Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/121382/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Liu, Zhijin, Wen, Huamei, Hailer, Frank ORCID: <https://orcid.org/0000-0002-2340-1726>, Dong, Fang, Yang, Zuomin, Liu, Tao, Han, Ling, Shi, Fanglei, Hu, Yibo and Zhou, Jiang 2019. Pseudogenization of Mc1r gene associated with transcriptional changes related to melanogenesis explains leucistic phenotypes in *Oreonectes cavefish* (Cypriniformes, Nemacheilidae). *Journal of Zoological Systematics and Evolutionary Research* 57 (4) , pp. 900-909.
10.1111/jzs.12286 filefile

Publishers page: <https://doi.org/10.1111/jzs.12286>
<<https://doi.org/10.1111/jzs.12286>>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies.

See

<http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

1 **Title:** Pseudogenization of *Mc1r* gene associated with transcriptional
2 changes related to melanogenesis explains leucistic phenotypes in
3 *Oreonectes* cavefish (Cypriniformes, Nemacheilidae)

4

5 Zhijin Liu^{1,2,†,*}, Huamei Wen^{1,†}, Frank Hailer^{3,†}, Fang Dong^{2,4,†}, Zuomin Yang⁵, Tao
6 Liu¹, Ling Han¹, Fanglei Shi², Yibo Hu^{2,*}, Jiang Zhou^{1,2,*}

7 ¹ Guizhou Normal University / State Key Laboratory Cultivation for Karst Mountain
8 Ecology Environment of Guizhou Province, Guiyang, Guizhou, 550001, China.

9 ² CAS Key Laboratory of Animal Ecology and Conservation Biology, Institute of
10 Zoology, Chaoyang District, Beijing, 100101, China.

11 ³ School of Biosciences, Cardiff University, Sir Martin Evans Building, Museum
12 Avenue, Cardiff, CF10 3AX, United Kingdom.

13 ⁴ Institute of Physical Science and Information Technology, Anhui University, Hefei,
14 Anhui, 230601, China.

15 ⁵ School of Life Sciences, Qufu Normal University, Qufu, Shandong, 273165, China.

16 † Contributed equally

17 * Correspondence: Jiang Zhou, zhoujiang@ioz.ac.cn; Yibo Hu, huyb@ioz.ac.cn;
18 Zhijin Liu, liuzj@ioz.ac.cn

19

20 Short running title: Leucistic *Oreonectes* cavefish

21 Abstract

22 Organisms that have colonized underground caves encounter vastly different selective
23 pressures than their relatives in above-ground habitats. While disruption of certain
24 pigmentation genes has been documented in various cave-dwelling taxa, little is
25 known about wider impacts across pigmentation and other gene pathways. We here
26 study the timeframe and transcriptional landscape of a leucistic and blind cypriniform
27 fish (*Oreonectes daqikongensis*, Nemacheilidae) that inhabits karst caves in Guizhou,
28 China. Based on data from the mitochondrial *ND4*, *ND5* and *Cytb* genes we show that
29 the divergence between *O. daqikongensis* and its most closely related pigmented
30 species occurred ca. 6.82 million years ago (95% HPD, 5.12–9.01), providing ample
31 time for widespread phenotypic change. Indeed, we found that the DNA sequence of
32 *Mclr* (melanocortin-1 receptor), a key gene regulating the biosynthesis of melanin in
33 most vertebrates, is pseudogenized in *O. daqikongensis*, caused by a 29-bp deletion in
34 the protein-coding region. Furthermore, 99,305 unigenes were annotated based on the
35 transcriptome of skin tissue of *Oreonectes* fish. Among the differentially expressed
36 unigenes, 7,326 (7.4% of the total unigenes) had decreased expression and 2,530
37 (2.5% of the total unigenes) had increased expression in *O. daqikongensis* skin. As
38 predicted, the expression of *Mclr* and 18 additional genes associated with melanin
39 biosynthesis were significantly down-regulated in the skin tissue of *O. daqikongensis*,
40 but not in its congener. Our results, integrating with other studies on cavefishes,
41 suggest that loss of pigmentation was caused by coding region loss-of-function
42 mutations along with widespread transcriptional changes, resulting from extended
43 evolutionary time as a cave-dwelling form.

44 **Keywords:** *Mclr*, *Oreonectes daqikongensis*, cavefish, frameshift mutation, leucism,
45 transcriptome.

46 **Introduction**

47 Uncovering the speciation process and genetic basis of phenotypic adaptations of
48 animals to a specific environment is a key goal in evolutionary and comparative
49 zoology. Underground rivers and caves are a globally widespread ecosystem that have
50 been independently colonized by numerous animals (McGaugh et al., 2014;
51 Bilandžija et al., 2013). Since light is lacking year-round, many animals inhabiting
52 such habitats show degeneration of skin pigmentation (Gross & Wilkens, 2013).
53 However, the timeframes on which cave dwellers have evolved their specialized
54 phenotypes, and the molecular mechanisms underlying such phenotypes, remain
55 largely unknown.

56 The melanocortin-1 receptor (MC1R) gene plays a key role in the regulation of
57 melanin biosynthesis (Majerus & Mundy, 2003). The MC1R receptor protein encoded
58 by *Mclr* gene belongs to the G-protein coupled receptors family, which has seven
59 transmembrane α -helices. α -MSH (α -Melanocyte-Stimulating hormone) binds to and
60 activates MC1R receptor at the membrane of melanophore cells and promotes the
61 biosynthesis of cAMP (cyclic adenosine monophosphate), followed by the synthesis
62 of melanin (Barsh & Cotsarelis, 2007). High levels of basal *Mclr* signaling cause
63 increased expression of the microphthalmia-associated transcription factor (MITF),
64 tyrosinase (TYR), tyrosinase-related protein 1 (TYRP1) and other melanogenesis-
65 related genes, which leads to increased eumelanin synthesis (Nishimura, 2011).

