GENETIC VARIATION IN FAST-EVOLVING EAST AFRICAN CICHLID FISHES: AN EVOLUTIONARY PERSPECTIVE

A Dissertation
Presented to
The Academic Faculty

By

Yong-Hwee Eddie Loh

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Biology

Georgia Institute of Technology

August 2011

GENETIC VARIATION IN
FAST-EVOLVING EAST AFRICAN CICHLID FISHES:
AN EVOLUTIONARY PERSPECTIVE

Approved by:

Dr. Todd Streelman, Advisor
School of Biology
Georgia Institute of Technology

Dr. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Soojin Yi, Co-Advisor
School of Biology
Georgia Institute of Technology

Dr. James Thomas
Department of Human Genetics
Emory University

Dr. John McDonald
School of Biology
Georgia Institute of Technology

Date Approved: June 14, 2011

To my dear parents Eric and Florence...

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my academic advisor Dr Todd Streelman for his guidance over these years. His infectious motivation and enthusiam spurred me to continually push beyond established boundaries, allowing me to grow both professionally and personally. I would like to thank my co-advisor Dr Soojin Yi for her unwavering support, advice and encouragement, which started from the very first days of my arrival at Georgia Tech. I would also like to thank my committee members Dr King Jordan, Dr John McDonald and Dr James Thomas for their valuable advice and constructive feedback on my research. Also to Dr Byrappa Venkatesh and Dr Sydney Brenner, I am grateful for the early opportunites to work with them, which inspired me onto this great career.

I am grateful to my friends spread all around the world, and especially those here from the Streelman and Yi Labs, for all the fun, laughter, and of course science, without which this journey wouldn't have been as eventful and fulfilling.

Most of all, I am thankful to my family, from both back in Singapore and here with me in Atlanta. Their love, understanding, support and encouragement made all these work possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CpG	Cytosine immediately followed by Guanine in 5' to 3' direction
DAF	Derived Allele Frequency
DE	*Docimodus evelynae*
LF	*Labeotropheus fuelleborni*
MA	*Melanochromis auratus*
MAF	Minor Allele Frequency
MC	*Mchenga conophorus*
MZ	*Maylandia zebra*
miRNA	MicroRNA
NP	*Nimbochromis polystigma*
PQS	Polymorphic Quality Score
RE	*Rhamphochromis esox*
SNP	Single Nucleotide Polymorphism
TM	*Tyrranochromis maculiceps*
UTR	Un-Translated Region

SUMMARY

Cichlid fishes from the East African Rift lakes Victoria, Tanganyika and Malawi represent a preeminent example of replicated and rapid evolutionary radiation. In this single natural system, numerous morphological (eg. jaw and tooth shape, color patterns, visual sensitivity), behavioral (eg. bower-building) and physiological (eg. development, neural patterning) phenotypes have emerged, much akin to a mutagenic screen. This dissertation encompasses three studies that seek to decipher the underpinnings of such rapid evolutionary diversification, investigated via the genetic variation in East African cichlids.

We generated a valuable cichlid genomic resource of five low-coverage Lake Malawi cichlid genomes, from which the general properties of the genome were characterized. Nucleotide diversity of Malawi cichlids was low at 0.26%, and a sample genotyping study found that biallelic polymorphisms segregate widely throughout the Malawi species flock, making each species a mosaic of ancestrally polymorphic genomes. A second genotyping study expanded our evolutionary analysis to cover the entire East African cichlid radiation, where we found that more than 40% of single nucleotide polymorphisms (SNPs) were ancestral polymorphisms shared across multiple lakes. Bayesian analysis of genetic structure in the data supported the hypothesis that riverine species had contributed significantly to the genomes of Malawi cichlids and that Lake Malawi cichlids are not monophyletic. Both genotyping studies also identified interesting loci involved in important sensory as well as developmental pathways that were well differentiated between species and lineages. We also investigated cichlid genetic variation in relation to the evolution of microRNA regulation, and found that divergent

selection on miRNA target sites may have led to differential gene expression, which contributed to the diversification of cichlid species.

