   genome size differences within bats, with assembly sizes ranging from 1.78 Gb for *Pipistrellus*

210   to 2.32 Gb for *Molossus* (Fig. 1g). As genome size is often correlated with transposable element

211   (TE) content and activity, we focused on the genomes of the six bats and seven other

212   representative Boreoeutherian mammals (Laurasiatheria + Euarchontoglires), selected for the

213   highest genome contiguity, and used a previously-described workflow and manual curation to

214   annotate TEs[31]. This showed that TE content generally correlates with genome size (Fig. 1g).

215   Next, we compared TE copies to their consensus sequence to obtain a relative age from each

216   TE family. This revealed an extremely variable repertoire of TE families with evidence of

217   recent accumulation (defined as consisting of insertions with divergences<6.6% from the

218   relevant consensus sequence). For example, while the 1.89 Gb *Rousettus* genome exhibits few

219   recent TE accumulations, ~0.38%, while ~4.2% of the similarly sized 1.78 Gb *Pipistrellus*

220   genome is derived from recent TE insertions (Fig. 1g-h). The types of TE that underwent recent

221   expansions also differ substantially in bats compared to other mammals, particularly in regards

222   to the evidence of recent accumulation by rolling-circle and DNA transposons in the

223   vespertilionid bats (Fig. 1g-h). These two TE classes have been largely dormant in most

224   mammals for the past ~40 million years and recent insertions are essentially absent from other

225   Boreoeutherian genomes[32]. These results add to previous findings revealing a substantial

226   diversity in TE content within bats, with some species exhibiting recent and ongoing

227   accumulation from TE classes that are extinct in most other mammals while other species show

228   negligible evidence of TE activity[33].

229

The Origin of Chiroptera within Laurasiatheria

231   Identifying the evolutionary origin of bats within Laurasiatheria is a key prerequisite for

232   comparative analyses aimed at revealing the genomic basis of traits shared by bats. However,

233   the phylogeny of Laurasiatheria and, in particular, the position of bats has been a long-standing,

234   unresolved phylogenetic question[14,34]. This is perhaps the most challenging interordinal

235   problem in placental mammal phylogenetics, as multiple phylogenetic and systematic

236   investigations using large nucleotide and genomic scale datasets or transposable element

237   insertions support alternative topologies[35]. These incongruent results have been attributed to

238   the challenge of identifying two, presumably short, internal branches linking four clades

239   (Chiroptera, Cetartiodactyla, Perissodactyla, Carnivora + Pholidota) that diverged in the Late

240   Cretaceous[35].

241

242   We revisited this question leveraging the high completeness of our gene annotation. First, we

243   extracted a comprehensive set of 12,931 orthologous protein-coding genes from the 48

7

244   mammalian genomes, resulting in a dataset comprising 21,471,921 aligned nucleotides in

245   length, which contained 7,913,054 parsimony-informative sites. The best-fit model of sequence

246   evolution for each of the 12,931 gene alignments was inferred using ModelFinder[36] (Table S9).

247   The species tree was then estimated by maximum likelihood using the model-partitioned

248   dataset with IQTREE[37] and rooted on Atlantogenata[38]. Branch-support values were obtained

249   by UFBoot[39] with 1000 bootstrap pseudoreplicates. These analyses led to 100% bootstrap

250   support across the entire tree (Fig. 2a) and seemingly identified the origin of bats within

251   Laurasiatheria. The basal split is between Eulipotyphla and other laurasiatherians (i.e.,

252   Scrotifera). Within Scrotifera, Chiroptera is the sister clade to Fereuungulata (Cetartiodactyla

253   + Perissodactyla + Carnivora + Pholidota). This tree disagrees with the Pegasoferae

254   hypothesis[40], which groups bats with Perissodactyla, Carnivora and Pholidota, but agrees with

255   concatenation analyses of phylogenomic data[41]. Evolutionary studies based on 102

256   retroposons, including ILS-aware analyses, also support a sister-group relationship between

257   Chiroptera and Fereuungulata, but differ from the present analyses in supporting a sister-group

258   relationship between Carnivora and Cetartiodactyla[34,35]. However, as the number of

259   homologous sites increases in phylogenomic datasets, so too does bootstrap support[42], even

260   sometimes for an incorrect phylogeny[43], and as non-coding sequences can produce a different

261   topology than coding sequences [44], we further explored the phylogenomic signal within our

262   genomes.

263

264   To assess whether the tree inferred from the concatenated dataset (Fig. 2a) is also supported by

265   the non-coding part of the genome, we estimated a phylogeny using the models of best fit

266   (Table S9) for a dataset comprising 10,857 orthologous conserved non-coding elements

267   (CNEs), which contained 5,234,049 nucleotides and 1,225,098 parsimony-informative sites

268   (Table S10), using methods as described above. The result of this analysis (Fig. 2b) supports a

269   tree similar but not identical to that inferred from the protein-coding sequences (Fig. 2a),

270   including a sister-group relationship between Chiroptera and Fereuungulata, but with

271   Perissodactyla more closely related to Carnivora + Pholidota than to Cetartiodactyla. The CNE

272   tree also recovered a different position for *Tupaia* (Scandentia) within Euarchontoglires.

273

274   Given that two very short branches at the base of Scrotifera define relationships between its

275   four major clades (Carnivora + Pholidota, Cetartiodactyla, Chiroptera, Perissodactyla), this

276   region of the placental tree may be in the "anomaly zone", defined as a region of tree space

277   where the most common gene tree(s) differs from the species tree topology[45]. In the case of

278   four taxa and a rooted pectinate species tree, anomalous gene trees should be symmetric rather

279   than pectinate. To explore this, we estimated the maximum-likelihood support of each protein-

280   coding gene (n=12,931) for the 15 possible bifurcating topologies involving four clades, in our

281   case with Eulipotyphla as the outgroup (Fig. S6), and with the sub-trees for the relevant clades

282   identical to those in Fig. 2a. Based on the log-likelihood scores, 2,104 gene alignments

283   supported more than one tree, so these genes were excluded from further analysis. The

284   remaining 10,827 genes supported one fixed tree topology over the other 14 (Table S11), with

285   the number of genes supporting each topology highlighted in Fig. 2c. The best-supported

286   topology was that of our concatenated dataset for protein-coding genes (Fig. 2a; Tree1 with

287   1007/10827 genes), showing a sister group relationship between Chiroptera and Fereuungulata,

288   which is also supported by the CNEs (Fig. 2b). This suggests that the majority of the genome

289   supports a sister relationship between Chiroptera and the other Scrotifera. That said, there were

290   four other topologies that had support from >800 genes (Tree14 883/10827; Tree04 865/10827;

291   Tree15 820/10827; Tree13 806/10827) (Fig. 2c). However, even with similar support levels

292   for several topologies, the phylogenetic position for Chiroptera is pectinate on the most

293   common gene tree and does not qualify as anomalous. If the base of Scrotifera is in the anomaly

294   zone, as suggested by coalescence analyses of retroposon insertions[35], then we may expect the

295   most common gene tree(s) to be symmetric rather than pectinate. We may also expect the

296   species tree based on concatenation to be symmetric instead of pectinate[45]. One explanation for

297   the absence of anomalous gene trees, and for a pectinate species tree based on concatenation,

298   is that both protein-coding genes and CNEs are generally under purifying selection, which

299   reduces both coalescence times and incomplete lineage sorting relative to neutrally evolving

300   loci[46,47].

301

302   Bias in phylogenetic estimates can also be due to model misspecification, which is an

303   inadequate fit between phylogenetic data and the model of sequence evolution used[48].

304   Misleading support for incorrect phylogenies can also be due to gene tree error arising from a

305   lack of phylogenetic informativeness amongst data partitions[49]. To overcome these biases, we

306   performed a series of compatibility analyses on each gene partition and across the supermatrix

307   at 1st, 2nd and 3rd codon sites; $1^{st} + 2^{nd}$ codon sites; $1^{st} + 2^{nd} + 3^{rd}$ codon sites; amino acids,

308   assuming a 4-state alphabet for nucleotides and a 20 state-alphabet for amino acids (see

309   supplementary methods section 4.2). We excluded all alignments for which evidence of

310   saturation of substitutions and thus decay of the historical signal was detected by SatuRation

311   1.0 (https://github.com/lsjermiin/SatuRationSatuRation). Furthermore, we excluded all

312 alignments for which model mis-specification due to evolution under non-homogeneous

313 conditions was detected by the matched-pairs test of symmetry[50] implemented in Homo 2.0

314 (https://github.com/lsjermiin/Homo2.0).

315

316 A total of 488 gene alignments, consisting of 1st and 2nd codon positions and containing all

317 48 taxa, were considered optimal for phylogenetic analysis (Table S12). We concatenated these

318 data into a supermatrix of 241,098 nucleotides in length with 37,588 informative positions and

319 completed all phylogenetic analyses using methods as described above. However, this reduced

320 data set did not provide an unambiguous phylogenetic estimate. Specifically, while the best-

321 supported topology differed from the best trees inferred using all protein-coding genes and

322 CNEs in its position of Chiroptera, which is now sister to Carnivora + Pholidota (Fig. 2d), this

323 node has low bootstrap support (58%; topology 13; Fig. 2d) and Approximately Unbiased (AU)

324 tests could not reject the topologies depicted in Fig. 2a and 2b. Furthermore, the phylogeny

325 inferred from the subset of 488 genes is also symmetric for the four major lineages of

326 Scrotifera, as may be expected if this node is in the anomaly zone and concatenation is

327 misleading. We further analysed these data using a single-site coalescence-based method,

328 SVDquartets[51,52], which provides an alternative to concatenation. The resulting optimal

329 topology also supported Chiroptera as sister taxa to Fereuungulata (Fig. S7, topology 1), which

330 is the most supported position from all of our analyses and data partitions.

