Published in final edited form as:

Genome. 2018 April ; 61(4): 254–265. doi:10.1139/gen-2017-0049.

A local duplication of the *Melanocortin receptor 1* locus in *Astyanax*

Joshua B. Gross¹, James Weagley^{2,4,5}, Bethany A. Stahl¹, Li Ma¹, Luis Espinasa³, and Suzanne E. McGaugh²

James Weagley: weagl001@umn.edu; Bethany A. Stahl: bethanyastahl@gmail.com; Li Ma: li.bluesky.05@gmail.com; Luis Espinasa: Luis.Espinasa@marist.edu; Suzanne E. McGaugh: smcgaugh@umn.edu

¹Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45221

²Department of Ecology, Evolution, and Behavior, University of Minnesota, St. Paul, Minnesota 55108

³School of Science, Marist College, Poughkeepsie, New York 12601

Abstract

In this study, we report evidence of a novel duplication of *Melanocortin receptor 1* (*Mc1r*) in the cavefish genome. This locus was discovered following the observation of excessive allelic diversity in a ~820 bp fragment of *Mc1r* amplified via degenerate PCR from a natural population of *Astyanax aeneus* fish from Guerrero, Mexico. The cavefish genome reveals the presence of two closely related *Mc1r* open reading frames separated by a 1.46 kb intergenic region. One open reading frame corresponds to the previously reported *Mc1r* receptor, and the other open reading frame (duplicate copy) is 975 bp in length, encoding a receptor of 325 amino acids. Sequence similarity analyses position both copies in the syntenic region of the single *Mc1r* locus in 16 representative craniate genomes spanning bony fish (including *Astyanax*) to mammals, suggesting we discovered tandem duplicates of this important gene. The two *Mc1r* copies share ~89% sequence similarity, and, within *Astyanax*, are more similar to one another compared to other *melanocortin* family members. Future studies will inform the precise functional significance of the duplicated *Mc1r* locus, and if this novel copy number variant may have adaptive significance for the *Astyanax* lineage.

Keywords

cavefish; *brown* mutation; Guerrero; copy number variation

¹Corresponding author: Joshua B. Gross, 312 Clifton Court, 711B Rieveschl Hall, Cincinnati, Ohio 45221; grossja@ucmail.uc.edu; 513-556-9708.

⁴Center for Genome Sciences and Systems Biology, Washington University in St. Louis, St. Louis, MO 63108 USA

⁵Center for Gut Microbiome and Nutrition Research, Washington University in St. Louis, St. Louis, MO 63108 USA

Consent/conflict of interest

The authors report no conflicts of interest.

Introduction

The gene *melanocortin receptor 1* (*Mclr*) encodes a seven-transmembrane domain G-protein coupled receptor expressed in melanin-producing cells (Tsatmali et al. 2002). This protein receptor binds-melanophore stimulating hormone, resulting in changes in intracellular cAMP that translate to melanin production (Slominski et al. 2004). In mammals, higher levels of intracellular cAMP are associated with eumelanin production (black/brownish hues), while lower levels of cAMP are associated with production of pheomelanin (orange/reddish hues; Hida et al. 2009). Owing to the penetrance of this molecular “switch”, variation in the coding sequence of *Mclr* causes dramatic effects on coloration across taxa. For instance, coding sequence changes impacting the structure of this receptor are associated with scores of coloration differences affecting mammals (Kijas et al. 1998), birds (Mundy 2005), snakes (Cox et al. 2013), lizards (Rosenblum et al. 2004), and fish (Gross et al. 2009).

In light of its role in diverse coloration schemes across wide taxonomic groups, the genetic and genomic structure of *Mclr* has remained remarkably stable across millions of years of vertebrate evolution (Logan et al. 2003). In every system examined thus far, *Mclr* is a simple locus typically comprising a single-exon open reading frame ~1kb in length (Makova and Norton 2005). In humans, a splice form has been reported, however its functional significance is unknown (Tan et al. 1999). Although *Mclr* is among a family of five structurally similar melanocortin receptors (Switonski et al. 2013), copy number variants of this gene have not been reported in the literature.

In this study, we investigated the sequence structure of a natural population of *Astyanax aeneus* cavefish collected from the Granadas cave, in the state of Guerrero, in southern Mexico (Espinasa et al. 2001). These fish are closely related to the blind cave-dwelling *Astyanax mexicanus* characins that populate the Sierra de El Abra ~1100 km to the northeast (Avisé and Selander 1972; Ornelas-García et al. 2008; Oliveira et al. 2011). Of ~20 specimens that were sampled, >2 alleles were discovered in several individuals. Members of the *Astyanax* genus are diploid (Morelli et al. 1983), with a karyotypic number of 25 (Kirby et al. 1977), and therefore we reasoned that this unexpected level of divergence may reflect copy number variation of the *Mclr* gene.

To summarize our results, mapping of the *Mclr* amplicons to the draft *Astyanax* genome (McGaugh et al. 2014) resulted in their alignment to two discrete positions on the same genomic scaffold. Bioinformatic analyses revealed the duplicate locus has an open reading frame, and BLAST analyses found both copies returned their strongest hits to a single genomic *Mclr* locus in every vertebrate examined, in an ancient syntenic region of each genome. Both copies of *Mclr* were highly similar to one another (differing in sequence by only 10.4%), but much less similar to any other *melanocortin* family members (differing in sequence by >30%). From a broader perspective, *melanocortins* can be regarded as a stable gene family with few reported paralogs. Two exceptions are *Mc5r*, which harbors an “a” and “b” form in *Danio rerio* (Logan et al. 2003), as well as *Mc4r* which harbors copy number variants associated with reproduction in *Xiphophorus multilineatus* (Smith et al. 2015). Interestingly, only *Mc5r* has an annotated paralog in the *Astyanax* genome, and *Mc5ra* and

Mc5rb are also present in zebrafish, which is 250My diverged from *Astyanax*. *Mc5ra* and *Mc5rb* were more divergent in sequence than were the two copies of *Mc1r*, consistent with a more ancient duplication of *Mc5r* than *Mc1r*. In sum, we report a putative duplication event affecting the gene *Mc1r* in two species of *Astyanax* fish – *Astyanax mexicanus* and *Astyanax aeneus* – both of which have a surface and cave morph. Future studies will clarify if and how this gene impacts melanocortin functioning in the *Astyanax* lineage, in particular for lineages that have adapted to the cave environment.

Materials and methods

Guerrero *Astyanax aeneus* cavefish analyses through degenerate cloning

We cloned a ~820 bp sub-fragment of the *Mc1r* open reading frame using degenerate PCR from 11 wild-caught, cave-dwelling individuals and 12 wild-caught, surface-dwelling individuals (*A. aeneus*) from the Granadas cave and stream localities in Guerrero, Mexico (Espinasa et al. 2001). We utilized the following degenerate primers to successfully clone a fragment of *Mc1r* from *Astyanax aeneus* specimens: degen*Mc1r*_F1: 5'-GGGCCTGATCTCCTGGTNGARAAYAT-3'; degen*Mc1r*_R1: 5'-GGGGGTCGATCAGAGTTRCADATDAT-3'; degen*Mc1r*_R2: 5'-CAGGATCAGGATCGCARRAARAANGG-3'. A gradient PCR approach using annealing temperatures of ~48°C, ~59°C and ~70°C was used. All three annealing temperatures produced amplified fragments (Fig. 1A). PCR products were directly sequenced or re-amplified in a second reaction using 1µl of the first reaction as the template. This resulted in a substantially improved yield that was isolated, gel purified, subcloned and Sanger-sequenced. Each fragment was subcloned to the pGEM-T Easy vector (Promega; Madison, WI) and multiple clones were Sanger-sequenced in forward and reverse directions using M13 universal primer sites.

Nucleotide and protein sequence alignments

For Guerrero samples, all *Mc1r* amplicons were aligned separately for cave- and surface-dwelling experiments using SeqMan Pro (v. 13.0.2; DNASTar, Madison, WI). Initial default assembly parameters were implemented (minimum match percentage = 80%; match size = 12; minimum sequence length = 100; maximum added gaps per kb in sequence/contig = 70) yielding a single contig. These parameters resulted in many sequence conflicts in the resulting consensus, indicating several mismatched sequences. To test whether this sequence diversity resulted from multiple alleles present in the same sequencing reaction, we performed a second series of assemblies using a more stringent minimum match percentage of 95% (other parameters remained unchanged). This resulted in both cave- and surface-dwelling fish datasets producing multiple contigs. Routine BLASTX searches (<https://blast.ncbi.nlm.nih.gov>) identified each contig with varying degrees of sequence similarity (ranging from ~89%–100%) to the gene *Mc1r* in *Astyanax mexicanus*.

