

This is the peer reviewed version of the following article: *Pokrovac I, Rohner N, Pezer Ž. The prevalence of copy number increase at multiallelic copy number variants associated with cave colonization. Mol Ecol. 2024;33(9):e17339*, which has been published in final form at <https://onlinelibrary.wiley.com/doi/10.1111/mec.17339>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.

Title Page

The prevalence of copy number increase at multiallelic CNVs associated with cave colonization

Running title: CNVs associated with cave colonization

Ivan Pokrovac¹, Nicolas Rohner², Željka Pezer^{1†}

¹ Ruđer Bošković Institute, Zagreb, Croatia

² Stowers Institute for Medical Research, Kansas City, MO, USA

† corresponding author: zpezer@irb.hr

18 Abstract

19 Copy number variation is a common contributor to phenotypic diversity, yet its involvement in
20 ecological adaptation is not easily discerned. Instances of parallelly evolving populations of the
21 same species in a similar environment marked by strong selective pressures present
22 opportunities to study the role of copy number variants (CNVs) in adaptation. By identifying CNVs
23 that repeatedly occur in multiple populations of the derived ecotype and are not (or are rarely)
24 present in the populations of the ancestral ecotype, the association of such CNVs with adaptation
25 to the novel environment can be inferred. We used this paradigm to identify CNVs associated
26 with recurrent adaptation of the Mexican tetra (*Astyanax mexicanus*) to cave environment. Using
27 a read-depth approach, we detected CNVs from previously re-sequenced genomes of 44
28 individuals belonging to two ancestral surface and three derived cave populations. We identified
29 102 genes and 292 genomic regions that repeatedly diverge in copy number between the two
30 ecotypes and occupy 0.8% of the reference genome. Functional analysis revealed their
31 association with processes previously recognized to be relevant for adaptation, such as vision,
32 immunity, oxygen consumption, metabolism, and neural function and we propose that these
33 variants have been selected for in the cave or surface waters. The majority of the ecotype-
34 divergent CNVs are multiallelic and display copy-number increases in cave fish compared to
35 surface fish. Our findings suggest that multiallelic CNVs - including gene duplications, and
36 divergence in copy number provide a fast route to produce novel phenotypes associated with
37 adaptation to subterranean life.

38

39 Keywords: copy number variation, cave colonization, parallel evolution, rapid adaptation, gene
40 duplication, genomic structural variation

41 Introduction

42

43 Biological evolution operates on phenotypic and genetic variation. Structural variation is a
44 dominant form of genetic variation in terms of genome proportion, population frequency, and
45 taxonomic ubiquity. It refers to the inter-individual variation in the orientation, position, or copy
46 number of a genomic sequence. The latter covers deletions and amplifications of sequences,
47 collectively termed copy number variants (CNVs). CNVs affect both coding and noncoding
48 regions and thus may alter phenotype in various ways. For example, by affecting complete genes
49 CNVs may cause changes in gene dosage (Maron et al. 2013; Handsaker et al. 2015). They can
50 modulate the expression of individual genes by affecting enhancers or promoters, or reorganize
51 whole regulatory networks by striking at a single critical transcription factor (Vickrey et al. 2018;
52 Yuste-Lisbona et al. 2020). CNVs in exons cause changes in gene structure, consequently
53 changing the structure and efficiency of protein products (Boettger et al. 2016). With such
54 prevalence and prolific influence on phenotype, CNVs are considered to have a large impact on
55 phenotypic diversity.

56 Despite their pervasiveness among genomes and taxa, it is not clear to what extent
57 CNVs contribute to adaptation. Even in the most extensively studied species - humans, no
58 consensus has been reached yet and the assumptions range from neutral evolutionary processes
59 acting on the majority of CNVs to the significant contribution of adaptive evolution (Iskow et al.
60 2012; Saitou et al. 2022). Multiple factors complicate the reconciliation between studies such as
61 the heterogeneity of CNVs in terms of type, size, genomic context, and mutation rate, as well as
62 technical difficulties and limitations pertaining to the choice of methodology (reviewed in Pokrovac
63 and Pezer 2022). Moreover, experimental evolution as a tool is generally not feasible for more
64 complex multicellular organisms. The dynamics of CNVs at macro- and microevolutionary scales
65 are therefore inferred from comparative genomics and population-scale data, respectively, by
66 examining evolutionary signatures, such as divergence patterns and CNV frequencies. In this
67 respect, the instances of multiple populations with similar traits in similar environments represent
68 precious opportunities for studying the evolution by natural selection, because they provide an
69 element of reproducibility: if the trait repeatedly occurs in parallel, it is less likely to have been
70 driven by genetic drift, but rather it (and the underlying genetic variant) evolved multiple times as
71 a response to a common selection pressure associated with habitat similarity (Rundle et al.
72 2000). CNVs are surprisingly understudied in the context of parallel evolution and recurrent
73 adaptation. The only system that has received somewhat more attention in this regard is
74 freshwater colonization by marine fish (Hirase et al. 2014; Lowe et al. 2018; Ishikawa et al. 2022).
75 These studies identified CNVs, including changes in gene copy number (CN), that underly
76 adaptation to freshwater environments, by analyzing data from multiple ancestral and derived
77 populations.

78 We here use population-scale genomic data from Mexican tetra (*Astyanax mexicanus*) to
79 investigate the contribution of CNVs to adaptation to cave environments. *A. mexicanus* is a fish
80 species that exists in the form of two ecotypes: surface-dwelling, which inhabits lakes and rivers
81 throughout Mexico and southern Texas, and cave-dwelling, which can be found in waters of
82 multiple caves in northeastern Mexico (Gross 2012). The molecular data suggests that there are
83 two lineages of *A. mexicanus* - new and old lineage, which separated 150,000 - 300,000 years
84 ago (Herman et al. 2018). The cave form is considered derived from ancestral surface
85 populations that invaded caves on more occasions 10,000 - 100,000 years ago and these
86 transitions occurred in both lineages (Fumey et al. 2018). Despite some degree of gene flow from
87 surface populations and evidence of reticulate evolution, troglomorphic traits are maintained in
88 cave populations (Herman et al. 2018) indicating the presence of strong selection pressures to
89 cope with environmental challenges such as constant darkness, reduced food availability, low
90 oxygen level, and low parasite diversity. These are the key forces that drive the evolution of cave-
91 derived traits, such as eye degeneration (Moran et al. 2015), loss of pigment (Bilandžija et al.
92 2013), sleep loss (Duboué et al. 2011; Jaggard et al. 2018), increase in appetite and starvation
93 resistance (Aspiras et al. 2015), increased fat accumulation (Xiong et al. 2018), insulin resistance
94 (Riddle et al. 2018), enlarged hypothalamus (Menuet et al. 2007), changes in behavior
95 (Yoshizawa et al. 2010; Elipot et al. 2013; Kowalko et al. 2013), larger number and size of

96 erythrocytes (Boggs et al. 2022; van der Weele and Jeffery 2022), and shift in the immune
97 investment strategy (Peuß et al. 2020). The strong driving selective factors and distinctive derived
98 traits, a well-known population history, and the existence of multiple ancestral and derived
99 populations make *A. mexicanus* a convenient system for studying the role of CNVs and other
100 structural variants in recurrent adaptation.

101 We use available genomic data from 44 fish belonging to three cave and two surface
102 populations (Herman et al. 2018) to detect CNVs by a read-depth based approach. By identifying
103 genes and genomic regions whose copy numbers diverge in parallel between cave and surface
104 populations we discover signatures of natural selection associated with the transition to
105 underground waters. Our findings highlight the role of multiallelic CNVs and the copy number
106 increase in rapid adaptation to cave environment.

107 Materials and Methods

108

109 *Data acquisition*

110

111 The reference genome AstMex3_surface in FASTA format as well as annotation
112 information and assembly report were downloaded from NCBI database, under RefSeq assembly
113 accession GCF_023375975.1 (Warren et al. 2023). Illumina reads from *A. mexicanus* population
114 resequencing data described in Herman et al. (2018) were downloaded from European
115 Nucleotide Archive (SRA accession: PRJNA260715). Briefly, we obtained data for 28 cave fish (9
116 from Molino, 9 from Pachón, and 10 from Tinaja population), and 15 surface fish (6 from Rascon,
117 and 9 from Río Choy population). Information about samples, including SRA run accession
118 numbers is provided in Table S1. Sequencing data corresponding to the previous *A. mexicanus*
119 genome assembly *Astyanax_mexicanus*-1.0.2 (GenBank accession GCA_004802775.1) was
120 included as an additional sample of Pachón population (SRA accession: PRJNA533584). The
121 data is based on 15 runs of Illumina HiSeq2000 on DNA isolated from heart, gill, and liver of a
122 single female Pachón cave fish. Runs were joined and subsequently trimmed and cleaned using
123 Trimmomatic (Bolger et al. 2014) and cutadapt (Martin, 2011) so that they pass quality control.

124

125 *Preprocessing and mapping*

126

127 FASTQ files were quality checked with FastQC package (Andrews, 2010). Reads were
128 mapped to the reference genome using Bowtie 2 (Langmead and Salzberg 2012) with default
129 parameters. SAMtools (Li et al. 2009) was used to fix mate-pair information, sort the data, identify
130 and mark duplicate reads, index the BAM files, and calculate the mean of per-base coverage
131 (Table S2).

132

133 *CNV calling*

134

135 To discover CNVs from NGS data, we employ CNVpytor (Suvakov et al. 2021), a Python
136 extension of CNVnator and a tool that identifies deletions and duplications from regions with a
137 lack or excess of mapped reads, respectively (Abyzov et al. 2011). Such a read-depth-based
138 approach is highly accurate at estimating diploid CNs from short-read sequencing data, is robust
139 to interindividual differences in genome coverage, and performs well in repetitive regions (Abyzov
140 et al. 2011; Pezer et al. 2015; Kosugi et al. 2019; Garg et al. 2021). These features make it
141 suitable for comparing genome-wide patterns of CN between populations.

142 The optimal bin size was set for each sample individually, such that the ratio of global RD
143 mean to global RD standard deviation was between 4 and 5, corresponding to the relative
144 standard deviation of global RD of 0.25 and 0.2, respectively. The bin size ranged from 500 to
145 800 bp. Copy numbers (CNs) were determined with CNVpytor by using the *-genotype* option.

146

147 *Downstream analyses*

148

149 Principal component analysis and hierarchical agglomerative clustering were performed
150 in Python programming language using *SciPy* and *sklearn* packages.

151 Permutations were performed in Python *NumPy* package. Coordinates of CNV calls were
152 shuffled randomly on the same chromosomes, while ensuring that CNV size distribution and
153 duplications-to-deletions ratio matched those of the true data and that the annotated assembly
154 gaps were avoided. CNV calls were intersected with specific features using *bioframe* package
155 (Open2C et al. 2022).

156 Functional analysis of genes was performed by using the DAVID tool (Sherman et al.
157 2022). To assign biological processes and pathways, annotations from
158 UP_KW_BIOLOGICAL_PROCESS, GOTERM_BP_DIRECT, and KEGG_PATHWAY were used.
159 The most frequently occurring words from the DAVID output were extracted and compiled
160 manually into a list of terms that encompass words or word roots. Such terms were used to
161 analyze DAVID output by using the *grep* command in Linux. For example, the term "nerv*" was

162 used to count all instances of the words *nerve*, *nervous*, and *innervation*. *Bedtools closest* was
163 used to find the nearest genes to the noncoding CNVRs (Quinlan and Hall 2010).