331

332 Taken together, multiple lines of evidence suggest that the majority of the genome supports

333 Chiroptera as sister to all other scrotiferans. However, different regions of the genome can and

334 do reflect alternative evolutionary scenarios. This highlights the importance of generating

335 phylogenetic inferences from multiple genomic regions and the importance of screening these

336 regions for violations of phylogenetic assumptions and incongruent signals, especially when

337 dealing with short internal branches.

338

339 **Genome-wide screens for gene selection, losses and gains**

340 To study the genomic basis of exceptional traits shared by bats, we first performed three

341 unbiased genome-wide screens for gene changes that occurred in the six bats. First, we

342 screened 12,931 genes classified as 1:1 orthologs for signatures of positive selection on the

343 ancestral bat (stem Chiroptera) branch under the aBSREL model[53] using HyPhy[54] and the

344 best-supported phylogeny (Fig. 2a). For genes with significant evidence for selection after

345 multiple test correction (FDR<0.05), we manually inspected the underlying alignment to

10

346 ensure homology (supplementary methods section 4.3.1), and additionally required that the

347 branch-site test implemented in PAML codeml[55] independently verified positive selection

348 (P<0.05). This revealed 9 genes with a robust signal of positive selection at the bat ancestor

349 (Table S13). While these 9 genes have diverse functions, they included two genes with hearing-

350 related functions, which may relate to the evolution of echolocation. These genes, *LRP2* (low-

351 density lipoprotein receptor-related protein 2, also called megalin) and *SERPINB6* (serpin

352 family B member 6) are expressed in the cochlea and associated with human disorders

353 involving deafness. *LRP2* encodes a multi-ligand receptor involved in endocytosis that is

354 expressed in the kidney, forebrain and, importantly, is also expressed in the cochlear duct[56].

355 Mutations in this gene are associated with Donnai-Barrow Syndrome, an autosomal recessive

356 disease with symptoms including sensorineural deafness[57], and progressive hearing loss has

357 also been observed in *Lrp2* knockout mice[58]. Similarly, *SERPINB6* is associated with non-

358 syndromic hearing loss and this serine protease inhibitor is expressed in cochlear hair cells[59,60].

359 Sites identified as having experienced positive selection at the bat ancestor showed bat-specific

360 substitutions in both genes. Interestingly, the echolocating bats showed a specific asparagine

361 to methionine substitution in *LRP2*. In *Rousettus,* the only non-laryngeal echolocator in our six

362 bats, this site has been substituted for a threonine. Combined with analysis of 6 other publicly

363 available bat genomes (n=6), we confirmed the presence of a methionine in all laryngeal

364 echolocating bats (n=9) and a threonine residue in all non-echolocating pteropodids (n=3) (Fig.

365 S8).

366

367 We also initially identified positive selection in the bat ancestor in a third hearing-related gene,

368 *TJP2* (tight junction protein 2), that is expressed in cochlear hair cells and associated with

369 hearing loss[61,62]. However, manual inspection revealed a putative alignment ambiguity and the

370 manually-corrected alignment had a reduced significance (aBSREL raw P=0.009, not

371 significant after multiple test correction considering 12,931 genes). Interestingly, the corrected

372 alignment revealed a four amino acid microduplication found only in echolocating bats (n=9)

373 (Fig. S9), which may be explained by incomplete lineage sorting or convergence. It should be

374 noted that insertions and deletions may also affect protein function but are not considered by

375 tests for positive selections, however a phylogenetic interpretation of these events may uncover

376 functional adaptations. In general, experimental studies are required to test whether the pattern

377 of positive selection and bat-specific mutations on the stem Chiroptera branch affect hearing-

378 related functions of these three genes. If so, this would provide molecular support for laryngeal

11

379   echolocation as a shared ancestral trait of bats and subsequent loss in pteropodids, informing a

380   long-standing debate in bat biology of whether ancestral bats had the ability to echolocate[12].

381

382   In addition to hearing-related genes, our genome-wide screen revealed selection on immunity-

383   related genes, *CXCL13* (C-X-C motif chemokine ligand 13), *NPSR1* (neuropeptide S receptor

384   1) and *INAVA* (innate immunity activator), which may underlie bats' unique tolerance of

385   pathogens[9]. The CXCL13 (previously B-lymphocyte chemoattractant) protein is a B-cell

386   specific chemokine, which attracts B-cells to secondary lymphoid organs, such as lymph nodes

387   and spleen[63]. *NPSR1* expresses a receptor activated by neuropeptide S. Activation of NSPR1

388   induces an inflammatory immune response in macrophages and *NSPR1* polymorphisms have

389   been associated with asthma in humans[64]. *INAVA* encodes an immunity-related protein with a

390   dual role in innate immunity. In intestinal epithelial cells, this gene is required for intestinal

391   barrier integrity and the repair of epithelial junctions after injury[65,66]. Consistent with these

392   functions, mutations in human *INAVA* are associated with inflammatory bowel disease[67], a

393   disorder characterized by chronic inflammation of the gastrointestinal tract and an increased

394   susceptibility to microbial pathogens. In macrophages, *INAVA* amplifies an IL-1β-induced pro-

395   inflammatory response by enhancing NF-kB signalling[66].

396

397   While a genome-wide screen for significant signatures of positive selection is comprehensive,

398   considering 12,931 orthologous genes may reduce sensitivity due to the necessity to correct for

399   12,931 statistical tests. To increase the sensitivity in detecting positive selection in genes

400   relevant for prominent bat traits (i.e. longevity, immunity, metabolism[2]) we further performed

401   a screen considering 2,453 candidate genes (Table S14) associated with these terms according

402   to Gene Ontology (GO), AmiGO[68] and GenAge[69] annotations. This reduced gene set permitted

403   a screen for signatures of positive selection using both the aBSREL model and the branch-site

404   test implemented in codeml (supplementary methods section 4.3.1). Requiring significance by

405   both aBSREL and codeml (FDR<0.05), we found 10 additional genes with robust evidence of

406   positive selection in the ancestral bat lineage (Table S15, Fig. S10). These genes include *IL17D*

407   and *IL-1β*, which are involved in immune system regulation [70] and NF-kB activation (*IL-*

408   *1β*)[66,71], and *GP2* and *LCN2*, which are involved in the response to pathogens[72,73]. Interestingly,

409   selection was also inferred for *PURB*, a gene that plays a role in cell proliferation and regulates

410   the oncogene *MYC*[74], which was previously shown to be under divergent selection in bats[16] and

411   which exhibits a unique anti-ageing transcriptomic profile in long lived *Myotis* bats[8]. Overall,

412   combining genome-wide and candidate gene screens revealed robust patterns of selection in

12

413   stem Chiroptera on several genes involved in immunity and infection, which suggests that

414   ancestral bats evolved immunomodulatory mechanisms that enabled a higher tolerance to

415   pathogens.

416

417   Second, we used a previously developed approach[75] to systematically screen for gene loss. This

418   revealed 10 genes that are inactivated in our six bats but present in the majority of

419   Laurasiatheria (Table S16). Two of these genes again point to changes in immune function in

420   bats, having immune-stimulating and pro-inflammatory functions; *LRRC70* (leucine rich repeat

421   containing 70, also called synleurin) and *IL36G* (interleukin 36 gamma) (Fig. 3a). *LRRC70* is

422   expressed in a broad range of tissues and potentiates cellular responses to multiple cytokines[76]

423   and is well conserved among Laurasiatheria. Importantly, *LRRC70* strongly amplifies bacterial

424   lipopolysaccharide-mediated NF-kB activation[76]. Our finding of *LRRC70* loss in bats makes

425   this poorly characterized gene an interesting target for future mechanistic studies. *IL36G*,

426   encodes a pro-inflammatory interleukin belonging to the interleukin-1 family. Increased

427   expression of *IL36G* was detected in psoriasis and inflammatory bowel disease patients, and

428   *IL36G* is likely involved in the pathophysiology of these diseases by inducing the canonical

429   NF-kB pathway and other proinflammatory cytokines[77-79]. Further analysis of common

430   mutations between our assembled genomes and previously published bat genomes (n=9),

431   revealed these genes were in fact lost multiple times within Chiroptera (Fig. S11 and S12),

432   suggesting these genes came under relaxed selection in bats followed by with subsequent gene

433   losses. Together, genome-wide screens for gene loss and positive selection revealed several

434   genes involved in NF-kB signalling (Fig. 3b), suggesting that altered NF-kB signalling may

435   contribute to immune related adaptations in bats.