Overall, the patterns of cichlid genetic variation seem to be dominated by the phenomena of extensive sharing of ancestral polymorphisms. We thus believe that standing genetic variation in the form of ancestrally inherited polymorphisms, as opposed to variations arising from new mutations, provides much of the genetic diversity on which selection acts, allowing for the rapid and repeated adaptive radiation of East African cichlids.

CHAPTER 1

INTRODUCTION

The attempt to understand how and what makes organisms different as they originate from common descent has been a central aim of evolutionary biology. Since the dawn of evolutionary research, many animal systems that had displayed adaptive evolution, from Darwin's finches, to the Carribean *Anolis* lizards, to *Drosophila* flies, have been and are still being studied. These studies of genetics and evolution have progressed tremendously over the past century, but detailed knowledge of the forces and mechanisms that lead to the emergence of new species remains a central problem. As we move into the genomic era, advances in molecular technology, applied to the study of closely related taxa, promises to reveal even more into the subtleties of the genetic and mechanistic basis of evolutionary novelty and adaptation. Such studies, applied to the most spectacular extant group of vertebrate radiation, the East African cichlid fishes, would thus be highly informative.

Cichlid fishes from the East African Rift lakes Victoria, Tanganyika and Malawi represent a preeminent example of replicated and rapid evolutionary radiation. Almost 2000 unique species had evolved over a period of just 10 million years. The diversity of species currently observed in each of the major lakes was founded by just one or very few species that had undergone rapid adaptive radiations, leading to flocks of several hundred closely related but phenotypically diverse species. In this single natural system, numerous morphological (eg. jaw and tooth shape, color patterns, visual sensitivity), behavioral (eg. bower-building) and physiological (eg. development, neural patterning) phenotypes have emerged, much akin to a mutagenic screen. Moreover, the recency of this evolutionary radiation has retained high levels of genomic similarity between

species. This background expectation of similarity presents us with a unique opportunity to more efficiently and successfully study and understand basic evolutionary processes and mechanisms by which new species are generated, plus to identify outliers of genetic variation from which we can initiate further studies into the genes and mechanisms that makes organisms distinct.

In Chapter 2, I describe a novel genome sequencing strategy, the generation of low-coverage genomic sequences of five Lake Malawi cichlid species and the identification of single nucleotide polymorphisms (SNPs) among them, performed for the study of genetic variation and diversity in cichlids. This genomic resource, which before then was sorely-lacking and much anticipated by cichlid researchers worldwide, allowed us to obtain a more comprehensive look into the genomic content and structure, as well as the level of genetic variation in cichlids. We successfully genotyped a small test sample of SNPs in Lake Malawi cichlids, which revealed not only the genetic structure differences and inter-relationships between species and lineages, but also identified genes that were well-differentiated between species and lineages. Building upon this successful proof-of-concept study, Chapter 3 describes the extension of genotyping studies to include more SNP and cichlid samples from throughout Africa, from which we obtained further insight into the origins of genetic variation in Lake Malawi cichlids, as well as the genetic relationships and interactions among the entire East African cichlid assemblage. We also identified more well-differentiated genes that should be further investigated in future studies.

In Chapter 4, a different perspective was chosen to study cichlid genetic variation and differentiation, this time concentrating the focus on the evolution of a particular molecular mechanism, microRNA riboregulation. MicroRNAs are an integral class of gene regulators implicated in a diverse range of biological processes and diseases, such as development, cellular proliferation and differentiation, neurogenesis and

neurodegeneration, and many forms of cancer. We hypothesized that divergence of microRNAs or their target sequences might have contributed to phenotypic evolution in Lake Malawi cichlids, and found that indeed, divergent selection had been acting on microRNA target sequences that could lead to differential gene expression.