We evaluated if these contigs localized to different regions of the genome to determine if they represented different copies of the gene *Mc1r*. We tested this by performing BLAST analyses of each contig to the cavefish genome, AstMex1.0.2 (see below). All *Mc1r*

amplicons from the Guerrero population localized to two regions of the same contig on Scaffold KB882081.1.

We realigned our Guerrero *Mc1r* amplicons to this scaffold using SeqMan Pro. Owing to the level of sequence divergence between *Astyanax aeneus* from Guerrero and the reference cavefish genome (Pachón *Astyanax mexicanus* from El Abra region), we relaxed the stringency for read alignment. Specifically, we lowered the minimum match percentage value from 80% (default) to 75%. The resultant assembly yielded two discrete alignment positions on scaffold KB882081.1, providing the first evidence of two putative copies of *Mc1r* in the cavefish genome. All subsequent nucleotide alignments were carried out in SeqMan Pro using the cavefish genome scaffold KB882081.1 as reference sequence.

BLAST analyses

All BLAST analyses were performed using the Ensembl BLAST search tool (<http://www.ensembl.org/Multi/Tools/Blast?db=core>). We tested several search tools, and found highly similar results with BLAT, BLASTN and BLASTX (NCBI). We principally utilized the BLAT search tool since it provided rapid and reliable results. All search analyses involved nucleotide queries against the ‘Genomic sequence’ DNA database, retaining all other default parameters.

To determine sequence similarity between both *Mc1r* and Duplicate *Mc1r* across taxa, we performed BLAT against the following genomes: *Astyanax mexicanus* (AstMex102), *Danio rerio* (GRCz10), *Gadus morhua* (gadMor1), *Gallus gallus* (Gallus_gallus-5.0), *Gasterosteus aculeatus* (BROAD S1), *Latimeria chalumnae* (LatCha1), *Lepisosteus oculatus* (LepOcu1), *Mus musculus* (CL57BL6 reference; GRCm38.p5), *Oreochromis niloticus* (Orenil1.0), *Oryzias latipes* (HdrR), *Petromyzon marinus* (Pmarinus_7.0), *Poecilia formosa* (PoeFor_5.1.2), *Takifugu rubripes* (FUGU 4.0), *Tetraodon nigroviridis* (TETRAODON 8.0), *Xenopus tropicalis* (JGI 4.2), *Xiphophorus maculatus* (Xipmac4.4.2). All reported BLAT searches were carried out in January and February, 2017. All sequence similarity studies reported the top returned hit for each genome based on the highest BLAST score. We recorded the following information: genomic location, identity of overlapping genes, orientation, query start/end, sequence length, BLAST score, E-value, and the percent ID value (see Table 1). BLAST searches facilitated analyses of neighboring genomic regions for syntenic studies (Table 2). All syntenic studies were performed in the Ensembl genome browser following identification of the strongest hit for *Mc1r*/Duplicate *Mc1r* in sixteen representative genomes.

El Abra fish rearing, tissue isolation and sequence analysis

Representative surface-dwelling fish and Pachón, Tinaja, and Chica cavefish were maintained under the normal rearing conditions in a satellite fish facility at the University of Cincinnati, as previously described (Gross et al. 2016). To obtain genomic DNA, individuals were briefly anesthetized using buffered tricaine (MS-222), under sterile conditions, to excise a small portion of the caudal fin. Genomic DNA was extracted from fin clips of four individuals from each of the following *Astyanax* populations from the Sierra de El Abra: surface fish, Pachón cavefish, Tinaja cavefish, and Chica cavefish using the DNeasy blood &

Tissue Kit (Qiagen, Valencia, CA). We performed standard PCR to amplify a ~3749 bp fragment, inclusive of the duplicate *Mc1r* genomic interval, from each population using the forward primer 5'-TCTGCCATTAGAGGAGGCC-3' and reverse primer 5'-AGAGTTCCGGCTGGTGT-3' using Taq polymerase (Roche; Basel, Switzerland). The following PCR cycling conditions were implemented: denaturation (94°C for 2 min), 35 cycles of the following: denaturation (94°C for 30 sec), annealing (60°C for 45 sec), and elongation (at 68°C for 3 min and 30 sec); and an elongation step (72°C for 7 min). PCR products were purified and subcloned into the pGEM-T Easy vector and sequenced using universal M13R and M13F sites. Additional Sanger sequence analysis was carried out using the following forward sequencing primers: 5'-TGAGATTGACTGTGAACAACCAT-3'; 5'-CCAAAGAAGGCAACAAGGCAC-3'; and the following reverse sequencing primers: 5'-TGAGAAAGCCATTTGAACGCT-3'; 5'-CTGGACAACGTCATTGATGTGA-3'. Sanger sequencing reads were trimmed to remove vector contamination (pGEM-T Easy), and aligned to the scaffold KB882081.1 (by population) using the following alignment parameters: match size: 12; minimum match percentage: 90; and minimum sequence length: 100. All other default parameters were retained. All protein analyses were carried out in MegAlign (v.13.0; DNASTar, Madison, WI) and aligned using the ClustalV method (see Fig. 4).

All live specimens were collected and generously provided to us by Dr. Richard Borowsky (New York University). This study was performed in accordance with the Guide for the Care and Use of Laboratory Animals of the NIH. The Institutional Animal Care and Use Committee (IACUC) at the University of Cincinnati approved this Animal Use Protocol (Number 10-01-21-01).

Alignment and BLAST analysis of *Astyanax mexicanus* transcriptome

Previously published Trinity-assembled transcriptome data (McGaugh et al., 2014) were used to assay expression of the two *Mc1r* copies. Tissue specific expression was analyzed in brain, heart, kidney, liver, muscle, nasal and skin tissue from adult Pachón cavefish (*A. mexicanus* from El Abra region), as well as whole eyes of surface fish from Texas. Whole-embryo time series transcriptome data collected over the course of early development (10 hours post fertilization (hpf) through 72 hpf) from *Astyanax mexicanus* surface and cavefish were also analyzed for expression of the two *Mc1r* copies. Tissue-specific blast databases were created for previously published (McGaugh et al. 2014) tissue-specific transcriptomes. Surface-specific and cave-specific early-development transcriptome BLAST databases were created by concatenating data across time points for each ecotype, creating a Trinity-assembled transcriptome and identifying the longest open reading frame. The *Astyanax Mc1r* and Duplicate *Mc1r* sequences from the reference genome were used as a query against the database for each tissue and ecotype-specific developmental data using BLASTn with an e-value cut-off of 0.01. Since we conducted within-species BLAST searches, if a copy was expressed in a given tissue at the time of sampling, then we would expect high-quality blast hits returned for that tissue or developmental stage. Since paralogues are often difficult to distinguish with short-reads, it is possible that transcripts from *Mc1r*, Duplicate *Mc1r*, and other members of the *melanocortin* gene family could be returned as BLAST hits in these searches. Therefore, we required at least 90% identity of the raw read to *Mc1r*

and/or Duplicate *Mc1r*, and examined the length of the open reading frame covered by raw transcriptome reads.

We visualized the mapping of RNA-seq unique reads to each of the two *Mc1r* copies by aligning the Illumina reads directly to the reference genome using the BWA-mem algorithm (Li, 2013) in *bwa-0.7.1* (Li, 2009). Samtools and the Integrative Genomics Viewer (IGV) were used to visualize the abundance of RNA-seq reads from eye, skin, and cavefish developmental reads that aligned to each copy of *Mc1r* on genomic scaffold KB882081.1 (Figure S1–S4; Li et al. 2009; Robinson et al. 2011).

Results

Degenerate PCR cloning reveals multiple *Mc1r* alleles in *Astyanax aeneus* cavefish and surface fish from the Guerrero locality

We implemented a standard degenerate PCR approach to amplify a fragment of the *Mc1r* gene in wild-caught individuals from the Guerrero locality (Espinasa et al. 2001). We initially sought to determine if *Mc1r* coding mutations were present in southern Guerrero cavefish populations, as had been discovered in closely-related cavefish species of the Sierra de El Abra region to the north (Gross et al. 2009). Similar to El Abra populations, the Granadas locality harbors both cave- and surface-dwelling morphs, and the cave-dwelling forms are depigmented (Espinasa et al. 2001). Because *Mc1r* has only been reported as a single-exon gene in fish (Logan et al. 2003), we cloned this fragment from genomic DNA using a gradient series of three annealing temperatures ranging from ~48°C – ~70°C (Fig. 1A). In addition, because Guerrero fish (*Astyanax aeneus*) do not have a sequenced genome, we employed degenerate PCR primers in hopes of amplifying a conserved region of the *Mc1r* open reading frame. Our primers successfully amplified a weak band in a first round of gradient PCR in each of two individuals (Fig. 1A), so we seeded a second reaction with 1 µl of the first reaction as the template. This resulted in a robust amplification of a ~820 bp fragment that was cut, gel-purified and Sanger sequenced either directly, or following subcloning into the pGEM-T Easy vector. Once we validated this approach, we used the same primers to amplify the same fragment from cave-collected Guerrero fish (n = 11) and surface-collected Guerrero fish (n = 12).