436

437   Third, we investigated changes in gene family size, which revealed 35 cases of significant gene

438   family expansions and contractions at the bat ancestor (Table S17). Among these, we inferred

439   an expansion of the *APOBEC* gene family. Expansion involved *APOBEC3*-type genes (Fig.

440   3c) and supported a small expansion in the ancestral bat lineage, followed by up to 14

441   duplication events within Chiroptera. The *APOBEC3* locus is highly-dynamic, with a complex

442   history of duplication, loss and fusion in Mammalia[80]. Our analysis of this locus in Chiroptera

443   adds to previous evidence of a genus specific expansion in the flying foxes (genus *Pteropus*)[81],

444   showing this locus has undergone many independent expansions in bats. *APOBEC* genes are

445   DNA and RNA editing enzymes with roles in lipoprotein regulation and somatic

446   hypermutation[82]. *APOBEC3*-type genes have been previously associated with restricting viral

13

447   infection, transposon activity[82] and may also be stimulated by interferon signalling[83].

448   Expansion of *APOBEC3* genes in multiple bat lineages suggests these duplications may

449   contribute to viral tolerance in these lineages.

450

451   **Integrated Viruses in Bat Genomes**

452   There is mounting evidence to suggest that bats are major zoonotic reservoir hosts, as they can

453   tolerate and survive viral infections (e.g. Ebola and MERs), potentially due to adaptations in

454   their immune response[84], consistent with our findings of selection in immune-related genes

455   (e.g. *INAVA*) and expansions of the viral-restricting *APOBEC3* gene cluster. We screened our

456   high-quality genomes to ascertain the number and diversity of endogenous viral elements

457   (EVEs), considered as 'molecular fossil' evidence of ancient infections. Given their retroviral

458   life cycle endogenous retroviruses (ERVs) are the largest group found among all EVEs in

459   vertebrate genomes[85,86] (making up ~10% of the mouse[87] and 8% of the human genome[88]),

460   while non-retroviral EVEs are far less numerous in animal genomes[86].

461

462   Using reciprocal BLAST searches and a custom comprehensive library of viral protein

463   sequences we first screened our six bat genomes and seven mammalian outgroups

464   (supplementary methods section 3.4) for the presence of EVEs, including ERVs and non-

465   retroviral EVEs. We identified three predominant non-retroviral EVE families: *Parvoviridae*,

466   *Adenoviridae* and *Bornaviridae* (Fig. 4a). Parvovirus and bornavirus integrations were found

467   in all bats except for *Rousettus* and *M. molossus* respectively. A partial filovirus EVE was

468   found to be present in the Vespertilionidae (*Pipistrellus & Myotis*), but absent in the other bat

469   species, suggesting that vespertilionid bats have been exposed in the past to and can survive

470   filoviral infections, corroborating a previous study[89].

471

472   Next, we identified retroviral proteins from all ERV classes within the bat genomes. Consistent

473   with other mammals, the highest number of integrations came from beta- and gamma-like

474   retroviruses[90,91], with beta-like integrations most common for *pol* and *gag* proteins and gamma-

475   like integrations most common for *env* proteins in most of the bats (Fig. 4b & Fig. S13).

476   Overall, the highest number of integrations was observed in *M. myotis* (n=630), followed by

477   *Rousettus* (n=334) with *Phyllostomus* containing the lowest (n=126; Fig. 4b, Table S18).

478   Additionally, we detected ERV sequences with hits for alpha- and lenti-retroviruses in

479   reciprocal BLAST searches. Until now, alpharetroviruses were considered as exclusively

480   endogenous avian viruses[92]. Thus, our discovery of endogenous alpharetroviral-like elements

14

481   in bats is the first record of these sequences in mammalian genomes, widening the known

482   biodiversity of potential hosts for retrovirus transmission. We detected several alpha-like *env*

483   regions in *Phyllostomus, Rhinolophus*, and *Rousettus* (Fig. 4b), showing that multiple and

484   diverse bat species have been and possibly are being infected by alpharetroviruses. We also

485   detected lentivirus *gag*-like fragments in *Pipistrellus*, which are rarely observed in endogenized

486   form[93].

487

488   To identify historical ancestral transmission events, we reconstructed a phylogenetic tree from

489   our recovered ERVs with the known viral protein 'probe' sequences for all six bat genomes

490   and seven mammalian outgroups (Fig. S14). The majority of sequences group as single bat-

491   species clusters, suggesting that relatively recent integration events, more than ancestral

492   transmission (Fig. S14) govern the ERV diversity. While, most ERVs are simple retroviruses,

493   consisting of *gag*, *pol* and *env* genes, we found an unusual diversity of complex retroviruses in

494   bats, which are generally rare in endogenous form[93-95] (Fig. S14). We detected a clade of 5

495   *Rhinolophus pol* sequences clustered together with reference foamy retroviruses – Feline

496   Foamy Virus (FFV) and Bovine Foamy Virus (BFV). Foamy retroviruses in bats were detected

497   before from metagenomic data from *Rhinolophus affinis*[96], however, until now it was unclear

498   whether these sequences represented exogenous or endogenous viruses[97]. With the detection

499   of these sequences, we can now confirm the presence of endogenous spumaretroviruses in the

500   *R. ferrumequinum* genome, which furthers our understanding of the historical transmission

501   dynamics of this pathogen. We also detected *pol* sequences in the *Molossus* genome clustering

502   closely with reference delta sequences (Bovine Leukemia virus – BLV, Human T-

503   lymphotropic Virus – HTLV). *Pol* regions for delta retroviruses in bats have not been detected

504   before, with only partial gag and a single LTR identified previously in *Miniopterus* and

505   *Rhinolophus* species[94,98].

506

507   Overall these results show that bat genomes contain a surprising diversity of ERVs, with some

508   sequences never previously recorded in mammalian genomes, confirming interactions between

509   bats and complex retroviruses, which endogenize exceptionally rarely. These integrations are

510   indicative of past viral infections, highlighting which viruses bat species have co-evolved with

511   and tolerated, and thus, can help us better predict potential zoonotic spillover events and direct

512   routine viral monitoring in key species and populations. In addition, bats, as one of the largest

513   orders of mammals, are an excellent model to observe how co-evolution with viruses can shape

514   the mammalian genome over evolutionary timescales. For example, the expansion of the

15

515   *APOBEC3* genes in bats shown herein, could be a result of a co-evolutionary arms race shaped

516   by ancient retroviral invasions, and could contribute to the restriction in copy number of

517   endogenous viruses in some bat species. Given that these findings were generated from only

518   six bat genomes we can be confident that further cross-species comparison with similar quality

519   bat genomes will bring even greater insight.

520

521   **Changes in Non-Coding RNAs**

522   In addition to coding genes, changes in non-coding (nc)RNAs can be associated with

523   interspecific phenotypic variation and can drive adaptation[99,100]. We used our reference-quality

524   genomes to comprehensively annotate non-coding RNAs and search for ncRNA changes

525   between bat species and other mammals. To annotate different classes of conserved non-coding

526   RNA genes, we used computational methods that capture characteristic sequence and structure

527   features of ncRNAs (Fig. 5a; supplementary methods section 5.1). We found that a large

528   proportion of non-coding RNA genes were shared across all six bats (Fig. S15), and between

529   bats and other mammals (e.g. 95.8% ~ 97.4% shared between bats and human).

530

531   Within ncRNAs, we next investigated microRNAs (miRNA), which can serve as

532   developmental and evolutionary drivers of change[101]. We employed a strict pipeline to annotate

533   known miRNAs in our six bat genomes and in the 42 outgroup mammal taxa (Table S19,

534   supplementary methods section 5.1) and investigated how the size of miRNA families evolved

535   using CAFÉ[102]. We identified 286 miRNA families present in at least one mammal and

536   observed massive contractions of these miRNA families (Fig. S16) with an estimated overall

537   rate of 'death' 1.43 times faster than the rate of 'birth' (see supplementary methods section

538   5.1). There were 19 families that significantly (FDR<0.05) contracted in the ancestral bat

539   branch, with no evidence of expansions, and between 4 and 35 miRNA families were

540   contracted across bats (Fig. 5b, Fig. S16). We also inferred the miRNA families lost in each

541   bat lineage using a Dollo parsimony approach, which revealed 16 miRNA families that were

542   lost in the bat ancestor (Fig. S17 and S18). Interestingly, the oncogenic miR-374 was lost in all

543   bat species but was found in the other examined orders (Table S19). Since miR-374 promotes

544   tumour progression and metastasis in diverse human cancers[103], this bat specific loss may

545   contribute to low cancer rates in bats[16].

546

547   Next, we investigated the evolution of single-copy miRNA genes to determine if sequence

548   variation in these miRNAs may be driving biological change. Alignments of 98 highly

16

549   conserved, single-copy miRNA genes identified across all 48 mammal genomes revealed that

550   one miRNA, miR-337-3p, had unique variation in the seed region in bats compared to all other

551   42 mammals (Fig. 5c). miR-337-3p was pervasively expressed in brain, liver, and kidney across

552   all six bat species (Fig. S19). Given that the seed sequences of microRNAs represent the

553   strongest determinant of target specificity, these changes are expected to alter the repertoire of

554   sequences targeted by miR-337-3p in bats.

