Comparative and population genomics analyses of eared pheasants inhabiting highly varying altitudes

Siwen Wu[2†], Kun Wang[1,3†], Xuehai Ge[1†], Sisi Yuan[2], Dong-Dong Wu[3], Changrong Ge[1*], Junjing Jia[1*], Zhengchang Su[2*], Tengfei Dou[1*]

1. Faculty of Animal Science and Technology, Yunnan Agricultural University, Kunming 650201, Yunnan, China.

2. Department of Bioinformatics and Genomics, College of Computing and Informatics, the University of North Carolina at Charlotte, Charlotte, NC 28223, USA.

3. State Key Laboratory of Genetic Resources and Evolution/Key Laboratory of Healthy Aging Research of Yunnan Province, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China.

†These authors contributed equally to this work.

*Corresponding author(s): Changrong Ge (gcrzal@126.com), Junjing Jia (junjingli2009@hotmail.com), Zhengchang Su (zcsu@uncc.edu), Tengfei Dou (tengfeidou@sina.com)

Abstract

Oxygen pressure varies dramatically with altitudes on Earth; however, humans and animals thrive at almost all altitudes. To better understand genetic basis underlying adaptation of closely related species to varying altitudes, we annotated and compared the genome of a white eared pheasant (WT) (*Crossoptilon crossoptilon*) inhabiting high altitudes and the genome of a brown eared pheasant (BR) (*C. mantchuricum*) inhabiting low altitudes, and found that the WT genome harbors 967 more genes but 457 fewer pseudogenes than the BR genome. Moreover, we compared genetic variations in populations of WT and BR as well as of blue eared pheasants (BL) (*C. auritum*) inhabiting intermediate altitudes, and identified thousands of selective sweeps in each species. Intriguingly, the unique genes and pseudogenes in the two genomes converge on the same set of altitude adaptation-related pathways of four functional categories as genes in selective sweeps in each species. Thus, these species appear to adapt to highly varying altitudes by diverging selection on the same traits via loss-of-function mutations and fine-tuning genes in common pathways.

Introduction

Humans are naturally adapted to lowland environments where oxygen is abundant. However, some human populations such as Tibetans in Asia, Andeans in Americas, and Ethiopians in Africa, have adapted to highland environments with an altitude above 4,000 m since they exited from Africa more than 100,000 years ago[1-3]. Likewise, populations of many animal species that also are naturally adapted to lowlands thrive in highland environments, such as Tibetan chickens[4-6], Tibetan pigs[7,8], Tibetan goats[9], Tibetan sheep[10,11], Tibetan mastiffs[12], Tibet cattle[13,14], ground tits[15], and yaks[16], to name a few. The major challenge faced by these populations of humans and animals in highland environments is the low oxygen pressure-induced stress on their physiological systems[17-19]. The long-term adaptation to the stress has resulted in genetic changes in these human[1-3,14,17,20-27] and animal[4-7,10-16,19,28-31] populations as a result of long-term adaptation to the low oxygen pressure environments. Adaptive mutations of many genes of humans[2,14,17,18,26,32,33] and animals[16,19,30,31,33,34] have been linked to highland environments. Moreover, introgression has been shown as a major mechanism for human[1,3,27] and domesticated animals[9,10,13] to adapt to highland environments. Despite these great progresses, the understanding of human and animal adaptation to highly varying altitudes is still limited[33]. Particularly, some closely related animal species with little introgression since their divergence can adapt to highly varying altitudes, thus, it is interesting to understand the genetic basis of their adaptation to from low- to intermediate- and high-altitude environments.

Crossoptilon, belonging to the Phasianidae family in the Galliformes order, is a rare but important genus endemic in China[35]. There are four species in the *Crossoptilon* genus, including Tibetan eared pheasant (TB) (*C. harmani*), white eared pheasant (WT) (*C. crossoptilon*), blue

eared pheasant (BL) (*C. auritum*) and brown eared pheasant (BR) (*C. mantchuricum*)[36,37]. These species are found in coniferous forests, mixed broadleaf-conifer forests and alpine scrubs in various parts of China with very different altitudes ranging from 20 m to 5,000 m[38]. TB and WT are believed to be conspecifics and sympatrically inhabit muntane forest at a high altitude of 3,000~5,000 m. TB is distributed in southeastern Tibet and adjacent northern India (3,000~5,000 m)[39], while WT is found in Qinghai, Sichuan, Yunnan and Tibet (3,000~4,300 m)[36]. BL and BR are closely related, but allopatrically inhabit different areas. BL is only encountered in the mountains of Qinghai, Gansu, Sichuan and Ningxia, at an intermediate altitude of 1,500~3,000 m[36], while BR is mainly distributed in mountains in Shanxi and Hebei and near Beijing, at a low altitude of 20~1,000 m[36]. As we show later in this study, WT, BL and BR have little introgression due to their niche isolation, and thus are excellent models to study the genetic basis for closely-related species to adapt to highly varying altitudes.

Recently, Wang et al.[40] assembled a BR genome and generated re-sequencing data of three BR subpopulations (n=40) and a BL population (n=12). The three BR subpopulations were collected from sites in Shannxi province, Shanxi province, and Hebei province & Beijing metropolitan, respectively[40], with an altitude of 20-1,000 m, while the BL population were sampled from sites in Gansu province[40] with an altitude of 1,000~1,500 m (Figure S1). Although the previous study provides new insights into genomic changes during the course of BR's long-term population decline[40], genetic basis underlying the adaptation of the BR and BL to low and intermediated altitudes, respectively, were not investigated. Moreover, with a small contig N50 (0.11 Mb), this BR genome assembly has limitations as a reference genome to study the *crossoptilon* species. To fill the gap, we recently sequenced and assembled the genome of a WT

individual at chromosome-level with a contig N50 of 19.6 Mb, and re-sequenced a population of 10 WT individuals collected in Diqing Tibetan Autonomous Prefecture, Yunnan province and a BL individual from Guangdong province (Figure S1)[41]. Combining our data in WT and BL[41] with those of Wang et al. in BL and BR[40], we were well positioned to investigate the genetic bases of these closely-related avian species to adapt to highly varying altitudes using comparative and population genomics approaches. Interestingly, our results from both approaches suggest that these closely related avian species adapt to highly varying altitudes by diverging selection on the four functional categories via loss-of-function mutations and fine-tuning gene functions or expression levels in common biological pathways. Moreover, our annotated genes in the WT and BR genomes and predicted selective sweeps in the three species can be good resources to study the biology and to design preservation strategies to better understand and protect these endangered birds.

Results

Annotation of protein-coding genes in the WT and BR genomes

To see whether gene contents are related to adaption of closely related avian species to different altitudes, we first annotated the assembled WT[41] and BR[40] genomes using a combination of reference- and RNA-based methods (Materials and Methods). Using the coding DNA sequences (CDSs) of 53 well-represented Aves (Table S1) as the templates and a total 23 RNA-seq datasets from various tissues of the WT and BR individuals (Materials and Methods), we annotated 16,377 and 15,410 protein-coding genes as well as 1,519 and 1,976 pseudogenes in the WT and BR genomes, respectively (Table 1). The vast majority (15,565 and 14,727) of

these genes and all the pseudogenes (1,519 and 1,976) in the WT and BR genomes were predicted based on the CDSs in the reference genomes (reference-based), while a small portion (812 and 683) of the genes were predicted based on the RNA-seq datasets in the two species (RNA-based) (Table 1). Most of the reference-based genes (14,815 and 14,518) in both species have an intact open reading frame (ORF) (intact genes), while the remaining small numbers (750 and 209) contain premature stop codon or ORF-shift mutations, which are not supported by the corresponding short DNA reads, we thus called them partially supported genes, as the mutations might be caused by sequencing errors, particularly, in long reads. Vast majority of the intact genes (99.01%), partially supported genes (97.73%), and pseudogenes (95.13%) in WT were transcribed in at least one of the 20 tissues examined (Table S2). These percentages (96.54%, 93.78% and 90.84%) are smaller in BR (Table S3) as RNA-seq data from only three tissues are available[40]. Most (751 and 631) of the RNA-based genes in both species can be mapped to the NT database (NT-supported new genes), suggesting that they are likely true genes. The remaining small numbers (61 and 52) that cannot be mapped to the NT database are likely novel genes (Table 1).

We identified 450 pairs of RNA-based genes in the WT (55.4%) and BR (65.9%) assemblies, with an identity > 96.8%. Thus, they are likely true orthologous genes, as the same patterns of transcriptional noise are unlikely to occur in two different species. Moreover, the expression levels of the RNA-based genes in both species (Tables S2 and S3) were highly tissue-specific (Figures 1a, 1b, 1c and 1d), indicating again that they might be authentic and functional. In addition, we randomly selected 17 of the 812 RNA-based genes in WT and measured their expression levels in 16 tissues using RT-qPCR. We found that 15 (88.24%) of them were

121 transcribed in at least one of the tissues and five of which were putative novel genes (Table S4,
122 Figure 1e), further suggesting that most of the RNA-based genes at least in WT are likely
123 authentic, although the expression patterns of some genes are different from those seen in the
124 RNA-seq data due probably to the different sensitivity of RT-qPCR and RNA-seq methods
125 (Figure 1f). Moreover, we annotated 12 and eight in WT and BR, respectively, of the 274
126 presumed missing genes in avian species[42] (Table S5).

