Gene Duplication in an African Cichlid Adaptive Radiation

Heather E Machado[1,3], Domino A Joyce[2], Christian RL Reilly[1,4], Ginger Jui[1,5], David H Lunt[2], Suzy CP Renn[1§]

[1]Department of Biology, Reed College, Portland, OR 97202, USA
[2]Department of Biological Sciences, University of Hull, Hull HU6 7RX, UK
[3]Currently at Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA, 94305, USA
[4]Currently at Santa Catalina School, Monterey, CA 93940, USA
[5]Currently at Department of Plant and Microbial Biology, University of California, Berkeley, California 94720-3102, USA

[§]Corresponding author

Email addresses:
HEM: machadoheather@gmail.com
DAJ: d.joyce@hull.ac.uk
CRLR: Christian_Reilly@santacatalina.org
GJ: ginger.jui@gmail.com
DHL: d.h.lunt@hull.ac.uk
SCPR: renns@reed.edu

Abstract

Background

Gene duplication is a source of evolutionary innovation and can contribute to the divergence of lineages; however, the relative importance of this process remains to be determined. The explosive divergence of the African cichlid adaptive radiations provides both a model for studying the general role of gene duplication in the divergence of lineages and also an exciting foray into the identification of genomic features that underlie the dramatic phenotypic and ecological diversification in this particular lineage. We present the first genome-wide study of gene duplication in African cichlid fishes, identifying gene duplicates in three species belonging to the Lake Malawi adaptive radiation (*Metriaclima estherae*, *Protomelas similis*, *Rhamphochromis* "chilingali") and one closely related species from a non-radiated lineage (*Astatotilapia tweddlei*).

Results

Using *Astatotilapia burtoni* as reference, microarray comparative genomic hybridization analysis of 5689 genes reveals 134 duplicated genes among the four cichlid species tested. Between 51 and 55 genes were identified as duplicated in each of the three species from the Lake Malawi radiation, representing a 38% – 49% increase in number of duplicated genes relative to the non-radiated lineage (37 genes). Duplicated genes include several that are involved in immune response, ATP metabolism and detoxification.

Conclusions

These results contribute to our understanding of the abundance and type of gene duplicates present in both radiated and non-radiated cichlid fish lineages. The duplicated genes identified in this study provide candidates for the analysis of functional relevance with regard to phenotype and divergence. Comparative sequence analysis of gene duplicates can address the role of positive selection and adaptive evolution by gene duplication, while further study across the phylogenetic range of cichlid radiations (and more generally in other adaptive radiations) will determine whether the patterns of gene duplication seen in this study consistently accompany rapid radiation.

Background

Adaptive radiation, the evolution of genetic and ecological diversity leading to species proliferation in a lineage, is thought to be the result of divergent selection for resource specialization [1-3]. Differential selection in heterogeneous environments can result in adaptive radiation when there is a genetic basis for variability in organisms' success in exploiting alternative resources [1-5]. Examples of such radiations include the Cambrian explosion of metazoans [6], the diversification of Darwin's finches in the Galapagos [7], variations in amphipods and cottoid fishes in Lake Baikal [8], the Caribbean anoles [9], the Hawaiian Silverswords [10] and the explosive speciation of the cichlid fishes in the African Great Lakes [11].

The cichlid fishes are the product of an incredible series of adaptive radiations in response to the local physical, biological and social environment. While cichlids can be found on several continents [12], the most dramatic radiations are those of the haplochromine cichlids in the great lakes of East Africa. This speciose clade exhibits unprecedented diversity in morphological and behavioral characteristics [13] and accounts for ~10% of the world's teleost fish. Interestingly, this clade also includes lineages that have remained in a riverine environment and have not radiated [14].

Classic work by Ohno [15] proposed a prominent role for gene duplication events in evolutionary expansion, despite their frequent loss due to drift [16]. Duplication makes extra gene copies available for dosage effects, subfunctionalization, or neofunctionaliztion [17], with the resultant phenotype potentially contributing to an organism's fitness [for review see 18]. Current genomic research [e.g. primates: 19, 20] supports this, but the ability to compare closely related cichlid lineages that have and have not undergone an evolutionary radiation provides a critical tool for testing the association of gene duplication with adaptive radiation.

We used array-based comparative genomic hybridization (aCGH) to identify gene duplications among 5689 genes for three Lake Malawi radiation species, which began accumulating molecular diversity approximately 5 million years ago [21] (*Metriaclima estherae*, *Protomelas similis*, *Rhamphochromis* "chilingali") and one closely related riverine species from a non-radiated lineage (*Astatotilapia tweddlei*) (Figure 1). This is the first genome-wide study of gene duplication among haplochromine cichlids.

Results

aCGH identification of duplicated genes

A total of 5689 microarray features passed quality control measures in all four test species. Among these, 145 array features (representing 134 genes) were determined to have an increased genomic content (*i.e.* copy number) for one or more heterologous species relative to *A. burtoni* ($P < 0.1$ FDR corrected) (Tables 1, 2). This included duplications of 54 genes in *M. estherae*, 51 in *P. similis*, and 55 in *R.* "chilingali", compared to only 37 in *A. tweddlei*, the species from the non-radiated lineage (Figure 2). The number of duplicated genes identified for the species from the radiated lineage represents a 38% – 49% increase relative to the number of duplicated genes

identified in *A. tweddlei*. Consistent with their shared evolutionary history, shared duplications were prevalent among the three Lake Malawi species, with 11 duplications shared among all three and 16 duplications shared between two of the three species (Figure 2). Five genes had greater gene copy number in all four species relative to *A. burtoni*. Genes found duplicated in only one of the four species were also identified. This included 27 genes in *M. estherae*, 20 in *P. similis*, 24 in *R.* "chilingali" and 27 in *A. tweddlei*. BLAST comparison of array feature sequence similarity to the nucleotide database allows annotation and predicted function for discussion of possible adaptive processes. Based on these annotations, several candidate genes was identified as duplicated in and among lineages. Repeated similarity of functional annotations was noticed, particularly for genes involved in immune response, ATP metabolism and detoxification.