In totality, this dissertation studied genetic variation at different levels of biological organization. From a broad system-wide perspective, genome-wide variation trends revealed insights into the evolutionary history of the East African cichlid radiation. On the level of molecular mechanisms, which are crucial organism-wide processes affecting proper biological function, we found evidence suggesting that evolution of microRNA regulation had played a role in cichlid diversification. From the gene-specific level of functional genomics, we discovered well-differentiated genes that could possibly affect important phenotypic outcomes. These different perspectives allowed us to gain more comprehensive understanding into genetic variation and it's role in organismal diversification and evolution.

CHAPTER 2

COMPARATIVE ANALYSIS REVEALS SIGNATURES OF DIFFERENTIATION AMID GENOMIC POLYMORPHISM IN LAKE MALAWI CICHLIDS[1]

2.1 Abstract

Cichlid fishes from East Africa are remarkable for phenotypic and behavioral diversity on a backdrop of genomic similarity. In 2006, the Joint Genome Institute completed low coverage survey sequencing of the genomes of five phenotypically and ecologically diverse Lake Malawi species. We report a computational and comparative analysis of these data that provides insight into the mechanisms that make closely related species different from one another.

We produced assemblies for the five species ranging in aggregate length from 68 – 79 Mb, identified putative orthologs for over 12,000 human genes, and predicted more than 32,000 cross-species single nucleotide polymorphisms (SNPs). Nucleotide diversity was lower than that found among laboratory strains of the zebrafish. We collected around 36,000 genotypes to validate a subset of SNPs within and among populations and across multiple individuals of about 75 Lake Malawi species. Notably, there were no fixed differences observed between focal species nor between major lineages. Roughly 3 to 5% of loci surveyed are statistical outliers for F_{ST} within species, between species and between major lineages. Outliers for F_{ST} are candidate genes that may have experienced a history of natural selection in the Malawi lineage.

We present a novel genome sequencing strategy, useful when evolutionary diversity is the question of interest. Lake Malawi cichlids are phenotypically and behaviorally

[1] Loh YH, Katz LS, Mims MC, Kocher TD, Yi SV, Streelman JT. 2008. Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biol*. 9(7):R113.

diverse, but appear genetically like a subdivided population. The unique structure of Lake Malawl cichlid genomes should facilitate conceptually new experiments, employing SNPs to identity genotype-phenotype association, using the entire species flock as a mapping panel.

2.2 Background

Cichlid fishes from the East African Rift lakes Victoria, Tanganyika and Malawi represent a preeminent example of replicated and rapid evolutionary radiation (Kocher 2004). This group of fishes is a significant model of the evolutionary process and the coding of genotype to phenotype, largely because tremendous diversity has evolved in a short period of time among lineages with similar genomes (Won *et al.* 2005, Won *et al.* 2006, Hulsey *et al.* 2007). Recently evolved cichlid species segregate ancestral polymorphism (Moran and Kornfield 1993, Nagl *et al.* 2998) and may exchange genes (Smith *et al.* 2003, Seehausen 2004). Numerous genomic resources have been developed for East African cichlids (many of which are summarized in www.cichlidgenome.org). These include: genetic linkage maps for tilapia (Albertson *et al.* 2003, Kocher *et al.* 1998, Carleton *et al.* 2002) and Lake Malawi species (Albertson *et al.* 2003, Streelman and Albertson 2006); fingerprinted bacterial artificial chromosome libraries (Katagiri *et al.* 2005); EST sequences for Lake Tanganyika and Lake Victoria cichlids (The Gene Index Project; compbio.dfci.harvard.edu/tgi); and first-generation micro-arrays (Kijimoto *et al.* 2005, Renn *et al.* 2004). Many studies have used these resources to study cichlid population genetics, molecular ecology, and phylogeny (reviewed in Kornfield and Smith 2000, Genner and Turner 2005). Recent reports have capitalized on the diversity among East African cichlids to study the evolution and genetic basis of many traits, including behavior (Aubin-Horth *et al.* 2007), olfaction (Blais *et al.* 2007), pigmentation (Streelman *et al.* 2003, Allender *et al.* 2003, Lee *et al.* 2005),

vision (Spady *et al.* 2005, Parry *et al.* 2005), sex determination (Lee *et al.* 2004, Lee *et al.* 2005), the brain (Huber *et al.* 1997) and craniofacial development (Albertson *et al.* 2003, Albertson *et al.* 2005, Streelman and Albertson 2006).