555

556   To test this hypothesis, we used reporter assays[104,105] to determine if the bat and human versions

557   of miR-337-3p were functionally active and if they showed species-specific regulation of an

558   "ideal" predicted target sequence (Table S20). While bat miR-337-3p strongly repressed the

559   expression of its cognate bat target sequence, it had no effect on the human site, and *vice versa*

560   (Fig. 5d). This result demonstrated that the miR-337-3p seed changes found in bats alter its

561   binding specificity. To explore whether this difference in binding specificity changes the set of

562   target genes regulated by bat miR-337-3p, we used our raw Iso-seq data to identify 3'UTRs of

563   coding genes in bats (n=6,891-16,115) and determined possible target genes of miR-337-3p

564   using a custom *in silico* pipeline (Table S21; supplementary methods section 5.3.5). We also

565   obtained the equivalent human 3'UTRs for all predicted bat 3'UTRs and identified the human

566   miR-337-3p gene targets (supplementary methods section 5.3.5). In bats, miR-337-3p was

567   predicted to regulate a distinct spectrum of gene targets compared to humans (Table S22). GO

568   enrichment analysis of these target gene sets suggests a shift towards regulation of

569   developmental, rhythmic, synaptic and behavioural gene pathways by miR-337-3p in bats (Fig.

570   5e), pointing to a dramatic change in processes regulated by miR-337-3p in bats compared to

571   other mammals.

572

573   In addition to losses and changes in miRNAs, continuous miRNA innovation is observed in

574   eukaryotes, which is suggested as a key player in the emergence of increasing organismal

575   complexity[99]. To identify any novel miRNAs that evolved in bats, we performed deep

576   sequencing of small RNA libraries from brain, liver and kidney for all six bats (Table S23),

577   analysed these data using a comprehensive custom analysis pipeline (see supplementary

578   methods section 5.3.3), and identified those miRNAs that possess seed regions not found in

579   miRBase (release 22). This screen revealed between 122 and 261 novel miRNAs across the six

580   bat genomes. Only a small number of these novel miRNAs were shared across the six bats,

581   supporting rapid birth of miRNAs on bat lineages (Fig. S20). We identified 12 novel miRNAs

582   that were found in all six bats but did not have apparent homologs in other mammals (Table

583   S24). Prediction of miRNAs from genomic sequences alone may result in false positives due

584   to the occurrence of short hairpin-forming sequences that are predicted to form hairpins but are

585   not processed or functionally active, emphasizing the need for experimental testing of these

586   miRNAs. Therefore, to test whether these candidates indeed function as miRNAs we selected

587   the top 3 candidates (bat-miR-4665, bat-miR-19125, bat-miR-6665) (Table S24) based on their

588   expression and secondary structures, and experimentally tested their ability to regulate an ideal

589   target sequence in reporter assays, as above (Table S20). Two of the three miRNAs (miR-

590   19125 and miR-6665) were able to regulate their targets, showing that they are actively

591   processed by the endogenous miRNA machinery, and able to be loaded onto the RISC complex

592   to repress target mRNAs (Fig. 5f). Thus, miR-19125 and miR-6665 represent true miRNAs

593   that are novel to bats. Taken together, these data demonstrate innovation in the bat lineage with

594   regard to miRNAs both in seed sequence variation as well as novel miRNA emergence.

595

596   In summary, our genomic screens and experiments revealed losses of ancestral miRNAs, gains

597   of novel functional miRNA and a striking case of miRNA seed change that alters the target

598   specificity. Changes in these miRNAs and their target genes point to a regulatory role in cancer,

599   development and behaviour in bats. Further detailed mechanistic studies will be crucial to

600   determine the role of these miRNAs in bat physiology and evolution.

601

602   **Discussion**

603   We have used a combination of state-of-the-art methods including long-read, short-read, and

604   scaffolding technologies to generate chromosome level, near-complete assemblies of six bats

605   that represent diversity within Chiroptera. These reference-quality genomes improve on all

606   published bat genomes and are on par with the best reference-quality genomes currently

607   generated for any eukaryotic species with a complex, multi-gigabyte genome. Compared to the

608   contiguity and completeness of previous bat genomes assembled with short reads, our

609   reference-quality genomes offer significant advances. First, while fragmented and incomplete

610   assemblies hamper gene annotation, reference-quality genomes allow comprehensive

611   annotations by integrating a variety of methods and evidence. In particular, reference-quality

612   genomes facilitate genome alignment, which provides a powerful way of transferring gene

613   annotations of related species to new assemblies and ensures that transcriptomic data can be

614   comprehensively mapped. Second, while fragmented and incomplete assemblies resulted in

615   countless efforts by individual labs to laboriously clone and re-sequence genomic loci

616   containing genes of interest, such efforts are not necessary with comprehensively annotated,

18

617   reference-quality assemblies. Third, reference-quality assemblies are a resource for studying
618   gene regulation by non-coding RNAs and cis-regulatory elements. The high completeness
619   enables a comprehensive mapping of functional genomics data such as miRNA read data and
620   epigenomic data (e.g. ChIP-seq, ATAC-Seq), and the high contiguity is crucial for assigning
621   regulatory regions to putative target genes and linking genotype to phenotype.

622

623   The six reference-quality assemblies coupled with methodological advances enabled us to
624   address the long-standing question of the phylogenetic position of bats within Laurasiatheria.
625   We used our comprehensive gene annotations to obtain the largest set of orthologous genes
626   and homologous regions to date, which enabled us to explore the phylogenetic signal across
627   different genomic partitions. Consistently, a variety of phylogenetic methods and data sets
628   estimate that bats are a sister clade to Fereuungulata and highlight the importance of
629   maximising the genetic coverage and ensuring that the appropriate models and data are used
630   when reconstructing difficult nodes.

631

632   Our comprehensive and conservative genome-wide screens investigating gene gain, loss and
633   selection provide candidates that are likely related to the unique immunity of bats. Furthermore,
634   our screens reveal selection in hearing genes in stem Chiroptera, which is consistent with the
635   hypothesis that echolocation evolved once in bats and was secondarily lost in Pteropodidae,
636   but inconsistent with the alternative hypothesis that echolocation evolved twice independently
637   within bats. As such, our analysis provides molecular evidence informing a long-standing
638   question of when echolocation evolved. We further show that bats have a long coevolutionary
639   history with viruses and identified unique mammalian viral integrations. Finally, we explored
640   the non-coding genome in bats, where we found miRNAs that were novel to bats, lost in bats,
641   or carried bat specific changes in their seed sequence. These important regulators of gene
642   expression point to ancestral changes in the bat genome that may have contributed to
643   adaptations related to the low incidence of cancer in bats, as well as developmental and
644   behavioural processes.

645

646   While the six bat genomes presented here are an excellent starting point to understand the
647   evolution of exceptional traits in bats, questions remain to be addressed in future studies,
648   particularly because bats as a group exhibit such an incredible diversity. To resolve the
649   phylogeny of the 21 currently-recognized bat families and to further understand the evolution
650   and molecular mechanisms of traits that vary among bat families, such as longevity, mode of

19

651   echolocation or diet, the Bat1K project aims at producing, in its next phase, reference-quality

652   assemblies for at least one member of each of the 21 bat families. To enable efficient use of

653   our reference-quality genomes, we provide all genomic and transcriptomic data together with

654   all annotation and genome alignment in an open access genome browser (https://genome-

655   public.pks.mpg.de) for download and visualization. These, and future bat genomes are

656   expected to provide a rich resource by which to address the evolution of the extraordinary

657   adaptations in bats and contribute to our understanding of key phenotypes including those

658   relevant for human health and disease.

References

1   Lazzeroni, M. E., Burbrink, F. T. & Simmons, N. B. Hibernation in bats (Mammalia: Chiroptera) did not evolve through positive selection of leptin. *Ecol Evol* **8**, 12576-12596, doi:10.1002/ece3.4674 (2018).

2   Teeling, E. C. *et al.* Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu Rev Anim Biosci* **6**, 23-46, doi:10.1146/annurev-animal-022516-022811 (2018).

3   Kacprzyk, J. *et al.* A potent anti-inflammatory response in bat macrophages may be linked to extended longevity and viral tolerance. *Acta chiropterologica* **19**, 219-228 (2017).

4   Ahn, M., Cui, J., Irving, A. T. & Wang, L. F. Unique Loss of the PYHIN Gene Family in Bats Amongst Mammals: Implications for Inflammasome Sensing. *Sci Rep* **6**, 21722, doi:10.1038/srep21722 (2016).

5   O'Shea, T. J. *et al.* Bat flight and zoonotic viruses. *Emerg Infect Dis* **20**, 741-745, doi:10.3201/eid2005.130539 (2014).

6   Brook, C. E. & Dobson, A. P. Bats as 'special' reservoirs for emerging zoonotic pathogens. *Trends Microbiol* **23**, 172-180, doi:10.1016/j.tim.2014.12.004 (2015).

7   Wilkinson, G. S. & Adams, D. M. Recurrent evolution of extreme longevity in bats. *Biol Lett* **15**, 20180860, doi:10.1098/rsbl.2018.0860 (2019).