164 To find genes with significant differences in CN between ecotypes or lineages, we
165 combined the results of two approaches: 1) population-pair approach, in which Welch T-Tests
166 were used in all combinations of a cave versus a surface population, or in all combinations of a
167 new versus an old-lineage population; 2) bulk-comparison approach, in which Mann-Whitney test
168 was used in comparison of all cave versus all surface individuals, or in comparison of all new-
169 lineage versus all old-lineage individuals. The nonparametric Mann-Whitney test was chosen for
170 bulk comparisons to account for multimodality in CN across the populations, given the population-
171 specific profile of CNVs. We further adjusted p-values using the FDR Benjamin-Hochberg
172 method and FWER Holm's method, setting the significance level at alpha 0.05. We considered
173 the difference to be significant when the significance level criterion was met in both approaches.

174 CNVpytor enables us to track two key metrics for genomic region analysis: total reads
175 mapped, and unique reads mapped. To estimate the fraction of zero mapping quality reads
176 (reads mapped to multiple regions), we calculate the difference between total and unique reads
177 mapped to a region and divide it by the number of all reads mapped to the same region.

178

179 *Validity considerations*

180

181 All CNVs in the dataset were detected and genotyped by CNVpytor, an extension of
182 CNVnator developed in Python (Suvakov et al. 2021; Abyzov et al. 2011). CNVpytor has
183 additional features and improved performance in terms of speed but retains CNVnator's high
184 sensitivity, low false-discovery rate, and high genotyping accuracy (Suvakov et al. 2021). The
185 genotyping accuracy of its algorithm reaches 96% and is robust to differences in coverage
186 (Abyzov et al. 2011; Pezer et al. 2015; Kosugi et al. 2019), rendering it one of the most commonly
187 used tools for CNV detection based on the read-depth approach. Moreover, a recent study
188 validated the genotyping performance of CNVnator in tandemly repetitive regions and found a
189 high correlation ($R^2 = 0.81$) between CNs deduced from short reads and long reads (Garg et al.
190 2021), demonstrating that CNVnator's algorithm is reliable even in such problematic regions in
191 which the mapping position of short reads cannot be determined with certainty. Our major
192 findings are based on copy numbers genotyped by the CNVpytor. We rely on the reports of high
193 accuracy described above as well as on the previously conducted validations in other studies that
194 reported high correlation and concordance with copy numbers determined by different
195 experimental and bioinformatic approaches (tan Nguyen et al. 2013; Yi et al. 2014; Shebanits et
196 al. 2019; Meng et al. 2022).

197 We considered the possibility that our major findings could be influenced by the high
198 proportion of mapped reads with low quality. For example, non-uniquely mappable reads could
199 pile up at one particular position of a genotyped region and consequently inflate the inferred copy
200 number of the whole region. We therefore assessed the proportion of reads with zero mapping
201 quality (MAPQ=0) at each of the 292 ecotype-divergent CNVRs and the 102 ecotype-divergent
202 CNV genes (Figure S6). At the largest majority of these genotyped loci, there were no low-quality
203 reads mapped, and the proportion is smaller than 5% at approximately 90% of analyzed the loci.
204 These numbers suggest that the effect of non-uniquely mapping reads on the genotyped copy
205 numbers of CNVRs and CNV genes is insignificant.

206 Results

207

208 *Population diversity*

209

210 By using CNVpytor we detected between 3,001 and 12,262 CNVs per animal (Table S2
211 in Supplementary Material). There was no significant difference between lineages or ecotypes in
212 the total number of detected CNVs. New lineage animals contained a higher proportion of
213 duplications than animals of the old lineage (Welch t-test, p-value 9×10^{-12} , Cohen's D 2.9):
214 approximately 8% of all calls in Choy and Molino fish are duplications, whereas they constitute on
215 average 4% of all calls in populations belonging to the old lineage. The higher duplication-to-
216 deletion ratio in new-lineage animals can be explained by the fact that the AstMex3 reference is
217 based on the fish from the Choy population, which belongs to the new lineage. Genomes are
218 expected to share more of their content within the same lineage than between lineages, *i.e.* fewer
219 regions in the reference genome are expected to be missing in the new-lineage genomes than in
220 the genomes of old-lineage fish. There was no major difference in the proportion of duplications
221 between surface and cave ecotypes (Welch t-test, p-value 0.36).

222 The extent of shared genetic variation between individuals can provide clues on genetic
223 diversity. In order to estimate relative genetic diversity within and between populations, we
224 analyzed the number of overlapping CNV calls between any two individuals in our dataset.
225 Predicted CNV call breakpoints are defined by genomic coordinates in the reference genome and
226 the distance between the breakpoints defines CNV length. If the genomic position of a CNV call in
227 one genome overlaps with the position of a call in another genome, and they do so over a
228 minimum of 50% of both lengths, these calls are defined as a shared CNV between the two
229 genomes. Based on such a definition, we find that any two samples share on average 1,946
230 CNVs, corresponding to 34% of all CNVs in a single pairwise comparison. Individuals are the
231 most similar to one another in the Molino population which belongs to the cave-dwelling new
232 lineage, where they share 3,448 CNVs on average (65%). This finding suggests the lowest
233 genetic diversity in Molino population, in accord with previous observations based on analyses of
234 SNP data (Bradic et al. 2013; Herman et al. 2018). Río Choy population (surface-dwelling new
235 lineage) is the most diverse, with individual pairs sharing on average 1,198 CNVs (36%). The
236 pattern of shared CNVs clusters populations according to their lineages (Figure 1A) but groups
237 the Rascon surface and Tinaja cave population together. This disagrees with the previously
238 proposed phylogeny based on SNP data in which cave populations Tinaja and Pachón form a
239 monophyletic sister group to Rascon surface population (Herman et al. 2018).

240 In order to identify CNV loci that are specific to populations, we defined copy number
241 variable regions (CNVRs) as regions enclosed by genomic coordinates of merged calls from all
242 individuals of the same population. This enabled us to determine population-private CNV loci, *i.e.*
243 calls present within CNVRs in at least one individual of a single population whereas absent from
244 all other populations (Table 1). We find 4,257 cave-specific and 4,728 surface-specific CNVRs.
245 The proportion of private CNVRs that are found in single animals is greater in surface fish (65%),
246 than in cave fish (41%). This is also evident in comparisons of individual surface populations
247 (72% in Río Choy and 68% in Rascon) with individual cave populations (45% - 62%). A single
248 private CNVR contains on average five CNVs in cave fish and only three in surface animals
249 (Table 1). These analyses suggest that cave fish more often share the same CNV, consistent
250 with pairwise similarity analysis based on the number of shared CNVs (Figure 1A). It was
251 proposed that most genetic variation in the caves results from standing genetic variation from the
252 ancestral surface stock and possible gene flow between the populations (Bradic et al. 2012).
253 Hence, genetic variation in cave animals is expected to largely represent a subset of surface
254 variation. The finding of a similar number of cave-specific CNVRs and surface-specific CNVRs is
255 therefore surprising. There was no significant difference between cave and surface in the
256 functional content of ecotype-specific CNVRs (Figure S1). Majority of the genes that overlapped
257 these CNVRs were associated with numerous signaling and metabolic pathways to similar extent
258 in both ecotypes.

259 We next analyzed the proportion of singleton CNVs, defined as CNVs detected in only
260 one animal and having no overlap with CNVs in other samples of the whole dataset. We detected

261 between 5 and 1,537 singletons per animal - corresponding to 0.12% - 12.53% of all CNVs. The
262 proportion of singletons, relative to the number of all detected CNVs within an individual, is higher
263 in surface populations than in cave populations (Figure 1B). This suggests that surface fish have
264 larger diversity compared to cave fish and agrees with previous analyses based on microsatellite
265 and SNP data (Bradic et al. 2012; Herman et al. 2018). The singleton proportion as well as the
266 CNV presence-absence pattern suggest low genetic diversity within the Pachón population
267 (Figure 1). This population is characterized by small size and stronger isolation, and our results
268 align with the low polymorphism observed from microsatellite data (Legendre et al. 2023).

269 To estimate the fraction of the AstMex3 genome that is copy number variable in the
270 whole set, we merged calls across all samples into CNVRs. We found in total 15,088 CNVRs on
271 assembled chromosomes, comprising cumulatively 260.2 Mbp of sequence. Compared to the
272 total sequence length of assembled chromosomes (1,321 Mbp), this translates into 19.7% of the
273 reference sequence being copy number variable in natural *A. mexicanus* populations. We
274 discover that 10,035 CNVRs overlap genes (including protein-coding genes, lncRNA, rRNA,
275 tRNA, and all other annotated gene classes) and occupy 213 Mbp in total, whereas the rest 5,053
276 (occupying 48 Mbp in total) can be considered purely noncoding CNVRs. The larger proportion of
277 CNVRs associated with genes may be explained by the high proportion of gene-encoding
278 sequences present in the reference assembly: as much as 63% of the assembled sequence is
279 annotated as genes in the AstMex3.

280

281 *Population-specific patterns of gene copy number*

282

283 We detected in total 2,819 unique protein-coding genes that are entirely spanned by a
284 CNV in at least one sample. We refer to this set of genes as CNV genes (Table S3 in
285 Supplementary_Tables.xlsx). The average length of a CNV gene is 9,263 bp (median 5,335 bp),
286 which is substantially shorter than the average length of all annotated protein-coding genes
287 (34,441 bp; median 13,069 bp). Over a third of CNV genes (978) are uncharacterized, *i.e.* of yet-
288 unknown function, which is a four-fold increase compared to the proportion of uncharacterized
289 protein-coding genes in the whole assembly (8.5%). This enrichment of uncharacterized genes in
290 the set of CNV genes could be a consequence of their generally short length (7.3 kbp on average
291 in the whole genome), such that smaller genes are more likely to entirely reside within CNVs than
292 longer genes. In order to test this, we performed permutation analyses. Coordinates of CNV calls
293 were shuffled randomly, keeping the distribution of CNV lengths as is in the true data. In each
294 permutation, the average length of CNV genes and the proportion of uncharacterized CNV genes
295 were calculated. Based on 100 such permutations, we would expect to find about 10% of
296 uncharacterized genes within the set of CNV genes, which is very close to the proportion seen in
297 the whole AstMex3. Moreover, we would expect their average length to be around 15.7 kbp,
298 which is lower than the genome average (34.4 kbp) yet substantially higher than the true CNV
299 genes set (9.3 kbp). CNV genes are therefore 3.5 times more likely to be uncharacterized and 1.7
300 times shorter than expected. A previous study on three-spined stickleback fish demonstrated that
301 CNVs are enriched for evolutionary new genes which are generally shorter and for which no
302 evidence of homology in other species has been found (Chain et al. 2014). Their function is often
303 unknown; hence these genes are described as “uncharacterized”. Notably, given the recency of
304 the AstMex3 assembly and the associated genome annotations, some of the genes labeled as
305 “uncharacterized” may not actually be true orphans or lineage-specific genes, but may instead
306 represent genes that have orthologs in other species, of yet undetermined function. However, the
307 lack of such characterization, short gene length, and the strong enrichment compared to
308 expectations based on permutations in our study, provide further support to the hypothesis that
309 CNVs are enriched for young genes (Chain et al. 2014).

310 By using the *-genotype* option in CNVpytor we determined the copy number of every
311 CNV gene in each sample (Table S3). Based on gene copy numbers, the samples cluster by their
312 geographic location (Figure 2A and 2B). Interestingly, the two surface populations, new-lineage
313 Río Choy and old-lineage Rascon, are closer than expected, given that the split between the two
314 lineages happened at least 200,000 years ago (Herman et al. 2018). Moreover, hierarchical
315 clustering suggests that Rascon surface and Tinaja cave fish are the most similar based on the

316 pattern of gene copy number and that the two populations form a sister clade with the Río Choy
317 surface population (Figure 2B).