66 *Mclr* has been identified as the key candidate gene for animal pigmentation, and
67 several *Mclr* mutations have been shown to explain the presence of leucistic/light-
68 colored phenotypes (Majerus & Mundy, 2003). Leucism describes a partial loss of
69 pigmentation, with phenotypes such as white or patchily white-colored skin, hair or

70 feathers, but pigmentation still being partially present, e.g. the pigment cells in the
71 eyes which are not affected by the leucistic condition. Albinism, on the other hand, is
72 a phenotype with complete absence of melanin, resulting in albinistic skin,
73 feathers/hair and eyes. Vertebrates with albinism are not only white (or sometimes
74 pale yellowish) in exterior coloration, but also possess pale eyes, often pink or red in
75 color due to blood vessels showing through. In mammals, the leucistic phenotype of
76 *Ursus americanus kermodei* is caused by the substitution of tyrosine by cysteine at the
77 298th amino acid (Y298C) of MC1R (Ritland, Newton & Marshall, 2001). The
78 substitution of arginine by cysteine at 18th amino acid and arginine by tryptophan at
79 the 109th amino acid (R18C and R109W) of MC1R also result in the light-coated
80 phenotype of *Chaetodipus intermedius* (Nachman, Hoekstra & D'Agostino, 2003). In
81 reptiles, the light-colored skins are also associated with mutations of *Mclr* in lizards
82 and geckos (Mendes et al., 2018; Machado et al., 2019). In birds, the leucistic
83 phenotype of the plumage in *Falco rusticolus* is caused by the substitution of valine
84 by isoleucine at the 128th amino acid of MC1R (Zhan et al., 2012; Johnson, Ambers &
85 Burnham, 2012). Furthermore, melanistic plumage polymorphism in lesser snow
86 geese (*Anser caerulescens*) and arctic skuas (*Stercorarius parasiticus*) is tightly
87 associated with sequence variation within the *Mclr* gene (Mundy et al., 2004). In fish,
88 two different *Mclr* alleles (963 and 969 bp long) were found in wild populations of
89 guppies (*Poecilia reticulata*) and individuals with the 963 bp *Mclr* allele tended to
90 show less black pigmentation than those with the 969 bp allele (Ayumi et al., 2011).

91 The genus *Oreonectes* (Cypriniformes, Nemacheilidae) was first described by
92 Günther (1868), with *O. platycephalus* as the type species, and a total of 18 species of
93 *Oreonectes* are currently considered as valid (Deng et al., 2016; Günther, 1868).
94 Besides *O. platycephalus*, *Oreonectes* fish are semi-cave-dwelling or cave-restricted

species, rendering the genus an interesting model to study evolutionary consequences
ocolonization of cave habitats (Deng et al., 2016). One particularly specialized species
is *O. daqikongensis*, which was only recently described (Deng et al., 2016). The
species exhibit numerous typical characteristics of cave dwellers (e.g., lack of eyes),
but is the only leucistic *Oreonectes* fish which lack pigmentation (Fig.1a and b) (Deng
et al., 2016). We here ask the following evolutionary questions: what is the timeframe
on which *O. daqikongensis* evolved the leucistic phenotype? Is the *Mclr* gene
involved in the occurrence of leucism? More broadly, how is the translational
landscape of the melanogenesis pathway changed in this species? To answer these
questions, we first reconstructed a dated phylogeny of five *Oreonectes* species based
on mitochondrial *Cytb*, *ND4* and *ND5* gene sequences. To determine whether the
Mclr gene is functional in *O. daqikongensis*, the locus was sequenced in *O.*
daqikongensis and another three *Oreonectes* species. Finally, to illuminate the
evolutionary impacts of long-term cave-dwelling on *Oreonectes* fishes, determine the
function of other melanogenesis-related genes in *O. daqikongensis*, we further
collected skin tissue samples of *O. daqikongensis* and its close-related species with
pigmentation and performed whole-transcriptome sequencing.

Materials and methods

Sample collection

Fifteen individuals of *O. daqikongensis*, 16 *O. jiarongensis* individuals, six *O.*
dongliangensis individuals and ten *O. shuilongensis* individuals were collected in
Libo and Sando county of Guizhou Province (Fig. 1). These samples have been
deposited in the Museum of Guizhou Normal University (for catalog numbers see
Table 1).

Phylogenetic analyses and divergence time estimation

DNA extraction was performed for all the collected samples according to the manufacturer's instructions (Qiagen Tissue Kit). To infer the phylogenetic relationships of *Oreonectes* species, the mitochondrial DNA genes (*ND4*, *ND5* and *Cyt b*) of each one individual of herein studied four *Oreonectes* species were amplified with primer sets for Nemacheilidae mitogenome (Rui et al. 2012; Wang et al. 2016; Table 2). Mitogenome sequences of the other 53 species in Nemacheilidae, two species in Cobitidae and two species in Balitoridae were obtained from GenBank (Supplementary Table S1). Sequences were aligned using ClustalX (Thomps et al., 1997) and rechecked by eye. MEGA v7.0 (Sudhir et al., 2016) was used to estimate genetic distances between the taxa in *Oreonectes* based on P-distance method. The Bayesian Inference (BI) of phylogenetic reconstruction was employed using MrBayes 3.1.2 (Ronquist & Huelsenbeck, 2003), using species from Cobitidae (*Cobitis sinensis* and *Misgurnus anguillicaudatus*) and Homalopteridae (*Jinshaia sinensis* and *Sinogastromyzon sichangensis*) as outgroups. MODELTEST 3.06 (Posada & Crandall, 1998) was run to determine the appropriate model of sequence evolution in a likelihood ratio test. For Bayesian phylogenetic inference, four Markov chain Monte Carlo (MCMC) runs were performed for 100,000 generations, sampling every ten generations. The initial 5% of trees were discarded as burn-in.