127 **Most pseudogenes found in the WT and BR genomes are unprocessed and unitary and thus**
128 **might lose gene functions**

129 Vast majority (for WT, 1,310 or 86.2%; for BR, 1,701 or 86.1%) of pseudogenes in both WT and
130 BR genomes are unprocessed, i.e., they raised due to direct mutations in the parent genes. We
131 failed to find any functional copy of these unprocessed pseudogenes; thus, they are also
132 unitary[43]. To see whether the pseudogenization alleles in the 1,519 and 1,976 pseudogenes in
133 the WT and BR genomes, respectively, are under selection, we mapped the short reads from 10
134 WT individuals and 41 BR individuals to the corresponding genomes (Materials and Methods).
135 We found that most of the first pseudogenization alleles along the parental genes were fixed in
136 the population of the two species (Figures 2a and 2b), suggesting that these pseudogenization
137 events might be under positive selection and thus contribute to the phenotypic traits of the two
138 species. Interestingly, the synonymous mutations in true genes in the two species, which are
139 generally assumed to be selectively neutral, are largely uniformly distributed long the CDSs,
140 except at the two ends, where the numbers of synonymous mutations decrease (Figures 2c and
141 2d), consistent with an earlier report in chickens[44]. The reduced synonymous mutations at the
142 two ends of CDSs suggest that they might harbor functional elements not related to their

143 coding functions, such as transcriptional and post-transcriptional regulatory elements[45]. In
144 contrast, pseudogenization mutations are strongly biased to the 5'-ends and 3'-ends in both
145 species (Figures 2c and 2d). The same phenomenon was also found in the other species such as
146 human[44,46] and chickens[44]. Therefore, the strongly biased pseudogenization mutations to the 5'-
147 end and 3'-end in the two species also suggest that they are under positive selection.

148 Finally, of the 1,519 pseudogenes in the WT genome, 727 (47.9%) have alternative
149 transcripts, but only 10 (1.4%) have functional alternative transcripts (Table S6). Of the 1,976
150 pseudogenes in BR, In BR, 1,712 have alternative transcripts, but only 17 (1.0%) have functional
151 alternative transcripts (Table S7). These results suggest that most pseudogenes in both species
152 do not have functional alternative transcripts, further suggesting that most of them have lost
153 the functions of their parent genes.

154 **Unique genes and pseudogenes in WT and BR might be related to their adaptation to distinct**
155 **ecological niches**

156 We found that the WT and BR genomes shared 14,507 of their genes (88.6% and 94.1%) and
157 1,040 of their pseudogenes (68.5% and 52.6%) (Table 1). Meanwhile, the WT genome contains
158 1,870 unique genes and 479 unique pseudogenes, while the BR genome harbors 903 unique
159 genes and 936 unique pseudogenes (Tables S8~S11, Figure 3a). Of the 1,870 unique genes in
160 the WT genome, 658 (35.2%) are unique pseudogenes and 1,212 (64.8%) are missing in the BR
161 genome. Of the 936 unique pseudogenes in BR, 658 (70.3%) are unique genes and 278 (29.7%)
162 are missing in the WT genome (Figure 3a). Moreover, of the 903 unique genes in the BR
163 genome, 173 (19.2%) are unique pseudogenes and 730 (80.8%) are missing in the WT genome.
164 Of the 479 unique pseudogenes in the WT genome, 173 (36.1%) are unique genes and 306

165 (63.9%) are missing in the BR genome (Figure 3a). There are a total of 3,357 genes that either
166 are unique genes or become unique pseudogenes in the WT or BR genomes. To see whether
167 the unique genes and pseudogenes in each species are related to their unique traits and reflect
168 the evolutionary pressure received from their niches, we assigned GO terms to the unique
169 genes and pseudogenes in each species based on their homologs in chickens or humans. The
170 unique genes and pseudogenes in WT are involved in 69 and 25 biological pathways (Table S12),
171 respectively, while those in BR are involved in 27 and 59 biological pathways, respectively
172 (Table S12). Interestingly, the unique genes and pseudogenes with GO term assignments are
173 mainly involved in and affect, respectively, pathways related to four major physiological
174 functions: cardiovascular functions, energy metabolism, neuronal functions and immunity
175 (Table S12, Figure 3b). It has been shown that alterations in these functions are critical for
176 animals to adapt to different altitudes[33]. Thus, the unique genes and loss of functions of the
177 missing genes and pseudogenes in BR and WT might play a role in the adaption of the two
178 species to their strikingly different altitudes. Examples of such unique genes and pseudogenes
179 are detailed in Supplementary Note 1.

180 **There is little introgression among the three species since their divergences**

181 To further understand the genetic basis of altitude adaptation, we next compared the SNPs of
182 10 WT individuals, 41 BR individuals and 12 BL individuals (Table 2). Using our assembled WT
183 genome[47] as the reference, we identified SNPs in the BL, BR and WT populations (Table 2).
184 Based on the bi-allelic SNPs (16 million) in the populations of the three species (Table 2), we
185 performed principal component analysis (PCA) of the individuals of the three species, and found
186 that individuals of the same species form distinct clusters (Figure 4a). Notably, each of the three

BR subpopulations forms a distinct compact subcluster, which is consistent with the previous result[40]. We also run the admixture algorithm[48] on the SNPs to reveal their genetic structures. For K=2, individuals of BL and BR were grouped into one cluster, while WT individuals formed the other cluster (Figure 4b), which is consistent with previous phylogenetic analysis that BL and BR are more closely related to each other[39]. For K=3, individuals of each species formed a distinct cluster with little introgression among the three species, consistent with the result from PCA. These results suggest that the populations of the three species are strongly structured, and that they have adapted to highly varying altitudes with little introgression since their divergences.

Genes in selective sweeps of each species might be related to their distinct ecological niches

To identify genomic regions and genes that might be related to the adaptation of the three species to their distinct ecological niches, particularly, to different altitudes, we identified selective sweeps in each species by comparing its SNPs with those of the other two species using both the genetic differentiation (F_{ST}) and difference in nucleotide diversity ($\Delta\pi$) parameters. To rigorously evaluate the statistical significance of the F_{ST} and $\Delta\pi$ values, we normalized them (ZF_{ST} and $Z\Delta\pi$) using a permutation test[49] (Materials and Methods). Figure 5 shows the distribution of ZF_{ST} and $Z\Delta\pi$ as well as their values for each 40 kb genome windows with a 20 kb-step size for the three pairwise comparisons. We consider a window with ZF_{ST} > 2.33 (P=0.01) and $Z\Delta\pi$ > 1.64 (P=0.05) or $Z\Delta\pi$ < -1.64 (P=0.05) as a selection sweep. Since adjacent selective sweeps can overlap with one another, we merged the overlapping ones as a discrete selection sweep (DSS) in each species. For the BL vs BR comparison, we identified 208 and 638 selective sweeps, and 128 and 463 DSSs containing 143 and 474 genes in BL and

209 BR, respectively (Tables 3 and S13). For the BL vs WT comparison, we identified 92 and 274
210 selective sweeps, and 72 and 181 DSSs containing 67 and 167 genes in BL and WT, respectively
211 (Tables 3 and S14). For the BR vs WT comparison, we identified 306 and 359 selective sweeps,
212 and 184 and 188 DSSs containing 196 and 167 genes in BR and WT, respectively (Tables 3 and
213 S15).

214 Genes in the DSSs of each comparison might be related to adaptation to their different
215 ecological niches. About 55% of genes in these selective sweeps identified in each comparison
216 do not harbor fixed nonsynonymous mutations, thus, natural selection might act upon their
217 regulatory sequences, thereby changing their expression levels. The remaining about 45% of
218 the genes contain fixed nonsynonymous mutations that might alter the functions of encoded
219 proteins. For example, gene *MYO3B* is located in a selective sweep in the BR population based
220 on the BL vs BR comparison (Figure 5a). *MYO3B* in BR contains a fixed G-to-A nonsynonymous
221 mutation in exon 7, leading to an Arg-to-Lys substitution. Since the AGA codon for Arg at this
222 position in *MYO3B* is shared by WT, BL, quail (*Coturnix japonica*) and human (*Homo sapiens*)
223 (Figure 5d), the G-to-A mutation is BR lineage specific. Thus, the Arg-to-Lys substitution in
224 *MYO3B* in BR might be related to their unique traits. It is reported that gene *MYO3B* is required
225 for normal cochlear hair bundle development and hearing[50]. As BR has been hunted by human
226 for hundreds of years[51], it is highly likely that the mutation might increase their auditory
227 sensation, thereby helping them to escape earlier when perceiving hunters approaching. It is
228 interesting to investigate the effects of the Arg-to-Lys mutation on the auditory functions of BR
229 and its implication for natural selection and adaptation. Moreover, gene *PTPRC* is located in a
230 selective sweep in the WT populations based on both the BL vs WT and the BR vs WT

231 comparisons. *PTPRC* in WT contains a fixed A-to-G nonsynonymous mutation in exon 7,
232 resulting in a Lys-to-Arg substitution. As the AAA codon for Lys at the position in *PTPRC* is
233 shared by BL, BR, quail and human (Figure 5d), the mutation is WT lineage specific. Interestingly,
234 we found that the same mutation existed in the *PTPRC* genes of Lance-tailed manakin
235 (*Chiroxiphia lanceolata*) and White-ruffed manakin (*Corapipo altera*) (Figure 5d), both are
236 altitudinal migrants living in subtropical Andes highlands[52], due probably to convergent
237 evolution. It has been reported that *PTPRC* is related to the T cell activation[53]. Therefore, it is
238 highly likely that the nonsynonymous mutation in the *PTPRC* gene in WT and manakins might
239 help them adapt to high altitudes where there are fewer pathogens, and thus, the pressure on
240 the immune system in these species might be relaxed.