318 To find genes with significant differences in CN between ecotypes, we performed
319 appropriate statistical tests for all combinations of cave-surface population pairs, as well as for
320 comparison of all cave versus all surface individuals (see Methods section for details). Based on
321 this approach, we found that 102 out of 2,819 CNV genes were significantly different in copy
322 number between cave and surface ecotypes (Table S3). Over half (65) are predicted genes of
323 unknown function among which 89% (58 genes) have an average copy number higher in cave
324 than in surface fish. For example, we find between 4 and 24 copies of the *LOC125782174* in cave
325 individuals, whereas the same gene exists in 1-3 copies in the surface fish (Figure 2C). Similarly,
326 we find up to 4 copies of *LOC125799116* in genomes sampled in surface waters and 6-40 copies
327 in cave genomes. Among the genes that are characterized in this set, some have either lower or
328 higher average copy number in cave fish genomes compared to surface, and play a part in
329 different processes such as (innate) immunity (*LOC111188594* - fucoselectin-1-like;
330 *LOC111194616* - polymeric immunoglobulin receptor-like; *LOC111195410* - E3 ubiquitin-protein
331 ligase DTX3L; *LOC125785782*, *LOC125785783*, *LOC111188451* and *LOC125785616* - B-cell
332 receptor CD22-like; *LOC111197671* - C-type lectin domain family 4 member E-like;
333 *LOC125784871*, *LOC125785663* and *LOC125784856* - NLR family CARD domain-containing
334 protein 3-like), oxygen transport (*LOC111191628*, *LOC111196759*, *LOC111191630* and
335 *LOC103027764* - encoding hemoglobin subunits; *LOC125787068* - scavenger receptor cysteine-
336 rich type 1 protein M130-like), and lipid metabolism (*LOC103026892* - 60 kDa lysophospholipase;
337 *LOC125799429* - phospholipase B-like 1; *LOC111195147* - apolipoprotein L3-like). For example,
338 gene *LOC125784890* is predicted to encode trace amine-associated receptor 13c on
339 chromosome 20. This gene belongs to a family of vertebrate olfactory receptor genes and its
340 copy number is reduced in cave genomes compared to surface fish (Figure 2C). Similarly, gene
341 *LOC111195147* predicted to encode apolipoprotein L3-like seems to be completely deleted in
342 cave genomes and is present mainly in one or two copies in surface populations (Figure 2C).
343 Apolipoprotein L3 in humans is implicated in the movement of lipids (including cholesterol) within
344 cytoplasm, and the binding of lipids to organelles (Gene database - National Library of Medicine;
345 Gene ID: 80833). Gene *LOC125801369* which encodes a protein similar to zinc finger protein 501
346 on chromosome 4 is present in 3-9 copies in most cave fish samples, whereas it is present in 1-2
347 copies in the majority of surface genomes. Similarly, the copy number of gene *si:dkey-93h22.7* is
348 amplified in cave fish populations (Figure 2C). This gene encodes golgin subfamily A member 6-
349 like protein 22 and its expression is restricted to testis in humans (Gene database - National
350 Library of Medicine; Gene ID: 440243).

351 Similarly, we found 157 genes with significant differences in their CN between lineages
352 (Table S3). Among the 74 uncharacterized genes in this set, 40 had an average copy number
353 higher in new lineage compared to old lineage animals. Several annotated genes stand out as
354 being specifically amplified or deleted in only one lineage. For example, we find a substantially
355 higher copy number of the *ERVFC1* gene annotated on chromosome 16 in old-lineage animals.
356 This gene encodes endogenous retroviral envelope protein and is present in 1-3 copies in Choy
357 and Molino fish, and in 6-27 copies in genomes of Rascon, Tinaja, and Pachón fish (Figure S2).
358 Instances of gene *PGBD4* annotated on chromosomes 1, 2, 10, 19, and 20 are present mainly as
359 one copy per diploid in old-lineage animals and as two or three copies in the new-lineage
360 animals. This gene belongs to the family of piggyBac transposable element-derived (*PGBD*)
361 genes that are found in diverse animals (Sarkar et al. 2003). In humans, its expression is
362 enhanced in skeletal muscle, spermatids, and immune cells (Human Protein Atlas; Uhlén et al.
363 2015). Similarly, homologs of gene *SMAD3* annotated at different genomic locations on
364 chromosomes 4, 5, 11, 15, 17 and 22, are reduced to a single copy per diploid on average in fish
365 of the old lineage, compared to on average three copies in the new lineage. In humans, this gene
366 encodes a widely expressed transcription factor with roles in many cellular processes such as cell
367 proliferation, cell movement, and apoptosis (Human Protein Atlas; Uhlén et al. 2015).

368
369 *Ecotype-divergent CNVs at genes*
370

371 Differences in copy number between populations can result in phenotypic differences
372 between them. If genes affected by divergent CNVs are associated with particular biological
373 processes, this may indicate that these processes have also diverged between the populations.
374 To identify such processes that might be specifically targeted for copy number divergence
375 between surface and cave fish, we explored the functional context of genes that are frequently
376 affected by CNVs in single ecotypes.

377 We find a total of 9,187 protein-coding genes that are overlapped by CNVs. At 1,653
378 (18%) of the genes, we detected CNVs in only one sample, and at 476 (5%) genes CNVs are
379 found in all 44 analyzed genomes (Figure S3). Of the 9,187 genes, 15% (1,407) were affected by
380 CNVs exclusively in cave (Molino, Pachón, Tinaja) populations (Table S4) and 19% (1,726) in
381 surface (Rascon, Río Choy) populations (Table S5). Within this set of ecotype-specific events, the
382 largest proportion of the genes were affected in only one animal (42% or 589 genes in the cave,
383 and 62% or 1,064 genes in the surface). A much smaller fraction overlapped CNVs in multiple
384 animals and in at least one animal per population, *i.e.* 4% (58 genes) in the cave and 9% (163
385 genes) in the surface fish. Genes that are the most frequently affected by ecotype-specific events
386 are genes that are associated with immune response. For example, genes such as
387 *LOC125785663* (NLR family CARD domain-containing protein 3-like), *LOC111196508*
388 (scavenger receptor cysteine-rich type 1 protein M130), *LOC103031898* (deleted in malignant
389 brain tumors 1 protein), and *pikfyve* (phosphoinositide kinase, FYVE finger containing) are
390 associated with innate immunity, inflammatory response and antiviral defense. CNVs in these
391 genes are detected in 13-28 cave individuals (of the 29 in total; Table S4). Similarly, genes
392 implicated in these processes are found to overlap the most frequent surface-specific events:
393 *LOC103031476* (NACHT, LRR and PYD domains-containing protein 3), *mf41* (ring finger protein
394 41), *LOC125801137* (E3 SUMO-protein ligase ZBED1-like) and *LOC125782699* (scavenger
395 receptor cysteine-rich type 1 protein M130-like) are affected by CNVs in 9-13 out of 15 surface
396 individuals (Table S5). Within the set of cave-specific events with the highest frequency (>30%) in
397 analyzed populations, we find genes associated with processes such as visual perception (*rgrb* -
398 retinal G protein-coupled receptor b; *LOC111194948* - TOG array regulator of axonemal
399 microtubules protein 1; *map2* - microtubule-associated protein 2; *opn8a* - opsin 8 group member
400 a; *LOC107197208* - complement C1q-like protein 3), genes encoding hemoglobin subunits
401 (*LOC111191630* - hemoglobin subunit beta-2-like; *LOC111191631* - hemoglobin embryonic
402 subunit alpha; *LOC111191628* - hemoglobin embryonic subunit alpha), and genes implicated in
403 neurological functions (*plppr3a* - phospholipid phosphatase related 3a; *LOC103030484* - ras-
404 related protein Rab-26; *rab3c* - RAS oncogene family member). Genes associated with these
405 processes are either not or are not frequently affected by CNVs in surface populations (Tables S4
406 and S5).

407 Ecotype-divergent CNVs can affect genes along their whole length, as described for CNV
408 genes above. For example, gene *LOC125785663* appears to be either entirely deleted or
409 reduced to a single copy in cave individuals (Figure 3, Table S3). An ecotype-divergent CNV
410 gene can be affected by both deletions and duplications in different individuals, such as the gene
411 *LOC111194948*, which appears to be either deleted or amplified in the cave genomes, while
412 present mainly in two copies in surface individuals (Figure 3, Table S3). CNVs can affect only a
413 part of a gene at high frequency, such as in the case of *rgrb* and *mf41*. In all analyzed cave
414 genomes, a region from exon 4 to 5 of *rgrb* is deleted. Similarly, a region encompassing the first
415 intron and the first two exons of *mf41* is amplified in all analyzed surface fish (Figure 3).

416

417 *Ecotype-divergent CNVRs*

418

419 Of the 15,088 CNVRs detected in the whole set, 292 showed significant differences in
420 copy number between cave and surface animals (Wilcoxon rank sum test, adjusted pval < 0.01),
421 in all combinations of surface-cave comparisons as well as in comparison of all surface fish with
422 all cave fish (Table S6). These constitute in total 10.47 Mbp, equaling 0.8% of the reference
423 genome. Almost a third of these CNVRs (87) are noncoding, whereas the majority (205) overlap
424 or contain one or more genes (Table S6). Average copy numbers of these genomic regions range
425 much wider in cave fish than in surface fish. Of the CNVRs that show significant differences in

426 average CN between ecotypes, the largest majority varies only slightly in surface fish, existing
427 mainly in one or two copies per diploid. The same genomic regions in cave fish genomes are
428 amplified up to 34 copies on average (Figure 4A). Two-thirds of these regions (195/292) exist at
429 higher copy numbers in cave fish compared to surface, whereas only 97 CNVRs are estimated to
430 have lower copy numbers in cave fish.

431 We analyzed functional annotations of genes overlapping CNVRs with significant
432 differences in their copy number between ecotypes. Of the 460 genes in the 205 regions, only 83
433 had associated annotations for biological processes or pathways (Table S7). About half of these
434 genes were associated with numerous signaling and metabolic pathways, including MAPK, Toll-
435 like and C-type lectin signaling pathways (Figure 4B and Table S8). Other highly represented
436 categories included transport of substances, regulation of different processes, and processes
437 related to nervous system functioning such as neuroactive ligand-receptor interaction and
438 neurotransmitter transport. These categories were also present at similarly high proportions in
439 genomic regions that did not significantly differ in copy number between ecotypes (Figure 4B).
440 However, some processes seem to be more represented in genomic regions with divergent copy
441 number between cave and surface animals. They include categories associated with cell
442 adhesion, apoptosis, glucose metabolism, as well as metabolisms of cytochrome, retinol and
443 nicotinate/nicotinamide, and the degradation of branched-chain amino acids (Figure 4B and Table
444 S8). Notably, among divergent CNVRs, we find many that intersect genes potentially involved in
445 biological processes that have been shown to change upon adaptation to darkness, such as
446 vision, development, and behavior. For example, genes *LOC111197051* and *LOC103025981* are
447 annotated in the AstMex3 assembly as two copies of the same gene that encodes guanylyl
448 cyclase-activating protein 2, which is involved in phototransduction. These genes overlap a ~66
449 kb long genomic region that is affected by duplications only in cave fish (Figure 4C). Similarly, a
450 31 kb genomic region is amplified in cave fish that overlaps gene *aldh1a2* (Figure 4C). This gene
451 encodes retinaldehyde dehydrogenase 2, which is an enzyme essential for proper embryonal
452 morphogenesis (Niederreither et al. 2002). A small, ~1 kb region within gene *nrb* appears to exist
453 in cave fish as a single copy per diploid (Figure 4C). This gene encodes neuropeptide B, which is
454 expressed in the central nervous system and implicated in regulating feeding behavior (Singh and
455 Davenport 2006).