A Bayesian tree for estimating the divergence time was reconstructed in BEAST 1.61 (Drummond & Rambaut, 2007) using the GTR +I+G model, which provided the best fit. Trees were produced based on two independent runs of 10 million MCMC steps each sampling every 2,000th iteration, discarding the initial 25% of values as burn-in. We utilized the calibrated Yule model, as recommended by Heled &

Drummond (2012) for analysis of sequences from different species. We employed calibration points obtained from the dated Nemacheilidae phylogenetic tree by Wang et al. (2016) to (1) place a log normally distributed prior on the age of the root of the tree containing all samples of *Oreonectes* species and outgroups, based on estimated divergence time between Cobitidae and Nemacheilidae at approximately 30 million years ago (Mya; Frickhinger, 1991), and (2) a prior for the origination of the genus *Cobitis* as 13.8–15.9 Mya (Zhou, 1992). An uncorrelated lognormal model of lineage variation with a constant population size tree prior was employed (Drummond et al., 2012). Convergence of two independent MCMC runs was assessed in Tracer v.1.6 (available at <http://tree.bio.ed.ac.uk>), as was convergence of model parameter values, i.e. effective sample size [ESS] values being >200. The tree and posterior distribution were summarized with TreeAnnotator v.1.5.3 and visualized by FigTree v.1.4.3 (available at <http://tree.bio.ed.ac.uk>).

DNA extraction, primer design, polymerase chain reaction and DNA sequencing of *Mc1r* gene

The candidate *Mc1r* gene was amplified by polymerase chain reaction (PCR). Primers used in this study are MC1R-F (5'-GAATATCAGAGGTGTGCTGAAGC-3') and MC1R-R (5'-TCCTTGAGAGTCTTGCGCAG-3') which were designed based on alignments of the flanking regions of *Mc1r* in *Astyanax mexicanus*, *Carassius auratus* and *Triplophysa rosa*. The *Mc1r* coding region was amplified for two individuals each of *O. jiarongensis*, *O. dongliangensis* and *O. shuilingensis*, and three individuals of *O. daqingensis*. PCR was carried out in 50 µl reaction mixtures containing 50 mM KCl, 10 mM Tris-HCl, 1.5 mM Mg²⁺, 200 µmol of each dNTP, 0.2 µmol of each primer, 1.5 U Hotstart Taq DNA polymerase (QIAGEN), 1 µg/µL BSA and ≤10 ng of genomic DNA. The PCR cycling conditions were as follows: initial denaturation at

168 95°C for 5 min, followed by 32 cycles each of denaturation (95°C for 30 sec),
169 annealing (at 58°C for 30 sec), and elongation (at 72°C for 60 sec) with a final
170 elongation step at 72°C for 10 min. The PCR products were examined for specificity
171 following electrophoresis on 1.5% agarose gels, then sequenced on an ABI 3130 xl
172 DNA Sequencer (Applied Biosystems).

173 **DNA sequence analysis of the *Mc1r* gene**

174 DNA sequencing data was read and assembled by SeqMan II (DNASTAR Lasergene,
175 version 6). The open reading frame (ORF) analysis and amino acid sequence
176 inference was completed by using the search ORF program of EditSeq 6.1
177 (DNASTAR Lasergene, version 6.1). Clustalw2
178 (<http://www.ebi.ac.uk/Tools/msa/clustalo>) was applied to perform nucleotide
179 alignment. The protein transmembrane structure was predicted and functionally
180 analyzed using TMHMM v.2.0 (<http://www.cbs.dtu.dk/services/TMHMM/>), and
181 MODELLER (<https://swissmodel.expasy.org/>; Eswar et al., 2008) was used to
182 construct the three-dimensional structure of the MC1R protein.

183 **Total RNA isolation, library preparation, and sequencing**

184 Skin tissues were isolated from the rear of three *O. daqikongensis* and three *O.*
185 *jiarongensis* individuals immediately after capture in the wild. The isolated tissues
186 were immediately put into liquid nitrogen, frozen and stored at -80°C. Total RNA was
187 isolated from each sample of tissue using TRIzol reagent (Life Technologies, CA, US)
188 according to the manufacturer's instructions. The purity and content of each RNA
189 sample were measured using the Qubit RNA Assay Kit in a Qubit 2.0 Fluorometer
190 (Life Technologies, CA, USA) and confirmed by running total RNA samples on 1 %
191 agarose gels.

One paired-end (PE) cDNA library was generated for each sample. Illumina sequencing was conducted at Beijing Novogene Biological Information Technology Co., Ltd., Beijing, China (<http://www.novogene.com/>) using the Illumina TruSeqTM RNA Sample Preparation Kit (Illumina, San Diego, CA, USA) following the manufacturer's recommendations (Kircher Stenzel & Kelso, 2009). Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under an elevated temperature in the proprietary Illumina fragmentation buffer. First-strand cDNA was synthesized using random oligonucleotides and Super Script II (Life Technologies). Second-strand cDNA synthesis was subsequently performed using DNA polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities, and enzymes were then removed. After adenylation of 3' ends of DNA fragments, Illumina PE adapter oligo nucleotides were ligated to prepare for hybridization. To preferentially select cDNA fragments of 300 bp in length, the library fragments were purified with the AMPure XP system (Beckman Coulter, Beverly, MA, USA). DNA fragments with ligated adaptor molecules on both ends were selectively enriched using Illumina PCR Primer Cocktail in a 10-cycle PCR (polymerase chain reaction). Products were purified (AMPure XP system) and quantified using the Agilent High-Sensitivity DNA assay on the Agilent Bioanalyzer 2100 system (Agilent Technologies Co. Ltd., Beijing, China). Clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq 4000 platform and 2x150 bp paired-end reads were generated.