Curiously, when we began analyzing our results we noted that the “same” fragment of *Mc1r* from a single individual returned BLAST results between ~89% – 100% similar to the accessioned *Mc1r* sequence for *Astyanax mexicanus* (Gross et al. 2009). These results implied that a single individual harbored more than two alleles. This was surprising since members of the *Astyanax* genus exhibit a karyotypic number of 25 diploid chromosomes (Kirby et al. 1977). Therefore, if *Mc1r* represents a single-copy genomic locus, one would only predict a maximum of two allelic forms per individual. We next considered that there might have been a duplication of the *Mc1r* locus that led to a copy number variant in two *Astyanax* species (*aeneus* and *mexicanus*) that has not been reported elsewhere. Towards this end, we re-sequenced the ~820 bp fragment in our Guerrero specimens and aligned them to the recently-sequenced *Astyanax mexicanus* genome (McGaugh et al. 2014).

Two Mc1r alleles are situated ~1.4 kb apart in a cis chromosomal orientation on the same genomic scaffold

The *Astyanax* genome was published as a set of >10,000 genomic scaffolds (McGaugh et al. 2014). Our BLAST analyses localized all *Mc1r* amplicons from Guerrero fish to a ~3.4 kb region of genomic scaffold KB882081.1 (4,742,310 bp in length; Fig. 2). When we analyzed the positional distribution of our *Mc1r* amplicons, we found that the reads aligned to two discrete regions of the scaffold, very near one another (Fig. 2; green). Further, both “copies” of *Mc1r* resided on the same contig, APW001074105.1 (12,988 bp in length). Based on the current annotation of the genome, both copies are directed in the same orientation, on the reverse strand, of KB882081.1 at positions 3150542 – 3151511 (*Mc1r*) and 3148107 – 3149081 (Duplicate *Mc1r*; Fig. 2). Because a duplication event yielded two copies of *Mc1r*, both loci must be considered paralogous. In this report, we refer to the first-discovered copy as “*Mc1r*”. This gene is 970 bp in length and represents the form of *Mc1r* previously characterized and described (Gross et al., 2009). The second copy is referenced as “Duplicate *Mc1r*” and is 975 bp in length. As more is learned about this duplicated region of the *Astyanax* genome in the future, the nomenclature may be revised.

These loci were discovered based on *Mc1r* amplicons from *Astyanax aeneus* that were aligned to the cavefish reference genome (*Astyanax mexicanus*; Pachón cave locality). The sequence fragments we analyzed were predicted to only cover ~84% of the annotated open reading frame of *Mc1r*, therefore we sought to determine if the structure of both copies were intact. We evaluated ~300 bp of sequence directly 5′ and 3′, in the reference genome, based on the originally cloned fragments. We found that both copies harbor an intact open reading frame. In the Pachón cavefish reference genome, the intergenic region between each copy is 1460 bp (Fig. 2). Thus, the entire genomic interval inclusive of both *Mc1r* copies and the intergenic region is 3404 bp.

We performed BLAST searches of *Mc1r* and Duplicate *Mc1r* and identified two very strong hits (i.e., BLAST score > 1500). The highest sequence similarity for *Mc1r* (970 bp) was recovered at the “*Mc1r*” position (100%; Fig. 2, blue region) and a lower degree of similarity at the “Duplicate *Mc1r*” position (89.61%; Fig. 2, orange region). Conversely, when we performed routine BLAST similarity searches of Duplicate *Mc1r* (975 bp), we recovered the highest sequence similarity at the “Duplicate *Mc1r*” position (100%; Fig. 2, orange region) and a lower degree of similarity at the “*Mc1r*” position (89.61%; Fig. 2, blue region).

We next reasoned that if this was indeed a “duplication” of *Mc1r*, we should see a substantial degree of sequence similarity between both copies. Towards this end, we performed a pairwise sequence alignment. We translated the nucleotide sequence for both copies and found that the open reading frame for Duplicate *Mc1r* produces a predicted protein of 325 amino acids. The Pachón cavefish genome reference of *Mc1r* includes a 2 bp deletion at positions 22 and 23 that leads to a frameshift and premature stop codon, truncating the protein at 105 amino acids (described in Gross et al. 2009). To better compare the structure of the *Mc1r* receptor to the duplicate form of the *Mc1r* receptor, we compared nucleotide and protein structures across surface-dwellers and Pachón cavefish. Both populations are members of the *Astyanax mexicanus* species. We aligned the Pachón form

of *Mc1r* (105 amino acids), the wild type surface fish form of *Mc1r* (324 amino acids), and the Duplicate *Mc1r* protein (325 amino acids; Fig. 3). The alignment of wild type (surface) *Mc1r* to the Duplicate *Mc1r* protein structure revealed 38 amino acid substitutions (Fig. 3).

Both *Astyanax* *Mc1r* copies localize to a single *Mc1r* locus in the genomes of other craniates

To test the assumption that this duplication of the *Mc1r* locus is not widely present across other genomes, we performed a series of comprehensive BLAST searches in sixteen annotated craniate genomes (see Methods). When we BLAST analyzed the “*Mc1r*” sequence, we retrieved two strong hits exclusively from the *Astyanax* genome. Of the 16 genomes assessed, the top hit from 11 genomes was *Mc1r* (Table 1). The only organism for which a match was not returned was in the (most distantly related) lamprey, *Petromyzon marinus*. The top hit from the chicken, *Xenopus*, mouse and coelacanth genomes were *tcf25*, *tubb3*, *Mc3r*, and (unannotated transcript) ENSLACG00000007844, respectively.

When we BLAST analyzed the “Duplicate *Mc1r*” sequence, we again retrieved two strong hits from the *Astyanax* genome at each locus on scaffold KB882081.1. Of the 16 genomes evaluated, the top hit from 12 genomes (including all of the fish genomes) was *Mc1r* (Table 1). Notably, the “Duplicate *Mc1r*” sequence returned a top hit in lamprey (*Petromyzon marinus*) for *Mc1r*. Again, the top returned hits were identical for the chicken, *Xenopus*, mouse and coelacanth genomes of *tcf25*, *tubb3*, *Mc3r*, and ENSLACG00000007844.

Despite the fact that “*Mc1r*” did not return a hit for lamprey, the originally annotated *Mc1r* in *Astyanax* exhibited a higher mean BLAST score to the homologous single copy *Mc1r* in other vertebrates (mean BLAST score = 785.7), whereas the Duplicate *Mc1r* was considerably more divergent to the single copy *Mc1r* of other species (mean BLAST score = 559.6).

We next evaluated the positions of 11 genes surrounding the *Mc1r* locus in *Astyanax*. In the genomic region 5′ of *Mc1r*, we identified the following six annotated genes (most distant to most proximal): *urah*, *pdcd5*, *mvda*, *cyba*, *gas8*, and *tcf25*. In the 3′ region of the *Mc1r* locus, we identified the following annotated genes (most proximal to most distant): *tubb3*, *def8*, *mthfsd*, *foxf1*, *irf8*. When we evaluated the surrounding genomic region to the *Mc1r* locus in our top BLAST hits from 15 additional craniate genomes spanning lamprey to mammals, we discovered many of the same genes present in the same interval (Table 2). No evidence of synteny was recovered in the lamprey genome, but our power to evaluate synteny was limited because *Mc1r* was near the 3′ end of a scaffold.