241 **Genes under selection tend to be involved in the same altitude adaptation-related pathways**
242 **in the three species**

243 There are a total of 1,106 genes that are in a selective sweep predicted based on at least one of
244 the three comparisons. They share only 105 genes with the 3,357 unique genes and
245 pseudogenes in the WT and BR genomes (Figure 6a). Intriguingly, the genes with GO term
246 assignments in the selective sweeps in each species tend to be involved in the same pathways
247 in the four functional categories (Tables S13-S15) as are the unique genes and unique
248 pseudogenes with GO term assignments in the WT and BR genomes (Table S12), although the
249 two sets only share four genes (Figure 6b). For example, selection on cardiovascular function-
250 related genes *MAPK1* and *PRKAR2B* in BL might be related to its adaptation to intermediate
251 altitudes, while selection on neuronal function related gene *PRKAR2B* might be beneficial for BL
252 to survive in its ecological niches (Figure 6b). Moreover, selection on cardiovascular (*PLA2G4A*),

253 immunity (*ITPR2*), energy metabolism (*PRKCD*) and neuronal function (*SLC17A6*) related genes
254 might be beneficial for BR to adapt to low altitude (Figure 6b). Furthermore, selection on the
255 hypoxia response (*PPARG*) related genes might help WT to adapt to the high-altitude niches,
256 while selection on immunity (*PPP3CB*, *MAPK8* and *PTPRC*) and neuronal function (*GRIN3A*)
257 related genes might be beneficial for WT to adapt to their unique ecological niches (Figure 6b).
258 More examples of genes in the selective sweeps, which are involved in altitude-adaptation
259 related pathways, are detailed in Supplementary Note 1.
260

261 **Discussion**

262 To better understand the genetic bases of adaptation of closely related eared pheasant species
263 to highly varying altitudes, we combined our own published data in WT and BL[41] with those in
264 BR and BL published by Wang et al.[40]. We first annotated the assembled WT[41] and BR[40]
265 genomes, and compared their gene compositions. We identify similar numbers of proteins-
266 coding genes in the WT (16,447) and BR (15,410) genomes to those annotated in other avian
267 species[54], but surprisingly large numbers of pseudogenes in them (1,519 in WT and 1,976 in BR),
268 not previously reported in other avian genomes, to our best knowledge. We provide evidence
269 that most pseudogenization mutations are under positive selection and the resulting
270 pseudogenes might have lost gene functions.

271 We found that many unique genes in one species (WT or BR) are often either missing or
272 pseudogenized in the other species (BR or WT). Interestingly, these unique genes and genes
273 with loss-of-function mutations (pseudogenization or missing) are mainly involved in and affect,
274 respectively, pathways related to four functional categories: cardiovascular, neuronal, energy

275 metabolic and immune functions. One way of alteration of these functions due to the presence
276 or loss of function of the genes in WT might be related to its adaptation to high altitude, as
277 have been reported for other high-altitude adapted species and populations of humans and
278 animals[4,7,11,15,18,19,22,23,31]. The opposite way of alteration of these functions due to the loss of
279 function and presence of the genes in BR might be related to its adaptation to low altitude. For
280 example, we found that gene *PIK3C2G* in BR is pseudogenized in WT. This gene is involved in
281 many pathways, including the HIF pathway, endothelin signaling pathway, angiogenesis
282 pathway, VEGF signaling pathway, T cell activation pathway and inflammation mediated by
283 chemokine and cytokine signaling pathway (Figure 3b). It is likely that functional *PIK3C2G* in BR
284 plays a role in BR's demands for these cardiovascular and immune functions to adapt to low
285 altitudes, while the loss-of-function of *PIK3C2G* in WT might better meet WT's demands for
286 these functions to adapt to high altitudes. Moreover, gene *CACNB1* in WT is pseudogenized in
287 BR. This gene is involved in the beta1/2 adrenergic receptor signaling pathway, metabotropic
288 glutamate receptor group III pathway, nicotinic acetylcholine receptor signaling pathway and
289 thyrotropin-releasing hormone receptor signaling pathway (Figure 3b). It is likely that functional
290 *CACNB1* in WT might play a role in WT's demands for these neuronal and energy metabolic
291 functions to adapt to high altitudes or other aspects of its unique niches, while the loss-of-
292 function of *CACNB1* in BR might better meet BR's demands for these functions to adapt to low
293 altitudes or other aspects of its unique niches.

294 Next, we compared the SNPs in the WT, BL and BR populations. Interestingly, we found
295 that genes in the selective sweeps in each species are mainly involved in the same GO pathways
296 of the four functional categories as are the unique genes and pseudogenes in the WT and BR

297 genomes, even though the two sets of genes have little overlaps (Figure 6a). Most genes in the
298 selective sweeps do not have amino acid-altering mutations, although some may have missense
299 mutations such as *MYO3B* in BR and *PTPRC* in WT (Figure 5d), become pseudogenes such as
300 *CACNB2*, or be missing such as *PRKCD* in BR. These results suggest that natural selection might
301 act on their *cis*-regulatory sequences, thereby altering their expression levels in relevant tissues.
302 It appears that natural selection on these genes might be related to the adaptation of the
303 species to different niches, particularly different altitudes. For example, in the BL vs BR
304 comparison, gene *MAPK1* in BL is included in a selective sweep. This gene is involved in the
305 VEGF signaling pathway, angiogenesis pathway, endothelin signaling pathway, angiotensin II-
306 stimulated signaling through G proteins and beta-arrestin pathway, insulin/IGF pathway-
307 mitogen activated protein kinase kinase/MAP kinase cascade pathway and many pathways of
308 the immunity (Figure 6b). Thus, it is likely that selection on *MAPK1* might be related to the
309 adaptation of BL to intermediate altitude. In the WT vs BL comparison, gene *PPARG* is included
310 in a selective sweep, and the gene is involved in the HIF pathway[55] (Figure 6b), thus selection on
311 the gene might be related to the adaptation of WT to low oxygen pressure at high altitudes.
312 It is not surprising that the unique genes/pseudogenes in WT and BR and genes in the
313 selective sweeps of BL, BR and WT converge on the same sets of pathways involved in
314 cardiovascular, energy metabolic, neuronal and immune functions. On the one hand, high
315 altitude niches with low oxygen pressure, low temperature and less availability of food would
316 pose a direct pressure on WT's cardiovascular system and energy metabolism functions, while
317 such pressure would be somewhat relaxed on BL and even more relaxed on BR. On the other
318 hand, other altitude related ecological factors might exert different pressures on the three

species for their neuronal functions for food foray and escaping for predators, and for immune functions for different pathogens. Indeed, it has been shown that high altitude niches can have effects on the neuronal functions such as sensory perception in domestic yaks (*Bos grunniens*)[16], and olfaction in Tibetan pig[7], wild boars[7] and ground tit (*Parus humilis*)[15]. In this study, we found that a nonsynonymous mutation in *MYO3B* in the BR population might alter their auditory sensation. It has also been shown that high altitude niches can have effects on immune functions in ground tit[15]. In this study, we found that a nonsynonymous mutation in *PTPRC* in the WT population and Manakins might alter their T-cell functions. Thus, it appears that the three eared pheasant species might adapt to highly varying altitudes or other aspects of their distinct ecological niches by loss-of-function mutations and fine-tuning genes involved in the same set of pathways to meet their varying demands for the four major functions. In other words, genes in this set of pathways might play roles in the adaptation of the three pheasant species to different niches, particularly, different altitudes by complete deletion (gene missing), pseudogenization, alteration of CDSs, or alteration of *cis*-regulatory sequences.

Materials and Methods

Datasets

We download our following previously published data[41], including the assembled genome of the WT individual from GenBank with accession number GCA_036346035.1, re-sequencing data of 10 WT individuals from NCBI SRA with accession number SRP433016, re-sequencing data of a BL individual from NCBI SRA with accession number SRP471294, and RNA-seq data from 20 tissues of the WT individual from NCBI SRA with accession number SRP432961. We downloaded

341 the following published datasets[40], including the assembled genome of the BR individual from
342 China National Genomics Data Center (CNGDC) with accession number GWHAOPW00000000,
343 re-sequencing data of 41 BR individuals from CNGDC with project number PRJCA003284, RNA-
344 seq data from the BR individual's blood, developing primaries and developing tail feathers
345 tissues from CNGDC with project number PRJCA003284, and re-sequencing data of 12 BL
346 individuals from CNGDC with project number PRJCA003284.

347 **Ethics approval**

348 All the experimental procedures were approved by the Animal Care and Use Committee of the
349 Yunnan Agricultural University (approval ID: YAU202103047). The care and use of animals fully
350 complied with local animal welfare laws, guidelines, and policies.