The assembly of transcriptome data, gene annotation and analysis

Raw sequence data reads in FASTA format were first processed using in-house Perl scripts. In this step, clean reads were obtained by removing adapters, poly-N, and low-quality sequences from the raw read data. To generate a common and nonredundant *Oreonectes* transcriptome database for further RNA-seq analyses, transcriptome assembly was carried out based on clean data from *O. jiarongensis* using Trinity software (v.2.0.6), with all parameters set as default (Grabherr et al., 2011). The resulting transcriptome assembly was trimmed with a custom perl script to remove contigs under 500 base pairs (Maytin et al., 2018). Before annotation, unigenes were picked from the transcriptome with CD-hit (Li & Godzik, 2006). Intactness of the assembled *O. shuilingensis* transcriptome was assessed with the software tool BUSCO (Benchmarking Universal Single-Copy Orthologs) that is based on evolutionarily informed expectations of gene content, with default settings (Simão et al., 2015; Waterhouse et al. 2018). The unigenes were then annotated with BLASTx (BLAST + v.2.2.25) by querying these to the following databases: NCBI non-redundant protein sequences (Nr), NCBI non-redundant nucleotide sequences (Nt), the Protein Family database (Pfam), Swiss-Prot, Gene Ontology (GO), the Eukaryotic Orthologous Groups database (KOG) and the Kyoto Encyclopedia of Genes and Genomes (KEGG). The E-value cutoff was set at 1×10^{-5} .

Gene expression quantification and identification of differentially expressed genes (DEGs)

We mapped trimmed Illumina reads from the skin tissue from *O. daikongensis* and *O. jiarongensis* to *Oreonectes* transcriptome using Bowtie2 with default parameters (Langmead, 2010). Gene expression levels were estimated by RSEM (v.1.2.15) for each sample (Li & Dewey, 2011). First, the read count for each gene was obtained from the result of clean data mapped back onto the assembled transcriptome

reference. Subsequently, the read count of each sequenced library was adjusted with the edgeR program package to ‘fragments per kb per million mapped (FPKM). The edgeR package (<http://bioconductor.org/packages/release/bioc/html/edgeR.html>) was used to obtain the “base mean” value for identifying DEGs. A false discovery rate (FDR) ≤ 0.05 and the absolute value of $\log_2$ ratio ≥ 1 (two-fold change difference) were set as the thresholds for the significance of the gene expression difference between two groups. In addition, information for DEGs was collected from unigenes annotations, and these genes were subjected to GO and KEGG significant enrichment analyses to identify biological functions and metabolic pathways in which these genes participate. Furthermore, we collected information for candidate genes involved in melanogenesis.

Results

Phylogenetic analysis and divergence time estimation

To infer the phylogenetic relationships of *Oreonectes* species, the phylogeny of five *Oreonectes* species and other 53 species in Nemacheilidae was reconstructed using two species in Cobitidae and two species in Balitoridae as outgroups. Five mitochondrial fragments were amplified for each one individual of herein studied four *Oreonectes* species and the continuous mitochondrial sequence were aligned, including the complete sequences of *ND4* (1,382 bp), *ND5* (1,837 bp) and *Cytb* (1,141 bp) (Table 2). In the concatenated alignment (*ND4*: 1st -1,382nd bp, *ND5*: 1,383rd - 3,219th bp, *Cytb*: 3,220th -4,360th bp) that comprised 4,360 bp across 61 species, we found 2290 variable sites out of the alignment (Supplementary data 1). The genetic distances among *Oreonectes* species ranged from 0.010 to 0.149 (Table 3). GTR+I+G model was the best fitting model for each of these three mitochondrial genes

individually, and the same substitution model was also identified as the best-fitting model for the concatenated sequence, with a gamma shape parameter of 1.0419, a transition/transversion ratio of 6.23 and base frequencies of A =0.3226, C= 0.2148, G= 0.1366 and T= 0.3260. Phylogenetic analyses resulted in virtually identical tree topologies (Fig. 1d), with a basal divergence among *Oreonectes* species at ca. 17.53 Mya (95% HPD, 13.93–22.13 Mya), and a split between *O. daqikongensis* and its most closely-related species (*O. dongliangensis* and *O. jiarongensis*) at 6.82 Mya (95% HPD, 5.12–9.01 Mya; Fig. 1d).

Pseudogenization of *Mclr* gene in *O. daqikongensis*

The amplicon length of *Mclr* gene ranged from 971–1,004 bp. The protein coding region of the *Mclr* gene of *O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis* was 988 bp, while it was only 955bp in *O. daqikongensis* (Supplementary Fig. S1 and Supplementary data 2). The *Mclr* sequences from 15 *O. daqikongensis* individuals all had a 29 bp deletion (starting at nucleotide position 236 of the alignment), a six bp insertion (starting at nucleotide position 819) and a nine bp deletion (starting at nucleotide position 867). The first 29 bp deletion caused a frame-shift mutation of *Mclr* gene of *O. daqikongensis*, resulting in a premature STOP codon (TGA; starting at nucleotide position 279–281), while *Mclr* sequences were intact in *O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis* (Fig. 1e). Thus, *Mclr* is pseudogenized in *O. daqikongensis*, showing truncation of the protein within the intracellular loop. The MC1R receptor protein sequence of *O. daqikongensis* was predicted to be incapable of generating a functional transmembrane domain through TMHMM analysis, while the *Mclr* receptor proteins of other *Oreonectes* species with pigmented skin (*O. jiarongensis*, *O. dongliangensis* and *O. shuilongensis*) were predicted to contain an intact transmembrane domain (Fig. 2a and 2b). Similarly, the three-dimensional

structure of the *Mclr* receptor protein predicted by MODELLER exhibited disrupted transmembrane domain in *O. daqikongensis* (Fig. 2c and 2d), but not in the other species.

Sequencing, *de novo* assembly, and annotation of the *Oreonectes* Transcriptome

After trimming adapters and removing low quality reads, 67,824,194 – 44,390,242 pair-end reads were retained for three individuals of *O. daqikongensis* and three individuals of *O. jiarongensis*. Then *de novo* assembly of *O. jiarongensis* transcriptome was performed with all the clean data of three individuals of *O. jiarongensis*. We got 99,305 unigenes and the total length, average length and N50 value of the unigenes are 207,947,183 bp, 2,039 bp, and 3,571 bp, respectively (Fig. 3a). Of the 4,584 Benchmarking Universal Single-Copy Orthologs (BUSCOs) for Actinopterygii, 89.60% were complete and a further 5.14% were fragmented in *O. jiarongensis* transcriptome. It suggested that the *de novo* assembly was relatively successful comparing to several fish transcriptomes assembly (Supplementary Fig. S2 and Supplementary Table S2). We then annotated the unigenes using seven functional databases. Finally, 71,877 (NR: 72.38%), 77,110 (NT: 77.65%), 66,773 (Swissprot: 67.24%), 42,045 (KOG: 42.34%), 57,845 (KEGG: 58.25%), 71,542 (GO: 72.24%), and 63,813 (Pfam: 64.26%) of the unigenes were annotated (Fig. 3b).