This finding suggests an ancient syntenic block surrounds the *Mc1r* locus in jawed vertebrates (Table 2). Our synteny analysis, however, helps explain two top returned hits from chicken and *Xenopus* (which were not *Mc1r*). In chicken, the top returned hit was *tcf25*; and in *Xenopus*, the top returned hit was *tubb3*. Interestingly, both genes immediately flank the *Mc1r* locus in *Astyanax* – *tcf25* is in the adjacent 5′ region, *tubb3* is in the adjacent 3′ region. A neighboring gene was returned in both organisms as the top BLAST hit providing support that the homologous region of the genome was correctly identified in every annotated genome, with the exception of mouse. For mouse BLAT searches, *Mc3r* was

returned as the top hit for both *Astyanax Mc1r* and Duplicate *Mc1r*, albeit with a weak similarity score (BLAST score = 65; Table 1) and 40 bp of alignment. Upon further investigation, though, Ensembl annotation clearly indicates that synteny is conserved within mouse for the annotated *Mc1r* and a subsequent BLASTn search using NCBI Genome BLAST indicated the top hit for *Astyanax Mc1r* and Duplicate *Mc1r* was mouse *Mc1r*.

The duplicated *Mc1r* locus is not a different member of the melanocortin gene family

We evaluated whether the Duplicate *Mc1r* locus may be a different member of the *melanocortin* family by examining the top returned BLAT hits within the *Astyanax* genome for both copies of *Mc1r*. Our rationale for this analysis is based in part on the finding teleost genomes that two different melanocortin genes, *Mc2r* and *Mc5r*, are physically linked to one another (Logan et al. 2003). Therefore, it is formally possible that this duplication event did not solely involve the *Mc1r* locus. Our BLAT analysis retrieved 12 hits for *Mc1r*, and 9 hits for Duplicate *Mc1r* (Table 3). *Mc1r* returned some level of sequence similarity to all of the annotated melanocortin family members and one unannotated transcript (ENSAMXG00000018244). However, results showed the highest sequence similarity in the *Astyanax* genome is to both copies of *Mc1r* (Table 3). Moreover, the genomic positions of every other returned hit was quite different from the *Mc1r* copies, i.e., no other returned hit landed on the same genomic scaffold, KB882081.1 (Table 3). We retrieved similar results when we analyzed the 9 returned BLAST hits for Duplicate *Mc1r*, and the *melanocortin* family members *Mc2r* and *Mc5rb* were not returned in this analysis. However, the strongest returned hits were to both *Mc1r* copies on KB882081.1, with much weaker support for other *melanocortin* family members – none of which were localized to the same genomic scaffold.

To further evaluate the degree of similarity between the *Mc1r* copies and other *melanocortin* genes, we compared sequence similarity scores from this analysis (Fig. 4A,B). Our results showed that *Mc1r* yielded high BLAT scores for both *Mc1r* and Duplicate *Mc1r* (1500+), while all of the remaining *melanocortin* genes had BLAST scores between 45 – 282 (Fig. 4A). Similarly, Duplicate *Mc1r* yielded high scores for Duplicate *Mc1r* and *Mc1r* (1500+), with weak BLAST scores for other *melanocortin* genes (BLAST scores between 45 – 240; Fig. 4B).

To visualize the relative sequence similarity across *Astyanax melanocortin* family members, we aligned nucleotide sequences for *Mc1r*, Duplicate *Mc1r*, *Mc2r*, *Mc3r*, *Mc4r*, *Mc5ra* and *Mc5rb* using the ClustalV method (Fig. 4C). Interestingly, the resulting phylogram based on nucleotide sequences demonstrated that both copies of *Mc1r* (Fig. 4C, green) were far more similar to one another than either was to other *melanocortin* family members. Interestingly, the only paralogous melanocortin family member, *Mc5ra* and *Mc5rb*, demonstrates more sequence divergence (Fig. 5C, blue arrowhead) than *Mc1r* and Duplicate *Mc1r*. This finding is consistent with the *Mc5r* duplication being more ancient (and present in distantly-related zebrafish). When we aligned amino acid sequences from the melanocortin family, the topology of the phylogram changed slightly, however the results are quite similar to the nucleotide alignment. Namely, *Mc1r* and Duplicate *Mc1r* are most similar to one another in amino acid sequence, at the exclusion of other melanocortin receptors (Fig. 4D, green).

Additionally, Mc5ra and Mc5rb are more divergent than are the Mc1r and Duplicate Mc1r receptors (Fig. 4D, blue arrowhead).

The genomic interval inclusive of both copies of Mc1r is present in surface and cavefish populations from the Sierra de El Abra

To further validate the predictions of the *Astyanax* genome, we amplified the entire ~3.4 kb interval from representative individuals from the Sierra de El Abra region of northeastern Mexico. This geographic region is home to 29 different populations of cave-dwelling *Astyanax*, as well as the surface-dwelling form (Mitchell et al. 1977). We successfully amplified the entire interval from four representative individuals from the Pachón, Tinaja, and Chica cave localities, as well as the surface fish lineage. All fragments were subcloned and Sanger sequenced using custom sequencing primers (Methods). This analysis confirmed that the ~3.4 kb interval predicted from the cavefish reference genome contains two copies of *Mc1r* (Fig. 5).

When we aligned sequencing reads from Tinaja cavefish, we found a single point mutation (A208G) present in this group. However, compared to the *Mc1r* in the Pachón reference genome, both Tinaja *Mc1r* and Tinaja Duplicate *Mc1r* are predicted to produce a normal full-length protein (972 bp and 975 bp, respectively). The intergenic region (1.46 kb), and the Duplicate *Mc1r* sequence did not differ between Tinaja and the Pachón cavefish reference genome. In commercially-derived Chica cavefish, we found the predicted copy of *Mc1r* (972 bp), however the intergenic region (1458 bp) harbored a 2 bp deletion (Fig. 5, blue arrowhead – Chica) that was not present in the reference genome. Further, the 5' region of the Duplicate *Mc1r* locus harbored a point mutation (T87C; Fig. 5, green – Chica) not present in Tinaja or the reference genome. These results resembled those of our surface fish alignment, wherein we observed a normal *Mc1r* locus (972 bp), and smaller intergenic region (1458 bp) inclusive of two indels (Fig. 5, blue arrowheads – surface fish), one of which was identical to the deletion in Chica. In surface fish, we observed the same T87C point mutation in Duplicate *Mc1r* (T87C; Fig. 5, green – Surface fish), also present in Chica cavefish, which was uniformly present in all sequencing reads.

We aligned the original sequencing reads from our Guerrero analysis, however these reads were not collected across the entire genomic interval. Rather, these reads amplified ~84% of the *Mc1r* open reading frames since they were isolated using degenerate PCR. When we evaluated sequence variation, we found several variable SNPs in the Guerrero cavefish samples (Fig. 5, grey arrowheads – Guerrero cavefish), at positions 247, 335, 464, 573 and 652 in *Mc1r*, and positions 345 and 396 in Duplicate *Mc1r*. Interestingly, these SNPs were variable within the population and always had one allele in common with the *A. mexicanus* cavefish genome reference. In addition, we discovered one Guerrero cave individual that had a 7 bp deletion in 3' region of the Duplicate *Mc1r* open reading frame (Fig. 5., red arrowhead, 723–729 – Guerrero cavefish). This mutation is predicted to cause a frameshift, causing a premature truncation of the Duplicate *Mc1r* protein, shortening it from 325 amino acids to 293 amino acids (Fig. 5B). At present, the size and sequence composition of the intergenic region remains unknown in Guerrero cavefish.

We analyzed our Guerrero surface fish reads and similarly found a number of variable SNPs in the sequenced portion of the *Mc1r* open reading frame, at positions 335, 464 and 573 (Fig. 5, grey arrowheads – Guerrero surface fish). These SNPs were also present in the Guerrero cavefish population. Similarly, we identified variable SNPs in the Duplicate *Mc1r* open reading frame at positions 345 (shared with Guerrero cavefish), 370, 381 and 707 (Fig. 5, grey arrowheads – Guerrero surface fish). No other obvious sequence alterations were present from our analyses. In sum, this cloning analysis revealed that the 3.4 kb genomic interval predicted from the cavefish genome is accurately annotated and present across multiple *Astyanax* species.

Transcriptomic profiling suggests the *Mc1r* duplicate is expressed

We queried previously published, assembled transcriptomes for evidence of *Mc1r* and Duplicate *Mc1r* via BLASTn. We only found hits to the eye transcriptome (which was derived from a Texas surface individual) and skin transcriptome (derived from a Pachón cave individual). In the eye transcriptome, we found a single transcript that spanned the majority of Duplicate *Mc1r* and another that spanned almost the entire open reading frame of *Mc1r*. In the skin transcriptome, two transcripts covered the majority of Duplicate *Mc1r* but only the first half of *Mc1r* yielded a BLAST match. These data were supported by alignment of RNA-seq raw reads to the genome and evaluation in IGV (Figures S1–S4). Nearly the entire open reading frame for Duplicate *Mc1r* is covered by raw reads from the eye and skin, while *Mc1r* is covered by raw reads from the eye but only the 5′ half of *Mc1r* is covered by raw reads in the skin.