Quantitative PCR verification

Four loci found to be duplicated in one or more test species according to aCGH were chosen for quantitative PCR (qPCR) validation for their observed duplication patterns- one duplicated in all species relative to *A. burtoni*, two duplicated in all three Lake Malawi radiation species and one species-specific duplication (Table 2). Primer pairs that were designed to *A. burtoni* sequence successfully amplified product with a similar or slightly reduced efficiency in each heterologous species tested (Table 2). We estimated the copy number relative to *A. burtoni* for these loci based on the array hybridization ratio, and compared that to the copy number estimated from the qPCR results. Each species with a duplication of a given locus as identified by the microarray analysis also showed significantly increased copy number of that locus according to the qPCR analysis (Figure 3). In addition, the pattern of relative copy number among test species observed in the qPCR analysis, reflected, with few exceptions, the pattern of relative copy number observed in the microarray analysis.

Discussion

Gene duplication is an important source of functional novelty and has a demonstrated role in adaptive evolution [18]. Such adaptations can allow for niche diversification, as has been suggested for thermal adaptation [plants: 22, Antarctic ice fish: 23] and for metabolic novelty [C-4 photosynthesis: 24]. The adaptive radiations of the African cichlid fishes exhibit remarkable niche exploitation in the presence of low levels of sequence divergence [reviewed by 13, 21]. However, little is known regarding the relative number of duplicated genes, nor the identity of duplicated genes, within this group. If there is an increased rate of gene duplication or gene duplicate retention in radiated lineages, or if particular duplications are associated with these lineages, then their pattern and identity could provide insight into the processes facilitating the rapid expansion of the African cichlids. The patterns reported and validated here indicate shared and increased gene duplication within the Lake Malawi radiation compared to a close non-radiating lineage. Based on individual gene names and functional annotations, several candidate genes, including those that are involved in immune response, ATP metabolism and detoxification, are identified as duplicated in and among lineages (Table 1). Some of these gene duplicates may underlie adaptive phenotypic change.

Immune response

The evolution of immune response is a potent factor contributing to the divergence of lineages, resulting from strong selection on certain loci [25-27]. Several genes associated with immune response are found to be duplicated in the Lake Malawi species, including two finTRIM genes

(one duplicated in *P. similis* and the other in both *P. similis* and *R.* "chilingali"). This gene family is known to play a role in immunity against viral infection, and several finTRIM paralogs have been found in teleost fishes, resulting from duplication and positive selection (70 in trout, 84 in zebrafish) [28]. Five major histocompatibility complex (MHC) genes- two MHC class I, two MHC class II, and kinesin-like protein 2- are also found duplicated in one or more of the species from the radiated lineage. The MHC gene family, in addition to being involved in immunity [salmon: 29], has a history of expansion and contraction through duplication and deletion [30]. MHC gene families vary in size among teleosts, with particularly large families in cichlids [31-34]. Additional immune related genes duplicated in the Lake Malawi radiation include an immunoglobulin light chain, small inducible cytokine [associated with the MHC region in stickleback: 35], and sestrin 3. In *A. tweddlei*, the test species from the non-radiated lineage, two immune genes, kallikrein-8 and natural killer cell lecin-type receptor, are also found to be duplicated. The identification of several duplicated immune function genes is consistent with previous work documenting size variability and rapid expansion of immune function gene families [Drosophila: 25, silkworm: 36] that may allow species to invade new niches.

ATP metabolism

ATP metabolism and function is critical to many physiological processes. Two ATP synthases and one ATP transporter are found duplicated among the four species. Subunits G and E of vacuolar ATPases, which couple the energy of ATP hydrolysis to proton transport across intracellular and plasma membranes, are duplicated in *A. tweddlei* and *M. estherae*, respectively. In *R.* "chilingali", the adenine nucleotide translocator (ANT) s598 is found duplicated. This mitochondrial transmembrane protein is the most abundant mitochondrial protein and is integral in the exchange of ADP and ATP between the mitochondria and the cytoplasm. Increased expression of mitochondrial ATP synthase has been found in cold acclimated carp [37] and ANT genes are being studied for their potential adaptive role in thermal acclimation [fugu: 38]. The ATP synthase and transport genes found duplicated in this study could also be associated with acclimation to ecological variation in Lake Malawi or could be associated with other differential metabolic demands.

Detoxification

Selection on duplicated detoxification genes (those involved in the breakdown of toxic compounds) can determine survival in particular environments or can contribute to expansion into new niches. One example is seen in plant-herbivore interactions, where gene duplication has been implicated in the ability of herbivores to detoxify plant defense compounds and prevent exclusion of the herbivore from that food source [39, 40]. We detect duplication of detoxification genes in all three species from the radiated lineage. In *P. similis* and *R.* "chilingali", the sulfotransferase (SULT) gene cytosolic sulfotransferase 3 is found duplicated. SULT genes are detoxifying enzymes that catalyze the transfer sulfonate groups to endogenous compounds and xenobiotics. Once sulfated, compounds may become more easily excreted from the body. In zebrafish, ten SULT proteins have been cloned, two of which show strong activity towards environmental estrogens [41]. Zebrafish SULTs have also been found to act on other xenobiotics [42]. In Atlantic cod, a SULT gene was found to be upregulated in response to polluted water [43]. In *R.* "chilingali", two other genes involved in detoxification, arsenic methyltransferase and ferritin (heavy subunit), are found duplicated. Arsenic methyltransferase converts inorganic arsenic into less harmful methylated species, and ferritin is an iron storage protein that is essential for iron

homeostasis, keeping iron concentrations at non-toxic levels. Another iron-related protein, the iron-sulfur cluster assembly enzyme, was also duplicated in *R.* "chilingali". It is possible that some of these gene duplicates have been retained due to a selective advantage for metabolic breakdown of environmental compounds and toxins.