Identification of differentially expressed genes (DEGs)

The mapping rate for transcriptome of *O. daqikongensis* and *O. jiarongensis* to *O. jiarongensis* transcriptome assembly was $72.34\% \pm 3.1 \times 10^{-3}$ and $75.26\% \pm 2.8 \times 10^{-3}$, respectively, suggesting that the transcriptome was well-assembled. We identified 9,856 unigenes as differentially expressed, based on a >2-fold change between *O. daqikongensis* and *O. jiarongensis* with an FDR corrected *P*-value (*q*-value) less than

0.05 ($\log_2$ -fold changes $[\log_2FC] > 1$, $[-\log_{10}q\text{-value}] > 1.3$). Among the differentially expressed unigenes, 7,326 (7.4% of the total unigenes) had decreased expression and 2,530 (2.5% of the total unigenes) had increased expression in *O. daqikongensis* skin (Fig. 4a). Nineteen genes belonging to the GO term “melanin biosynthetic process” and KEGG pathway “melanogenesis” (04916) were down-regulated, in which the expression level of twelve genes (including *Mclr*) were low in *O. daqikongensis* (Fig. 4b and 4c).

Discussion

The divergence between *O. daqikongensis* and its most closely related species with pigmentation (*O. dongliangensis* and *O. jiarongensis*) was estimated to be ~ 6.82 Mya (Fig 1d and Supplementary Fig 1), and our data suggested that the pseudogenization of *Mclr* occurred in the leucistic *O. daqikongensis* after its divergence from non-leucistic congeners. Given the lack of informative variable sites in the *Mclr* alignment between *O. daqikongensis* and its closest relative, estimation of the *Mclr* pseudogenization time is infeasible (Supplementary Fig. S1; Zhao et al., 2010). Nevertheless, the presence of several indels in *O. daqikongensis Mclr*, along with the widespread and drastic transcriptional changes in this lineage, it appears that *O. daqikongensis* represents a relatively ancient cave-dwelling species in which widespread adaptations and loss-of function mutations have accumulated. This renders our study species an interesting system to understand long-term evolutionary consequences of cave dwelling.

As fish living in caves, *O. daqikongensis*, *A. mexicanus*, and *Sinocyclocheilus anshuiensis* all have the leucistic phenotype (McGaugh et al., 2014; Bilandžija et al., 2013; Gross, Borowsky & Tabin, 2009; Gross & Wilkens, 2013; Jeffery et al., 2015;

Stahl & Gross, 2015; Yang et al., 2016). Similar selective pressures can lead to the parallel evolution of identical or similar traits in distantly related species, often referred to as adaptive phenotypic convergence (Christin, Weinreich & Besnard, 2010; Stern, 2013; Storz, 2016). A critical mechanism underlying phenotypic convergence is genetic convergence, including the same metabolic and regulatory pathways, protein-coding genes, or even identical amino acid substitutions in the same gene (Christin, Weinreich & Besnard, 2010; Stern, 2013; Storz, 2016; Mundy, 2005). Compared to the cavefish *A. mexicanus*, the mechanism of leucism in *Oreonectes* cavefish was also caused by pseudogenized *Mclr*, representing a remarkable scenario of genetic convergence (Gross, Borowsky & Tabin, 2009). The pseudogenization of *Mclr* in *O. daqikongensis* could directly or indirectly led to the low-level expression of dozens of melanogenesis related genes. As a result, the biosynthesis of melanin is blocked, leading to the albino phenotype.

In *A. mexicanus*, the reduced melanin phenotype of *A. mexicanus* is also due to the *Oca2* gene mutation, and the molecular mechanism and the degree of degeneration in skin pigmentation vary in different populations and circumstances. In Molino cave, *A. mexicanus* lost the 21st exon of *Oca2*, while in Pachón cave it was the deletion of exon 24-3'UTR which resulted in leucism (Protas et al., 2006). In the case of *S. anshuiensis*, genome analysis discovered that a mutation of glycine to arginine (G420R) in exon 4 of the *Tyr* gene encoding tyrosinase might result in leucism of the skin (Yang et al., 2016). Tyrosinase is the rate-limiting oxidase for controlling the production of melanin. It can catalyze the production of melanin and other pigments from tyrosine by oxidation (Kumar et al., 2011). Furthermore, a deletion of *Mpv17* might have also played a role in the leucism of *S. anshuiensis*. In the case of *O. daqikongensis*, besides the genes directly regulated by MC1R, other melanogenesis-

related genes involved in the Wnt signaling pathway and MAPK signaling pathway were also down-regulated (Fig. 4b). Our results, integrating with above-mentioned studies on other cavefishes, suggest that loss of pigmentation was caused by various kinds of coding region loss-of-function mutation along with widespread transcriptional changes, resulting from extended evolutionary time as a cave-dwelling form.

Oreonectes fishes could be chosen as representatives of three key nodes of the evolutionary process from a surface-dwelling lifestyle, to semi-cave dwelling, and finally to a permanent cave existence. Almost all *Oreonectes* species show some cave-related traits, such as part or complete eye degeneration and leucism, which makes this genus a good study system of micro-evolution. Therefore, additional genome information of *Oreonectes* species (*de novo* genome assembly, e.g.) ought to be an indispensable resource for a more comprehensive understanding of the adaptive evolutionary mechanism of cavefishes.

379 **Declarations**

380 **Acknowledgments**

381 We are further grateful to Kai Gao, Jian Chen, Zhonglong Cen and Gang Li for their
382 help in sampling and laboratory work. We thank Martin Burrows and Wenhua Xiong
383 for assistance in writing and data analysis, and two anonymous reviewers for their
384 helpful feedback.