Next, we queried raw RNA-seq reads from this dataset with BLASTn to ensure that transcripts were missing in other tissues, the developmental data, and the 3′ end of *Mc1r* in skin. No RNA-seq reads passed ‘percent identity’ or ‘e-value’ thresholds for brain, muscle, or liver tissue. For RNA-seq reads which passed percent identity or e-value thresholds, the open reading frame was covered only in part for either *Mc1r* copy in heart (< 12 % of the open reading frame for either copy), kidney (< 7%), and nasal tissue (< 56%). In contrast in eyes, RNA-seq reads which passed filtering thresholds spanned approximately 90% of the open reading frame length for both copies. Consistent with our other RNA results, it appears that in skin Duplicate *Mc1r* is expressed but *Mc1r* may not be. For example, RNA-seq reads span 98.7% of the open reading frame for Duplicate *Mc1r* but only 83.5% of the *Mc1r*. Because this was a BLASTn search of Pachón cavefish RNA-seq to Pachón cavefish genomic DNA, we reasoned we could increase the percent identity threshold to 98% to remove paralogous hits. When we apply this more stringent filter, RNA-seq reads from skin still span 98.7% of the open reading frame for Duplicate *Mc1r* but only 71.9% of the *Mc1r*, suggesting that some of the originally detected RNA-seq coverage of *Mc1r* in the skin was actually RNA-seq reads from paralogs.

In contrast in the developmental transcriptome assembly data, we found some coverage of both *Mc1r* copies in the developmental RNA-seq reads for Pachón cavefish (98% ID threshold indicates reads cover 85.5% of the Duplicate *Mc1r* open reading frame and 77.8% of the *Mc1r* open reading frame), but less coverage for surface fish (70% coverage of either open reading frame even with most permissive percent identity threshold). The

incongruence between transcriptome assembly and RNA-seq reads may arise from the different read lengths from the developmental dataset (50 bp single-end reads) versus the tissue-specific datasets (100 bp paired-end reads). Thus, the transcriptome assemblies from the tissue-specific datasets may be a more accurate representation of existing transcripts than the ones from the developmental time series, which likely have a less complete transcriptome assembly.

Investigation of a tandem duplicate *Mc1r* in closely related genomes

Since the evolutionary distance between zebrafish and *Astyanax* spans over 250My of divergence (Nakatani et al. 2011), we investigated two other taxa with draft genome sequences to more precisely identify the timing of this tandem duplicate. First, we used BLASTn to query the channel catfish draft genome for a match to zebrafish *Mc1r*. We localized *Mc1r* on scaffold NC_030419.1 with an identity of 79%. Additional hits to zebrafish *Mc1r* in the channel catfish genome were either short (< 250) or lower identity (65%). Thus, we felt confident we identified the locality of *Mc1r* in the catfish genome. Next we queried the two copies of *Mc1r* from *Astyanax* and found that the best hit mapped to the identical location as the zebrafish *Mc1r* with the *Astyanax* Duplicate *Mc1r* exhibiting a slightly lower bit score and slightly lower percent identity than the *Astyanax Mc1r* (76.6% versus 80% identity). This strongly suggests presence of a single genomic copy of *Mc1r* in channel catfish.

We repeated this protocol with a closer relative to *Astyanax*, the long fin white skirt tetra, *Gymnocorymbus ternetzi* (Oliveira et al. 2011). We used a draft SOAPdenovo2 assembly (Li et al. 2009) that was based on 43.16 Gb of data (after filtering for quality and adaptors) generated by 126 bp paired end reads from an Illumina HiSeq2500 in High Output run mode using v3 reagents (raw reads: SAMN03734669). This animal was obtained dead from Aqualand Aquarium Center in Minneapolis, MN in Spring 2015, and DNA was extracted from body tissue. A single contig and single scaffold provided hits to both *Astyanax* Duplicate *Mc1r* and *Astyanax Mc1r* separated by 2462 nucleotides. The scaffold and contig were not sufficiently long to contain the entire length of both copies of *Astyanax Mc1r*. We interpret this result as preliminary evidence as this contig contains only the 5' most 493 nucleotides of *Astyanax* Duplicate *Mc1r* and 3' most 450 nucleotides of *Astyanax Mc1r*. However, this data suggests that additional Characidae, especially among clade 52 from Oliveira et al. (2011), be examined for this tandem duplicate to further narrow the phylogenetic distribution of the *Mc1r* duplication.

Discussion

A novel copy number variant of *Mc1r*

In this report, we describe the novel discovery, to our knowledge, of a copy number variant for the *Mc1r* gene. We discovered this variant through attempts to clone a subfragment of this gene in a natural population of *Astyanax aeneus* fish from the Guerrero state in southern Mexico. We utilized a degenerate PCR approach, which necessarily relaxed the amplification stringency to the targeted region of the gene (Wilkie and Simon 1991). This resulted in the amplification of a second (duplicate) copy alongside the amplification of the

intended copy of *Mc1r*. Subsequent sequence analyses revealed unexpected allelic diversity in our amplified fragments.

With the release of the *Astyanax mexicanus* cavefish draft genome (McGaugh et al. 2014) came an ability to analyze the position and similarity of each allele across the genome. At this point, we discovered that our sequences localized to two distinct genomic regions. This indicated that our interpretation of allelic diversity was actually two different genes occupying separate positions in the genome. Strikingly, these genes are predicted to be very near to one another – separated by an intergenic distance of only 1.4 kb. Each *Mc1r* gene copy shows the highest degree of similarity to the other – and low similarity to other *melanocortin* family members. This indicates that this *Mc1r* duplication is not a genomic translocation of a different *melanocortin* family member, but a tandem duplicate.

Following comprehensive sequence analyses, we found no evidence for a second copy of *Mc1r* in any other published craniate genome. Furthermore, the top returned hit for both copies of the *Astyanax* *Mc1r* was a single locus *Mc1r* in 16 craniate genomes. This locus resided in an ancient syntenic block of genes detected in fifteen of the sixteen genomes we queried.

How did this duplicated copy of *Mc1r* arise over the course of evolution? It is believed that copy number variation occurs much more frequently than other forms of mutagenesis (Hastings et al. 2009). In humans, copy number variants of chromosomal segments and genes are believed to be quite common (Iafrate et al. 2004; Korb et al. 2007), and are presumed to have played key roles in the rapid evolution of the human and great ape lineage (Hastings et al. 2009; Stankiewicz et al. 2004). The duplication of *Mc1r* in *Astyanax* is a short-range repeat (~1.4 kb), implying that this copy number variant arose through an unequal crossing over event (i.e., non-homologous recombination). This duplication was present in *Astyanax aeneus* from Guerrero and *Astyanax mexicanus* from the Pachón cave. When we sequenced across the 3.4 kb interval inclusive of both *Mc1r* copies, the same genomic interval was recovered in *A. mexicanus* surface-dwelling fish, as well as *A. mexicanus* cavefish from the Tinaja and Chica localities. This implies that the *Mc1r* duplication evolved prior to the divergence of Guerrero fish (*Astyanax aeneus*) and extant populations (*Astyanax mexicanus*) of the northern Sierra de El Abra.

Copy number variation has been reported for a few pigmentation genes in natural and domesticated animal systems. For instance, Pielberg et al. (2002) reported high allelic diversity associated with the *KIT* receptor locus in domesticated and wild pigs. Interestingly, variation at these loci, which included triplication of the *KIT* gene in one lineage, appeared to influence the white coat color phenotype in pigs. Similarly, it has been reported that a domesticated canine breed, standard poodles, demonstrate increased risk for ‘squamous cell carcinoma of the digit’ (SCCD). This risk is associated with a ~28 kb-long copy number variant composed of multiple enhancer elements at the *KIT Ligand* locus (Karyadi et al. 2013). Standard poodles with white fur do not harbor the increased risk for SCCD found in standard poodles with black fur. Interestingly, the difference in risk between different colored members of the same breed may be conferred through protective allelic variation at the *Mc1r* locus (Karyadi et al. 2013).

Fontanesi et al. (2009) first reported copy number variation at the pigmentation locus *Agouti*. The Girgentana and Saanen breeds of domesticated goats harbor a copy number variation of >100 kb inclusive of *Agouti*, which is associated with white coat coloration (Fontanesi et al. 2009; Fontanesi et al. 2010). This gene encodes the agouti signaling protein that antagonizes the Mc1r receptor, leading to a pigment switch from eumelanin to pheomelanin. Future analyses will illuminate if and how a *Mc1r* tandem duplicate impacts the pigmentation (or alternate) physiology of the *Astyanax* lineage.