Gene family membership

Gene families by their very nature reveal a propensity for duplication and duplicate retention of certain genes. One study estimated that 38% of known human genes can be assigned to gene families, based on amino acid sequence similarity [44]. These gene families typically consist of two genes, but the largest gene families can have more than 100 members. In the present study, several of the genes found to be duplicated were members of large gene families, comprised of multiple known genes. These include 40S and 60S ribosomal proteins (duplicated in *R.* "chilingali" and *M. estherae*), claudin 29a (*M. estherae*), GTPase IMAP family member 7 (*P. similis*), C-type lectin domain family 4 (*M. estherae*), high-mobility group 20B (HMG20B) from HMG-box superfamily (*A. tweddlei*), and hox gene cluster genes (all species). Hox genes are important in the regulation of development, and have been found to be associated with differential jaw development in cichlid fishes [45]. An immunoglobulin light chain gene belonging to the largest gene family represented in this study was found duplicated in *P. similis*. Since large gene families are comprised of multiple paralogs and may possess a greater tendency for expansion, it is not surprising that large gene families are well represented in our list of duplicated regions.

qPCR verification

The robust validation of aCGH results using quantitative PCR not only verifies the increased genomic content for all four loci analyzed in test species relative to *A. burtoni,* it also provides a complementary approach that may prove to be a more efficient means to survey candidate loci in future population level analyses. For each locus, the pattern of copy number among the four test species relative to *A. burtoni* is similar to that found by aCGH. However, the absolute copy number estimated by qPCR differs from that estimated with array results. This is particularly true of the DY626766 and DY632057 loci, which showed greater qPCR copy number than predicted, despite the underestimation bias possible for those loci. This discrepancy is likely due to the fact that aCGH will produce an underestimate of true copy number when there is sequence divergence of the heterologous species relative to the platform or that qPCR, like microarray hybridization, provides more accurate relative measures than absolute measures. Nonetheless, even for the two instances in which reduced primer efficiency in the tested heterologous species would have been expected to result in an underestimate rather than an overestimate of copy number, the pattern identified by aCGH was upheld. Regardless of discrepancies in magnitude, our quantitative PCR results demonstrate the validity of this technique for estimation of relative copy number in heterologous species. Therefore, this technique may provide an efficient means to assess copy number variation (CNV) of candidate loci within a larger population in order to illuminate the role of gene duplication on a microevolutionary scale.

Technical considerations

The use of aCGH was initially developed for cancer studies and has been applied to several within species studies, but has less frequently been used to assess between species patterns of gene duplication. Careful consideration of the technical biases and conservative interpretation of the results are warranted [46, 47]. Here, because genomic content for each gene has been assessed

relative to the array platform species *A. burtoni*, those genes that appear to be duplicated in all heterologous species may actually represent a reduction in genomic content in *A. burtoni* due to gene deletion events. We identify five such genes, two annotated as Hox gene cluster genes, one as a Ras-related C3 botulinum toxin substrate gene and two that lack annotation, that appear to be duplicated in all four test species, but which may in fact be deleted in *A. burtoni*. In our study we do not attempt to distinguish between these two scenarios.

The hybridization bias due to sequence divergence of the heterologous species from the platform species is another important consideration for the interpretation of aCGH results. Diverged sequences will hybridize less well to the array feature than *A. burtoni* DNA. Therefore, it follows that duplicated genes for which the paralog is highly diverged will be less likely to be detected as duplicated than duplicated genes with paralogs that are less diverged from the platform species, as found by Machado and Renn [47]. Therefore, older gene duplication events, those with very little purifying selection pressure, and those with strong positive selection in the gene region represented on the array are less likely to be identified, while recent duplication events are more likely to be identified. In this study, we use a recent adaptive radiation so that, whilst strong positive selection on duplicates might be overlooked, the majority of duplications are likely to be identified. We find a pattern of increased gene duplication in these Lake Malawi haplochromines, with 38-49% more genes duplicated than in the non-radiated lineage. Care must be taken in interpreting this increase in the context of adaptive radiation, with three primary considerations. First, only a subset of genes (i.e. those present on the array with available sequence) was tested. Second, gene duplicates may have become fixed in ancestral populations due to neutral processes such as founder events, genetic bottlenecks or drift during the relatively recent evolutionary past. Sequence data from multiple species will be necessary to distinguish neutral vs. adaptive evolutionary processes. Third, due to the shared evolutionary history of the three Lake Malawi species, they cannot be considered independent, as such the tantalizing results of our single comparison of radiated versus non-radiated lineages requires further support before general patterns associated with adaptive radiation can be rigorously discussed. Fortunately, the African cichlids provide such a system with which to undertake this [14].

Conclusions

Only recently have studies begun to examine the patterns of gene duplication and copy number polymorphism across species in natural systems, beyond primates [e.g. 23, 48]. We present the largest analysis thus far of patterns of gene duplication across lineages of the African cichlid radiations. We identify several candidate gene duplicates in four cichlid species and find a pattern of increased gene duplication within the Lake Malawi radiation. While our inference regarding the adaptive value of candidate gene duplicates must be tempered, the results of this study support the hypothesis that gene duplication, particularly of genes related to immune response, ATP metabolism and detoxification, is a characteristic of the Lake Malawi adaptive radiation. Assessment across a greater phylogenetic range of cichlid radiations will identify consistent patterns of gene duplication associated with radiated and non-radiated lineages, and comparative sequence analysis will reveal the potential contribution of natural selection to gene duplicate evolution.