385 **Funding**

386 This project was supported by the key project of Science-technology basic condition
387 platform from The Key Project of Science and Technology Program of Guizhou
388 Province “Study of biological and ecological values for Fanjingshan World Natural
389 Heritage Nomination of Tongren City (3052 2015 Qiankehe SY)”, The Ministry of
390 Science and Technology of the People’s Republic of China (Grant No.
391 2005DKA21402), Provincial Governor Fund of Guizhou (Qian-Sheng-Zhuan-He-Zi
392 (2010) 14) project and the Youth Innovation Promotion Association of Chinese
393 Academy of Sciences (2011077).

394 **Authors’ contributions**

395 Zhijin Liu, Huamei Wen and Fang Dong performed the experiments, analysed the
396 data, wrote the paper, and participated in the design of the study. Frank Hailer, Fanglei
397 Shi, Zuomin Yang, Tao Liu and Ling Han participated in data analysis, interpretation,
398 and drafting the manuscript. Jiang Zhou and Yibo Hu conceived the study,
399 participated in its design and coordination, and helped to draft the manuscript. All
400 authors read and approved the final manuscript.

401 **Ethics approval and consent to participate**

Our experimental procedures complied with the current laws on animal welfare and research in China, and were specifically approved by the Animal Research Ethics Committee of the Institute of Zoology, Chinese Academy of Sciences.

Competing interest

The authors declare that they have no competing interests.

Consent for publication

All authors gave final approval for publication.

422 **Table 1.** Summary of sample information.

Species	Museum catalog number & GenBank accession number	Sampling location
<i>O. daqikongensis</i>	GZNU20110128002-2011012806	N 25°17'05.1"
	GZNU20151225001-20151225007	E 107°44'54.3"
	mtDNA (NC_039131, GZNU20151225002)	
<i>O. jiarongensis</i>	GZNU20120128001-20120128006	N 25°27'58.82"
	GZNU20151221001-20151221007	E 108°06'47.49"
	mtDNA (KU987437, GZNU20151221001)	
<i>O. dongliangensis</i>	GZNU20160815012-20160815017	N 25°19'48.95"
	mtDNA (KY626008, GZNU20160815012)	E 108°01'18.53"
<i>O. shuilongensis</i>	GNUG20090723002-20090723003	N 25°50'41.87"
	GNUG20160627001-20160627008	E 107°52'57.76"
	mtDNA (KF640641, GZNU20160627001)	

423

424

425

426

427

428

429

430

431

432

433

434 **Table 2.** Primers used for the mtDNA and *MC1R* gene amplifications.

Gene	Primers	Sequence (5' to 3')	Amplicon length
mtDNA	L14724	GAC TTG AAA AAC CAC CGT TG	1192 bp
	H15915	CTC CGA TCT CCG GAT TAC AAG AC	
	L11008	TTG ACT ACC CAA AGC CCA	1249 bp
	H12256	CTT AGA GGG CAA TAG GTG TAA	
	L12166	TGA CAC TGA ATA AAT ACA GCC CT	1022 bp
	H13187	TGA CAC TGA ATA AAT ACA GCC CT	
	L13003	GGT TCC ATT ATT CAC AGC CT	1270 bp
	H14272	TAG GGT TAG TTG CTG TGG C	
	L14080	AAT GGC TCA GCA GCT AAA G	1355 bp
	H15431	TAA TAA ATG GGT GTT CTA CTG G	
<i>MC1R</i>	MC1R-F	GAATATCAGAGGTGTGCTGAAGC	971-1004 bp
	MC1R-R	TCCTTGAGAGTCTTGCGCAG	

435

436

437 **Table 3.** Genetic distances based on P-distance method among *Oreonectes* species.

<i>O. daqikongensis</i>				
<i>O. dongliangensis</i>	0.103			
<i>O. jiarongensis</i>	0.101	0.010		
<i>O. shuilongensis</i>	0.106	0.094	0.091	
<i>O. platycephalus</i>	0.149	0.148	0.147	0.142

438

439 **References**

- 440 **Ayumi T, Hiroaki Y, Jun Y, Cock van O, Masakado K. 2011.** The MC1R gene in
441 the guppy (*Poecilia reticulata*): Genotypic and phenotypic polymorphisms. *BMC*
442 *Research Notes* **4**:31.
- 443 **Barsh G, Cotsarelis G. 2007.** How Hair Gets Its Pigment. *Cell* **130**: 779–781.
- 444 **Bilandžija H, Ma L, Parkhurst A, Jeffeary WR. 2013.** A potential benefit of
445 leucism in *Astyanax* cavefish: down-regulation of the *Oca2* gene increases
446 tyrosine and catecholamine levels as an alternative to melanin synthesis. *PLoS*
447 *One* **8**: e80823.
- 448 **Christin PA, Weinreich DM, Besnard G. 2010.** Causes and evolutionary
449 significance of genetic convergence. *Trends in Genetics* **26**: 400–405.
- 450 **Deng HQ, Wen HM, Xiao N, Zhou J. 2016.** A new blind species of the cave genus
451 *Oreonectes* from Guizhou, China (Nemacheilinae). *ZooKeys* **637**: 47–59.
- 452 **Drummond AJ, Rambaut A. 2007.** BEAST: Bayesian evolutionary analysis by
453 sampling trees. *BMC Evolutionary Biology* **7**: 214.
- 454 **Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012.** Bayesian phylogenetics
455 with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**: 1969–
456 1973.
- 457 **Eswar N, Eramian D, Webb B, Shen MY, Sali A. 2008.** Protein structure modeling
458 with MODELLER. *Methods in Molecular Biology* **426**: 145–159.
- 459 **Frickhinger KA. 1991.** Fossilien atlas fishes. Mergus, Melle, Germany. 1088 pp.
- 460 **Günther A. 1868.** Catalogue of the Fishes in the British Museum. Volume seventh
461 .London: Trustees of the British Museum, 512.
- 462 **Gross JB, Borowsky R, Tabin CJ. 2009.** A novel role for Mc1r in the parallel
463 evolution of depigmentation in independent populations of the cavefish *Astyanax*
464 *mexicanus*. *PLoS Genetics* **5**: e1000326.
- 465 **Gross JB, Wilkens H. 2013.** Leucism in phylogenetically and geographically distinct
466 populations of *Astyanax* cavefish arises through the same loss-of-function *Oca2*
467 allele. *Heredity* **111**:122–130.