456 Noncoding genomic regions that are significantly different in copy number between
457 ecotypes may be relevant for adaptation if they contain regulatory elements. Such alterations may
458 cause differences in the regulation of specific processes associated with nearby genes. To
459 explore this idea, we extracted the closest gene to each of the 87 divergent noncoding CNVRs
460 and analyzed their functions. Although only 27 genes had associated annotations for biological
461 processes or pathways, they mainly agree with the analysis of gene-encoding CNVRs described
462 above (Table S9). Many of these genes can be associated with processes that are known to
463 change upon adaptation to the cave environment. For example, a gene that encodes cytochrome
464 c oxidase subunit 4 (*LOC103039037*) is located ~350 bp downstream of a ~1.4 kb region that is
465 deleted in cave fish, whereas the same region exists as two copies per diploid in surface fish
466 (Figure 4C). This gene is associated with oxidative phosphorylation and cardiac muscle
467 contraction, and the cave-specific deletion of the CNVR next to this gene may reflect differences
468 in its regulation as an adaptation to hypoxic conditions in caves. Another alteration that might be
469 associated with adaptation to reduced oxygen levels in a subterranean environment is the
470 amplification of a ~8 kb genomic segment, about 5 kb upstream of *ssbp1*. This gene is important
471 for mitochondrial biogenesis and it is tempting to speculate that 9-17 copies of the CNVR near
472 this gene in cave fish might be responsible for the altered regulation of mitochondria production
473 as a possible route to compensate for hypoxia (Gutsaeva et al. 2008; Gamboa and Andrad 2009).
474 Interestingly, a small region (~2 kb) located ~27 kb upstream of *trh* gene is found at lower copy
475 number in surface fish compared to cave fish (Figure 4C). This gene is expressed in
476 hypothalamic neurons as thyrotropin-releasing hormone that has major roles in many biological
477 processes including metabolic activity, thermoregulation, locomotor activity, pain perception, and
478 sleep regulation (Wozniak and Quinnell 2015).

479
480

Evolutionary implications

481

482 On average, 57% of all detected CNVs overlap protein-coding genes (Figure 5A), 12% of
483 which affect whole genes. The second most prominent group of genes that are affected are
484 genes that encode long noncoding RNAs, with approximately 10% of all detected events
485 overlapping them. About a third of all CNVs have no overlap with any of the categories related to
486 gene features (Figure S4).

487 Permutations of calls in analyzed samples can provide clues to the mode of evolution that
488 acts on CNVs affecting particular gene categories. For example, if the detected proportion of
489 CNVs affecting a particular group of genes is similar to the expected proportion, those CNVs can
490 be considered to represent neutral variation. In contrast, if the expected proportion is higher or
491 lower than the detected proportion, CNVs affecting such category of genes may be under
492 purifying or positive selection, respectively. In order to see if the detected proportions of CNVs
493 overlapping different gene features are expected by chance, or if there is some bias for or against
494 CNVs, we permuted the calls and analyzed their overlap with each of the annotated gene
495 categories (Figure 5, Figure S5). Analysis of permuted data suggests that between 69% and
496 78% of all CNVs would affect protein-coding genes by random chance (Figure 5A). Comparison
497 with the detected proportions that range from 55% to 61% suggests that CNVs are generally
498 biased away from protein-coding genes in all samples. This is further corroborated when the
499 number of affected genes is considered: on average 18% of all protein-coding genes are
500 expected to be affected by CNVs based on permutations, compared to the detected 12% of
501 genes in the true data (Figure 5B).

502 The bias against CNVs in protein-coding genes is contributed by events affecting parts of
503 genes, whereas complete genes appear to be duplicated or deleted in a neutral fashion (Figure
504 5C). For the most part, this is in agreement with previous findings in humans that suggested
505 purifying selection acts against all types of structural variants that affect protein-coding genes,
506 except complete duplications (Collins et al. 2020). Similarly, it was demonstrated that duplication
507 events, rather than deletions, are more likely to include entire protein-coding genes in the
508 stickleback fish genomes (Lowe et al. 2018). The apparently neutral variation of gene copy
509 number in our dataset could explain the stratification of samples based on CNV genes (Figure 2),
510 which roughly reflects the demographic history of these populations. Interestingly, we find two
511 exceptions: Molino population, in which fewer gene duplications are detected than expected -
512 suggesting purifying selection, and Rascon population, in which gene deletions seem to be
513 subject to positive selection (Figure 5C, upper panels). To investigate if the observed bias against
514 CNVs affecting part of genes is a function of the unusually large intron sizes in the Mexican tetra
515 (Jakt et al. 2022), we analyzed the overlap of CNVs with exons and introns. The comparison of
516 permuted and true data suggests that CNVs affecting a part of intron or exon are subject to
517 neutral evolution (Figure 6). Interestingly, complete intron or exon duplications seem to be better
518 tolerated than deletions. Therefore, the general bias against CNVs at protein-coding genes is
519 contributed predominantly by selective constraints against deletions of complete exons or introns,
520 whereas neutral evolutionary forces seem to shape variation in gene copy number.

521 Based on comparisons of detected CNVs with permuted data (Figures 5, 6, and S5), we
522 compiled the inferred mode of mutation of CNVs intersecting all major annotated gene categories
523 in the AstMex3 assembly (Table 2). Interestingly, duplications of whole genes, regardless of gene
524 category, seem to generally evolve neutrally, whereas deletions of complete genes are either
525 neutral or under positive selection, such as in genes encoding lncRNAs, rRNAs, tRNAs, and
526 pseudogenes. Surprisingly, only partial duplications and partial deletions of protein-coding genes
527 seem to be subject to purifying selection (Figure 5B).

528

529

530 Discussion

531

532 Over the course of evolution, taxonomically diverse animals have successfully
533 transitioned from surface to subterranean environments and thus converged to a set of adaptive
534 traits such as loss of vision and pigment, and decline in metabolic rate. These species thus
535 provide a natural setting for studying the molecular basis of adaptation to constant darkness and

536 temperature, food scarcity, and low oxygen levels. Copy number variation as a form of genetic
537 variation has not been sufficiently studied in this context, especially given that it has been
538 recognized as a major contributor to phenotypic variation and rapid adaptation to novel and
539 extreme environments (Kondrashov et al. 2002; Kondrashov 2012; Lye and Purugganan 2019;
540 Rinker et al. 2019). A recent study addressed the role of CNVs in adaptation to subterranean
541 environment at a macroevolutionary scale (Balart-García et al. 2023), yet a comprehensive
542 exploration at the level of population is missing. We here considered the role of CNVs in the
543 adaptation of the Mexican tetra to cave. By exploiting a set of genomic data previously generated
544 from specimens collected at five distinct localities, we infer events and biological processes that
545 may be under selection. The same genomes were previously studied by Warren et al. (2021), but
546 only deletions were considered, and the study did not perform systematic analysis to detect
547 divergence between ecotypes. Therefore, our study here represents the first comprehensive
548 analysis of duplications and deletions from a micro-evolutionary perspective that identifies
549 divergent CNVs associated with parallel cave colonization.

550 Considering all calls in analyzed individuals, we estimate that one-fifth of the reference
551 sequence is subject to variation in copy number in natural *A. mexicanus* populations, the majority
552 of which is contributed by CNVs affecting genes. However, it is important to note that only a third
553 of the assembled AstMex3 consists of sequences outside genes. Previous studies in humans
554 demonstrated that repetitive regions are particularly rich in structural variants, including CNVs
555 (Huddleston et al., 2017; Audano et al., 2019; Ebert et al., 2021). By approaching the completion
556 of the *A. mexicanus* reference genome assembly, especially with respect to repetitive sequences,
557 we expect to see an increase in the proportion of CNVs in noncoding regions.

558 We observe strong stratification of CNVs between populations, in line with numerous
559 studies in various species (Sudmant et al. 2015; Pezer et al. 2015; Xu et al. 2016; Dorant et al.
560 2020; Zhu et al. 2020; Jang et al. 2021; Solé et al. 2019; Yang et al. 2023). The profile of CNVs
561 based on presence-absence patterns follows the previously established phylogenetic relationship
562 between new and old lineage *A. mexicanus* populations based on SNPs (Herman et al. 2018).
563 However, within the old lineage, cave population Pachón is less similar to the Tinaja cave
564 population, and the latter shares a larger proportion of CNVs with the Rascon surface population.
565 This departure from SNP-based phylogenetic inference is even more obvious when CNV genes
566 are considered: the hierarchical clustering based on gene copy number positioned the old lineage
567 Pachón closer to the new lineage Molino population. These differences can be explained in the
568 light of the elevated mutation rate of CNVs compared to SNPs (Zhang et al. 2009). Additionally,
569 other factors may contribute to the observed patterns, such as the population size and the degree
570 of migration between populations. Recent findings suggest that the Pachón population consists of
571 only a few hundred individuals inhabiting a relatively isolated area, the direct consequences of
572 which are low genetic polymorphism and limited gene flow from surface or other cave populations
573 (Legendre et al. 2023). Accordingly, among the studied populations in our analyses, Pachón
574 shows low diversity at the level of CNVs. Therefore, factors such as small population size and
575 isolation of the Pachón population combined with the high CNV mutation rate may have
576 increased the effect of genetic drift on CNVs genome-wide, resulting in an unexpected position of
577 Pachón on the phylogenetic tree.

578 We used multiple approaches to estimate the relative CNV diversity of analyzed
579 populations, and the results consistently indicate lower genetic diversity of cave populations
580 compared to surface. Such finding agrees with previous studies based on SNPs and
581 microsatellite data and can be explained by a combination of small effective population size,
582 limited nutrient and space availability in caves, as well as possible bottleneck events (Bradic et al.
583 2012; Bradic et al. 2013; Herman et al. 2018). Most of the genetic variation in the caves is
584 proposed to represent a subset of standing genetic variation from the ancestral surface stock
585 (Bradic et al. 2012). However, we detected a high and comparable number of genomic regions
586 that show copy number variation in an ecotype-specific manner, *i.e.* thousands of CNVs are
587 detected in either cave or surface genomes but not both. Such observation raises a possibility
588 that a great deal of copy number variation arises independently in both ecotypes. Nevertheless,
589 these results are heavily dependent on the sample size, and many more genomes per ecotype
590 would need to be screened for more reliable numbers and a firm interpretation of this finding.

591 To infer events that are likely under selection, we searched for genomic loci that satisfy
592 all of the following criteria: 1) the region is divergent in copy number between surface and cave
593 populations, 2) the divergence is significant in all cave-surface population-pair comparisons, and
594 3) the copy number change proceeds in the same direction in all population-pair comparisons.
595 Such stringent criteria allowed us to identify almost three hundred genomic regions containing
596 CNVs that may have been selected for in the cave or surface waters. These regions cumulatively
597 account for nearly 1% of the genome. This proportion is well in line with the estimate obtained by
598 a study that analyzed populations of three-spined sticklebacks in the context of freshwater
599 colonization (Lowe et al. 2018). Within this set, we identified approximately one hundred genes
600 with ecotype-divergent CNs, most of which show properties of young genes and are mainly
601 amplified in cave populations. Similarly, two-thirds of all detected ecotype-divergent genomic
602 regions have elevated copy numbers in cave (derived) genomes. Interestingly, gene copy number
603 increase, rather than decrease, was also found to be dominant in the derived freshwater
604 populations of sticklebacks (Hirase et al. 2014; Lowe et al. 2018; Ishikawa et al. 2022). These
605 studies together with our findings suggest that larger gene-copy numbers may generally confer
606 higher adaptive potential upon colonization of novel and extreme environments. An earlier
607 analysis of human CNVs suggested that duplications are under less stringent evolutionary
608 constraints than deletions, thus representing a larger target for adaptive selection (Sudmant et al.
609 2015). Moreover, duplications are more likely to show higher mutation rates due to the
610 susceptibility to nonallelic homologous recombination between directly oriented duplicated
611 sequences. This enables them to frequently change their copy-number state over a short time
612 (Sudmant et al. 2015) and thus persist as multiallelic CNVs within a population. The majority of
613 ecotype-divergent regions in our study are present at multiple copy-number states in the dataset
614 and we show that many encompass complete genes. These multiallelic CNV genes may be
615 particularly relevant in an evolving lineage and may increase fitness from the moment of their
616 origin, most likely as a protein dosage effect in response to a changing environment (Kondrashov
617 et al. 2002; Handsaker et al. 2015). Although we find a considerable proportion of cave-specific
618 CNVs (detected only in the cave fish) in the whole dataset, the majority of the identified divergent
619 regions show some copy number variation in the surface populations as well. Hence, the
620 selection seems to draw from the pre-existing standing genetic variation in the ancestral
621 populations, as the fastest route for adaptation to occur (Jones et al. 2012; Lai et al. 2019; Zong
622 et al. 2021).