351 **Real-time quantitative PCR (RT-qPCR) analysis**

352 Three female adult WT individuals were collected from Diqing Tibet Autonomous Prefecture,
353 Yannan Province, China. The birds were killed by electric shock to unconsciousness followed by
354 neck artery bleeding. One to two grams of relevant tissues were aseptically collected from each
355 individual chicken in a centrifuge tube within 20 mins after sacrifice and immediately frozen in
356 liquid nitrogen, then stored at -80°C until use. RT-qPCR was performed using the Bio-Rad CFX96
357 real-time PCR platform (Bio-Rad Laboratories. Inc, America) and SYBR Green master mix (iQTM
358 SYBRGreen ® Supermix, Dalian TaKaRa Biotechnology Co. Ltd. Add). We randomly selected 17
359 putative new genes of validation. The primers of the 17 putative new genes and their annealing
360 temperatures are listed in Supplementary Note 2. The β-actin gene was used as a reference.
361 Primers were commercially synthesized (Shanghai Shenggong Biochemistry Company P.R.C).
362 Each PCR reaction was performed in 25 μl volumes containing 12.5 μl of iQ™ SYBR Green

363 Supermix, 0.5 µl (10 mM) of each primer, and 1 µl of cDNA. Amplification and detection of
364 products was performed with the following cycle profile: one cycle of 95 °C for 2 min, and 40
365 cycles of 95 °C for 15 s, the annealing temperature (Supplementary Note 2) for 30 s, and 72 °C
366 for 30 s, followed by a final cycle of 72 °C for 10 min. The specificity of the amplification product
367 was verified by electrophoresis on a 0.8% agarose gel and DNA sequencing. The $2^{-\Delta Ct}$ method
368 was used to analyze mRNA abundance. All tissues were analyzed with three biological replicates
369 and each biological replicate with five technical replicates.

370 **Protein-coding gene annotation**

371 To annotate the protein-coding genes and pseudogenes in the WT and BR genomes, we used a
372 combination of reference-based and RNA-based method as previously described {ref.}. Briefly,
373 for the reference-based method, we use the CDSs of 53 well-represented avian genomes in
374 NCBI (Table S1) as the templates. We map all the CDSs isoforms of genes in the 53 genomes to
375 an assembly using Splign (2.0.0)[56] using default settings. For each template gene whose CDSs
376 can be mapped to the assembly, we concatenate all the mapped parts on the assembly. If the
377 CDSs of the template gene can be mapped to multiple loci on the assemblies, we choose the
378 locus with the highest mapping identity and check whether the resulting concatenated
379 sequence forms an intact ORF (it has a length of an integer time of three and contains no
380 premature stop codon). If yes, we called the locus of the concatenated sequence an intact
381 gene. Otherwise, we mapped the Illumina DNA short reads from the same individual to the
382 locus in the assembly allowing no mismatch and gaps using bowtie (2.4.1)[57]. If each nucleotide
383 position in the locus can be mapped by at least 10 short reads, we consider the locus as a
384 pseudogene; otherwise, we call it a partially supported gene since the pseudogenization

385 mutation might be caused by the sequencing errors in long reads that cannot be corrected by
386 DNA short reads.

387 For the RNA-based method, we first map all the RNA-seq reads from the 23 tissues of
388 WT and BR to the rRNA database SILVA_138[58] and filtered out the mapped reads. We then map
389 the unmapped reads of each pheasant to its corresponding assembly using STAR (2.7.0c)[59] using
390 default settings. Based on the mapping results, we assemble the mapped reads for each
391 pheasant into transcripts using genome-guided method in Trinity (2.8.5)[60] using default settings.
392 Then we map the assembled transcripts of each pheasant to their corresponding assemblies
393 using Splign (2.0.0)[56] using default settings and remove those that partially overlap the genes
394 and pseudogenes predicted by the reference-based method or the non-coding RNA genes (see
395 below). For each remaining assembled transcripts, we find the longest ORF in it, and if the ORF
396 is at least 300 bp long, we call it a protein-coding gene.

397 For completeness, we also annotate tRNA, miRNA, rRNA, snoRNA, telomerase RNA and
398 SRP RNA genes in each genome using infernal (1.1.2)[61] with the Rfam (v.14) database[62] as the
399 reference using default settings.

400 **Single nucleotide variants calling**

401 To calculate the fixation rate of the pseudogenes in the population of each species, we called
402 the SNPs in the pseudogenes of each species using the DNA short reads from the species'
403 population (n=10 for WT and n=40 for BR) using GATK (4.1.6) using default settings[63]. To call the
404 variants in the populations of BL, BR and WT, we first mapped the short reads of each individual
405 to the WT genome using BWA (0.7.17)[64] and SAMtools (1.9)[65] using default settings, and then
406 called the variants for each individual using GATK-HaplotypeCaller (4.1.6)[63] using default

407 settings and merged the variants of each individual from the same species using GATK-
408 CombineGVCFs (4.1.6)[63] using default settings. We removed the SNPs with Quality by depth
409 (QD) < 2, Fisher strand (FS) > 60, Root mean square mapping quality (MQ) < 40, Strand odd ratio
410 (SOR) > 3, Rank Sum Test for mapping qualities (MQRankSum) < -12.5 and Rank Sum Test for
411 site position within reads (ReadPosRankSum) < -8 and indels with QD < 2, FS > 200, SOR > 10,
412 Likelihood-based test for the consanguinity among samples (InbreedingCoeff) < -0.8 and
413 ReadPosRankSum < -20.

414 **Population structure and PCA analyses**

415 Principal component analysis (PCA) was performed using plink (1.90)[66] with default settings
416 using called biallelic SNPs in the three species. The same SNPs were used to infer the population
417 structure using ADMIXTURE (1.3.0)[48] with default settings.

418 **Selective sweeps detection**

419 Selective sweeps in a species relative to another were detected using two methods including
420 genetic differentiation (F_{ST})[67] and difference in nucleotide diversity ($\Delta\pi$). We estimated F_{ST} for
421 each pair of comparison using VCFtools (0.1.16)[68] with a sliding window of 40 kb and a step size
422 of 20 kb with default settings. We estimated π for each species using VCFtools (0.1.16)[68] with a
423 sliding window of 40 kb and a step size of 20 kb with default settings, and calculated the
424 difference in nucleotide diversity ($\Delta\pi$) in each window for each pair of comparison. To evaluate
425 the statistical significance of the F_{ST} and π values for a comparison, we generated a Null model
426 by shuffling the allele frequency data for 100 times while keeping the SNP positions fixed[49]. We
427 then computed F_{ST} and $\Delta\pi$ for the permuted windows as well as their means ($\mu(F_{ST_{Null}})$ and

$\mu(\Delta\pi_{Null})$) and standard deviations ($\sigma(F_{ST_{Null}})$ and $\sigma(\Delta\pi_{Null})$). We computed the Z value for each F_{ST} and $\Delta\pi$ value for the comparison by using the following formulas:

$ZF_{ST}(i) = (F_{ST}(i) - \mu(F_{ST_{Null}}))/\sigma(F_{ST_{Null}})$ and

$Z\Delta\pi(i) = (\Delta\pi(i) - \mu(\Delta\pi_{Null}))/\sigma(\Delta\pi_{Null})$.

We consider a window with $ZF_{ST} > \alpha$ and $Z\Delta\pi > \beta$ or $Z\Delta\pi < -\beta$ ($\alpha>0$, $\beta>0$) as a selection sweep. For the selective sweeps in each pair of comparison, we used the value of $Z\Delta\pi$ to help identify the selective sweeps of each species. Windows with $Z\Delta\pi$ scores < -β represent selective sweeps of the minuend species, and windows with $Z\Delta\pi$ scores > β represent selective sweeps of the subtrahend species.

Data Availability

The WT genome is available at GenBank with accession number GCA_036346035.1; the BR genome is available at CNGDC with accession number GWHAOPW00000000; re-sequencing data of 10 WT individuals are available at NCBI SRA with accession number SRP433016; re-sequencing data of 41 BR individuals are available at CNGDC with project number PRJCA003284; re-sequencing data of 12 BL individuals are available at CNGDC with project number PRJCA003284 and at NCBI SRA with accession number SRP471294; RNA-seq data from 20 tissues of the WT individual are available at NCBI SRA with accession number SRP432961; RNA-seq data from the BR individual's blood, developing primaries and developing tail feathers tissues are available at CNGDC with project number PRJCA003284. Our gene annotation results of the WT and BR genomes are available at FigShare (https://doi.org/10.6084/m9.figshare.24438295.v1).

Code Availability

All gene annotation code and the corresponding pipeline description are available at https://github.com/zhengchangsulab/A-genome-assebmly-and-annotation-pipeline.

Author Contributions

CG, JJ, ZS and TD supervised and conceived the project; KW, XG and DW collected tissue samples and conducted molecular biology experiments; SW and SY assembled and annotated the genomes; SW and ZS performed data analysis; and SW and ZS wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (U2002205 and U1702232), Yunling Scholar Training Program of Yunnan Province (2014NO48), Yunling Industry and Technology Leading Talent Training Program of Yunnan Province (YNWR-CYJS-2015-027), Natural Science Foundation of Yunnan Province (2019IC008 and 2016ZA008), and Department of Bioinformatics and Genomics of the University of North Carolina at Charlotte.