Methods

aCGH identification of duplicated genes

Genomic DNA, extracted from ethanol-preserved field tissue samples by standard ProteinaseK/Phenol protocol, was size reduced by Hydroshear (Genome Solutions/Digilab) to 1 – 5 Kb. DNA (4μg) and labeled with Alexa-Fluors conjugated dCTP by Klenow polymerization (Invitrogen, Bio-Prime). Each species was hybridized twice (in dye swap) against a reference pool of *A. burtoni* genomic DNA using the *A. burtoni* cDNA microarray (GEO platform GPL6416). After a 16 hour hybridization (67.5°C, 3.4X SSC, 0.15% SDS, 1 mM DTT, Cot-1DNA), arrays were washed and scanned (Axon 4100B, Genepix).

Microarray data (GEO series GSE19368) were filtered by omitting features with a lack of sequence information, known ribosomal content, or that had faint array signal (<2 SD above background). Only features that survived this quality control for all eight microarrays were analyzed. Data were corrected for background intensity ("minimum") and were loess normalized within array using 250 conserved features [49]. This corrects for bias introduced by sequence divergence under standard normalization [50]. Duplicated genes were identified as those with increased fluorescence according to the "lmFit" statistical model with "eBayes" correction and FDR adjustment for $P < 0.1$ significance level [51]. The reported results are underestimates of duplication levels, due to the fact that diverged duplicates are less likely to be detected [47]. GEL_{50} measurements [52] indicated that experiments were of similar statistical power (*M. estherae*: 1.80, *P. similis*: 1.95, *R.* "chilingali": 1.61, *A. tweddlei*: 1.89).

Quantitative PCR

Genomic content was validated for four genes using qPCR (Table 3). gDNA concentration was quantified with 1.5X SYBR Green I (Roche Applied Science) on a Nanodrop 3300 (Thermosavant). Triplicate qPCR reactions (Opticon MJ Research) contained 0.75x SybrGreen, 1x Immomix (Biolabs), 200-500 nM primers and 0.2 ng sample DNA in 10 μl reactions (95 °C- 10 min; 35 cycles of: 94 °C- 2 min, 60 °C- 20 sec, 72 °C- 15 sec, and 2 min extension). Copy number relative to *A. burtoni* was calculated as CT, the cycle number at a set threshold relative to the *A. burtoni* standard curve, standardized to an *A. burtoni* copy number of 1. Primer efficiency was calculated with a dilution series for *A. burtoni* DNA and one test species (supp. table S2).

Authors' contributions

SCPR, DHL, DJ conceived of the project. HEM, CRLR, GJ performed the experiments. HEM conducted the analyses. SCPR, HEM, DHL prepared the manuscript.

Acknowledgements

Funded by Murdock Charitable Life Trust and NSF-OIS 0818957. Thanks to Martin J Genner for *Rhamphochromis* "chilingali" samples.

Figures

Figure 1 - Phylogenetic positions of experimental (stars) and reference (circle) taxa
The maximum likelihood tree is based on 1785 bp mitochondrial ND2. Nodes not supported by 50% maximum likelihood SH values are collapsed.

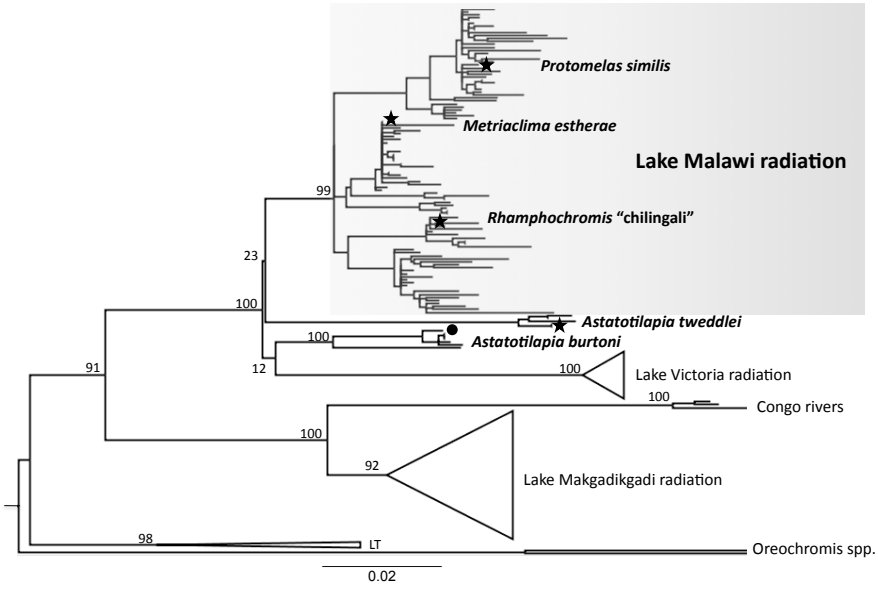

Figure 2 - Genes identified as duplicated among test species (P < 0.1 FDR)
A. twe: *A. tweddlei*; *M. est*: *M. estherae*; *P. sim*: *P. similis*; *R. chi*: *R.* "chilingali". Shared: genes found duplicated in multiple species; Specific: genes found duplicated in only one species; lake: species belonging to the Lake Malawi radiation (*M. estherae*, *P. similis*, *R.* "chilingali"); river: the river species *A. tweddlei*.

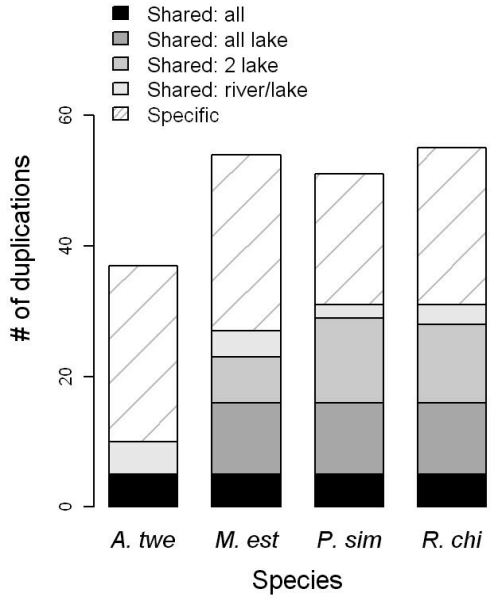

Figure 3 - qPCR validates gene copy number determined by aCGH
Abbreviations are genus and species initials. Primer loci are named for the Genbank number of the *A. burtoni* array feature sequence. ** P <0.1 FDR, * P <0.2 FDR found by array analysis.