In 2006, under the auspices of the Community Sequencing Program, the Joint Genome Institute completed low coverage survey sequencing of the genomes of five Lake Malawi species. Species were chosen to maximize the morphological, behavioral and genetic diversity among the Malawi species flock. This represents a novel genome project. Low coverage sequencing is now a routine strategy to uncover functional or 'constrained' genomic elements (Margulies and Birney 2008). The rationale is as follows: one compares genome sequence of distantly related organisms (e.g., shark, diverse mammals) to a reference (e.g., human, mouse) and outliers of similarity will be observed against the background expectation of divergence (Kirkness *et al.* 2003, Margulies *et al.* 2005, Venkatesh *et al.* 2007, Pontius *et al.* 2007). Our interests in diversity suggest a conceptually similar, but logically reversed research objective. When the background expectation is similarity, how does one use low coverage genome sequencing to detect that which makes organisms distinct?

Here, we report computational and comparative analyses of survey sequence data to address the question of diversity. We had four major goals: (i) to produce a low coverage assembly for each of the five Lake Malawi species, (ii) to identify orthologs of vertebrate genes in these data, (iii) to predict single nucleotide polymorphisms (SNPs) segregating between species, and (iv) to use SNPs to evaluate the degree of genomic polymorphism and divergence at different evolutionary scales. Consequently, we produced assemblies for the 5 species ranging in aggregate length from 68 – 79 Mb, identified putative orthologs for over 12,000 human genes, and predicted more than 32,000 cross-species segregating sites (with about 2700 located in genic regions). We genotyped a set of these SNPs within and between Lake Malawi cichlid lineages and demonstrate

signatures of differentiation on the background of similarity and polymorphism. Our work should facilitate further understanding of evolutionary processes in the species flocks of East African cichlids. Moreover, the approach we outline should be broadly applicable in other lineages where phenotypic and behavioral diversity has evolved in a short window of evolutionary time.

2.3 Results

2.3.1 Sequence assembly

Trace sequences of five Lake Malawi cichlid species, *Mchenga conophorus* (MC; formerly genus *Copadichromis*), *Labeotropheus fuelleborni* (LF), *Melanochromis auratus* (MA), *Maylandia zebra* (MZ; formerly genus *Metriaclima*) and *Rhamphochromis esox* (RE), were downloaded from the GenBank Trace Archive and assembled into contiguous (contig) sequences. The average cichlid genome is 1.1×10^9 bases (Gregory *et al.* 2007) so the traces represent a sequence coverage of 12 to 17% for each of the five species (see Appendix A Table A1). Through several quality filtering and assembly steps (Methods), the resultant genomic assemblies of the five cichlid species yielded an average of 60,862 contigs with a mean length of 1193 bases per contig. The total first-pass assembly sequence length for each species ranged from 68,238,634 bases (MA) to 79,168,277 bases (MZ), or about 7% of an average cichlid genome. Assembly statistics are shown in Table 2.1.

We noted that these first-pass assemblies were 'over-assembled' by roughly a factor of 2 when compared to theoretical expectations (Lander and Waterman 1988). Theory suggests that random shotgun sequencing of single copy DNA, at 15% coverage of a 1.1 Gb genome, will result in an assembly length of about 153 Mb. We reasoned that our assemblies might be shorter than expected because multi-copy elements were grouped as if they were single copy sequence. Given the theoretical expectation (again for 15%

Table 2.1. First-pass genomic assembly statistics for five Lake Malawi cichlid species.