References

1 Zhang, P. *et al.* Denisovans and Homo sapiens on the Tibetan Plateau: dispersals and adaptations. *Trends Ecol Evol* **37**, 257-267 (2022). https://doi.org:10.1016/j.tree.2021.11.004

2 Peng, Y. *et al.* Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Mol Biol Evol* **28**, 1075-1081 (2011). https://doi.org:10.1093/molbev/msq290

3 Hu, H. *et al.* Evolutionary history of Tibetans inferred from whole-genome sequencing. *PLoS Genet* **13**, e1006675 (2017). https://doi.org:10.1371/journal.pgen.1006675

4 Wang, M. S. *et al.* Genomic Analyses Reveal Potential Independent Adaptation to High Altitude in Tibetan Chickens. *Mol Biol Evol* **32**, 1880-1889 (2015). https://doi.org:10.1093/molbev/msv071

5 Li, S. *et al.* A non-synonymous SNP with the allele frequency correlated with the altitude may contribute to the hypoxia adaptation of Tibetan chicken. *PLoS One* **12**, e0172211 (2017). https://doi.org:10.1371/journal.pone.0172211

6 Yuan, J. *et al.* Genome-wide run of homozygosity analysis reveals candidate genomic regions associated with environmental adaptations of Tibetan native chickens. *BMC Genomics* **23**, 91 (2022). https://doi.org:10.1186/s12864-021-08280-z

7 Li, M. *et al.* Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. *Nat.Genet* **45**, 1431-1438 (2013). https://doi.org:10.1038/ng.2811

8 Zhang, B. *et al.* Comparative transcriptomic and proteomic analyses provide insights into the key genes involved in high-altitude adaptation in the Tibetan pig. *Sci Rep* **7**, 3654 (2017). https://doi.org:10.1038/s41598-017-03976-3

9 Li, C. *et al.* Markhor-derived Introgression of a Genomic Region Encompassing PAPSS2 Confers High-altitude Adaptability in Tibetan Goats. *Mol Biol Evol* **39** (2022). https://doi.org:10.1093/molbev/msac253

10 Hu, X. J. *et al.* The Genome Landscape of Tibetan Sheep Reveals Adaptive Introgression from Argali and the History of Early Human Settlements on the Qinghai-Tibetan Plateau. *Mol Biol Evol* **36**, 283-303 (2019). https://doi.org:10.1093/molbev/msy208

11 Liu, J. *et al.* Genetic signatures of high-altitude adaptation and geographic distribution in Tibetan sheep. *Sci Rep* **10**, 18332 (2020). https://doi.org:10.1038/s41598-020-75428-4

12 Li, Y. *et al.* Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Mol Biol Evol* **31**, 1200-1205 (2014). https://doi.org:10.1093/molbev/msu070

13 Wu, D. D. *et al.* Pervasive introgression facilitated domestication and adaptation in the Bos species complex. *Nature ecology & evolution* **2**, 1139-1145 (2018). https://doi.org:10.1038/s41559-018-0562-y

14 Simonson, T. S., McClain, D. A., Jorde, L. B. & Prchal, J. T. Genetic determinants of Tibetan high-altitude adaptation. *Hum Genet* **131**, 527-533 (2012). https://doi.org:10.1007/s00439-011-1109-3

15 Qu, Y. *et al.* Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nat Commun* **4**, 2071 (2013). https://doi.org:10.1038/ncomms3071

16 Qiu, Q. *et al.* The yak genome and adaptation to life at high altitude. *Nat.Genet* **44**, 946-949 (2012). https://doi.org:10.1038/ng.2343

17 Simonson, T. S. *et al.* Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72-75 (2010). https://doi.org:10.1126/science.1189406

18 Ji, L. D. *et al.* Genetic adaptation of the hypoxia-inducible factor pathway to oxygen pressure among eurasian human populations. *Mol Biol Evol* **29**, 3359-3370 (2012). https://doi.org:10.1093/molbev/mss144

19 Xiong, Y. *et al.* Physiological and genetic convergence supports hypoxia resistance in high-altitude songbirds. *PLoS Genet* **16**, e1009270 (2020). https://doi.org:10.1371/journal.pgen.1009270

20 Peng, Y. *et al.* Down-Regulation of EPAS1 Transcription and Genetic Adaptation of Tibetans to High-Altitude Hypoxia. *Mol Biol Evol* **34**, 818-830 (2017). https://doi.org:10.1093/molbev/msw280

21 Xin, J. *et al.* Chromatin accessibility landscape and regulatory network of high-altitude hypoxia adaptation. *Nat Commun* **11**, 4928 (2020). https://doi.org:10.1038/s41467-020-18638-8

22 Xu, D. *et al.* A single mutation underlying phenotypic convergence for hypoxia adaptation on the Qinghai-Tibetan Plateau. *Cell Res* **31**, 1032-1035 (2021). https://doi.org:10.1038/s41422-021-00517-6

23 Zhu, X. *et al.* Divergent and parallel routes of biochemical adaptation in high-altitude passerine birds from the Qinghai-Tibet Plateau. *Proc Natl Acad Sci U S A* **115**, 1865-1870 (2018). https://doi.org:10.1073/pnas.1720487115

24 Guo, Y. B. *et al.* GCH1 plays a role in the high-altitude adaptation of Tibetans. *Zool Res* **38**, 155-162 (2017). https://doi.org:10.24272/j.issn.2095-8137.2017.037

25 Heinrich, E. C. *et al.* Genetic variants at the EGLN1 locus associated with high-altitude adaptation in Tibetans are absent or found at low frequency in highland Andeans. *Annals of human genetics* **83**, 171-176 (2019). https://doi.org:10.1111/ahg.12299

26 Xiang, K. *et al.* Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol Biol Evol* **30**, 1889-1898 (2013). https://doi.org:10.1093/molbev/mst090

27 Jeong, C. *et al.* Admixture facilitates genetic adaptations to high altitude in Tibet. *Nat Commun* **5**, 3281 (2014). https://doi.org:10.1038/ncomms4281

28 Li, D. *et al.* Population genomics identifies patterns of genetic diversity and selection in chicken. *BMC Genomics* **20**, 263 (2019). https://doi.org:10.1186/s12864-019-5622-4

29 Liu, X., Wang, X., Liu, J., Wang, X. & Bao, H. Identifying Candidate Genes for Hypoxia Adaptation of Tibet Chicken Embryos by Selection Signature Analyses and RNA Sequencing. *Genes* **11** (2020). https://doi.org:10.3390/genes11070823

30 Tang, Q. *et al.* HIF-1 regulates energy metabolism of the Tibetan chicken brain during embryo development under hypoxia. *Am J Physiol Regul Integr Comp Physiol* **320**, R704-r713 (2021). https://doi.org:10.1152/ajpregu.00052.2020

31 Zhang, Y., Zhang, H., Zhang, B., Ling, Y. & Zhang, H. Identification of key HIF-1α target genes that regulate adaptation to hypoxic conditions in Tibetan chicken embryos. *Gene* **729**, 144321 (2020). https://doi.org:10.1016/j.gene.2019.144321

32 Scheinfeldt, L. B. & Tishkoff, S. A. Living the high life: high-altitude adaptation. *Genome Biol* **11**, 133 (2010). https://doi.org:10.1186/gb-2010-11-9-133

33 Pamenter, M. E., Hall, J. E., Tanabe, Y. & Simonson, T. S. Cross-Species Insights Into Genomic Adaptations to Hypoxia. *Front Genet* **11**, 743 (2020). https://doi.org:10.3389/fgene.2020.00743

34 Friedrich, J. & Wiener, P. Selection signatures for high-altitude adaptation in ruminants. *Anim Genet* **51**, 157-165 (2020). https://doi.org:10.1111/age.12900

35 Li, X., Huang, Y. & Lei, F. Comparative mitochondrial genomics and phylogenetic relationships of the Crossoptilon species (Phasianidae, Galliformes). *BMC Genomics* **16**, 42 (2015). https://doi.org:10.1186/s12864-015-1234-9

36 Xin, L., Guangmei, Z. & Binyuan, G. A preliminary investigation on taxonomy, distribution and evolutionary relationship of the eared pheasants, Crossoptilon. *Dong wu xue bao.[Acta zoologica Sinica]* **44**, 131-137 (1998).