The potential role of Duplicate Mc1r in pigmentation and the brown mutation

Gross et al. (2009) first described a role for *Mc1r* in causing the *brown* mutation in *Astyanax* cavefish. Different *A. mexicanus* cave populations harbor distinct structural alterations to the *Mc1r* coding sequence. For instance, Pachón cavefish demonstrate a 2 bp deletion in the extreme 5' end of the open reading frame that leads to a truncated, non-functional receptor. Two other cave populations (Yerbaniz and Japonés) harbor a hypomorphic point mutation (C490T) that impacts the second intracellular loop of the receptor. Interestingly, functional validation of this receptor (using a morpholino directed against the first 25 bp of *Mc1r*) yielded a highly penetrant, hypomelanic phenotype in zebrafish. This hypomelanic phenotype was rescued upon co-injection of *in vitro*-transcribed surface fish (wild type) *Mc1r* mRNA.

The presence of a duplicated *Mc1r* locus raises the intriguing possibility that the second *Mc1r* copy may influence pigmentation in *Astyanax* cavefish. For instance, the Duplicate Mc1r receptor may perhaps rescue, to some extent, the loss-of-function allele in Pachón and the hypomorphic alleles reported in Yerbaniz and Japonés. If the Duplicate Mc1r receptor is expressed in the same tissues, and exerts a similar function to Mc1r, it may be predicted to rescue the effects of morpholino knockdown in *Astyanax mexicanus*.

A second possibility is that, owing to the local proximity of each *Mc1r* copy to one another, the same promoter sequence(s) may co-regulate expression of both copies. Previous work demonstrated that *Mc1r* expression is reduced in multiple cavefish populations (Pachón, Tinaja, and Chica), some of which harbor sequence mutations (Stahl and Gross 2015). Therefore, assuming that both *Mc1r* copies are functional, in certain caves (such as Pachón) with both coding and presumed regulatory mutations, reduced expression of the receptor may be the most relevant mechanism contributing to reduced pigmentation. This works under the assumption that both copies of *Mc1r*, whether or not they are structurally normal, are expressed at levels so low that they result in the *brown* phenotype.

An additional possibility is that the local duplication has led to two (paralogous) gene copies that have adopted novel functions. This process of gene duplication, neofunctionalization, and evolution has been characterized in a variety of contexts (Lynch and Conery 2000; Monson 2003; He and Zhang 2005). Although copy number variants have not been reported at the *Mc1r* locus to our knowledge, this form of variation has been described at the *Mc4r* locus in *Xiphophorus*, which harbor A, B1 and B2 alleles (Lampert et al. 2010). The wild type, functional copy of *Mc4r* is the A allele, while B1 and B2 alleles which harbor 6 bp and 10 bp deletions near the carboxy terminus. These altered *Mc4r* copies impact the Mendelian *P* locus, which determines when sexual maturity is reached in males, by acting as dominant-

negative mutations serving to delay puberty (Lampert et al. 2010). This mechanism serves to regulate male reproduction through variation in male body size (Smith et al. 2015), while conserving the normal roles governed by *Mc4r* in regulation of body weight and appetite (Lampert et al. 2010).

Perhaps the *Mc1r* locus, as originally described by Gross et al. (2009), mediates pigmentation function while the Duplicate *Mc1r* copy has evolved novel adaptive roles in the *Astyanax* genus. Further, *Mc1r* is known in other systems to have pleiotropic effects on behavior and physiology (Ducrest et al. 2008; Reissmann and Ludwig, 2013), and such phenotypic consequences may have contributed to the divergence or retention of duplicates. We note that Duplicate *Mc1r* appears to be evolving more rapidly than the first-discovered “*Mc1r*” gene since Duplicate *Mc1r* is less similar to *Mc1r* in other species (Table 1). At present, the reason for this potentially accelerated evolution is unclear, but it may be revealed as pleiotropic functions of the duplicate form. Alternatively, this sequence divergence may reflect accumulated neutral changes if this copy is non-functional. Future studies, aimed at functional characterization and spatial expression of both *Mc1r* copies, will inform the precise roles of these novel *Mc1r* paralogues, and inform their evolutionary relevance in the *Astyanax* lineage.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank members of the Gross lab for helpful discussions of this project. We also gratefully acknowledge the help Amanda Powers who assisted with earlier phases of this project. In addition, we are grateful to Marist College undergraduate students enrolled in Genetics (BIOL320) who performed the first direct PCR sequencing of Guerrero fish in 2010. We appreciate the resources provided by the Minnesota Supercomputing Institute, without which this work would not be possible.

Funding sources

JBG is supported by grants from the National Science Foundation (DEB-1457630), and the National Institutes of Dental and Craniofacial Research (NIH-DE025033).

Literature cited

- Awise JC, Selander RK. Evolutionary genetics of cave-dwelling fishes of the genus *Astyanax*. *Evolution*. 1972; 26:1–19. [PubMed: 28555764]
- Ducrest AL, Laurent K, Roulin A. Pleiotropy in the melanocortin system, coloration and behavioural syndromes. *Trends Ecol Evol*. 2008; 23(9):502–510. [PubMed: 18644658]
- Espinasa L, Rivas-Manzano P, Pérez HE. A new blind cave fish population of genus *Astyanax*: Geography, morphology and behavior. *Environ Biol Fishes*. 2001; 62:329–344.
- Fontanesi L, Beretti F, Riggio V, Gómez González E, Dall’Olio S, Davoli R, Russo V, Portolano B. Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet Genome Res*. 2009; 126(4):333–347. [PubMed: 20016133]
- Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall’Olio S, Colombo M, Casadio R, Russo V, Portolano B. An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC Genomics*. 2010; 11(1):639. [PubMed: 21083884]

- Gross JB, Borowsky R, Tabin CJ. A novel role for Mc1r in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. *PLoS Genet*. 2009; 5(1):e1000326. [PubMed: 19119422]
- Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet*. 2009; 10(8):551–564. [PubMed: 19597530]
- He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*. 2005; 169(2):1157–1164. [PubMed: 15654095]
- Iafate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. *Nat Genet*. 2004; 36(9):949–951. [PubMed: 15286789]
- Karyadi DM, Karlins E, Decker B, Carpintero-Ramirez G, Parker HG, Wayne RK, Ostrander EA. A copy number variant at the *KITLG* locus likely confers risk for canine squamous cell carcinoma of the digit. *PLoS Genet*. 2013; 9(3):e1003409. [PubMed: 23555311]
- Kijas JMH, Wales R, Törnsten A, Chardon P, Moller M, Andersson L. *Melanocortin receptor 1 (MC1R)* mutations and coat color in pigs. *Genetics*. 1998; 150(3):1177–1185. [PubMed: 9799269]
- Kirby RF, Thompson KW, Hubbs C. Karyotypic similarities between the Mexican and blind tetras. *Copeia*. 1977; 1977(3):578–580.
- Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE. Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007; 318(5849):420–426. [PubMed: 17901297]
- Lampert KP, Schmidt C, Fischer P, Volff JN, Hoffmann C, Muck J, Lohse MJ, Ryan MJ, Schartl M. Determination of onset of sexual maturation and mating behavior by melanocortin receptor 4 polymorphisms. *Curr Biol*. 2010; 20(19):1729–1734. [PubMed: 20869245]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–1760. [PubMed: 19451168]
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. 2009; 25:2078–2079. [PubMed: 19505943]
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. 2013; 1303.3997.
- Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics*. 2009; 25:1966–1966. [PubMed: 19497933]
- Logan DW, Bryson-Richardson RJ, Pagán KE, Taylor MS, Currie PD, Jackson IJ. The structure and evolution of the melanocortin and MCH receptors in fish and mammals. *Genomics*. 2003; 81(2):184–191. [PubMed: 12620396]
- Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science*. 2000; 290(5494):1151–1155. [PubMed: 11073452]
- Makova K, Norton H. Worldwide polymorphism at the MC1R locus and normal pigmentation variation in humans. *Peptides*. 2005; 26(10):1901–1908. [PubMed: 15979202]
- McGaugh SE, Gross JB, Aken B, Blin M, Borowsky R, Chalopin D, Hinaux H, Jeffery WR, Keene A, Ma L, Minx P. The cavefish genome reveals candidate genes for eye loss. *Nat Commun*. 2014; 5:5307. [PubMed: 25329095]
- Mitchell, RW., Elliott, WR., Russell, WH. Mexican eyeless characin fishes, genus *Astyanax*: Environment, distribution, and evolution. Texas Tech Press; Lubbock, Texas: 1977.
- Monson RK. Gene duplication, neofunctionalization, and the evolution of C4 photosynthesis. *Int J Plant Sci*. 2003; 164(S3):S43–S54.
- Mundy NI. A window on the genetics of evolution: *MC1R* and plumage colouration in birds. *Proc Biol Sci*. 2005; 272(1573):1633–1640. [PubMed: 16087416]
- Nakatani M, Miya M, Mabuchi K, Saitoh K, Nishida M. Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaeon origin and Mesozoic radiation. *BMC Evol Biol*. 2011; 11:177. [PubMed: 21693066]
- Oliveira C, Avelino G, Abe K, Mariguela T, Benine R, Ortí G, Vari R, Ricardo CeC. Phylogenetic relationships within the speciose family Characidae (Teleostei: Ostariophysi: Characiformes)