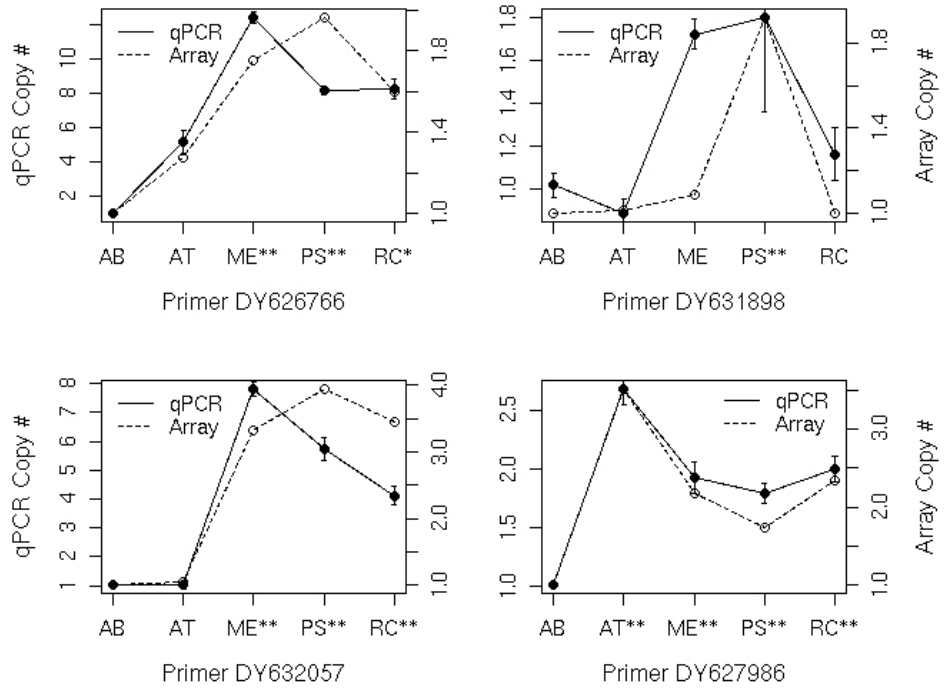

Table 1 - Genes duplicated relative to *A. burtoni* with informative BLAST hits

BitScore: the quality of the alignment for the annotated homology. *A.twe*: *A. tweddlei*; *M.est*: *M. estherae*; *P.sim*: *P. similis*; *R.chi*: *R.* "chilingali"; "ns": not significant; "*": the GenBank number is a representative for multiple array features for that gene.

GenBank	Homology	*A.twe*	*M.est*	*P.sim*	*R.chi*	BitScore
CN468828*	Adenine nucleotide translocator s598	ns	ns	ns	0.60	567
DY630000	Alcohol dehydrogenase Class VI	ns	ns	0.73	ns	379
DY630424	Alkylated DNA repair protein alkB homolog 7	ns	0.43	ns	ns	304
DY629046	Arsenic (+3 oxidation state) methyltransferase	ns	ns	ns	1.06	150
DY626788	ATPase, H+ transporting, lysosomal V0 subunit E	ns	0.76	ns	ns	87.8
DY628437	Claudin 29a (cldn29a) gene	ns	0.60	ns	ns	526
DY632040	Coiled-coil domain containing protein 80	ns	ns	1.19	2.13	434
DY629141	Crystallin gamma M2b	ns	ns	ns	0.43	829
DY626204	C-type lectin domain family 4 member C	ns	0.38	ns	ns	246
DY631088	Cystatin-B	0.45	ns	ns	ns	150
DY630353	Cytosolic sulfotransferase 3	ns	ns	0.62	0.64	713
CN470675	Dazl gene	ns	ns	ns	0.57	89.7
DY629967*	Ferritin heavy subunit	ns	ns	ns	0.82	1160
DY631817	Fish virus induced TRIM protein	ns	ns	0.59	ns	170
DY626596	Fish virus induced TRIM protein	ns	ns	0.41	0.44	145
DY628624	Gamma M7 crystallin	ns	ns	ns	0.42	169
DY630388	Glutamyl-tRNA(Gln) amidotransferase	0.48	ns	ns	ns	347
DY626115	GTPase IMAP family member 7	ns	ns	1.14	ns	370
CN471284	High-mobility group 20B	0.60	ns	ns	ns	163
CN469367	Hox gene cluster	1.34	1.16	0.86	1.11	183
DY627986	Hox gene cluster	1.81	1.12	0.80	1.22	95.1
DY629113	Immunoglobulin light chain	ns	ns	0.65	ns	482
CN468953	Iron-sulfur cluster assembly enzyme ISCU	ns	ns	ns	0.86	610
DY628151	Kallikrein-8 precursor	1.02	ns	ns	ns	102
DY627800	Kinesin-like protein 2 (knsl2)	ns	0.86	1.84	1.14	398
CN469578	KLR1 gene	1.04	ns	ns	ns	154
DY629760	LOC100150543, polyprotein	1.35	ns	0.65	0.79	141
CN468718	LOC100151545, similar to Protein KIAA0284	0.72	ns	ns	ns	145
DY629780	MHC class I	ns	0.84	1.26	1.05	161
DY630620	MHC class IA antigen	ns	ns	0.42	ns	120
DY630701	MHC class II alpha subunit	ns	ns	0.49	ns	764
DY631898	MHC class II antigen alpha chain	ns	ns	0.94	ns	87.8
DY631847	Mitotic spindle assembly checkpoint protein MAD2A	0.60	ns	ns	ns	374
DY627079	Muscle-type creatine kinase CKM2	ns	0.41	ns	ns	787
DY626009	Non-LTR retrotransposon Rex1a	0.70	ns	ns	ns	82.4
DY629391	Non-LTR retrotransposon Rex3_Tet	0.94	ns	ns	ns	122
CN469375*	Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1	ns	0.69	ns	ns	663
DY632057	Pituitary adenylate cyclase activating polypeptide receptor	ns	1.73	1.98	1.79	170
DY628779	Post-GPI attachment to proteins factor 2	ns	0.87	ns	ns	123
DY626114*	Ras association domain-containing protein 4	ns	0.82	ns	ns	1086
DY630104	Ras-related C3 botulinum toxin substrate 2	1.44	0.83	1.47	1.90	331
DY630508	Replication factor C subunit 5	1.04	ns	ns	ns	1234
DY628495	Ribosomal protein, large P2 (60S)	ns	ns	ns	1.01	161
DY630832*	Ribosomal protein S20 (40S)	ns	0.65	ns	ns	663
DY626643	Serine/threonine phosphatase gene	ns	0.57	0.57	0.54	87.8
CN470072	Sestrin 3	ns	1.30	1.61	1.70	116
DY629126	Short coiled-coil protein	ns	ns	ns	0.59	242
DY631649	SINE sequence	ns	0.78	ns	ns	138
DY630540	Small inducible cytokine SCYA102	ns	0.64	ns	ns	1204
CN471492	Solute carrier family 9 (sodium/hydrogen exchanger)	ns	ns	0.63	ns	197
CN471103	Ubiquitin	ns	ns	1.27	ns	985
DY629776	UDP glycosyltransferase 2 family, polypeptide A1	ns	0.92	ns	ns	304
CN469822	Vacuolar ATP synthase subunit G 1	0.79	ns	ns	ns	277