	MC	LF	MA	MZ	RE
Total number of contigs in assembly	61,923	58,245	63,297	65,094	55,751
Total length (bases)	73,425,564	70,858,381	68,238,634	79,168,277	71,295,074
Genome coverage[a] (%)	6.68	6.44	6.20	7.20	6.48
Mean trace length (bases)	1,055	1,092	991	1,145	1,153
Shortest contig length (bases)	50	50	50	50	50
Longest contig length (bases)	19,632	17,437	21,601	15,371	21,351
Mean contig length (bases)	1,186	1,217	1,078	1,216	1,279
Q25 contig length (bases)	759	846	783	805	934
Q50 (median) contig length (bases)	966	1,063	949	1,163	1,113
Q75 contig length (bases)	1,403	1,355	1,102	1,417	1,407
Total genic length (bases)	2,863,110 (3.9%)	2,841,933 (4.0%)	2,761,941 (4.0%)	2,851,968 (3.6%)	2,797,548 (3.9%)

[a] using an average cichlid genome size of 1.1×10^9 bases. LF, *Labeotropheus fuelleborni*; MA, *Melanochromis auratus*; MC, *Mchenga conophorus*; MZ, *Maylandia zebra*; RE, *Rhamphochromis esox*; Q25, 25th percentile; Q50, median or 50th percentile; Q75, 75th percentile.

coverage of a 1.1 Gb genome) that individual bases should only be sequenced a maximum of 4 to 5 times, we examined whether contigs were built from five or more trace sequences contributing overlapping bases. We observed that about 10 Mb of each first-pass assembly were derived from such contigs, and excluded these data from subsequent analyses (e.g., SNP prediction, see below). Notably, individual sequences contributing to these 'high trace number' contigs were not identified by RepeatMasker but did sometimes have Blast matches to putative repetitive elements (e.g., pol polyprotein, reverse transcriptase). Because of the keen interest in repetitive DNA families in cichlids (Takahashi and Okada 2002) and other organisms (Jordan *et al.* 2003), we have retained alignments of these 'high trace number' contigs and have marked them as such (see Appendix A Table A3 and A4).

2.3.2 Gene content and coverage

To establish the extent of gene content and coverage present in each assembly, we carried out BLASTX similarity searches (10^{-10} E-value cutoff) for each of the five

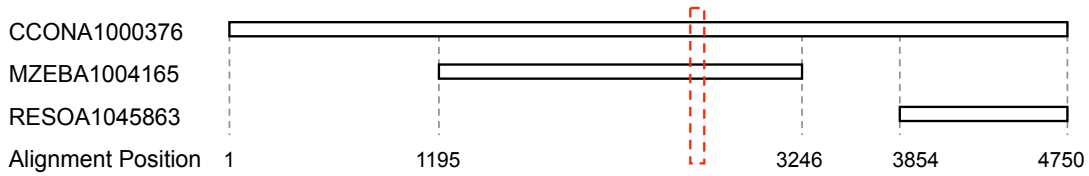
assemblies against a reference human proteome (RefSeq proteins). The average proportion of putative genic sequence amounted to 3.9% of the available genomes. The MZ assembly contained the highest gene coverage, possessing genic loci that were significantly similar to approximately 5,240 unique human proteins. The remaining four species yielded approximately similar numbers ranging from 5,020 to 5,170 genes. It must be noted however that most of these genes are highly fragmented and incomplete, due to the low coverage of the assembly. In all, a total of 36% (12,211 genes out of 34,180; see Appendix A Table A2) of the reference human proteome could be identified in one or more of the cichlid species.

2.3.3 Clustering and alignment

We obtained 25,458 clusters of putatively orthologous sequences, which were individually assembled into multi-species alignments for subsequent comparative analyses. Genic regions, as identified by similarity searches to known human and fish genes, were marked onto each alignment. Figure 2.1 illustrates a typical example of one such alignment.