37 Zheng, Z. *A complete checklist of species and subspecies of the Chinese birds*. (Science Press, 1994).

38 Lu, T. (Fuzhou: Fujian Science and Technology Press, 1991).

39 Wang, P. *et al.* The role of niche divergence and geographic arrangement in the speciation of Eared Pheasants (Crossoptilon, Hodgson 1938). *Mol Phylogenet Evol* **113**, 1-8 (2017). https://doi.org:10.1016/j.ympev.2017.05.003

40 Wang, P. *et al.* Genomic Consequences of Long-Term Population Decline in Brown Eared Pheasant. *Mol Biol Evol* **38**, 263-273 (2021). https://doi.org:10.1093/molbev/msaa213

41 Wu, S. *et al.* High-quality genome assembly of a C. crossoptilon and related functional and genetics data resources. *Scientific data* **11**, 247 (2024). https://doi.org:10.1038/s41597-024-03087-5

42 Lovell, P. V. *et al.* Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biol* **15**, 565 (2014). https://doi.org:10.1186/s13059-014-0565-1

43 Zhang, Z. D., Frankish, A., Hunt, T., Harrow, J. & Gerstein, M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol* **11**, R26 (2010). https://doi.org:10.1186/gb-2010-11-3-r26

44 Derks, M. F. L. *et al.* A survey of functional genomic variation in domesticated chickens. *Genet Sel Evol* **50**, 17 (2018). https://doi.org:10.1186/s12711-018-0390-1

45 Parmley, J. L. & Hurst, L. D. How do synonymous mutations affect fitness? *Bioessays* **29**, 515-519 (2007). https://doi.org:10.1002/bies.20592

46 Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet* **4**, e1000160 (2008). https://doi.org:10.1371/journal.pgen.1000160

47 Siwen Wu, K. W., Tengfei Dou, Sisi Yuan, Dong-Dong Wu, Changrong Ge, Junjing Jia,Zhengchang Su. High-quality genome assembly of a white eared pheasant individual and related functional and genetics data resources. *BioRxiv* (2023). https://doi.org:https://www.biorxiv.org/content/10.1101/2023.11.09.566452v1

48 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009). https://doi.org:10.1101/gr.094052.109

49 Churchill, G. A. & Doerge, R. W. Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963-971 (1994). https://doi.org:10.1093/genetics/138.3.963

50 Lelli, A. *et al.* Class III myosins shape the auditory hair bundles by limiting microvilli and stereocilia growth. *J Cell Biol* **212**, 231-244 (2016). https://doi.org:10.1083/jcb.201509017

51 Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20-d26 (2022). https://doi.org:10.1093/nar/gkab1112

52 Josep del Hoyo, A. E., Jordi Sargatal, David A Christie. *https://en.wikipedia.org/wiki/Handbook_of_the_Birds_of_the_World*, 1992).

53 Wu, L., Fu, J. & Shen, S. H. SKAP55 coupled with CD45 positively regulates T-cell receptor-mediated gene transcription. *Mol Cell Biol* **22**, 2673-2686 (2002). https://doi.org:10.1128/mcb.22.8.2673-2686.2002

54 Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737-746 (2021). https://doi.org:10.1038/s41586-021-03451-0

55 Graham, A. M. & McCracken, K. G. Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes EGLN1 and EPAS1 in high-altitude ducks. *Heredity (Edinb)* **122**, 819-832 (2019). https://doi.org:10.1038/s41437-018-0173-z

56 Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct* **3**, 20 (2008). https://doi.org:10.1186/1745-6150-3-20

57 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012). https://doi.org:10.1038/nmeth.1923

58 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596 (2013). https://doi.org:10.1093/nar/gks1219

59 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013). https://doi.org:10.1093/bioinformatics/bts635

60 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652 (2011). https://doi.org:10.1038/nbt.1883 [doi]

61 Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013). https://doi.org:10.1093/bioinformatics/btt509

62 Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**, D192-d200 (2021). https://doi.org:10.1093/nar/gkaa1047

63 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010). https://doi.org:10.1101/gr.107524.110

64 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009). https://doi.org:10.1093/bioinformatics/btp324

65 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). https://doi.org:10.1093/bioinformatics/btp352

66 Chen, Z. L. *et al.* A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat Commun* **10**, 3404 (2019). https://doi.org:10.1038/s41467-019-11337-z

67 Reynolds, J., Weir, B. S. & Cockerham, C. C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**, 767-779 (1983). https://doi.org:10.1093/genetics/105.3.767

68 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011). https://doi.org:10.1093/bioinformatics/btr330

Table 1: Gene annotation of the two species

Breeds	Reference-based			RNA-based		# Total genes	# Total pseudogenes
	# Intact	# Partial	# Pseudogenes	# NT-supported	# Novel		
WT	14,815	750	1,519	751	61	16,377	1,519
BR	14,518	209	1,976	631	52	15,410	1,976
Common	14,057		1,040	422	28	14,507	1,040

Table 2: Summary of the SNPs in each species

Breeds	# individuals	# Total SNPs
BL	12	10,147,889
BR	41	6,849,438
WT	10	9,004,245
All	63	16,459,558

Table 3: Summary of selective sweeps in different comparisons

Comparison	BL			BR			WT		
	# SS (# DSS)	# Genes in DSS	# Pathways involved	# SS (# DSS)	# Genes in DSS	# Pathways involved	# SS (# DSS)	# Genes in DSS	# Pathways involved
BL VS. BR	208 (128)	143	48	638 (463)	474	56	NA	NA	NA
BL VS. WT	92 (72)	67	17	NA	NA	NA	274 (181)	167	34
BR VS. WT	NA	NA	NA	306 (184)	196	45	359 (188)	167	36

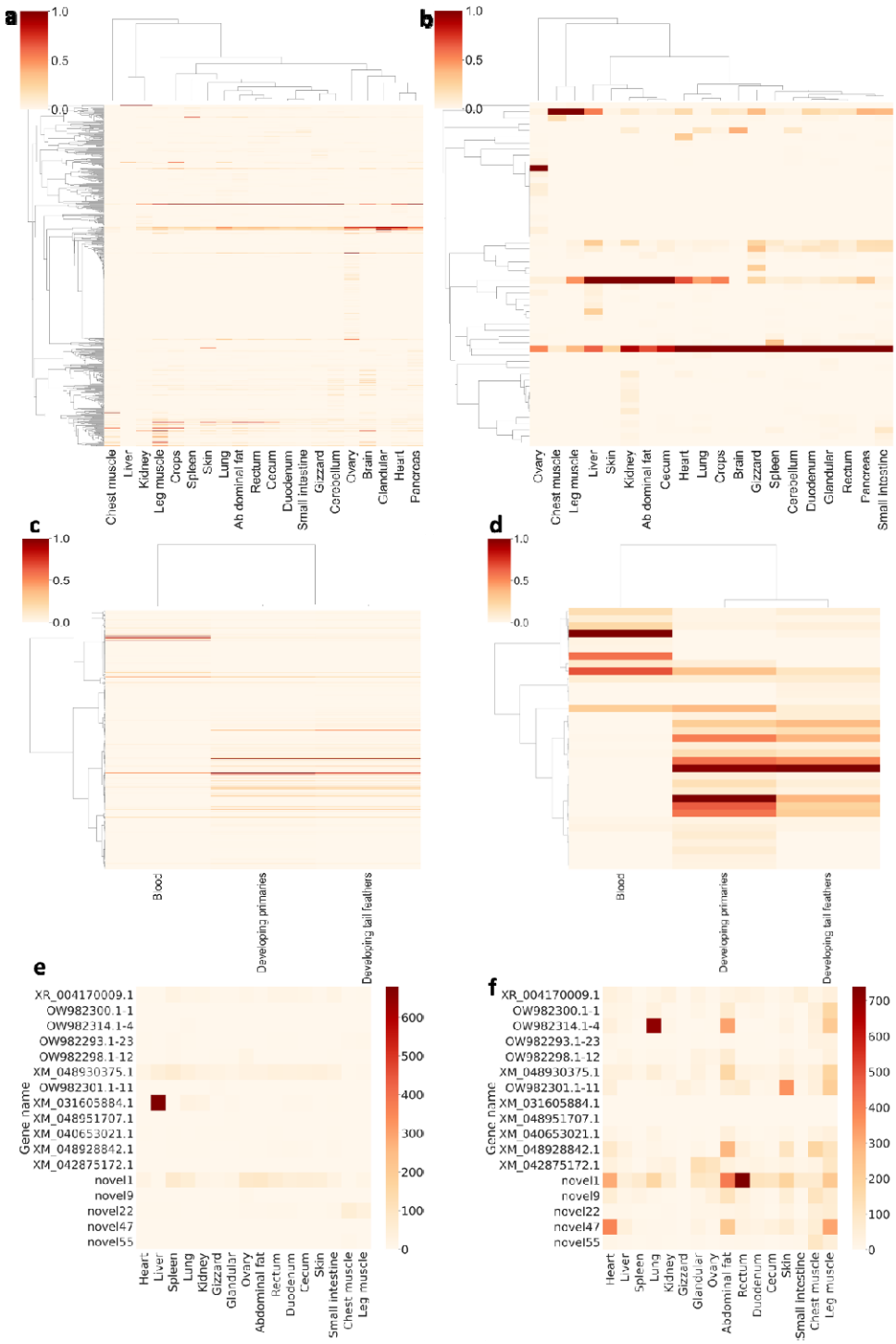

Figure 1. Heatmaps of new genes of the two species. **a.** Expression pattern of the NT-supported new genes in 20 tissues of WT. **b.** Expression pattern of the novel genes in 20 tissues of WT. **c.** Expression pattern of the NT-supported new genes in three tissues of BR. **d.** Expression pattern of the novel genes in three tissues in BR. **e.** Expression pattern of the 17 selected new genes in WT measured by RT-qPCR. **f.** Expression pattern of the 17 selected new genes from RNA-seq calculation in WT.