623 Many ecotype-divergent genomic regions in our study are associated with genes involved
624 in biological processes that have previously been identified as important for the adaptation of *A.*
625 *mexicanus* to life in caves. We find CNVs at or near various genes that are associated with the
626 functioning of the central nervous system, visual processing, metabolism, oxygen consumption,
627 and immune system, highlighting the involvement in environmental information processing as
628 their common feature (Kondrashov et al. 2002). The parallel divergence of these variants in cave
629 populations suggests their participation in physiological and behavioral responses to major
630 challenges such as constant darkness, low nutrient availability, low oxygen level, and differences
631 in parasite composition.

632 In conclusion, our findings support the notion that gene duplications and divergence in
633 copy number are important generators of evolutionary innovation associated with adaptation to
634 subterranean life (Balart-García et al. 2023). In line with previous observations based on studies
635 in other species, we suggest that CNVs contribute to phenotypic diversity and facilitate rapid
636 ecological adaptation (Kondrashov et al. 2002; Sudmant et al. 2015; Rinker et al. 2019).

637

638

639 Acknowledgments

640

641 This work was supported by the Croatian Science Foundation (grant UIP-2019-04-7898). NR is
642 funded by the National Institutes of Health (NIH, grant R24OD030214). Data analysis was

643 performed on the high-performance computing cluster at the University Computing Centre
644 (SRCE), University of Zagreb. We thank Branka Bruvo Mađarić for helpful discussions.

645 References

- 646
647 Aspiras AC, Rohner N, Martineau B, Borowsky RL, Tabin CJ. Melanocortin 4 receptor mutations
648 contribute to the adaptation of cavefish to nutrient-poor conditions. *Proc Natl Acad Sci U S A*.
649 2015 Aug 4;112(31):9668-73. doi: 10.1073/pnas.1510802112.
- 650 Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and
651 characterize typical and atypical CNVs from family and population genome sequencing. *Genome*
652 *Res*. 2011;21:974–84.
- 653 Andrews S. FastQC: a quality control tool for high throughput sequence data. Available online at:
654 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>. 2010.
- 655 Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty
656 ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK,
657 Eichler EE. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*. 2019
658 Jan 24;176(3):663-675.e19. doi: 10.1016/j.cell.2018.12.019.
- 659 Balart-García P, Aristide L, Bradford TM, Beasley-Hall PG, Polak S, Cooper SJB, Fernández R.
660 Parallel and convergent genomic changes underlie independent subterranean colonization across
661 beetles. *Nat Commun*. 2023 Jun 29;14(1):3842. doi: 10.1038/s41467-023-39603-1.
- 662 Bilandžija H, Ma L, Parkhurst A, Jeffery WR. A potential benefit of albinism in *Astyanax* cavefish:
663 downregulation of the *oca2* gene increases tyrosine and catecholamine levels as an alternative to
664 melanin synthesis. *PLoS One*. 2013 Nov 25;8(11):e80823. doi: 10.1371/journal.pone.0080823.
- 665 Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, Hirschhorn JN, McCarroll SA.
666 Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol
667 levels. *Nat Genet*. 2016 Apr;48(4):359-66. doi: 10.1038/ng.3510.
- 668 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
669 *Bioinformatics*. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170.
- 670 Bradic M, Teotónio H, Borowsky RL. The population genomics of repeated evolution in the blind
671 cavefish *Astyanax mexicanus*. *Mol Biol Evol*. 2013 Nov;30(11):2383-400. doi:
672 10.1093/molbev/mst136.
- 673 Bradic M, Beerli P, García-de León FJ, Esquivel-Bobadilla S, Borowsky RL. Gene flow and
674 population structure in the Mexican blind cavefish complex (*Astyanax mexicanus*). *BMC Evol Biol*.
675 2012 Jan 23;12:9. doi: 10.1186/1471-2148-12-9.
- 676 Chain FJ, Feulner PG, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, Lenz TL, Stoll M,
677 Bornberg-Bauer E, Milinski M, Reusch TB. Extensive copy-number variation of young genes
678 across stickleback populations. *PLoS Genet*. 2014 Dec 4;10(12):e1004830. doi:
679 10.1371/journal.pgen.1004830.
- 680 Dorant Y, Cayuela H, Wellband K, Laporte M, Rougemont Q, Mérot C, Normandeau E, Rochette
681 R, Bernatchez L. Copy number variants outperform SNPs to reveal genotype-temperature
682 association in a marine species. *Mol Ecol*. 2020 Dec;29(24):4765-4782. doi: 10.1111/mec.15565.
- 683 Duboué ER, Keene AC, Borowsky RL. Evolutionary convergence on sleep loss in cavefish
684 populations. *Curr Biol*. 2011 Apr 26;21(8):671-6. doi: 10.1016/j.cub.2011.03.020.
- 685 Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J,
686 Zhou W, Serra Mari R, Yilmaz F, Zhao X, Hsieh P, Lee J, Kumar S, Lin J, Rausch T, Chen Y,
687 Ren J, Santamarina M, Höps W, Ashraf H, Chuang NT, Yang X, Munson KM, Lewis AP, Fairley
688 S, Tallon LJ, Clarke WE, Basile AO, Byrka-Bishop M, Corvelo A, Evani US, Lu TY, Chaisson
689 MJP, Chen J, Li C, Brand H, Wenger AM, Ghareghani M, Harvey WT, Raeder B, Hasenfeld P,
690 Regier AA, Abel HJ, Hall IM, Flicek P, Stegle O, Gerstein MB, Tubio JMC, Mu Z, Li YI, Shi X,
691 Hastie AR, Ye K, Chong Z, Sanders AD, Zody MC, Talkowski ME, Mills RE, Devine SE, Lee C,
692 Korbelt JO, Marschall T, Eichler EE. Haplotype-resolved diverse human genomes and integrated

693 analysis of structural variation. *Science*. 2021 Apr 2;372(6537):eabf7117. doi:
694 10.1126/science.abf7117.

695 Elipot Y, Hinaux H, Callebert J, Rétaux S. Evolutionary shift from fighting to foraging in blind
696 cavefish through changes in the serotonin network. *Curr Biol*. 2013 Jan 7;23(1):1-10. doi:
697 10.1016/j.cub.2012.10.044.

698 Fumey J, Hinaux H, Noirot C, Thermes C, Rétaux S, Casane D. Evidence for late Pleistocene
699 origin of *Astyanax mexicanus* cavefish. *BMC Evol Biol*. 2018 Apr 18;18(1):43. doi:
700 10.1186/s12862-018-1156-7.

701 Gamboa JL, Andrade FH. Mitochondrial content and distribution changes specific to mouse
702 diaphragm after chronic normobaric hypoxia. *Am J Physiol Regul Integr Comp Physiol*. 2010
703 Mar;298(3):R575-83. doi: 10.1152/ajpregu.00320.2009.

704 Garg P, Martin-Trujillo A, Rodriguez OL, Gies SJ, Hadelia E, Jadhav B, Jain M, Paten B, Sharp
705 AJ. Pervasive cis effects of variation in copy number of large tandem repeats on local DNA
706 methylation and gene expression. *Am J Hum Genet*. 2021 May 6;108(5):809-824. doi:
707 10.1016/j.ajhg.2021.03.016.

708 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for
709 Biotechnology Information; [1988] – . Gene ID: 440243, GOLGA6L22 golgin A6 family like 22 [
710 Homo sapiens (human)]; [cited 2023 07 20]. Available from:
711 <https://www.ncbi.nlm.nih.gov/gene/440243/#gene-expression>

712 Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for
713 Biotechnology Information; [1988] – Gene ID: 80833, APOL3 apolipoprotein L3 [Homo sapiens
714 (human)]; [cited 2023 07 20]. Available from: <https://www.ncbi.nlm.nih.gov/gene/80833>

715 Gross JB. The complex origin of *Astyanax* cavefish. *BMC Evol Biol*. 2012 Jun 30;12:105. doi:
716 10.1186/1471-2148-12-105.

717 Gutsaeva DR, Carraway MS, Suliman HB, Demchenko IT, Shitara H, Yonekawa H, Piantadosi
718 CA. Transient hypoxia stimulates mitochondrial biogenesis in brain subcortex by a neuronal nitric
719 oxide synthase-dependent mechanism. *J Neurosci*. 2008 Feb 27;28(9):2015-24. doi:
720 10.1523/JNEUROSCI.5654-07.2008.

721 Herman A, Brandvain Y, Weagley J, Jeffery WR, Keene AC, Kono TJY, Bilandžija H, Borowsky
722 R, Espinasa L, O'Quin K, Ornelas-García CP, Yoshizawa M, Carlson B, Maldonado E, Gross JB,
723 Cartwright RA, Rohner N, Warren WC, McGaugh SE. The role of gene flow in rapid and repeated
724 evolution of cave-related traits in Mexican tetra, *Astyanax mexicanus*. *Mol Ecol*. 2018
725 Nov;27(22):4397-4416. doi: 10.1111/mec.14877.

726 Hirase S, Ozaki H, Iwasaki W. Parallel selection on gene copy number variations through
727 evolution of three-spined stickleback genomes. *BMC Genomics*. 2014 Aug 29;15(1):735. doi:
728 10.1186/1471-2164-15-735.

729 Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-
730 Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin CS, Korlach J,
731 Wilson RK, Eichler EE. Discovery and genotyping of structural variation from long-read haploid
732 genome sequence data. *Genome Res*. 2017 May;27(5):677-685. doi: 10.1101/gr.214007.116.

733 Ishikawa A, Yamanouchi S, Iwasaki W, Kitano J. Convergent copy number increase of genes
734 associated with freshwater colonization in fishes. *Philos Trans R Soc Lond B Biol Sci*. 2022 Jul
735 18;377(1855):20200509. doi: 10.1098/rstb.2020.0509.

736 Iskow RC, Gokcumen O, Lee C. Exploring the role of copy number variants in human adaptation.
737 *Trends Genet*. 2012 Jun;28(6):245-57. doi: 10.1016/j.tig.2012.03.002.

738 Jaggard JB, Stahl BA, Lloyd E, Prober DA, Duboue ER, Keene AC. Hypocretin underlies the
739 evolution of sleep loss in the Mexican cavefish. *Elife*. 2018 Feb 6;7:e32637. doi:
740 10.7554/eLife.32637.

741 Jakt LM, Dubin A, Johansen SD. Intron size minimisation in teleosts. *BMC Genomics*. 2022 Sep
742 1;23(1):628. doi: 10.1186/s12864-022-08760-w.

743 Jang J, Terefe E, Kim K, Lee YH, Belay G, Tijjani A, Han JL, Hanotte O, Kim H. Population
744 differentiated copy number variation of *Bos taurus*, *Bos indicus* and their African hybrids. *BMC*
745 *Genomics*. 2021 Jul 12;22(1):531. doi: 10.1186/s12864-021-07808-7.

746 Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody
747 MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT,
748 Summers BR, Knecht AK, Brady SD, Zhang H, Pollen AA, Howes T, Amemiya C; Broad Institute
749 Genome Sequencing Platform & Whole Genome Assembly Team; Baldwin J, Bloom T, Jaffe DB,
750 Nicol R, Wilkinson J, Lander ES, Di Palma F, Lindblad-Toh K, Kingsley DM. The genomic basis of
751 adaptive evolution in threespine sticklebacks. *Nature*. 2012 Apr 4;484(7392):55-61. doi:
752 10.1038/nature10944.