Table 2 - Genes duplicated relative to *A. burtoni* with no informative BLAST hit
A.twe: *A. tweddlei*; *M.est*: *M. estherae*; *P.sim*: *P. similis*; *R.chi*: *R.* "chilingali"; "ns": not significant; "*": the GenBank number is a representative for multiple array features for that gene.

GenBank	*A.twe*	*M.est*	*P.sim*	*R.chi*
DY631067	ns	0.78	0.80	1.02
DY626766	ns	0.81	0.98	0.68
DY629123	ns	0.87	0.83	1.41
DY630373	ns	0.89	1.00	1.23
DY632058*	ns	0.90	0.72	0.71
DY627641	ns	0.76	0.85	ns
DY630229	ns	0.88	0.83	ns
CN471811	ns	1.35	1.22	ns
DY631442	ns	1.40	1.16	ns
CN470857	ns	0.67	ns	0.96
DY632097	ns	0.79	ns	0.82
CN470988	ns	1.28	ns	1.34
CN470402	ns	ns	0.48	0.45
DY631821	ns	ns	0.61	0.60
DY626304	ns	ns	0.99	1.50
DY631315	ns	ns	1.57	1.16
DY629717	ns	ns	1.60	1.28
DY628642	ns	ns	1.62	1.13
DY629912	1.41	2.21	1.06	1.16
DY631507	0.67	0.69	0.78	0.61
DY627911	1.04	0.87	0.49	ns
DY632134*	0.94	0.86	ns	0.84
DY629482	1.39	0.71	ns	1.12
DY630867	0.97	0.64	ns	ns
CN470216	ns	0.39	ns	ns
DY631869	ns	0.39	ns	ns
DY632294	ns	0.41	ns	ns
DY627085	ns	0.44	ns	ns
DY630284	ns	0.54	ns	ns
DY628316	ns	0.64	ns	ns
DY630993	ns	0.67	ns	ns
DY631505	ns	0.72	ns	ns
DY626192	ns	0.75	ns	ns
DY631827	ns	0.86	ns	ns
DY626140	ns	1.05	ns	ns
DY632092	ns	1.09	ns	ns
DY625804	ns	1.23	ns	ns
DY627780	ns	1.51	ns	ns
CN470835	ns	1.55	ns	ns
DY628268	ns	ns	0.46	ns
CN471851	ns	ns	0.47	ns
DY631408	ns	ns	0.50	ns
DY626389	ns	ns	0.57	ns
CN469460	ns	ns	0.64	ns
CN470713	ns	ns	0.68	ns
DY626737	ns	ns	0.79	ns
CN471261	ns	ns	0.93	ns

DY631698	ns	ns	1.03	ns
DY629387	ns	ns	1.18	ns
DY632256	ns	ns	1.56	ns
DY626428	ns	ns	ns	0.39
DY628561	ns	ns	ns	0.42
DY628714	ns	ns	ns	0.48
CN469431	ns	ns	ns	0.50
DY628477	ns	ns	ns	0.58
CN470540	ns	ns	ns	0.60
CN469913	ns	ns	ns	0.63
CN470701	ns	ns	ns	0.65
DY628702	ns	ns	ns	0.67
CN472050	ns	ns	ns	0.70
DY627361	ns	ns	ns	0.74
DY629882	ns	ns	ns	0.77
DY630964	ns	ns	ns	0.95
DY631680	ns	ns	ns	1.02
DY629058	ns	ns	ns	2.41
DY626122	1.50	ns	ns	ns
DY628172	1.38	ns	ns	ns
CN469125*	1.32	ns	ns	ns
DY625919	1.18	ns	ns	ns
DY625845	1.16	ns	ns	ns
DY627087	1.15	ns	ns	ns
CN470724	1.02	ns	ns	ns
DY632007	0.99	ns	ns	ns
DY631850	0.85	ns	ns	ns
DY628517	0.76	ns	ns	ns
CN470646	0.73	ns	ns	ns
DY627338	0.72	ns	ns	ns
CN470597	0.65	ns	ns	ns
CN470781	0.58	ns	ns	ns
DY628148	0.50	ns	ns	ns
DY625884	0.49	ns	ns	ns

Table 3 - Oligonucleotide primers used for qPCR designed against GenBank sequence available for microarray features

Primer Efficiency: percent is based on 4-fold template dilutions for *A. burtoni* and one heterologous test species.