Roughly 1% of the alignments (294 alignments) showed percentages of variable sites above 2% (about tenfold higher than the average). It is impossible to know, given the low coverage of the sequenced genomes, whether these represent orthologous but divergent regions of cichlid genomes or the alignment of paralogous sequence. We therefore retained these alignments, and included a calculation of polymorphism for each alignment (see Appendix A Table A3), for the consideration of researchers using these data. For example, alignment 108866 contains sequence with similarity to asteroid homologue 1, with 8% of sites variable and a majority of replacement polymorphism. Given the lack of functional information about this novel signaling protein (first described in *Drosophila*; Kotarski *et al.* 1998), this alignment provides useful information even if

A.



B.

```
CC_BSXP22115.b1   ---------- ---------- ---------- ---------- ----------
CC_BSXP22115.g1   ---------- ---------- ---------- ---------- ----------
CC_BSXP25206.b1   ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP35532.b1   ---------- ---------- ---------- ---------- ----------
CC_BSXP36585.g1   ---------- ---------- ---------- ---------- ----------
CC_BSXP38321.b1   ---------- ---------- ---------- ---------- ----------
CC_BSXP4216.x1    ---------- ---------- ---------- ---------- ----------
CC_BSXP4216.y1    ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP46606.x1   nnnnnnnnnn nnnnnnnnnn CAGGCGAATG AAATGCCAGT GAATGTATAT
CC_BSXP46633.y1   ---------- ---------- ---------- ---------- ----------
CC_BSXP46680.x1   ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP5449.x1    accttgTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
CC_BSXP5449.y1    ---------- ---------- ---------- ---------- -annnccaGT
CC_BSXP60653.x2   ---------- ---------- ---------- ---------- ----------
CC_BSXP65585.x2   caggatctta gatcacttca gatcagtgct gcgttggngt nnnnnnnnnn
CC_BSXP78559.x2   ---------- ---------- ---------- ---------- ----------
MZ_BSXW1016.g1    ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW17626.y2   ACattgtgcg tttatatcGT CTggattaat ttggagCACt ggtggacAGT
MZ_BSXW24569.x2   ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW27546.y3   ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW42881.y2   accttgtgct ctta*ttcGT CTGGaTTAGT TTGCAGCACt ggtgCACag*
MZ_BSXW67708.y2   ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW68032.y2   ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
MZ_BSXW70307.g1   ---------- ---------- ---------- ---------- ----------
RE_BSYO72875.g1   ---------- ---------- ---------- ---------- ----------

CCONA1000376      ACATTGTGCT TTTATTTCGT CTGGATTAGT TTGCAGCACT GCTGCACAGT
MZEBA1004165      ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT
RESOA1045863      ---------- ---------- ---------- ---------- ----------

Consensus         ACATTGTGCT TTTATTTCGT CTGGGTTAGT TTGCAGCACT GCTGCACAGT  2701-2750
                  .......... .......... ....^..... .......... ..........
                                          8
```

Figure 2.1. Alignment of a typical cluster of orthologous sequences. (A) Overall alignment of assembly contigs from three different cichlid species with alignment positions indicated. (B) Expanded detail of nucleotide alignment. Filled pink block shows the expanded alignment corresponding to dotted red box in A. Filled blue block shows the alignment of corresponding species' traces that made up the assembly sequences. Lowercase nucleotides have base quality scores under 20. Dashes '-' represent sequence unavailability. Asterisks '*' represent gaps inserted into the sequences. Dots '.' represent identity in alignment. Cap '^' represents segregating site. Alignment positions shown after consensus sequence. Polymorphism quality score shown below A-G single nucleotide polymorphism site.

(and perhaps because) it includes paralogous loci. Another 12% of the alignments (2,119 total) contained individual species contigs that had consensus base positions derived from five or more trace sequences (see above).

For all subsequent analyses, we excluded 2,413 alignments that exhibited (i) a high percentage of variable sites and/or (ii) higher than expected coverage. More than 11.6 million bases of multiple species alignments remain, of which roughly 1.06 Mb were inferred as genic. This included 10,902,011 (986,506 genic) bases of two-species alignments, 721,049 (75,371 genic) bases of three-species alignments, 27,951 (2,898 genic) bases of four-species alignments and 877 (193 genic) bases of alignments containing all five species.