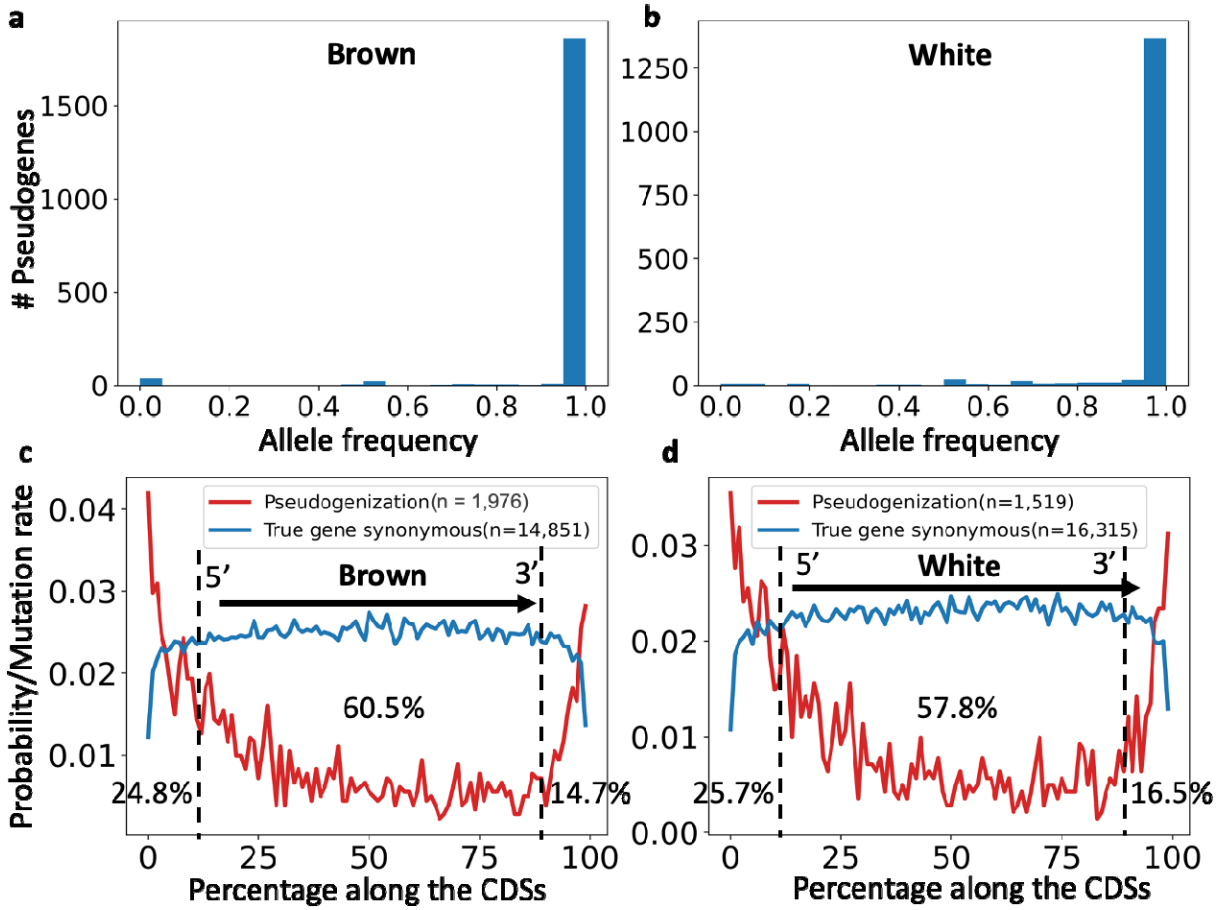

Figure 2. Fixation rate of pseudogenes in the WT and BR populations and distribution of the first pseudogenization alleles along the CDSs of parental genes. **a.** Number of pseudogenization alleles with indicated frequencies in the BR population. **b.** Number of the pseudogenization alleles with indicated frequencies in the WT populations. **c.** Distribution of the first pseudogenization alleles in the BR genome along the CDSs of parental genes. **d.** Distribution of the first pseudogenization alleles in the WT genome along the CDSs of parental genes.

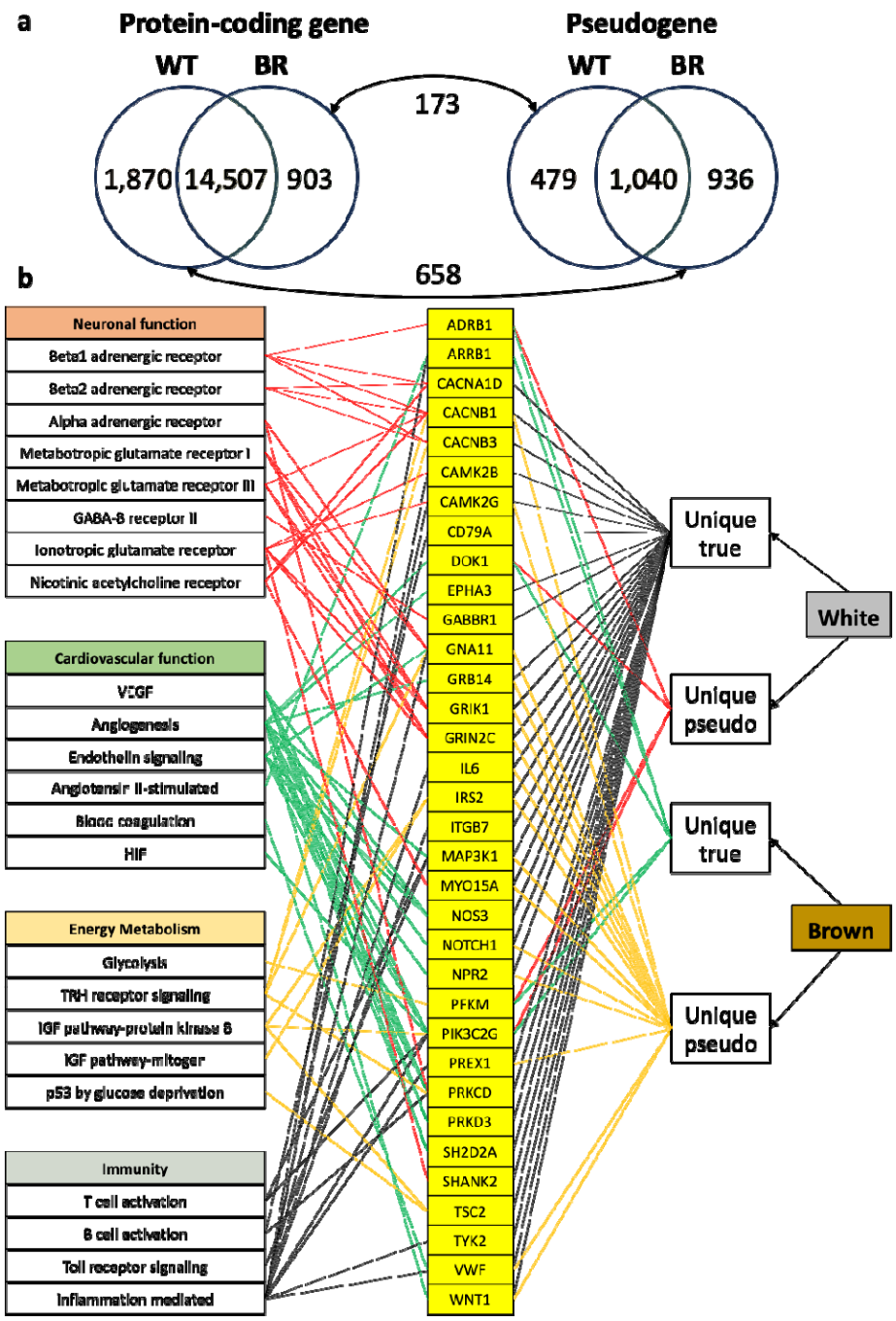

Figure 3. Functional analysis of the unique genes and pseudogene in the WT and BR genomes. **a.** Venn diagram showing shared and unique genes (left) and pseudogenes (right) in the two species. Of note, 658 WT genes are pseudogenized in BR and 173 BR genes are pseudogenized in WT, and there are a total of 3,357 non redundant unique genes and pseudogenes in the two genomes. **b.** The unique genes/pseudogenes in the WT and BR genomes that are involved in pathways related to four functional categories: neuronal functions, cardiovascular functions, energy metabolism, and immunity. Notably, unique genes in a species often became unique pseudogenes in the other species.

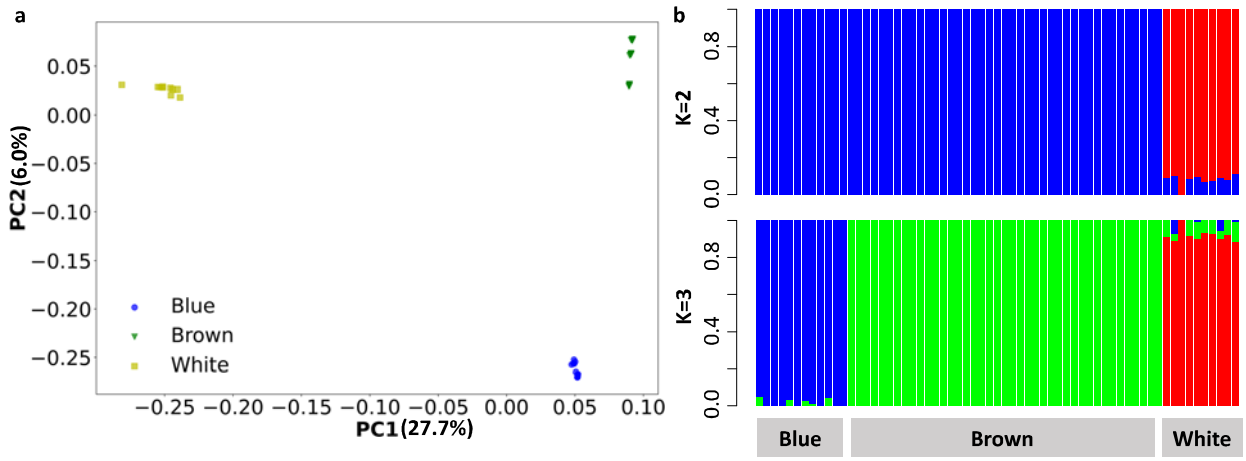

Figure 4. Population structure of WT, BL and BR. **a.** Principal component analysis of the populations of the three species based on their SNP spectrum. **b.** Genetic structure of the individuals of the three species estimated using ADMIXTURE.