				Primer Efficiency	
GenBank	Primer Sequence	Homology	Predicted Length	*A. burtoni*	Test Species
DY626766	F: TCGGTCTCCTTAACCGGATG	No Hit	193	86	74
	R: CTGAGTTTGGCTGCCCGTAA				*(P. similis)*
DY627986	F: ACGAACACCCGAACGGAAAC	Hox gene cluster	222	100	104
	R: GGTGCACGCACATGAACTGT				*(M. estherae)*
DY631898	F: CGTCCCAGTGAGGATGAGGA	MHC class II antigen	161	82	82
	R: TGATGCTGATCGGTTGATGC				*(R.* "chilingali"*)*
DY632057	F: ATTACTGCGAGTGCCGTCCA	Pituitary adenylate cyclase activating	150	91	78
	R: CTGCGCCCTGAAAGAACAGA	polypeptide receptor 1A			*(A. tweddlei)*

References

1. Dobzhansky T: **Genetics of the evolutionary process**: Columbia University Press; 1937.
2. Mayr E: **Animal species and evolution**. Cambridge, Massachusetts: Belknap Press of Harvard University Press; 1963.
3. Schluter D: **The Ecology of Adaptive Radiation**. Oxford, UK: Oxford University Press; 2000.
4. Slatkin M: **Ecological Character Displacement**. *Ecology* 1980, **61**:163-177.
5. Smith JM: **Sympatric speciation**. *Am Nat* 1966, **100**:637.
6. Gould SJ: **Wonderful Life: The Burgess Shale and the Nature of History**: W.W. Norton; 1989.
7. Darwin C: **The Origin of Species**: Bantam Books; 1859.
8. Fryer G: **Comparative aspects of adaptive radiation and speciation in Lake Baikal and the great rift lakes of Africa**. *Hydrobiologia* 1990, **211**:137-146.
9. Losos JB, Jackman TR, Larson A, de Queiroz K, Rodriguez S: **Contingency and determinism in replicated adaptive radiations of island lizards**. *Science* 1998, **279**:2115-2118.
10. Baldwin BG, Sanderson MJ: **Age and rate of diversification of the Hawaiian silversword alliance (Compositae)**. *Proc Natl Acad Sci U S A* 1998, **95**(16):9402-9406.
11. Fryer G, Iles TD: **The cichlid fishes of the Great Lakes of Africa: Their biology and evolution**: Oliver & Boyd, Croythorn House, 23 Ravelston Terrace, Edinburgh; 1972.
12. Farias IP, Orti G, Meyer A: **Total evidence: Molecules, morphology, and the phylogenetics of cichlid fishes**. *J Exp Zool* 2000, **288**(1):76-92.
13. Kocher TD: **Adaptive evolution and explosive speciation: The cichlid fish model**. *Nat Rev Genet* 2004, **5**(4):288-298.
14. Seehausen O: **African cichlid fish: a model system in adaptive radiation research**. *Proc R Soc Biol Sci Ser B* 2006, **273**(1597):1987-1998.
15. Ohno S: **Evolution by Gene Duplication**: Springer-Verlag; 1970.
16. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**(5494):1151-1155.
17. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations**. *Genetics* 1999, **151**(4):1531-1545.
18. Taylor JS, Raes J: **Duplication and divergence: The evolution of new genes and old ideas**. *Annu Rev Genet* 2004, **38**:615-643.
19. Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T *et al*: **Lineage-specific gene duplication and loss in human and great ape evolution**. *PLoS Biol* 2004, **2**(7):937-954.
20. Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang ZS, Baker C, Malfavon-Borja R, Fulton LA *et al*: **A burst of segmental duplications in the genome of the African great ape ancestor**. *Nature* 2009, **457**(7231):877-881.
21. Genner MJ, Seehausen O, Lunt DH, Joyce DA, Shaw PW, Carvalho GR, Turner GF: **Age of cichlids: New dates for ancient lake fish radiations**. *Mol Biol Evol* 2007, **24**(5):1269-1282.
22. Sandve SR, Rudi H, Asp T, Rognli OA: **Tracking the evolution of a cold stress associated gene family in cold tolerant grasses**. *BMC Evol Biol* 2008, **8**.
23. Chen ZZ, Cheng CHC, Zhang JF, Cao LX, Chen L, Zhou LH, Jin YD, Ye H, Deng C, Dai ZH *et al*: **Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish**. *Proc Natl Acad Sci U S A* 2008, **105**(35):12944-12949.
24. Monson RK: **Gene duplication, neofunctionalization, and the evolution of C-4 photosynthesis**. *Int J Plant Sci* 2003, **164**(3):S43-S54.
25. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG: **Dynamic evolution of the innate immune system in *Drosophila***. *Nat Genet* 2007, **39**(12):1461-1468.
26. Barreiro LB, Quintana-Murci L: **From evolutionary genetics to human immunology: how selection shapes host defence genes**. *Nat Rev Genet* 2010, **11**(1):17-30.
27. Lazzaro BP, Little TJ: **Immunity in a variable world**. *Philos Trans R Soc B Biol Sci* 2009, **364**(1513):15-26.
28. van der Aa LM, Levraud JP, Yahmi M, Lauret E, Briolat V, Herbomel P, Benmansour A, Boudinot P: **A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish**. *BMC Biology* 2009, **7**.
29. Lukacs MF, Harstad H, Grimholt U, Beetz-Sargent M, Cooper GA, Reid L, Bakke HG, Phillips RB, Miller KM, Davidson WS *et al*: **Genomic organization of duplicated major**