- based on multilocus analysis and extensive ingroup sampling. *BMC Evol Biol.* 2011; 11:275. [PubMed: 21943181]
- Pielberg G, Olsson C, Syvänen AC, Andersson L. Unexpectedly high allelic diversity at the *KIT* locus causing dominant *white* color in the domestic pig. *Genetics.* 2002; 160(1):305–311. [PubMed: 11805065]
- Reissmann M, Ludwig A. Pleiotropic effects of coat colour-associated mutations in humans, mice and other mammals. *Sem Cell Dev Biol.* 2013; 24(6):576–586.
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nature Biotech.* 2011; 29:24–26.
- Rosenblum EB, Hoekstra HE, Nachman MW. Adaptive reptile color variation and the evolution of the *Mc1r* gene. *Evolution.* 2004; 58(8):1794–1808. [PubMed: 15446431]
- Slominski A, Tobin DJ, Shibahara S, Wortsman J. Melanin pigmentation in mammalian skin and its hormonal regulation. *Physiol Rev.* 2004; 84(4):1155–1228. [PubMed: 15383650]
- Smith CC, Harris RM, Lampert KP, Scharl M, Hofmann HA, Ryan MJ. Copy number variation in the *melanocortin 4 receptor* gene and alternative reproductive tactics the swordtail *Xiphophorus multilineatus*. *Environ Biol Fishes.* 2015; 98(1):23–33.
- Stahl BA, Gross JB. Alterations in *Mc1r* gene expression are associated with regressive pigmentation in *Astyanax* cavefish. *Dev Genes Evol.* 2015; 225(6):367–375. [PubMed: 26462499]
- Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res.* 2004; 14(11):2209–2220. [PubMed: 15520286]
- Switonski M, Mankowska M, Salamon S. Family of melanocortin receptor (MCR) genes in mammals —mutations, polymorphisms and phenotypic effects. *J Appl Genet.* 2013; 54(4):461–472. [PubMed: 23996627]
- Tan CP, Kulju McKee K, Weinberg DH, MacNeil T, Palyha OC, Feighner SD, Hreniuk DL, Van Der Ploeg LH, MacNeil DJ, Howard AD. Molecular analysis of a new splice variant of the human melanocortin-1 receptor. *FEBS Lett.* 1999; 451(2):137–141. [PubMed: 10371153]
- Tsatmali M, Ancans J, Thody AJ. Melanocyte function and its control by melanocortin peptides. *J Histochem Cytochem.* 2002; 50(2):125–133. [PubMed: 11799132]
- Wilkie TM, Simon MI. Cloning multigene families with degenerate PCR primers. *Methods.* 1991; 2(1):32–41.

Figure 1.

Degenerate PCR amplification and Sanger sequencing in Guerrero cave and surface fish reveals the presence of two copies of *Mc1r* in *Astyanax*. Using a gradient PCR approach combined with degenerate primers, a ~820 bp fragment of *Mc1r* was amplified in each of two surface-dwelling Guerrero fish (A). Sequencing analyses revealed one cave-dwelling Guerrero fish harbored a 7 bp deletion in the Duplicate *Mc1r* open reading frame (723–729). This mutation is predicted to cause a frameshift, and premature truncation of the predicted receptor from 325 amino acids, to 293 amino acids in length (B).

Figure 2. A local duplication of *Mc1r* in the cavefish genome. The reference cavefish genome predicts two copies of *Mc1r* on Scaffold KB882081.1. This duplication was discovered based on the alignment of Sanger sequencing reads from a natural population of cave and surface Guerrero fish (green). The first *Mc1r* copy (blue), identical to that described by Gross et al. (2009), resides at KB882081.1:3150542–3151511 and the Duplicate *Mc1r* copy (orange) is directed in the same orientation and resides at position KB882081.1:3148107–3149081. The Duplicate *Mc1r* copy is predicted to produce an intact protein, with an open reading frame of 975 bp. Each *Mc1r* copy shares >89% sequence similarity, and identifies both loci via BLAST analyses (Table 1).

Alignment of Mc1r (Pachón reference), Mc1r and Duplicate Mc1r

Figure 3.

An alignment of predicted protein structures reveals substantial sequence similarity between two copies of Mc1r. The cavefish genome reference form of the Mc1r receptor harbors a 2 bp deletion that causes a frameshift and truncated receptor (105 amino acids in length). Therefore, we aligned this truncated form of Mc1r with the wild type form of Mc1r, and the duplicate form of Mc1r. The resultant ClustalV alignment of these receptor proteins reveals several sequence differences between the full length Mc1r and Duplicate Mc1r receptor (shaded in red).

Figure 4.

Blast analyses demonstrate that *Mc1r* and Duplicate *Mc1r* are much more similar to one another compared to other *melanocortin* family members. The distribution of BLAST scores following sequence analysis of the *Mc1r* open reading frame (A) and the Duplicate *Mc1r* open reading frame (B) indicate that each sequence is highly similar to one another (green), and far less similar to any of the other melanocortin family members (red dashed line). A phylogram depicting the relationships of all melanocortin genes (C) and proteins (D) confirms the similarity in sequence of *Mc1r* and Duplicate *Mc1r* (green asterisk). Note the paralogous forms of *Mc5ra* and *Mc5rb* are more divergent at the nucleotide (C, blue arrowhead) and protein levels (D, blue arrowhead), compared to the two copies of *Mc1r*.

Table 1

Distribution of the top returned BLAST hits of *McIr* and Duplicate *McIr* across multiple vertebrate genomes.

<i>McIr</i>									
Latin Name	Common Name	Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val
<i>Astyanax mexicanus</i>	Cavefish	KB882081.1:3150542-3151511	<i>McIr</i>	Reverse	1	970	970	1909	0
		KB882081.1:3148116-3149049	(Duplicate <i>McIr</i>)	Reverse	28	961	934	1536	0
<i>Danio rerio</i>	Zebrafish	18:31248410-31248693	<i>McIr</i>	Reverse	413	696	284	322	3.30E-88
<i>Gadus morhua</i>	Cod	GeneScaffold_3327:39683-40215	<i>McIr</i>	Forward	158	690	533	709	9.40E-205
<i>Gallus gallus</i>	Chicken	11:19084729-19085045	<i>TCTF25</i>	Forward	164	480	317	422	3.60E-118
<i>Gasterosteus aculeatus</i>	Stickleback	groupII:11981089-11981633	<i>McIr</i>	Forward	119	663	545	707	4.00E-204
<i>Latimeria chalumnae</i>	Coelacanth	JH127067.1:414698-414803	ENSLACG00000007844	Reverse	375	480	106	148	9.60E-36
<i>Lepisosteus oculatus</i>	Spotted Gar	LG23:1553570-1553962	<i>McIr</i>	Reverse	164	556	393	528	2.50E-150
<i>Mus musculus</i>	Mouse	2:172249290-172249329	<i>Mc3r</i>	Forward	441	480	40	65	7.50E-11
<i>Oreochromis niloticus</i>	Tilapia	GL831150.1:2017981-2018479	<i>McIr</i>	Reverse	119	617	499	622	1.30E-178
<i>Oryzias latipes</i>	Medaka	3:20858917-20859489	<i>McIr</i>	Forward	119	691	573	704	3.90E-203
<i>Petromyzon marinus</i>	Lamprey		No hit						
<i>Poecilia formosa</i>	Amazon Molly	KI519643.1:2928121-2928674	<i>McIr</i>	Reverse	119	672	554	768	1.80E-222
<i>Takifugu rubripes</i>	Fugu	scaffold_14:1824386-1824705	<i>McIr</i>	Forward	161	480	320	427	6.70E-120
<i>Tetraodon nigroviridis</i>	Tetraodon	5:4983993-4984312	<i>McIr</i>	Reverse	161	480	320	434	6.50E-122
<i>Xenopus tropicalis</i>	Xenopus	GL172939.1:1340118-1340386	<i>Tubb3</i>	Forward	212	480	269	219	3.20E-57
<i>Xiphophorus maculatus</i>	Platyfish	JH556788.1:140525-141078	<i>McIr</i>	Forward	119	672	554	762	1.60E-220
Duplicate <i>McIr</i>									
<i>Astyanax mexicanus</i>	Cavefish	KB882081.1:3148107-3149081	(Duplicate <i>McIr</i>)	Reverse	1	975	975	1916	0
		KB882081.1:3150551-3151484	<i>McIr</i>	Reverse	33	966	934	1536	0
<i>Danio rerio</i>	Zebrafish	18:31248317-31248404	<i>McIr</i>	Reverse	710	803	94	100	2.90E-21
<i>Gadus morhua</i>	Cod	GeneScaffold_3327:39683-40092	<i>McIr</i>	Forward	163	572	410	504	7.70E-143
<i>Gallus gallus</i>	Chicken	11:19084729-19085045	<i>TCTF25</i>	Forward	169	485	317	393	1.10E-109
<i>Gasterosteus aculeatus</i>	Stickleback	groupII:11981098-11981454	<i>McIr</i>	Forward	133	489	357	458	4.60E-129
<i>Latimeria chalumnae</i>	Coelacanth	JH127067.1:414725-414954	ENSLACG00000007844	Reverse	229	458	230	244	1.40E-64
									74.35