8   Huang, Z. *et al.* Longitudinal comparative transcriptomics reveals unique mechanisms underlying extended healthspan in bats. *Nat Ecol Evol* **3**, 1110-1120, doi:10.1038/s41559-019-0913-3 (2019).

9   Mandl, J. N., Schneider, C., Schneider, D. S. & Baker, M. L. Going to Bat(s) for Studies of Disease Tolerance. *Front Immunol* **9**, 2112, doi:10.3389/fimmu.2018.02112 (2018).

10   Jones, G., Teeling, E. C. & Rossiter, S. J. From the ultrasonic to the infrared: molecular evolution and the sensory biology of bats. *Front Physiol* **4**, 117, doi:10.3389/fphys.2013.00117 (2013).

11   Vernes, S. C. What bats have to say about speech and language. *Psychon Bull Rev* **24**, 111-117, doi:10.3758/s13423-016-1060-3 (2017).

12   Wang, Z. *et al.* Prenatal development supports a single origin of laryngeal echolocation in bats. *Nat Ecol Evol* **1**, 21, doi:10.1038/s41559-016-0021 (2017).

13   Brown, E. E., Cashmore, D. D., Simmons, N. B. & Butler, R. J. Quantifying the completeness of the bat fossil record. *Palaeontology* (2019).

14   Foley, N. M., Springer, M. S. & Teeling, E. C. Mammal madness: is the mammal tree of life not yet resolved? *Philos Trans R Soc Lond B Biol Sci* **371**, doi:10.1098/rstb.2015.0140 (2016).

15   Teeling, E. C. *et al.* A molecular phylogeny for bats illuminates biogeography and the fossil record. *Science* **307**, 580-584, doi:10.1126/science.1105113 (2005).

16   Foley, N. M. *et al.* Growing old, yet staying young: The role of telomeres in bats' exceptional longevity. *Sci Adv* **4**, eaao0926, doi:10.1126/sciadv.aao0926 (2018).

17   Hayman, D. T. Bats as Viral Reservoirs. *Annu Rev Virol* **3**, 77-99, doi:10.1146/annurev-virology-110615-042203 (2016).

18   Nowoshilow, S. *et al.* The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554**, 50-55, doi:10.1038/nature25458 (2018).

19   Tischler, G. Haplotype and Repeat Separation in Long Reads. *bioRxiv*, 145474, doi:10.1101/145474 (2017).

20   Tischler, G. & Myers, E. W. Non hybrid long read consensus using local de Bruijn graph assembly. *bioRxiv*, 106252 (2017).

21   Eckalbar, W. L. *et al.* Transcriptomic and epigenomic characterization of the developing bat wing. *Nat Genet* **48**, 528-536, doi:10.1038/ng.3537 (2016).

21

709 22 Pavlovich, S. S. *et al.* The Egyptian Rousette Genome Reveals Unexpected Features of
710   Bat Antiviral Immunity. *Cell* **173**, 1098-1110 e1018, doi:10.1016/j.cell.2018.03.070
711   (2018).
712 23 Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning.
713   *Nat Biotechnol*, doi:10.1038/nbt.4277 (2018).
714 24 Nature Biotechnology Editorial. A reference standard for genome biology. *Nat
715   Biotechnol* **36**, 1121, doi:10.1038/nbt.4318 (2018).
716 25 Waterhouse, R. M. *et al.* BUSCO applications from quality assessments to gene
717   prediction and phylogenomics. *Mol Biol Evol*, doi:10.1093/molbev/msx319 (2017).
718 26 Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321-
719   1325, doi:10.1126/science.1098119 (2004).
720 27 Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford)* **2016**,
721   doi:10.1093/database/baw093 (2016).
722 28 Sharma, V., Schwede, P. & Hiller, M. CESAR 2.0 substantially improves speed and
723   accuracy of comparative gene annotation. *Bioinformatics* **33**, 3985-3987,
724   doi:10.1093/bioinformatics/btx527 (2017).
725 29 Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in
726   eukaryotes with a generalized hidden Markov model that uses hints from external
727   sources. *BMC Bioinformatics* **7**, 62, doi:10.1186/1471-2105-7-62 (2006).
728 30 Kuo, R. I., Cheng, Y., Smith, J., Archibald, A. L. & Burt, D. W. Illuminating the dark
729   side of the human transcriptome with TAMA Iso-Seq analysis. *bioRxiv* (2019).
730 31 Platt, R. N., 2nd, Blanco-Berdugo, L. & Ray, D. A. Accurate Transposable Element
731   Annotation Is Vital When Analyzing New Genome Assemblies. *Genome Biol Evol* **8**,
732   403-410, doi:10.1093/gbe/evw009 (2016).
733 32 Pace, J. K., 2nd & Feschotte, C. The evolutionary history of human DNA transposons:
734   evidence for intense activity in the primate lineage. *Genome Res* **17**, 422-432,
735   doi:10.1101/gr.5826307 (2007).
736 33 Platt, R. N., 2nd, Mangum, S. F. & Ray, D. A. Pinpointing the vesper bat transposon
737   revolution using the Miniopterus natalensis genome. *Mob DNA* **7**, 12,
738   doi:10.1186/s13100-016-0071-y (2016).
739 34 Doronina, L. *et al.* Speciation network in Laurasiatheria: retrophylogenomic signals.
740   *Genome Res* **27**, 997-1003, doi:10.1101/gr.210948.116 (2017).
741 35 Springer, M. S. & Gatesy, J. An ABBA-BABA Test for Introgression Using
742   Retroposon Insertion Data. *bioRxiv*, doi:10.1101/709477 (2019).
743 36 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
744   ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*
745   **14**, 587-589, doi:10.1038/nmeth.4285 (2017).
746 37 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and
747   effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol
748   Biol Evol* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).
749 38 Tarver, J. E. *et al.* The Interrelationships of Placental Mammals and the Limits of
750   Phylogenetic Inference. *Genome Biol Evol* **8**, 330-344, doi:10.1093/gbe/evv261 (2016).
751 39 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2:
752   Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522,
753   doi:10.1093/molbev/msx281 (2018).
754 40 Nishihara, H., Hasegawa, M. & Okada, N. Pegasoferae, an unexpected mammalian
755   clade revealed by tracking ancient retroposon insertions. *Proc Natl Acad Sci U S A* **103**,
756   9929-9934, doi:10.1073/pnas.0603797103 (2006).

41   Tsagkogeorga, G., Parker, J., Stupka, E., Cotton, J. A. & Rossiter, S. J. Phylogenomic analyses elucidate the evolutionary relationships of bats. *Curr Biol* **23**, 2262-2267, doi:10.1016/j.cub.2013.09.014 (2013).

42   Jermiin, L. S., Poladian, L. & Charleston, M. A. Evolution. Is the "Big Bang" in animal evolution real? *Science* **310**, 1910-1911, doi:10.1126/science.1122440 (2005).

43   Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* **9**, e1000602, doi:10.1371/journal.pbio.1000602 (2011).

44   Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311-1320, doi:10.1126/science.1251385 (2014).

45   Degnan, J. H. & Rosenberg, N. A. Discordance of species trees with their most likely gene trees. *PLoS Genet* **2**, e68, doi:10.1371/journal.pgen.0020068 (2006).

46   Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10**, 195-205, doi:10.1038/nrg2526 (2009).

47   Hobolth, A., Andersen, L. N. & Mailund, T. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics* **187**, 1241-1243, doi:10.1534/genetics.110.124164 (2011).

48   Jermiin, L. S., Jayaswal, V., Ababneh, F. M. & Robinson, J. Identifying Optimal Models of Evolution. *Methods Mol Biol* **1525**, 379-420, doi:10.1007/978-1-4939-6622-6_15 (2017).

49   Dornburg, A., Su, Z. & Townsend, J. P. Optimal Rates for Phylogenetic Inference and Experimental Design in the Era of Genome-Scale Data Sets. *Syst Biol* **68**, 145-156, doi:10.1093/sysbio/syy047 (2019).

50   Ababneh, F., Jermiin, L. S., Ma, C. & Robinson, J. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics* **22**, 1225-1231, doi:10.1093/bioinformatics/btl064 (2006).

51   Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317-3324, doi:10.1093/bioinformatics/btu530 (2014).

52   Chou, J. *et al.* A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics* **16 Suppl 10**, S2, doi:10.1186/1471-2164-16-S10-S2 (2015).

53   Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol* **32**, 1342-1353, doi:10.1093/molbev/msv022 (2015).

54   Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679, doi:10.1093/bioinformatics/bti079 (2005).

55   Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586-1591 (2007).

56   Mizuta, K. *et al.* Ultrastructural localization of megalin in the rat cochlear duct. *Hear Res* **129**, 83-91, doi:10.1016/s0378-5955(98)00221-4 (1999).

57   Kantarci, S. *et al.* Mutations in LRP2, which encodes the multiligand receptor megalin, cause Donnai-Barrow and facio-oculo-acoustico-renal syndromes. *Nat Genet* **39**, 957-959, doi:10.1038/ng2063 (2007).

58   Konig, O. *et al.* Estrogen and the inner ear: megalin knockout mice suffer progressive hearing loss. *FASEB J* **22**, 410-417, doi:10.1096/fj.07-9171com (2008).