468 **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adicoins**
469 **X, Fan L, Raychowdhury R, Zeng QD, Chen ZH, Mauceli E, Hacohen N,**
470 **Gnirke A, Rhind N, Palma FD, Birren BW, Nusbaum C, Lindblad-Toh K,**
471 **Friedman N, Regev A. 2011.** Full-length transcriptome assembly from RNA-
472 Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

473 **Heled J, Drummond AJ. 2012.** Calibrated tree priors for relaxed phylogenetics and
474 divergence time estimation. *Syst Biol* **61**: 138-149.

475 **Jeffery WR, Ma L, Parkhurst A. Bilandžija H. 2015.** Pigment regression and
476 leucism in *Astyanax* cavefish. In: Keene AC, Yoshizawa M, McGaugh SE (eds)
477 *Biology and evolution of the Mexican cavefish*. Amsterdam: Elsevier, 155–171.

478 **Johnson JA, Ambers AD, Burnham KK. 2012.** Genetics of plumage color in the
479 Gyr Falcon (*Falco rusticolus*): analysis of the melanocortin-1 receptor gene. *The*
480 *Journal of Heredity* **103**: 315–321.

481 **Kaelin CB, Barsh GS. 2013.** Genetics of Pigmentation in Dogs and Cats. *Annual*
482 *Review Animal Bioscience* **1**: 125–156.

483 **Kumar CM, Sathisha UV, Dharmesh S, Rao AAG, Singh SA. 2011.** Interaction of
484 sesamol (3, 4-methylenedioxyphenol) with tyrosinase and its effect on melanin
485 synthesis. *Biochimie* **93**: 562–569.

486 **Kircher M, Stenzel U, Kelso J. 2009.** Improved base calling for the Illumina
487 Genome Analyzer using machine learning strategies. *Genome Biology* **10**: R83.

488 **Li WZ, Godzik A. 2006.** Cd-hit: a fast program for clustering and comparing large
489 sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

490 **Langmead B. 2010.** Aligning short sequencing reads with Bowtie. *Current Protocols*
491 *in Bioinformatics* **11**: 7.

492 **Li B, Dewey CN. 2011.** RSEM: accurate transcript quantification from RNA-seq data
493 with or without a reference genome. *BMC Bioinformatics* **12**: 323–338.

494 **Machado L, Šmíd J, Mazuch T, Sindaco R, Al Shukaili AS, Carranza S. 2018.**
495 Systematics of the Saharo- Arabian clade of the Palearctic naked- toed geckos
496 with the description of a new species of *Tropicolotes* endemic to Oman. *Journal*
497 *of Zoological Systematics and Evolutionary Research* **56**: 1–20.

- 498 **Majerus ME, Mundy NI. 2003.** Mammalian melanism: natural selection in black
499 and white. *Trends in Genetics* **19**: 585–588.
- 500 **Maytin AK, Davies SW, Smith GE, Mullen SP and Buston PM. 2018.** *De novo*
501 Transcriptome Assembly of the Clown Anemonefish (*Amphiprion percula*): A
502 New Resource to Study the Evolution of Fish Color. *Frontiers in Marine Science*
503 **5**:284.
- 504 **Mendes J, Salvi D, Harris DJ, Els J, Carranza S. 2018.** Hidden in the Arabian
505 Mountains: Multilocus phylogeny reveals cryptic diversity in the endemic
506 Omanosaura lizards. *Journal of Zoological Systematics and Evolutionary*
507 *Research* **56**:395–407.
- 508 **McGaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H,**
509 **Jeffery WR, Keene A, Ma L, Minx P, Murphy D, E.Q'Quin K, Rétaux S,**
510 **Rohner N, Searle SMJ, Stahl BA, Tabin C, Volff JN, Yoshizawa M, Warren**
511 **WC. 2014.** The cavefish genome reveals candidate genes for eye loss. *Nature*
512 *Communications* **5**: 5307.
- 513 **Mundy NI, Badcock NS, Hart T, Scribner K, Jansen K, Nadeau NJ. 2004.**
514 Conserved genetic basis of a quantitative trait involved in mate choice. *Science*
515 **303**: 1870–1873.
- 516 **Mundy NI. 2005.** A window on the genetics of evolution: *MC1R* and plumage
517 colouration in birds. *Proceedings of the Royal Society of London Series B-*
518 *Biological Sciences* **272**: 1633–1640.
- 519 **Nachman MW, Hoekstra HE, D'Agostino SL. 2003.** The genetic basis of adaptive
520 melanism in pocket mice. *Proceedings of the National Academy of Sciences of*
521 *The United State of America* **100**: 5268–5273.
- 522 **Nishimura EK. 2011.** Melanocyte stem cells: a melanocyte reservoir in hair follicles
523 for hair and skin pigmentation. *Pigment Cell Melanoma Res* **24**: 401–410.
- 524 **Protas ME, Hersey C, Kochanek D, Zhou Y, Wilkens H, Jeffery WR, Zon LI,**
525 **Borowsky R, Tabin CJ. 2006.** Genetic analysis of cavefish reveals molecular
526 convergence in the evolution of leucism. *Nature Genetics* **38**: 107–111.
- 527 **Posada D, Crandall, KA. 1998.** MODELTEST: Testing the model of DNA

528 substitution. *Bioinformatics* **14**: 817–818.

529 **Ritland K, Newton C, Marshall HD. 2001.** Inheritance and population structure of
530 the white-phased “Kermode” black bear. *Current Biology* **11**: 1468–1472.