753 Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. Selection in the evolution of gene duplications.
754 *Genome Biol*. 2002;3(2):RESEARCH0008. doi: 10.1186/gb-2002-3-2-research0008.

755 Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing
756 environment. *Proc Biol Sci*. 2012 Dec 22;279(1749):5048-57. doi: 10.1098/rspb.2012.1108.

757 Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of
758 structural variation detection algorithms for whole genome sequencing. *Genome Biol*. 2019 Jun
759 3;20(1):117. doi: 10.1186/s13059-019-1720-5.

760 Kowalko JE, Rohner N, Rompani SB, Peterson BK, Linden TA, Yoshizawa M, Kay EH, Weber J,
761 Hoekstra HE, Jeffery WR, Borowsky R, Tabin CJ. Loss of schooling behavior in cavefish through
762 sight-dependent and sight-independent mechanisms. *Curr Biol*. 2013 Oct 7;23(19):1874-83. doi:
763 10.1016/j.cub.2013.07.056.

764 Lai YT, Yeung CKL, Omland KE, Pang EL, Hao Y, Liao BY, Cao HF, Zhang BW, Yeh CF, Hung
765 CM, Hung HY, Yang MY, Liang W, Hsu YC, Yao CT, Dong L, Lin K, Li SH. Standing genetic
766 variation as the predominant source for adaptation of a songbird. *Proc Natl Acad Sci U S A*. 2019
767 Feb 5;116(6):2152-2157. doi: 10.1073/pnas.1813597116.

768 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012 Mar
769 4;9(4):357-9. doi: 10.1038/nmeth.1923.

770 Legendre L, Rode J, Germon I, Pavie M, Quiviger C, Policarpo M, Leclercq J, Père S, Fumey J,
771 Hyacinthe C, Ornelas-García P, Espinasa L, Rétaux S, Casane D. Genetic identification and
772 reiterated captures suggest that the *Astyanax mexicanus* El Pachón cavefish population is closed
773 and declining. *Zool Res*. 2023 Jul 18;44(4):701-711. doi: 10.24272/j.issn.2095-8137.2022.481.

774 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R;
775 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and
776 SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352.

777 Lowe CB, Sanchez-Luege N, Howes TR, Brady SD, Daugherty RR, Jones FC, Bell MA, Kingsley
778 DM. Detecting differential copy number variation between groups of samples. *Genome Res*. 2018
779 Feb;28(2):256-265. doi: 10.1101/gr.206938.116.

780 Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ, Buckler ES, Coluccio AE,
781 Danilova TV, Kudrna D, Magalhaes JV, Piñeros MA, Schatz MC, Wing RA, Kochian LV.
782 Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc Natl*
783 *Acad Sci U S A*. 2013 Mar 26;110(13):5241-6. doi: 10.1073/pnas.1220766110.

784 Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
785 *EMBnet. J*. 2011;17:10–12. doi: 10.14806/ej.17.1.200.

786 Meng G, Bao Q, Ma X, Chu M, Huang C, Guo X, Liang C, Yan P. Analysis of Copy Number
787 Variation in the Whole Genome of Normal-Haired and Long-Haired Tianzhu White Yaks. *Genes*
788 (Basel). 2022 Dec 18;13(12):2405. doi: 10.3390/genes13122405.

789 Menuet A, Alunni A, Joly JS, Jeffery WR, Rétaux S. Expanded expression of Sonic Hedgehog in
790 *Astyanax cavefish*: multiple consequences on forebrain development and evolution.
791 *Development*. 2007 Mar;134(5):845-55. doi: 10.1242/dev.02780.

792 Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S,
793 Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahsen S, Köster J. Sustainable
794 data analysis with Snakemake. *F1000Res*. 2021 Jan 18;10:33. doi:
795 10.12688/f1000research.29032.2.

796 Moran D, Softley R, Warrant EJ. The energetic cost of vision and the evolution of eyeless
797 Mexican cavefish. *Science Advances*. 2015;1:e1500363. doi: 10.1126/sciadv.1500363.

798 Niederreither K, Vermot J, Fraulob V, Chambon P, Dolle P. Retinaldehyde dehydrogenase 2
799 (RALDH2)- independent patterns of retinoic acid synthesis in the mouse embryo. *Proc Natl Acad
800 Sci U S A*. 2002 Dec 10;99(25):16111-6. doi: 10.1073/pnas.252626599.

801 Open2C, Nezar Abdennur, Geoffrey Fudenberg, Ilya Flyamer, Aleksandra A. Galitsyna, Anton
802 Goloborodko, Maxim Imakaev, Sergey V. Venev: Bioframe: Operations on Genomic Intervals in
803 Pandas Dataframes. *bioRxiv*. Preprint. 2022.02.16.480748; doi:
804 <https://doi.org/10.1101/2022.02.16.480748>

805 Peuß R, Box AC, Chen S, Wang Y, Tsuchiya D, Persons JL, Kenzior A, Maldonado E, Krishnan
806 J, Scharsack JP, Slaughter BD, Rohner N. Adaptation to low parasite abundance affects immune
807 investment and immunopathological responses of cavefish. *Nat Ecol Evol*. 2020 Oct;4(10):1416-
808 1430. doi: 10.1038/s41559-020-1234-2.

809 Pezer Ž, Harr B, Teschke M, Babiker H, Tautz D. Divergence patterns of genic copy number
810 variation in natural populations of the house mouse (*Mus musculus domesticus*) reveal three
811 conserved genes with major population-specific expansions. *Genome Res*. 2015 Aug;25(8):1114-
812 24. doi: 10.1101/gr.187187.114.

813 Riddle MR, Aspiras AC, Gaudenz K, Peuß R, Sung JY, Martineau B, Peavey M, Box AC, Tabin
814 JA, McGaugh S, Borowsky R, Tabin CJ, Rohner N. Insulin resistance in cavefish as an
815 adaptation to a nutrient-limited environment. *Nature*. 2018 Mar 29;555(7698):647-651. doi:
816 10.1038/nature26136.

817 Rinker DC, Specian NK, Zhao S, Gibbons JG. Polar bear evolution is marked by rapid changes in
818 gene copy number in response to dietary shift. *Proc Natl Acad Sci U S A*. 2019 Jul
819 2;116(27):13446-13451. doi: 10.1073/pnas.1901093116.

820 Rundle HD, Nagel L, Wenrick Boughman J, Schluter D. Natural selection and parallel speciation
821 in sympatric sticklebacks. *Science*. 2000 Jan 14;287(5451):306-8. doi:
822 10.1126/science.287.5451.306.

823 Saitou M, Masuda N, Gokcumen O. Similarity-Based Analysis of Allele Frequency Distribution
824 among Multiple Populations Identifies Adaptive Genomic Structural Variants. *Mol Biol Evol*. 2022
825 Mar 2;39(3):msab313. doi: 10.1093/molbev/msab313.

826 Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, Collins FH. Molecular
827 evolutionary analysis of the widespread piggyBac transposon family and related "domesticated"
828 sequences. *Mol Genet Genomics*. 2003 Nov;270(2):173-80. doi: 10.1007/s00438-003-0909-0.

829 Shebanits K, Günther T, Johansson ACV, Maqbool K, Feuk L, Jakobsson M, Larhammar D. Copy
830 number determination of the gene for the human pancreatic polypeptide receptor NPY4R using
831 read depth analysis and droplet digital PCR. *BMC Biotechnol*. 2019 Jun 4;19(1):31. doi:
832 10.1186/s12896-019-0523-9.

833 Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web
834 server for functional enrichment analysis and functional annotation of gene lists (2021 update).
835 *Nucleic Acids Res*. 2022 Mar 23;50(W1):W216–21. doi: 10.1093/nar/gkac194.

836 Singh G, Davenport AP. Neuropeptide B and W: neurotransmitters in an emerging G-protein-
837 coupled receptor system. *Br J Pharmacol.* 2006 Aug;148(8):1033-41. doi:
838 10.1038/sj.bjp.0706825.

839 Solé M, Ablondi M, Binzer-Panchal A, Velie BD, Hollfelder N, Buys N, Ducro BJ, François L,
840 Janssens S, Schurink A, Viklund Å, Eriksson S, Isaksson A, Kultima H, Mikko S, Lindgren G.
841 Inter- and intra-breed genome-wide copy number diversity in a large cohort of European equine
842 breeds. *BMC Genomics.* 2019 Oct 22;20(1):759. doi: 10.1186/s12864-019-6141-z.

843 Suvakov M, Panda A, Diesh C, Holmes I, Abyzov A. CNVpytor: a tool for copy number variation
844 detection and analysis from read depth and allele imbalance in whole-genome sequencing.
845 *Gigascience.* 2021 Nov 18;10(11):giab074. doi: 10.1093/gigascience/giab074.

846 tan Nguyen H, Merriman TR, Black MA. CNVrd, a read-depth algorithm for assigning copy-
847 number at the FCGR locus: population-specific tagging of copy number variation at FCGR3B.
848 *PLoS One.* 2013 Apr 30;8(4):e63219. doi: 10.1371/journal.pone.0063219.

849 Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf
850 C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigartyo CA, Odeberg J,
851 Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg
852 J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F,
853 Zwahlen M, von Heijne G, Nielsen J, Pontén F. Proteomics. Tissue-based map of the human
854 proteome. *Science.* 2015 Jan 23;347(6220):1260419. doi: 10.1126/science.1260419.

855 van der Weele CM, Jeffery WR. Cavefish cope with environmental hypoxia by developing more
856 erythrocytes and overexpression of hypoxia-inducible genes. *Elife.* 2022 Jan 5;11:e69109. doi:
857 10.7554/eLife.69109.

858 Vickrey AI, Bruders R, Kronenberg Z, Mackey E, Bohlender RJ, Maclary ET, Maynez R, Osborne
859 EJ, Johnson KP, Huff CD, Yandell M, Shapiro MD. Introgression of regulatory alleles and a
860 missense coding mutation drive plumage pattern diversity in the rock pigeon. *Elife.* 2018 Jul
861 17;7:e34803. doi: 10.7554/eLife.34803.

862 Warren WC, Boggs TE, Borowsky R, Carlson BM, Ferrufino E, Gross JB, Hillier L, Hu Z, Keene
863 AC, Kenzior A, Kowalko JE, Tomlinson C, Kremitzki M, Lemieux ME, Graves-Lindsay T,
864 McGaugh SE, Miller JT, Mommersteeg MTM, Moran RL, Peuß R, Rice ES, Riddle MR, Sifuentes-
865 Romero I, Stanhope BA, Tabin CJ, Thakur S, Yamamoto Y, Rohner N. A chromosome-level
866 genome of *Astyanax mexicanus* surface fish for comparing population-specific genetic differences
867 contributing to trait evolution. *Nat Commun.* 2021 Mar 4;12(1):1447. doi: 10.1038/s41467-021-
868 21733-z.

869 Warren WC, Carroll RA, Haggerty L, Keene AC, McGaugh SE, Ogeh D, Rice ES, Roback E,
870 Rohner N, Martin F, Maggs X. *Astyanax mexicanus* surface and cavefish chromosome-scale
871 assemblies for trait variation discovery. *bioRxiv* 2023.11.16.567450; doi:
872 <https://doi.org/10.1101/2023.11.16.567450>

873 Wozniak D, Quinnell T. Unmet needs of patients with narcolepsy: perspectives on emerging
874 treatment options. *Nat Sci Sleep.* 2015;7:51-61. <https://doi.org/10.2147/NSS.S56077>

875 Xiong S, Krishnan J, Peuß R, Rohner N. Early adipogenesis contributes to excess fat
876 accumulation in cave populations of *Astyanax mexicanus*. *Dev Biol.* 2018 Sep 15;441(2):297-304.
877 doi: 10.1016/j.ydbio.2018.06.003.