59   Sirmaci, A. *et al.* A truncating mutation in SERPINB6 is associated with autosomal-recessive nonsyndromic sensorineural hearing loss. *Am J Hum Genet* **86**, 797-804, doi:10.1016/j.ajhg.2010.04.004 (2010).

23

805 60   Tan, J., Prakash, M. D., Kaiserman, D. & Bird, P. I. Absence of SERPINB6A causes
806   sensorineural hearing loss with multiple histopathologies in the mouse inner ear. *Am J*
807   *Pathol* **183**, 49-59, doi:10.1016/j.ajpath.2013.03.009 (2013).
808 61   Walsh, T. *et al.* Genomic duplication and overexpression of TJP2/ZO-2 leads to altered
809   expression of apoptosis genes in progressive nonsyndromic hearing loss DFNA51. *Am*
810   *J Hum Genet* **87**, 101-109, doi:10.1016/j.ajhg.2010.05.011 (2010).
811 62   Hilgert, N. *et al.* Mutation analysis of TMC1 identifies four new mutations and suggests
812   an additional deafness gene at loci DFNA36 and DFNB7/11. *Clin Genet* **74**, 223-232,
813   doi:10.1111/j.1399-0004.2008.01053.x (2008).
814 63   Gunn, M. D. *et al.* A B-cell-homing chemokine made in lymphoid follicles activates
815   Burkitt's lymphoma receptor-1. *Nature* **391**, 799-803, doi:10.1038/35876 (1998).
816 64   Vendelin, J. *et al.* Downstream target genes of the neuropeptide S-NPSR1 pathway.
817   *Hum Mol Genet* **15**, 2923-2935, doi:10.1093/hmg/ddl234 (2006).
818 65   Mohanan, V. *et al.* C1orf106 is a colitis risk gene that regulates stability of epithelial
819   adherens junctions. *Science* **359**, 1161-1166, doi:10.1126/science.aan0814 (2018).
820 66   Luong, P. *et al.* INAVA-ARNO complexes bridge mucosal barrier function with
821   inflammatory signaling. *Elife* **7**, doi:10.7554/eLife.38539 (2018).
822 67   Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of
823   inflammatory bowel disease. *Nature* **491**, 119-124, doi:10.1038/nature11582 (2012).
824 68   Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics*
825   **25**, 288-289 (2008).
826 69   de Magalhaes, J. P. & Toussaint, O. GenAge: a genomic and proteomic network map
827   of human ageing. *FEBS letters* **571**, 243-247 (2004).
828 70   Saddawi-Konefka, R. *et al.* Nrf2 Induces IL-17D to Mediate Tumor and Virus
829   Surveillance. *Cell Rep* **16**, 2348-2358, doi:10.1016/j.celrep.2016.07.075 (2016).
830 71   Barker, B. R., Taxman, D. J. & Ting, J. P. Cross-regulation between the IL-1beta/IL-
831   18 processing inflammasome and other inflammatory cytokines. *Curr Opin Immunol*
832   **23**, 591-597, doi:10.1016/j.coi.2011.07.005 (2011).
833 72   Flo, T. H. *et al.* Lipocalin 2 mediates an innate immune response to bacterial infection
834   by sequestrating iron. *Nature* **432**, 917-921, doi:10.1038/nature03104 (2004).
835 73   Hase, K. *et al.* Uptake through glycoprotein 2 of FimH(+) bacteria by M cells initiates
836   mucosal immune response. *Nature* **462**, 226-230, doi:10.1038/nature08529 (2009).
837 74   Xu-Monette, Z. Y. *et al.* Clinical and Biologic Significance of MYC Genetic Mutations
838   in De Novo Diffuse Large B-cell Lymphoma. *Clin Cancer Res* **22**, 3593-3605,
839   doi:10.1158/1078-0432.CCR-15-2296 (2016).
840 75   Sharma, V. *et al.* A genomics approach reveals insights into the importance of gene
841   losses for mammalian adaptations. *Nat Commun* **9**, 1215, doi:10.1038/s41467-018-
842   03667-1 (2018).
843 76   Wang, W., Yang, Y., Li, L. & Shi, Y. Synleurin, a novel leucine-rich repeat protein that
844   increases the intensity of pleiotropic cytokine responses. *Biochem Biophys Res*
845   *Commun* **305**, 981-988, doi:10.1016/s0006-291x(03)00876-3 (2003).
846 77   Johnston, A. *et al.* IL-1F5, -F6, -F8, and -F9: a novel IL-1 family signaling system that
847   is active in psoriasis and promotes keratinocyte antimicrobial peptide expression. *J*
848   *Immunol* **186**, 2613-2622, doi:10.4049/jimmunol.1003162 (2011).
849 78   Nishida, A. *et al.* Increased Expression of Interleukin-36, a Member of the Interleukin-
850   1 Cytokine Family, in Inflammatory Bowel Disease. *Inflamm Bowel Dis* **22**, 303-314,
851   doi:10.1097/MIB.0000000000000654 (2016).
852 79   Bridgewood, C. *et al.* IL-36gamma has proinflammatory effects on human endothelial
853   cells. *Exp Dermatol* **26**, 402-408, doi:10.1111/exd.13228 (2017).

24

854   80   Munk, C., Willemsen, A. & Bravo, I. G. An ancient history of gene duplications,
855   fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evol Biol*
856   **12**, 71, doi:10.1186/1471-2148-12-71 (2012).
857   81   Hayward, J. A. *et al.* Differential Evolution of Antiretroviral Restriction Factors in
858   Pteropid Bats as Revealed by APOBEC3 Gene Complexity. *Mol Biol Evol* **35**, 1626-
859   1637, doi:10.1093/molbev/msy048 (2018).
860   82   Salter, J. D., Bennett, R. P. & Smith, H. C. The APOBEC Protein Family: United by
861   Structure, Divergent in Function. *Trends Biochem Sci* **41**, 578-594,
862   doi:10.1016/j.tibs.2016.05.001 (2016).
863   83   Roper, N. *et al.* APOBEC Mutagenesis and Copy-Number Alterations Are Drivers of
864   Proteogenomic Tumor Evolution and Heterogeneity in Metastatic Thoracic Tumors.
865   *Cell Rep* **26**, 2651-2666 e2656, doi:10.1016/j.celrep.2019.02.028 (2019).
866   84   Subudhi, S., Rapin, N. & Misra, V. Immune System Modulation and Viral Persistence
867   in Bats: Understanding Viral Spillover. *Viruses* **11**, doi:10.3390/v11020192 (2019).
868   85   Aswad, A. & Katzourakis, A. Paleovirology and virally derived immunity. *Trends Ecol*
869   *Evol* **27**, 627-636, doi:10.1016/j.tree.2012.07.007 (2012).
870   86   Katzourakis, A. & Gifford, R. J. Endogenous viral elements in animal genomes. *PLoS*
871   *Genet* **6**, e1001191, doi:10.1371/journal.pgen.1001191 (2010).
872   87   Mouse Genome Sequencing, C. *et al.* Initial sequencing and comparative analysis of
873   the mouse genome. *Nature* **420**, 520-562, doi:10.1038/nature01262 (2002).
874   88   Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**,
875   860-921, doi:10.1038/35057062 (2001).
876   89   Taylor, D. J., Dittmar, K., Ballinger, M. J. & Bruenn, J. A. Evolutionary maintenance
877   of filovirus-like genes in bat genomes. *BMC Evol Biol* **11**, 336, doi:10.1186/1471-2148-
878   11-336 (2011).
879   90   Hayward, A., Grabherr, M. & Jern, P. Broad-scale phylogenomics provides insights
880   into retrovirus-host evolution. *Proc Natl Acad Sci U S A* **110**, 20146-20151,
881   doi:10.1073/pnas.1315419110 (2013).
882   91   Skirmuntt, E. C. & Katzourakis, A. The evolution of endogenous retroviral envelope
883   genes in bats and their potential contribution to host biology. *Virus Res* **270**, 197645,
884   doi:10.1016/j.virusres.2019.197645 (2019).
885   92   Xu, X., Zhao, H., Gong, Z. & Han, G. Z. Endogenous retroviruses of non-
886   avian/mammalian vertebrates illuminate diversity and deep history of retroviruses.
887   *PLoS Pathog* **14**, e1007072, doi:10.1371/journal.ppat.1007072 (2018).
888   93   Katzourakis, A., Tristem, M., Pybus, O. G. & Gifford, R. J. Discovery and analysis of
889   the first endogenous lentivirus. *Proc Natl Acad Sci U S A* **104**, 6261-6265,
890   doi:10.1073/pnas.0700471104 (2007).
891   94   Farkasova, H. *et al.* Discovery of an endogenous Deltaretrovirus in the genome of long-
892   fingered bats (Chiroptera: Miniopteridae). *Proc Natl Acad Sci U S A* **114**, 3145-3150,
893   doi:10.1073/pnas.1621224114 (2017).
894   95   Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. & Pybus, O. G.
895   Macroevolution of complex retroviruses. *Science* **325**, 1512,
896   doi:10.1126/science.1174149 (2009).
897   96   Wu, Z. *et al.* Virome analysis for identification of novel mammalian viruses in bat
898   species from Chinese provinces. *J Virol* **86**, 10999-11012, doi:10.1128/JVI.01394-12
899   (2012).
900   97   Katzourakis, A. *et al.* Larger mammalian body size leads to lower retroviral activity.
901   *PLoS Pathog* **10**, e1004214, doi:10.1371/journal.ppat.1004214 (2014).
902   98   Hron, T. *et al.* Remnants of an Ancient Deltaretrovirus in the Genomes of Horseshoe
903   Bats (Rhinolophidae). *Viruses* **10**, doi:10.3390/v10040185 (2018).