531 **Rui M, Chen X, Yang J, Richard L. 2012.** Phylogenetic relationships of the genus
532 Homatula(Cypriniformes: Nemacheilidae), with special reference to the
533 biogeographic history around the Yunnan-Guizhou plateau. *Zootaxa* **3586**: 78–
534 94.

535 **Ronquist F, Huelsenbeck JP. 2003.** MrBayes 3: Bayesian phylogenetic inference
536 under mixed models. *Bioinformatics* **19**: 1572–1574.

537 **Stahl BA, Gross JB. 2015.** Alterations in *Mclr* gene expression are associated with
538 regressive pigmentation in *Astyanax* cavefish. *Development Genes and Evolution*
539 **225**: 367–375.

540 **Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015.**
541 BUSCO: assessing genome assembly and annotation completeness with single-
542 copy orthologs. *Bioinformatics* **31**: 3210–3212.

543 **Stern DL. 2013.** The genetic causes of convergent evolution. *Nature Reviews*
544 *Genetics* **14**: 751–764.

545 **Storz JF. 2016.** Causes of molecular convergence and parallelism in protein
546 evolution. *Nature Reviews Genetics* **17**: 239–250.

547 **Sudhir K, Glen S, Koichiro T. 2016.** MEGA7: molecular evolutionary genetics
548 analysis 7.0 version for bigger datasets. *Molecular Biology and Evolution*
549 **33**:1870-1874.

550 **Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997.** The
551 ClustalX windows interface: Flexible strategies for multiple sequence alignment
552 aided by quality analysis tools. *Nucleic Acids Research* **25**: 4876–4882.

553 **Toyofuku K, Valencia JC, Kushimoto T, Costin GE, Virador VM, Vieira WD,**
554 **Ferrans VJ, Hearing VJ. 2002.** The etiology of oculocutaneous leucism (OCA)
555 type II: the pink protein modulates the processing and transport of tyrosinase.
556 *Pigment Cell Research* **15**: 217–224.

557 **Wang Y, Shen YJ, Feng CG, Zhao K, Song ZB, Zhang YP, Yang LD, He SP. 2016.**
558 Mitogenomic perspectives on the origin of Tibetan loaches and their adaptation
559 to high altitude. *Scientific Reports* **6**: 29690.

560 **Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G,**
561 **Kriventseva EV, Zdobnov EM. 2018.** BUSCO Applications from Quality
562 Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and*
563 *Evolution*, 35: 543-548.

564 **Yang J, Chen XL, Bai J, Fang DM, Qiu Y, Jiang WS, Bian C, Lu J, He SY, Pan**
565 **XF, Zhang YL, Wang XA, You XX, Wang YS, Sun Y, Mao DQ, Liu Y, Fan**
566 **GY, Zhang H, Chen XY, Zhang XH, Zheng LP, Wang JT, Cheng L, Chen**
567 **JM, Ruan ZQ, Li J, Yu H, Peng C, Ma XY, Xu JM, He Y, Xu ZF, Xu P,**
568 **Wang J, Yang HM, Wang J, Whitten T, Xu X, Shi Q. 2016.** The
569 Sinocyclocheilus cavefish genome provides insights into cave adaptation. *BMC*
570 *Biology* **14**: 1.

571 **Zhan XJ, Dixon A, Fox NC, Bruford MW. 2012.** Missense SNP of the MC1R gene
572 is associated with plumage variation in the Gyrfalcon (*Falco rusticolus*). *Animal*
573 *Genetics* **43**: 460–462.

574 **Zhao HB, Yang JR, Xu HL, Zhang JZ. 2010.** Pseudogenization of the Umami Taste
575 Receptor Gene Tas1r1 in the Giant Panda Coincided with its Dietary Switch to
576 Bamboo. *Molecular Biology and Evolution* **27**: 2669–2673.

577 **Zhou JJ. 1992.** A new Cobitis from the middle Miocene of Shanwang, Shandong.
578 *Vertebrata Palasiatica* 30: 71-76.

579 **Figure legends**

580 **Figure 1.** (a) Map of sampling locations: 1. *O. shuilongensis*, 2. *O. jiarongensis*, 3. *O.*
581 *dongliangensis*, and 4. *O. daqikongensis*. (b) Photo of *O. daqikongensis*, and (c) *O.*
582 *jiarongensis*. (d) Phylogeny and divergence times among loaches (Nemacheilidae).
583 Red dots denote the utilized calibration points (see Methods). Blue bars show the 95%
584 posterior credibility intervals of divergence times. Number above on each node
585 indicates the Bayesian posterior probability. Red branches indicate the the genus
586 *Oreonectes*. (e) Schematic of MC1R receptor structure and ORF-disrupting mutations
587 of *Mclr* genes. The first line shows the disrupted sequence of *O. daqikongensis*, lines
588 below represent intact sequences of *O. jiarongensis*, *O. dongliangensis* and *O.*
589 *shuilongensis*. The codon containing an ORF-disrupting mutation (marked red and
590 underlined) is indicated by a box.

591 **Figure 2.** Two-dimensional model of MC1R of (a) *O. jiarongensis* and (b) *O.*
592 *daqikongensis*. (c) Three-dimensional structure of MC1R of *O. jiarongensis*, and (d)
593 and *O. daqikongensis*.

594 **Figure 3.** (a) Length distribution of assembled unigenes in *O. jiarongensis*
595 transcriptome. (b) Gene ontology classifications of unigenes from the *O. jiarongensis*
596 transcriptome.

597 **Figure 4.** (a) Differentially expressed genes for *O. daqikongensis* skin tissue vs. *O.*
598 *jiarongensis* skin tissue. Red dots: up-regulated genes: green: down-regulated genes.
599 (b) Schematic mechanisms of signaling pathways for melanogenesis. Down- regulated
600 genes in *O. daqikongensis* skin are shown in red. (c) Expression levels of
601 melanogenesis-related down-regulated genes of *O. daqikongensis* and *O. jiarongensis*
602 measured in FPKM, expression values are mean $\pm$ SD of at least three independent
603 experiments.