<i>McIr</i>										
Latin Name	Common Name	Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val	%ID
<i>Lepisosteus oculatus</i>	Spotted Gar	LG23:1553643-1553770	<i>McIr</i>	Reverse	361	488	128	172	3.80E-43	82.81
<i>Mus musculus</i>	Mouse	2:172249290-172249329	<i>Mc3r</i>	Forward	446	485	40	64	1.30E-10	90
<i>Oreochromis niloticus</i>	Tilapia	GL831150.1:2018142-2018386	<i>McIr</i>	Reverse	217	461	247	269	3.50E-72	77.33
<i>Oryzias latipes</i>	Medaka	3:20858926-20859280	<i>McIr</i>	Forward	133	487	355	434	5.90E-122	79.72
<i>Petronyazon marinus</i>	Lamprey	GL477446:365641-365782	<i>McIr</i>	Reverse	766	907	142	146	3.40E-35	76.76
<i>Poecilia formosa</i>	Amazon Molly	KI519643.1:2928255-2928665	<i>McIr</i>	Reverse	133	543	411	543	1.10E-154	82
<i>Takifugu rubripes</i>	Fugu	scaffold_14:1824386-1824672	<i>McIr</i>	Forward	166	452	287	331	5.60E-91	78.05
<i>Tetraodon nigroviridis</i>	Tetraodon	5:4983993-4984312	<i>McIr</i>	Reverse	166	485	320	380	9.40E-106	78.44
<i>Xenopus tropicalis</i>	Xenopus	GL172939.1:1340118-1340386	<i>Tubb3</i>	Forward	217	485	269	195	4.80E-50	65.43
<i>Xiphophorus maculatus</i>	Platyfish	JH556788.1:140533-140890	<i>McIr</i>	Forward	132	489	358	486	1.10E-137	82.96

Table 2

Co-distribution of genes in the syntenic genomic interval surrounding *Mc1r/Duplicate Mc1r* locus across diverse taxa.

Latin Name	Common Name	Genes in neighboring syntenic interval to <i>Mc1r/Duplicate Mc1r</i>
<i>Astyanax mexicanus</i>	Cavefish	<i>irf8, foxf1, mthfsd, def8, tubb3, tcf25, cyba, mvda, pdcd5, urah, gas8</i>
<i>Danio rerio</i>	Zebrafish	<i>def8, tcf25, pdcd5, cyba, mvda, gas8, urah</i>
<i>Gadus morhua</i>	Cod	<i>urah, pdcd5, mvda, cyba, tcf25, def8, mthfsd, foxf1, irf8</i>
<i>Gallus gallus</i>	Chicken	<i>gas8, tubb3, def8, tcf25</i>
<i>Gasterosteus aculeatus</i>	Stickleback	<i>irf8, foxf1, mthfsd, def8, cyba, mvda, tcf25, pdcd5, urah</i>
<i>Latimeria chalumnae</i>	Coelacanth	<i>tcf25, def8, gas8</i>
<i>Lepisosteus oculatus</i>	Spotted Gar	<i>mvda, cyba, tcf25, tubb3, def8, foxf1, mthfsd, gas8, irf8</i>
<i>Mus musculus</i>	Mouse	—
<i>Oreochromis niloticus</i>	Tilapia	<i>urah, mvda, pdcd5, cyba, tcf25, def8, mthfsd, foxf1</i>
<i>Oryzias latipes</i>	Medaka	<i>mthfsd, foxf1, def8, tcf25, cyba, mvda, pdcd5, urah</i>
<i>Petromyzon marinus</i>	Lamprey	<i>cfap20, nudt21</i>
<i>Poecilia formosa</i>	Amazon Molly	<i>urah, pdcd5, mvda, tcf25, def8, foxf1, mthfsd, irf8</i>
<i>Takifugu rubripes</i>	Fugu	<i>irf8, mthfsd, foxf1, def8, tcf25, mvda, pdcd5</i>
<i>Tetraodon nigroviridis</i>	Tetraodon	<i>urah, pdcd5, mvda, cyba, tcf25, def8, foxf1, mthfsd, irf8</i>
<i>Xenopus tropicalis</i>	Xenopus	<i>def8, tubb3, tcf25</i>
<i>Xiphophorus maculatus</i>	Platyfish	<i>urah, pdcd5, mvda, tcf25, def8, foxf1, mthfsd, irf8</i>

Table 3

Distribution of the top returned BLAST hits for *McIr* and Duplicate *McIr* against the draft *Astyanax* genome.

<i>McIr</i>							
Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val
KB882081.1:3150542-3151511	<i>McIr</i>	Reverse	1	970	970	1909	0
KB882081.1:3148116-3149049	(Duplicate <i>McIr</i>)	Reverse	28	961	934	1536	0
KB882286.1:27843-28149	ENSAMXG000000018244	Reverse	162	468	307	282	3.30E-76
KB882192.1:421492-421810	<i>Mc5ra</i>	Forward	152	470	319	251	1.10E-66
KB871694.1:294617-294880	<i>Mc4r</i>	Reverse	218	481	264	172	6.60E-43
KB871662.1:105472-105744	<i>Mc3r</i>	Forward	203	475	273	157	1.70E-38
KB882127.1:1237797-1238087	<i>Mc2r</i>	Reverse	211	501	291	146	2.60E-35
KB882127.1:1237541-1237587	<i>Mc2r</i>	Reverse	701	754	54	62	6.90E-10
KB882192.1:422051-422087	<i>Mc5ra</i>	Forward	708	744	37	59	5.80E-09
KB877145.1:3026-3064	<i>Mc5rb</i>	Forward	171	209	39	55	7.50E-08
KB871694.1:294912-294936	<i>Mc4r</i>	Reverse	162	186	25	47	2.90E-05
KB871662.1:105431-105455	<i>Mc3r</i>	Forward	162	186	25	45	1.10E-04
Duplicate <i>McIr</i>							
Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val
KB882081.1:3148107-3149081	(Duplicate <i>McIr</i>)	Reverse	1	975	975	1916	0
KB882081.1:3150551-3151484	<i>McIr</i>	Reverse	33	966	934	1536	0
KB882192.1:421492-421810	<i>Mc5ra</i>	Forward	157	475	319	240	2.00E-63
KB871694.1:294542-294880	<i>Mc4r</i>	Reverse	223	561	339	182	5.60E-46
KB882286.1:28056-28149	ENSAMXG000000018244	Reverse	167	260	94	151	8.00E-37
KB871662.1:106054-106195	<i>Mc3r</i>	Forward	766	907	142	135	5.80E-32
KB871662.1:105472-105502	<i>Mc3r</i>	Forward	208	238	31	53	2.40E-07
KB871694.1:294912-294936	<i>Mc4r</i>	Reverse	167	191	25	47	2.90E-05
KB871662.1:105431-105455	<i>Mc3r</i>	Forward	167	191	25	45	1.10E-04