99   Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* **12**, 846-860, doi:10.1038/nrg3079 (2011).

100   Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. & Peterson, K. J. MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci U S A* **105**, 2946-2950, doi:10.1073/pnas.0712259105 (2008).

101   Moran, Y., Agron, M., Praher, D. & Technau, U. The evolutionary origin of plant and animal microRNAs. *Nat Ecol Evol* **1**, 27, doi:10.1038/s41559-016-0027 (2017).

102   De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271, doi:10.1093/bioinformatics/btl097 (2006).

103   Zhang, J. *et al.* Upregulation of miR-374a promotes tumor metastasis and progression by downregulating LACTB and predicts unfavorable prognosis in breast cancer. *Cancer Med*, doi:10.1002/cam4.1576 (2018).

104   Devanna, P., van de Vorst, M., Pfundt, R., Gilissen, C. & Vernes, S. C. Genome-wide investigation of an ID cohort reveals de novo 3'UTR variants affecting gene expression. *Hum Genet* **137**, 717-721, doi:10.1007/s00439-018-1925-9 (2018).

105   Devanna, P. *et al.* Next-gen sequencing identifies non-coding variation disrupting miRNA-binding sites in neurological disorders. *Mol Psychiatry* **23**, 1375-1384, doi:10.1038/mp.2017.30 (2018).

Data availability statement

All data generated or analysed during this study are included in this published article and its supplementary information files. All genomic and transcriptomic data are publicly available for visualization and download via the open-access Bat1K genome browser (https://genome-public.pks.mpg.de). In addition, the assemblies have been deposited in the NCBI database and GenomeArk (https://vgp.github.io/genomeark/). Accession numbers for all data deposits can be found in the supplementary information files of this article.

Code availability statement

All code has been made available on github. Details of the location can be found in the supplementary information files of this article. Other custom software is available upon request.

Author contributions

MH, SCV, EWM and ECT conceived and supervised the project. MH, SCV, EWM and ECT provided funding. MLP, SJP, DD, GJ, RDR, AGL, ECT and SCV provided tissue samples for sequencing. ZH, JGR, OF and SW were responsible for nucleic acid extraction and sequencing. MP assembled and curated all genomes. DJ provided coding gene annotation and was responsible for coding gene evolutionary analysis. DJ provided multiple sequence and genome alignments. MH and DJ analysed UCE and genome completeness. DJ and MH established the

944   Bat1K genome browser. ZH provided non-coding gene annotation and was responsible for

945   non-coding gene evolutionary analysis. KL processed Iso-seq data and provided UTR

946   annotation. ZH, KL, PD and SCV conducted miRNA target prediction and gene ontology

947   enrichment. PD conducted miRNA functional experiments. GMH, LSJ, MS and ECT provided

948   phylogenomic analyses. GMH and LMD were responsible for codeml analysis. DJ, MH, ZH,

949   GMH, ECT, LMD and AP interpreted evolutionary analyses. BMK and MH developed the

950   TOGA gene projection tool and BMK provided projections for non-bat mammals. ECS, LBG

951   and AK provided EVE annotation and analysis. DR and KAMS provided TE annotation and

952   analysis. EDJ provided support for sequencing of *Phyllostomus* and *Rhinolophus* genomes. DJ,

953   ZH, MP, GH, MH, SCV, EWM and ECT wrote the manuscript. All authors provided edit and
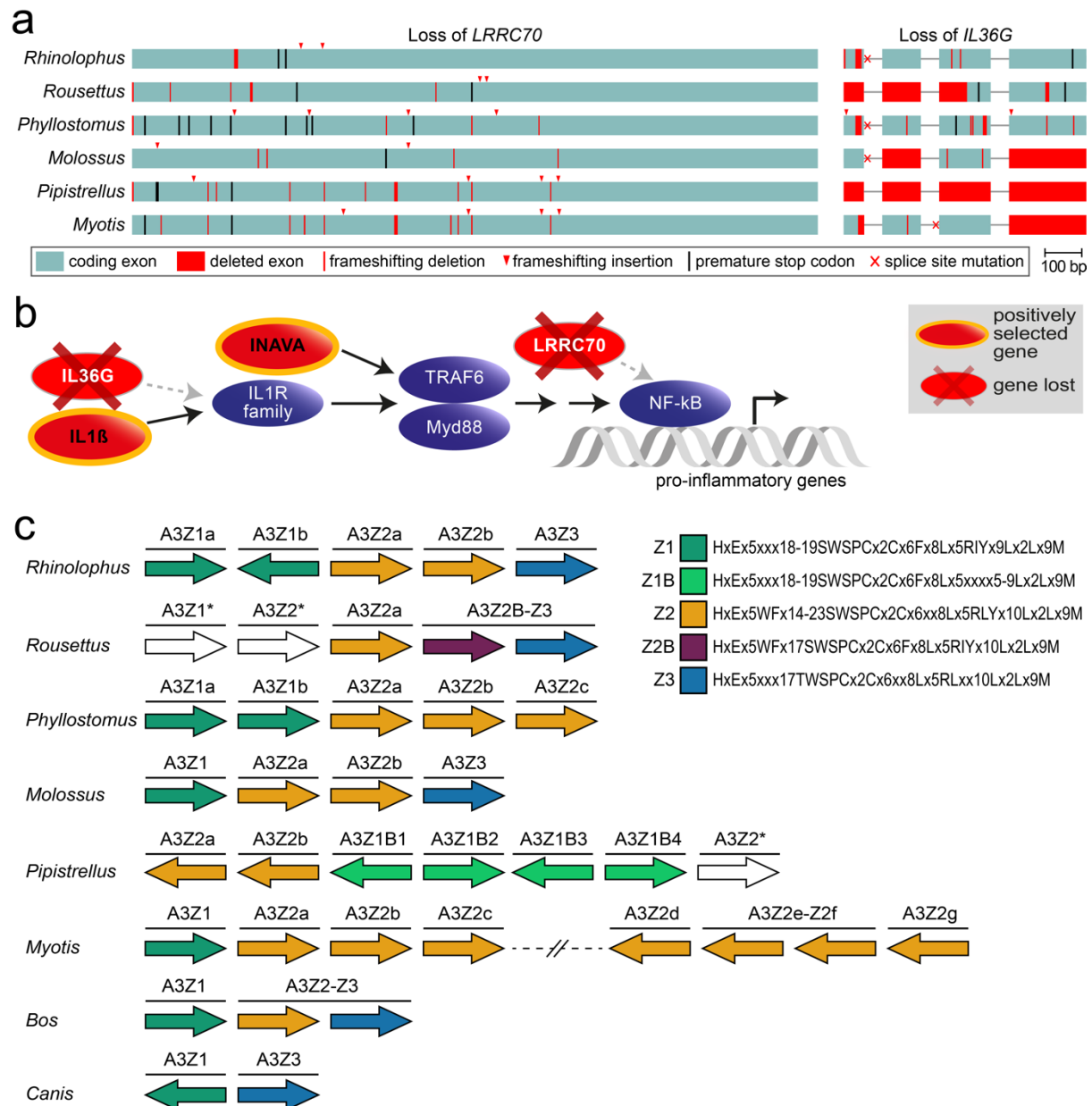
954   comment.

955

27

Figure 1: Assembly and annotation of the genomes of six bats. (a) Genome assembly strategy and the amount of data produced for assembling contigs and scaffolds. (b) Comparison of assembly contiguity. N(x)% graphs show the contig (left) and scaffold (right) sizes (y-axis), where x% of the assembly consists of contigs and scaffolds of at least that size. Dashed lines show contiguities of two recent bat assemblies, *Miniopterus* generated from short read data[21], and *Rousettus* generated from a hybrid of short and long read data[22]. (c) Comparison of coding gene completeness. Bar charts show the percent of 4104 highly-conserved mammalian BUSCO genes that are completely present, fragmented or missing in the assembly. (d) Comparison of completeness in non-exonic regions. Bar charts show the number of detected ultraconserved elements that align at stringent parameters. Ultraconserved elements not detected are separated into those that are missing due to assembly incompleteness and those that exhibit real sequence
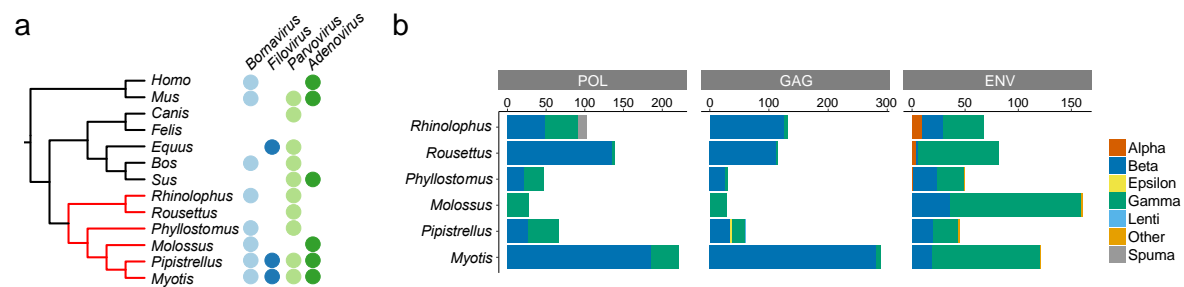
991   divergence. Note that human and mouse are not shown here because both genomes were used
992   to define ultraconserved elements[26]. (e) Our strategy to annotate coding genes combining
993   various types of gene evidences. (f) Comparison of the completeness of gene annotations, using
994   4101 BUSCO genes, and the number of annotated genes. (f) Bar charts compare genome sizes
995   and the proportion that consist of major transposon classes. (g) Fraction of the genome that
996   consists of recent transposon insertions, defined as transposons that diverged less than 6.6%
997   from their consensus sequence.
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030