878 Xu L, Hou Y, Bickhart DM, Zhou Y, Hay el HA, Song J, Sonstegard TS, Van Tassell CP, Liu GE.
879 Population-genetic properties of differentiated copy number variations in cattle. *Sci Rep.* 2016
880 Mar 23;6:23161. doi: 10.1038/srep23161.

881 Yang L, Han J, Deng T, Li F, Han X, Xia H, Quan F, Hua G, Yang L, Zhou Y. Comparative
882 analyses of copy number variations between swamp buffaloes and river buffaloes. *Anim Genet.*
883 2023 Apr;54(2):199-206. doi: 10.1111/age.13288.

884 Yi G, Qu L, Liu J, Yan Y, Xu G, Yang N. Genome-wide patterns of copy number variation in the
885 diversified chicken genomes using next-generation sequencing. *BMC Genomics*. 2014 Nov
886 7;15(1):962. doi: 10.1186/1471-2164-15-962.

887 Yoshizawa M, Goricki S, Soares D, Jeffery WR. Evolution of a behavioral shift mediated by
888 superficial neuromasts helps cavefish find food in darkness. *Curr Biol*. 2010 Sep 28;20(18):1631-
889 6. doi: 10.1016/j.cub.2010.07.017.

890 Yuste-Lisbona FJ, Fernández-Lozano A, Pineda B, Bretones S, Ortíz-Atienza A, García-Sogo B,
891 Müller NA, Angosto T, Capel J, Moreno V, Jiménez-Gómez JM, Lozano R. ENO regulates tomato
892 fruit size through the floral meristem development network. *Proc Natl Acad Sci U S A*. 2020 Apr
893 7;117(14):8187-8195. doi: 10.1073/pnas.1913688117.

894 Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and
895 evolution. *Annu Rev Genomics Hum Genet*. 2009;10:451-81. doi:
896 10.1146/annurev.genom.9.081307.164217.

897 Zhu C, Li M, Qin S, Zhao F, Fang S. Detection of copy number variation and selection signatures
898 on the X chromosome in Chinese indigenous sheep with different types of tail. *Asian-Australas J*
899 *Anim Sci*. 2020 Sep;33(9):1378-1386. doi: 10.5713/ajas.18.0661.

900 Zong SB, Li YL, Liu JX. Genomic Architecture of Rapid Parallel Adaptation to Fresh Water in a
901 Wild Fish. *Mol Biol Evol*. 2021 Apr 13;38(4):1317-1329. doi: 10.1093/molbev/msaa290.

902

903 Author Contributions

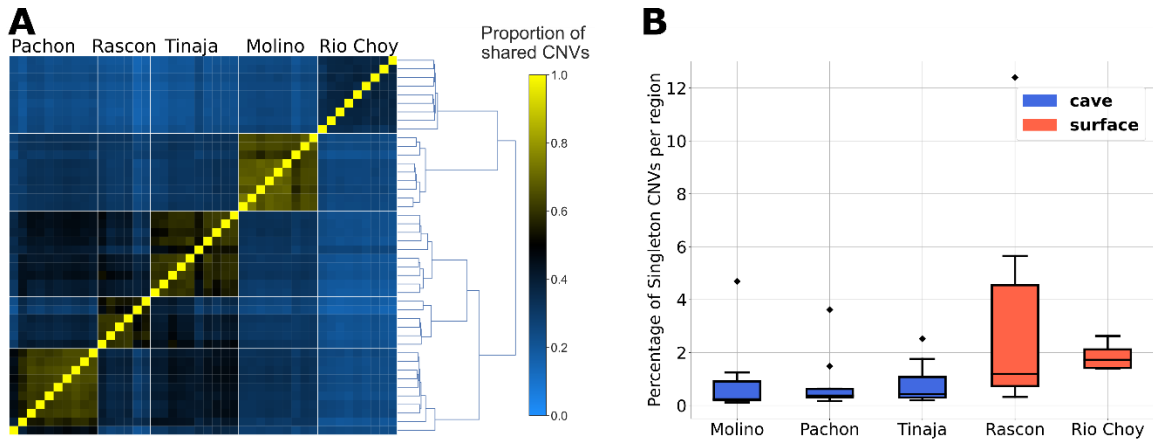
904 ZP designed the research. NR contributed the data. IP and ZP performed the research and
905 analyzed the data. ZP wrote the manuscript draft. All authors contributed to the final manuscript.

906

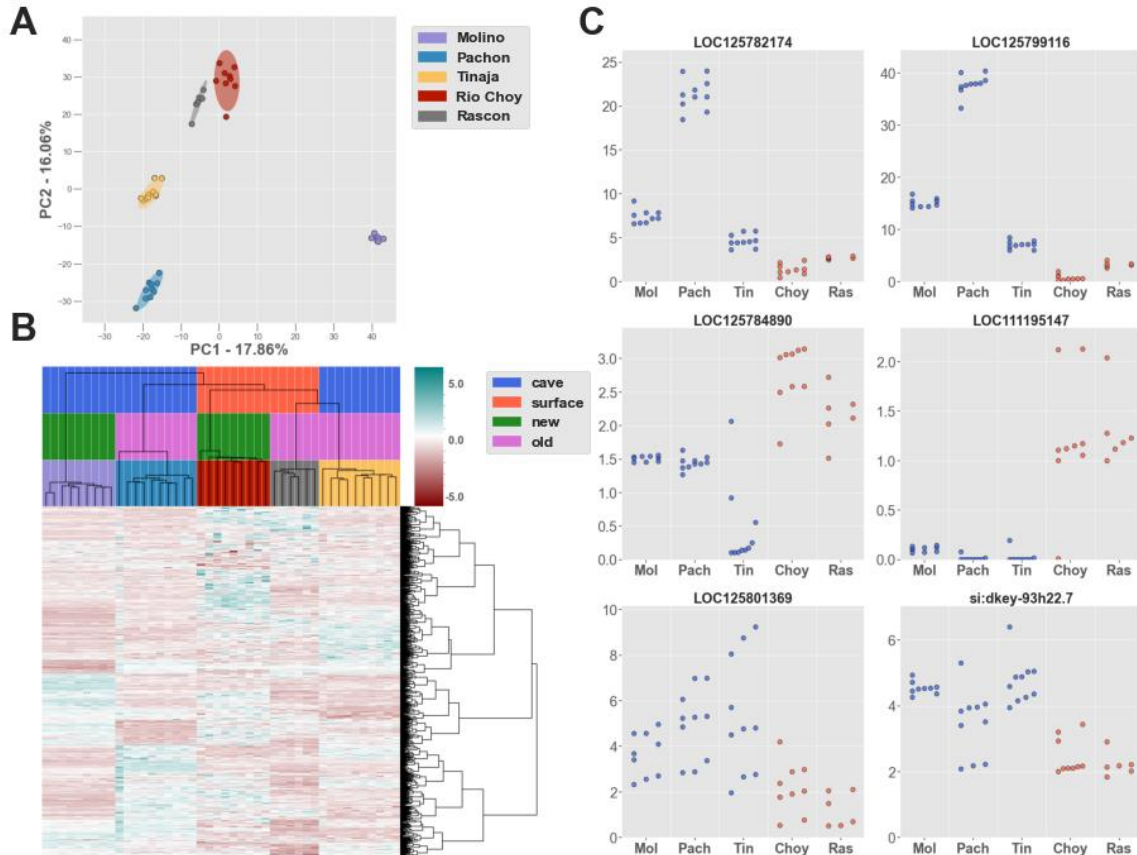
907 Data Accessibility and Benefit-Sharing

908 The data that supports the findings of this study are available in the supplementary material of
909 this article. Benefits from this research accrue from the sharing of our data and results.

910 Figures and Tables
911
912
913
914

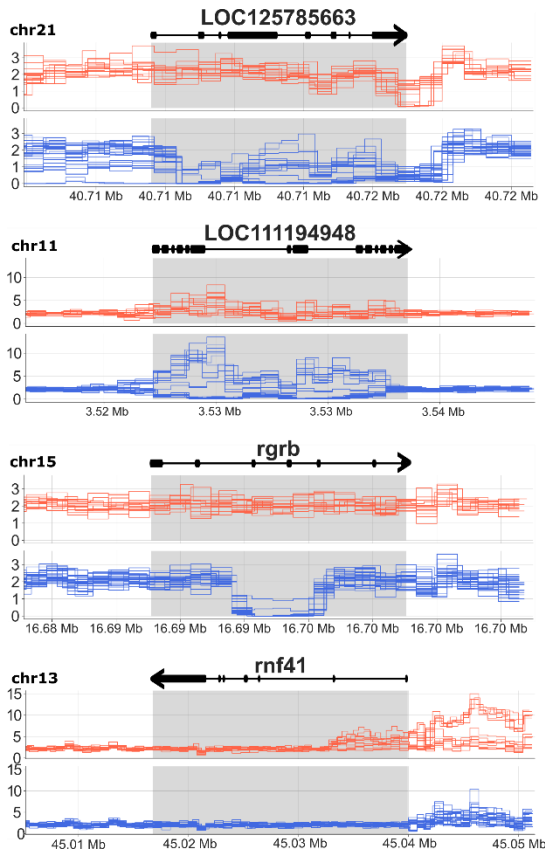


915
916
917
918 Figure 1. Genetic diversity analysis based on shared and singleton CNVs. A) Distance matrix
919 based on the average number of shared CNVs between two genomes. Samples are shown in the
920 same order from right to left as from top to bottom. Distance matrix was subjected to Ward's
921 method of hierarchical clustering. B) Distribution of singleton CNVs proportions by population.



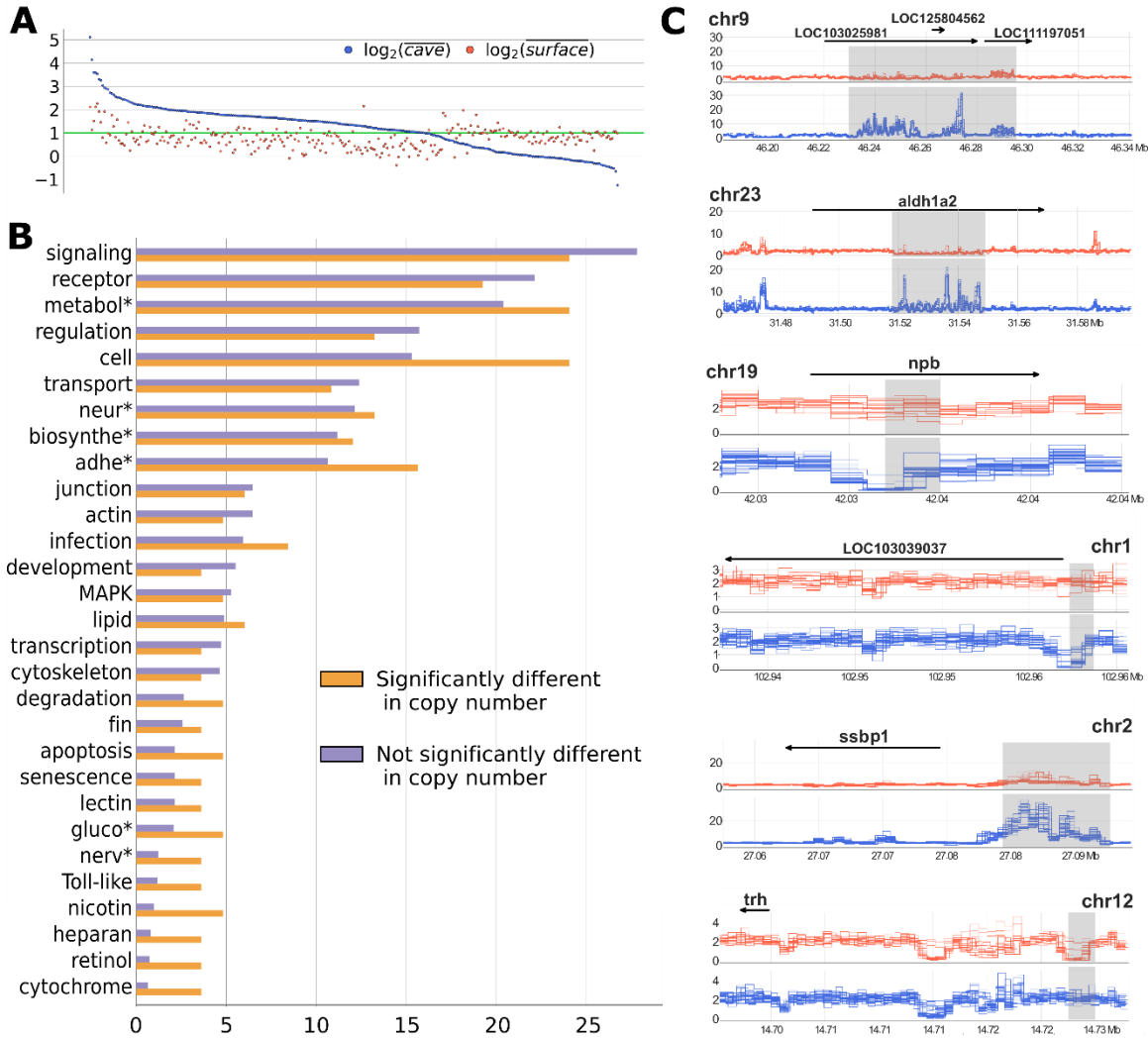
922
 923
 924
 925
 926
 927
 928
 929
 930
 931
 932
 933
 934
 935