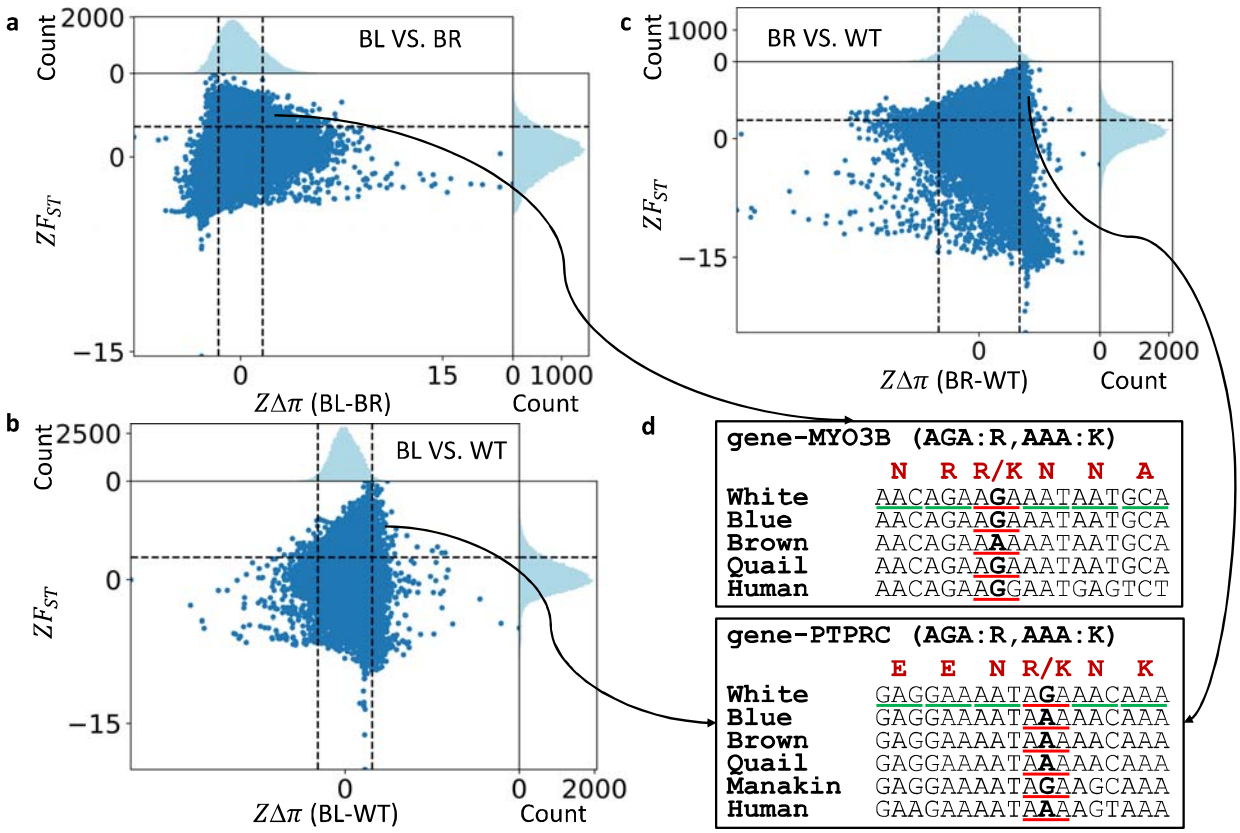

Figure 5. Distribution of the ZF_{ST} and $Z\Delta\pi$ values of each sliding window for the three pairwise comparisons. **a.** BL VS. BR comparison. **b.** BL VS. WT comparison. **c.** BR VS. WT comparison. **d.** Multiple alignments of part of CDSs of two genes in DSSs in different species. In subfigures a, b and c, the top panel represents the distribution of $Z\Delta\pi$ values and the bottom right one represents the distribution of ZF_{ST}values. Each dot in the bottom left panel represents a sliding window with its $Z\Delta\pi$ and ZF_{ST} values being its coordinates in the plot. The dashed lines represent the cut-offs for the ZF_{ST} and $Z\Delta\pi$ values.

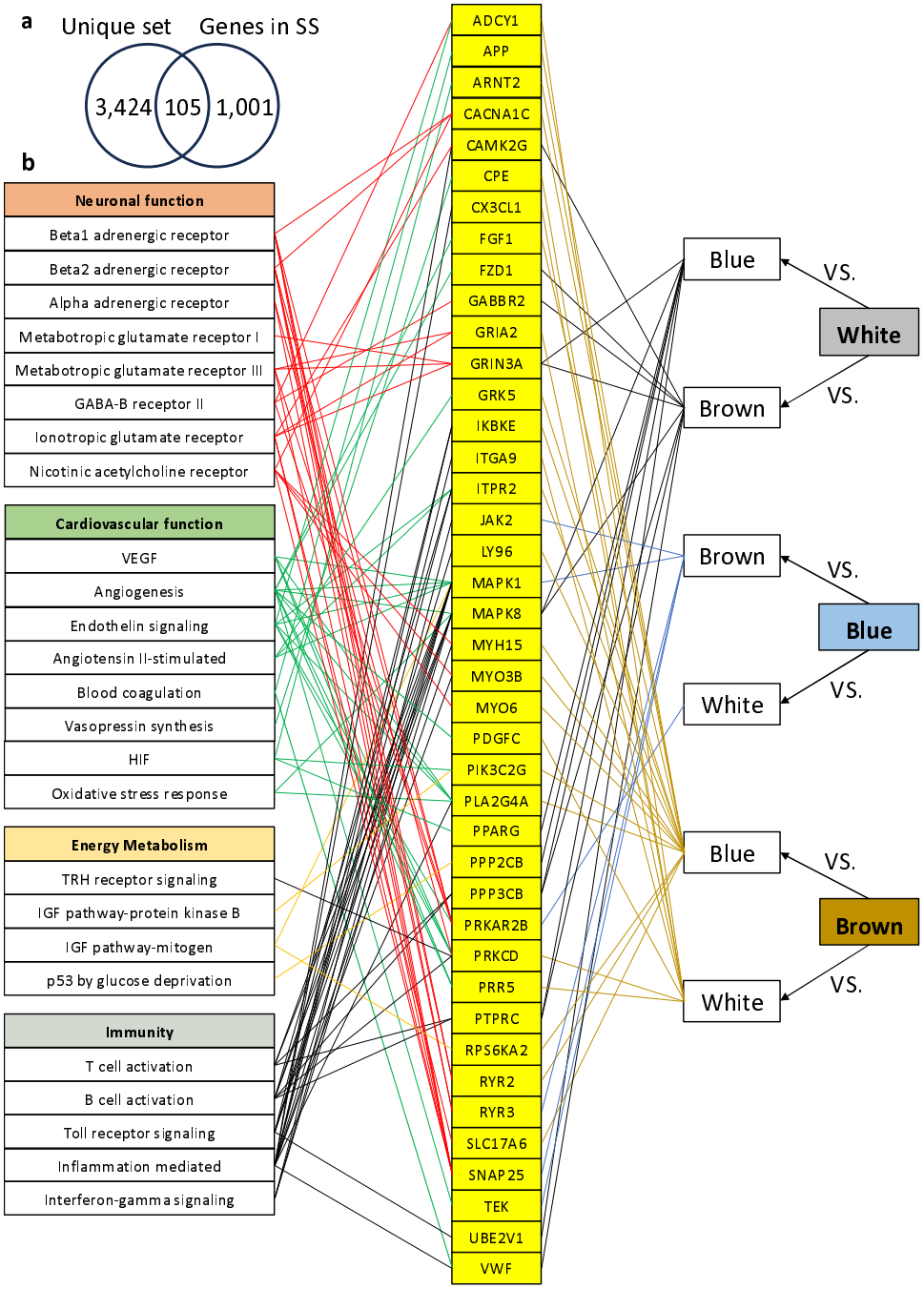

Figure 6. Comparison of unique genes and pseudogenes in the WT and BR genomes and genes in selective sweeps in the three species. **a.** Venn diagram of the set of unique genes and pseudogenes in the WT and BR genomes (unique set) and the set of genes in selective sweeps in the three species (genes in SS). **b.** The genes with GO term assignment in selection sweeps in each species that are involved in the same pathways of the same four functional categories as are the unique genes and unique pseudogenes with GO term assignment in the WT and BR genome, although the two sets only share four genes, i.e., *CAMK2G* and *VWF* are pseudogenized in BR, *PRKCD* is missing in BR, and *PIK3C2G* is pseudogenized in WT (Figure 3b).