histocompatibility complex class I regions in Atlantic salmon (*Salmo salar*). *BMC Genomics* 2007, **8**.
30. Miller KM, Kaukinen KH, Schulze AD: **Expansion and contraction of major histocompatibility complex genes: a teleostean example**. *Immunogenetics* 2002, **53**(10-11):941-963.
31. Malaga-Trillo E, Zaleska-Rutczynska Z, McAndrew B, Vincek V, Figueroa F, Sultmann H, Klein J: **Linkage relationships and haplotype polymorphism among cichlid Mhc class II B loci**. *Genetics* 1998, **149**(3):1527-1537.
32. Miller KM, Withler RE: **The salmonid class I MHC: limited diversity in a primitive teleost**. *Immunol Rev* 1998, **166**:279-293.
33. Persson AC, Stet RJM, Pilstrom L: **Characterization of MHC class I and beta(2)-microglobulin sequences in Atlantic cod reveals an unusually high number of expressed class I genes**. *Immunogenetics* 1999, **50**(1-2):49-59.
34. Sato A, Figueroa F, O'Huigin C, Steck N, Klein J: **Cloning of major histocompatibility complex (Mhc) genes from threespine stickleback, *Gusterosteus aculeatus***. *Mol Mar Biol Biotech* 1998, **7**(3):221-231.
35. Reusch TBH, Schaschl H, Wegner KM: **Recent duplication and inter-locus gene conversion in major histocompatibility class II genes in a teleost, the three-spined stickleback**. *Immunogenetics* 2004, **56**(6):427-437.
36. Tanaka H, Ishibashi J, Fujita K, Nakajima Y, Sagisaka A, Tomimoto K, Suzuki N, Yoshiyama M, Kaneko Y, Iwasaki T *et al*: **A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori***. *Insect Biochem Mol Biol* 2008, **38**(12):1087-1110.
37. Kikuchi K, Itoi S, Watabe S: **Increased levels of mitochondrial ATP synthase beta-subunit in fast skeletal muscle of carp acclimated to cold temperature**. *Fish Sci* 1999, **65**(4):629-636.
38. Itoi S, Misaki R, Hirayama M, Nakaniwa M, Liang CS, Kondo H, Watabe S: **Identification of three isoforms for mitochondrial adenine nucleotide translocator in the pufferfish *Takifugu rubripes***. *Mitochondrion* 2005, **5**(3):162-172.
39. Wen ZM, Rupasinghe S, Niu GD, Berenbaum MR, Schuler MA: **CYP6B1 and CYP6B3 of the black swallowtail (*Papilio polyxenes*): Adaptive evolution through subfunctionalization**. *Mol Biol Evol* 2006, **23**(12):2434-2443.
40. Fischer HM, Wheat CW, Heckel DG, Vogel H: **Evolutionary origins of a novel host plant detoxification gene in butterflies**. *Mol Biol Evol* 2008, **25**(5):809-820.
41. Liu TA, Bhuiyan S, Snow R, Yasuda S, Yasuda T, Yang YS, Williams FE, Liu MY, Suiko M, Carter G *et al*: **Identification and characterization of two novel cytosolic sulfotransferases, SULT1 ST7 and SULT1 ST8, from zebrafish**. *Aquat Toxicol* 2008, **89**(2):94-102.
42. Sugahara T, Yang YS, Liu CC, Pai TG, Liu MC: **Sulphonation of dehydroepiandrosterone and neurosteroids: molecular cloning, expression, and functional characterization of a novel zebrafish SULT2 cytosolic sulphotransferase**. *Biochem J* 2003, **375**:785-791.
43. Lie KK, Lanzen A, Breilid H, Olsvik PA: **Gene expression profiling in Atlantic cod (*Gadus morhua* l.) from two contaminated sites using a custom-made cDNA microarray**. *Environ Toxicol Chem* 2009, **28**(8):1711-1721.
44. Li WH, Gu ZL, Wang HD, Nekrutenko A: **Evolutionary analyses of the human genome**. *Nature* 2001, **409**(6822):847-849.
45. le Pabic P, Stellwag EJ, Scemama JL: **Embryonic Development and Skeletogenesis of the Pharyngeal Jaw Apparatus in the Cichlid Nile Tilapia (*Oreochromis niloticus*)**. *Anat Rec Adv Integr Anat Evo Biol* 2009, **292**(11):1780-1800.
46. Renn SCP, Machado HE, Jones A, Soneji K, Kulathinal RJ, Hofmann HA: **Using comparative genomic hybridization to survey genomic sequence divergence across species: a proof-of-concept from *Drosophila***. *BMC Genomics* 2010, **11**(271).
47. Machado HE, Renn SCP: **A critical assessment of cross-species detection of gene duplicates using comparative genomic hybridization**. *BMC Genomics* 2010, **11**(304).
48. Dopman EB, Hartl DL: **A portrait of copy-number polymorphism in *Drosophila melanogaster***. *Proc Natl Acad Sci U S A* 2007, **104**(50):19920-19925.
49. Salzburger W, Renn SCP, Steinke D, Braasch I, Hofmann HA, Meyer A: **Annotation of expressed sequence tags for the east African cichlid fish *Astatotilapia burtoni* and evolutionary analyses of cichlid ORFs**. *BMC Genomics* 2008, **9**(96):1-14.

50. van Hijum S, Baerends RJS, Zomer AL, Karsens HA, Martin-Requena V, Trelles O, Kok J, Kuipers OP: **Supervised Lowess normalization of comparative genome hybridization data - application to lactococcal strain comparisons**. *BMC Bioinf* 2008, **9**.
51. Smyth GK: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Stat Appl Genet Mol Biol* 2004, **3**:1-26.
52. Townsend JP: **Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays**. *BMC Bioinf* 2004, **5**.