Figure 2. Population specific pattern of gene copy number variation. A) Plot of the first two principal components of PCA based on copy numbers of 2,819 CNV genes. Confidence (95%) ellipses are shown around sample clusters. B) Heatmap of normalized copy numbers of the 2,819 CNV genes with hierarchical clustering based on genes (rows) and samples (columns) using Ward's method. Values are normalized by row (gene) such that the mean of every row is 0 and its standard deviation is 1. The heatmap visually represents relative differences in copy number of a particular gene between samples, ranging from lowest (red) to highest (green) values. The samples are colored as in A) by population, and by ecotype and lineage as indicated in the legend on the right. C) Copy numbers of several genes with significant differences between cave (blue) and surface (orange) individuals. Gene identifier is indicated on top of each plot. Mol - Molino; Pach - Pachón; Tin - Tinaja; Choy - Río Choy; Ras - Rascon.



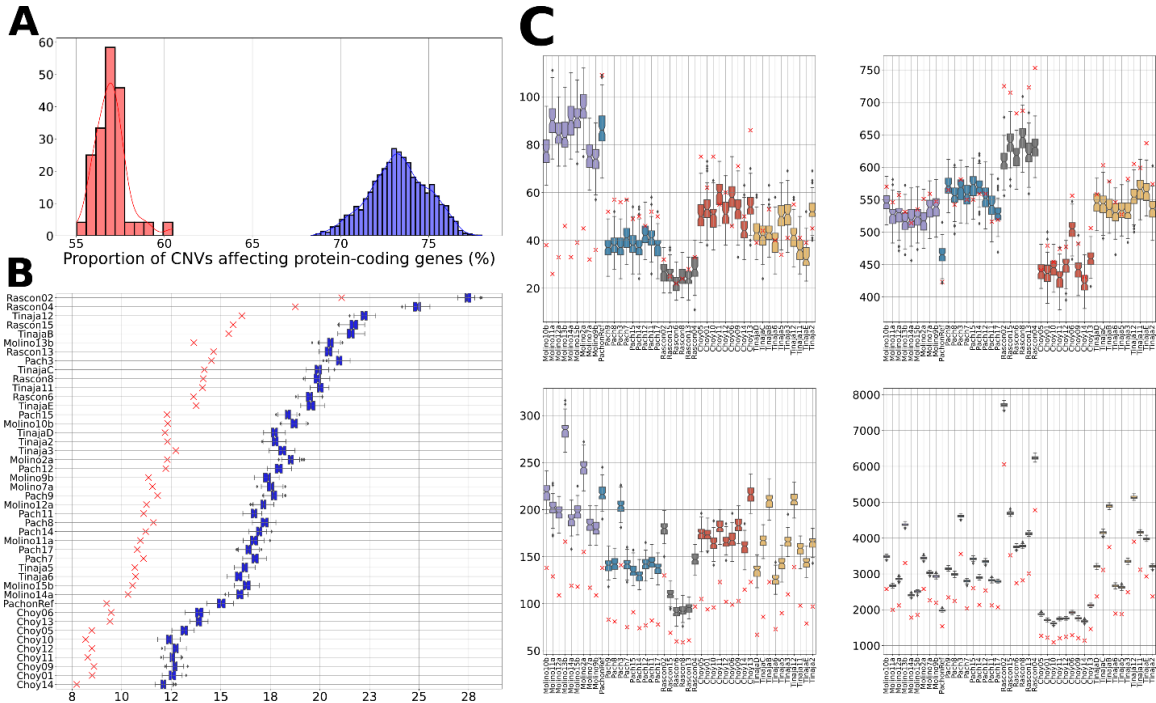
936
 937
 938
 939
 940
 941
 942

Figure 3. Read depth in the regions of several ecotype-divergent CNVs at genes. Read depth is shown per bin for cave (blue) and surface (orange) fish, as calculated by CNVpytor and normalized to represent the copy number per diploid. Gene structure and orientation are shown by arrows on top of which gene symbols are indicated. Gene position is highlighted in gray.



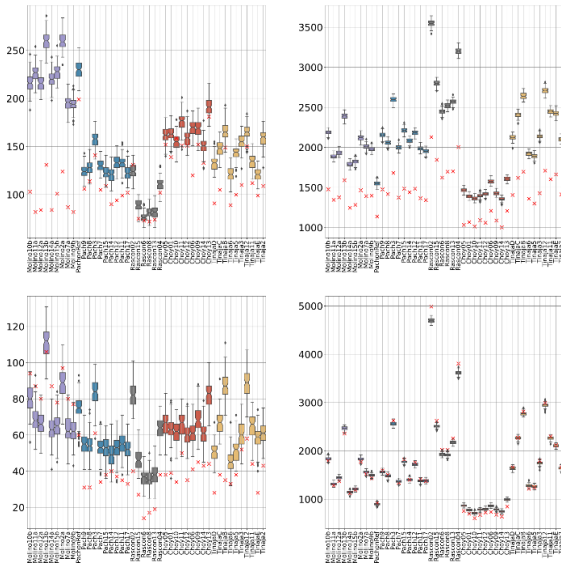
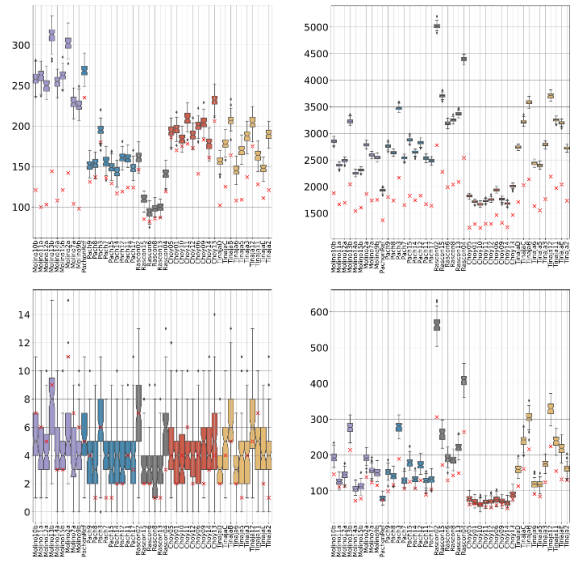
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962

Figure 4. Genomic regions divergent in copy number between surface and cave fish. A) Genomic regions that are significantly different in their copy number between the two ecotypes. All 292 CNVRs are shown on the X axis, where each position represents one CNVR. Every CNVR is genotyped with CNVpytor and the average CN per individual is shown as blue and orange dots for cave and surface fish, respectively. CNVRs are ordered by average copy number in cave fish from highest on the left to the lowest on the right. Copy numbers are shown as \log_2 -transformed values on y-axis; the green horizontal line depicts value of two copies per diploid. B) The most frequently occurring terms associated with genes that overlap CNVRs, analyzed separately for regions that significantly differ (orange) or do not differ (blue) in copy number between cave and surface fish. Terms are derived from biological process and pathway annotations in DAVID tool. Complete list of terms associated with words marked by asterisks is given in Table S8. C) Proportions are shown as percentages of the number of genes associated with a term relative to the total number of genes that have assigned annotations in DAVID. C) Read depth around several CNVRs (highlighted in gray) with divergent copy numbers between ecotypes. Gene positions and orientations are shown by arrows on top of which gene symbols are indicated. Read depth is plotted per bin for cave (blue) and surface (orange) fish, as calculated by CNVpytor and normalized to represent the copy number per diploid.



963
964
965
966
967
968
969
970
971
972
973

Figure 5. A) Distribution of percentage of CNVs affecting protein-coding genes across all 44 analyzed individuals, for the true (red) and permuted (blue) data. B) Proportion of protein-coding genes affected by CNV calls in true data (red crosses) and the distribution of expected proportions based on permuted data (blue boxplots). Data is shown by individuals, indicated by sample names on the left. C) Number of duplications (left plots) and deletions (right plots) that affect protein-coding genes entirely (upper plots) or partially (lower plots). Values are indicated per individual, as red crosses for true data and as boxplots for permuted data, representing the distributions of values based on 100 permutations. Boxplots are colored by population.

A**B**

974
 975
 976
 977
 978
 979
 980

Figure 6. Number of duplications (left plots) and deletions (right plots) that affect A) introns and B) exons entirely (upper plots) or partially (lower plots). Values are indicated per individual, as red crosses for true data and as boxplots for permuted data. Each boxplot represents a distribution of values based on 100 permutations. Boxplots are colored by population.

981 Table 1. Count of CNV regions that are private to population, ecotype and lineage
 982
 983

Group	Private CNVs	Private CNVRs	CNVRs [†] with single CNV	CNVs per CNVR [†] (average)
Molino	3,577	1,215	551 (45%)	5
Pachón	1,770	807	502 (62%)	4
Tinaja	3,142	1,266	623 (49%)	4
Río Choy	1,264	830	596 (72%)	3
Rascon	5,649	3,527	2,383 (68%)	3
cave	15,016	4,257	1,750 (41%)	5
surface	8,243	4,728	3,056 (65%)	3
new lineage	6,346	2,326	1,197 (52%)	5
old lineage	35,583	8,752	3,633 (42%)	6

984
 985 † Numbers refer to CNVRs that are private to population, ecotype, or lineage, as indicated by the first
 986 column

987 Table 2. Selection on duplications and deletions in gene categories inferred from comparisons of
 988 true and permuted data
 989
 990

Gene category	Category count ²	Deletion complete ¹	Duplication complete ¹	Duplication partial ¹	Deletion partial ¹
protein-coding	26,735	0 †	0†	-	-
pseudogene	1,376	+	0†	0	+
lncRNA	3,603	+	0†	0	0
snoRNA	297	0	0	N/A	N/A
snRNA	1,314	0	0	N/A	N/A
rRNA	9,252	+	0	N/A	N/A
tRNA	9,987	+	0	N/A	N/A

991
 992

993 ¹ neutral variation: “0”; variation under negative selection: “-”; variation under positive selection: “+”; not
 994 applicable: “N/A” (partial overlap cannot be determined due to considerably smaller gene length compared
 995 to bin size used for CNV detection)

996 ² number of annotated instances in the AstMex3 genome assembly

997 † categories with exceptions to the indicated inference (for details, see Figure S5)