Figure 2: Phylogenetic analysis of Laurasiatheria. (a) We inferred a mammalian phylogram using a supermatrix of 12,931 concatenated genes and the maximum likelihood method of tree reconstruction (topology 1, Fig. S6). (b) A total of 10,857 conserved non-coding elements (CNEs) were used to determine a mammalian phylogeny using non-coding regions (topology 2, Fig. S6). Bootstrap support values less than 100 are displayed, with internal nodes that differ to the protein-coding supermatrix highlighted in red. (c) All gene alignments were fit to the 15 laurasiatherian topologies (Fig. S6) explored to determine which tree had the highest likelihood score for each gene. The number of genes supporting each topology are displayed. (d) A supermatrix consisting of 1st and 2nd codon sites from 448 genes that are evolving under homogenous conditions, thus considered optimal 'fit' for phylogenetic analysis, was used to infer a phylogeny using maximum likelihood (topology13 Feig. S6). Bootstrap support values less than 100 are displayed, with internal nodes that differ to the protein-coding supermatrix phylogeny highlighted in red.
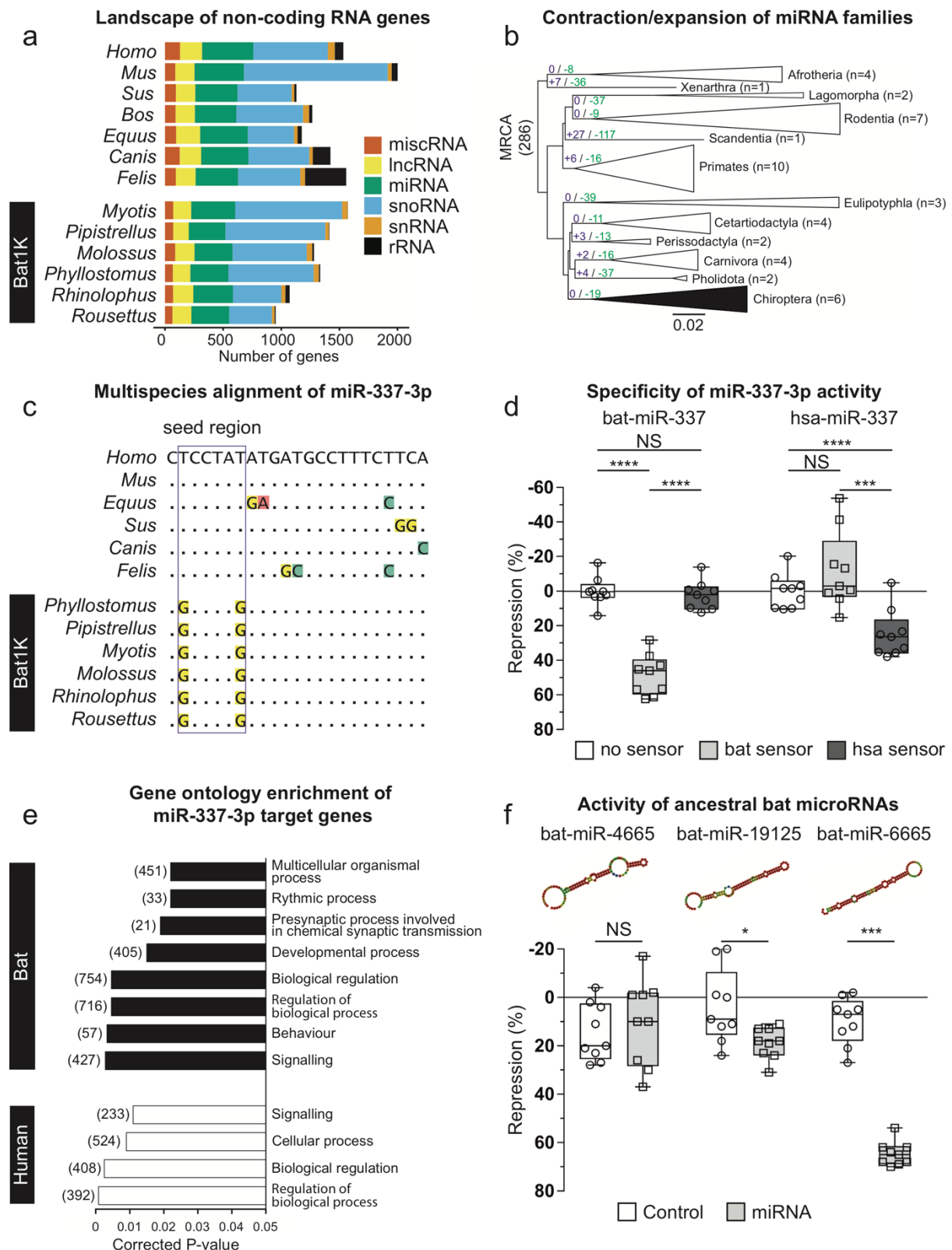
30

Figure 3: Genome-wide screens highlight changes in genes potentially involved in bat's unique immunity. (a) Inactivation of the immune genes *LRRC70* and *IL36G*. Boxes represent coding exons proportional to their size, overlaid with gene-inactivating mutations present in the six bats. (b) Diagram showing the canonical NF-kB signalling pathway (purple) and interacting proteins which have experienced positive selection or have been lost in bats. (c) Expansion of the *APOBEC3* gene locus in bats. Each arrow represents a cytidine deaminase domain, coloured by domain subtypes as defined by given motifs, with likely pseudogenes are in white. Genes containing multiple deaminase domains are shown as a single bar over more than one domain. A transposition event in *Myotis* has created two *APOBEC3* loci on different chromosomes, indicated by the broken line in this species. Cow and dog are shown as two Laurasiatheria outgroups, where cow also represents the likely, mammalian ancestral state.

31

Figure 4: Endogenous Viruses in Bat Genomes (a) Viral families identified in more than one genus mapped to phylogenetic tree of six bat species and seven additional mammals. Endogenous sequences identified as *Adenoviridae*, *Parvoviridae*, *Filoviridae* and *Bornaviridae* were represented across several mammalian genera. (b) Bar plot showing numbers of sequences found for each of the viral proteins in six species of bat and the representation in all seven *Retroviridae* genera.

32

Figure 5: The evolution of non-coding RNA genes in bats (a) The number of non-coding RNA genes annotated in six bat genomes and 7 reference mammalian genomes. (b) miRNA family expansion and contraction analyses in 48 mammalian genomes. The numbers highlighted on the branches designate the number of miRNA families expanded (purple, +) and contracted (green, -) at the order level. n indicates the number of species in each order used in the analysis. (c) The alignment of the mature miR-337-3p sequences across six bats and six

1078 reference species (miR-337-3p could not be found in *Bos taurus* genome). The box indicates
1079 the seed region of mature miR-337-3p, which is conserved across mammals, but divergent in
1080 bats. (d) Specificity of human (hsa) and bat miR-337-3p activity was shown using species
1081 specific sensors in luciferase reporter assays (n=9 per experiment; see supplementary methods
1082 section 5.4). Significance was calculated using two-way ANOVA test, followed by post-hoc
1083 Tukey calculation. Statistical significance is indicated as: ***p<0.001; ****p<0.0001. (e)
1084 Gene ontology enrichment (via DAVID) of targets predicted for human and bat miR-337-3p
1085 (f) Validation of the activity of ancestral bat miRNAs, absent in the other mammalian genomes.
1086 The predicted secondary structures for each novel miRNA are displayed. For each miRNA, the
1087 sensor was tested against a control unrelated miRNA that was not predicted to bind to the
1088 sensor (left) and the cognate miRNA (right) in luciferase reporter assays (n=9 per experiment;
1089 see supplementary methods section 5.4). Significance for each independent control-miRNA
1090 pair was calculated using pairwise t tests. Statistical significance is indicated as: *p<0.05;
1091 ***p<0.001. Box plots extend from the 25th to 75th percentiles, the central line represents the
1092 median value, and whiskers are drawn using the function "min to max" in GraphPad Prism7
1093 (GraphPad Software, La Jolla California USA, http://www.graphpad.com) and go down to the
1094 smallest value and up to the largest.
1095