2.3.4 Segregating sites

Further analysis of these 11.6 million bases of multiple alignments identified a total of 32,417 (0.28%) cross-species single nucleotide polymorphisms (SNPs). In order to classify the quality of an identified variable site, a polymorphism quality score (PQS) was defined, corresponding to the first digit of the lowest Phrap quality score among the nucleotides of the different species present at the polymorphic site (e.g., a polymorphic site between four species with base quality scores of 34, 45, 46 and 50 would be assigned a PQS of three). In total, 4,468 (13.8%) variable sites had a PQS of five or higher, 7,952 (24.5%) had a PQS of four, 8,236 (25.4%) a PQS of three, and the remaining 11,761 (36.3%) had a PQS of two. PQS for each variable site are provided on the alignments described in Appendix A Table A3 (also in cichlids.biology.gatech.edu). Nucleotide diversity (Watterson's θ_w) averaged over two-, three- and four-species alignments was 0.00257. Roughly 8% of all polymorphic sites (2,709) were located within the putative genic regions identified earlier. Alignments with fish and human proteins provided us with the phase information required to further classify these into

1,066 synonymous and 1,643 non-synonymous SNPs. Summaries of all alignments containing genic and non-genic polymorphisms are provided in Appendix A Table A3 and A4.

In order to investigate the pairwise differences between any two of the five species, all sequence alignment segments with two or more species were broken up into all possible pairwise alignments; this resulted in 1.06 – 1.55 Mb of alignment per pair. We then calculated the Jukes-Cantor distance between species pairs. The three shortest distances were between LF and MZ (0.229%), followed by MA/MZ (0.232%) and LF/MA (0.241%) and the greatest was between LF and RE (0.288%). These genetic distances include both within-species polymorphism and the fixed differences between species. Currently, there is no exhaustive estimate of within-species polymorphism for Malawi cichlids. Unpublished data from our own group (JT Streelman) indicates that for LF and MZ, within-species diversity (π) may be as high as 0.2%. Thus, the percentage of fixed genetic differences is likely to be extremely small in this assemblage (see following sections).

Finally, we calculated the ratio of replacement to synonymous substitutions (K_a/K_s) for concatenated genic alignments among all pairs of species. We used concatenated sequences because each segment represented only a small fraction of a gene, with only few nonsynonymous and synonymous sites. K_a/K_s ranged from 0.380 in MC/LF to 0.562 in LF/MA. These numbers are greater than the ratios found between *Fugu* and *Tetraodon* (0.127 – 0.144; Jaillon *et al.* 2004). Such high K_a/K_s values may indicate that positive selection, driven by adaptive radiation, is prevalent in cichlid fishes. However, given the expectation of few fixed differences between groups, this topic should be revisited with more data on the levels of segregating and fixed nucleotide substitutions among lineages.

2.3.5 Validation and generality of SNPs

We genotyped 96 SNPs in 384 Lake Malawi cichlid samples using Beckman Coulter SNPstream™ technology. The SNPs were partitioned into three categories to help us evaluate the comparative success rate of automated SNP prediction. First, we included 13 positive controls: genes previously sequenced by others (Spady *et al.* 2005, Won *et al.* 2006) and by us (JT Streelman, unpublished), with expected variation in Malawi cichlids. Positive controls included genes involved in morphogenesis (*otx1*, *otx2*, *pax9*), pigmentation (*mitf*, *ednrb*, *aim1*) and visual sensitivity (opsins *rh1*, *sws1*, *lws*, *sws2a*, *sws2b*). Next, we genotyped 59 SNPs identified using the automated procedure described in this report. We selected these SNPs to represent a range of PQS (from 2 to 5) and a variety of sequence types (genic, non-genic with a BLAST match < e^{-100} to *Tetraodon*, and non-genic with no BLAST match). Finally we wanted to compare our automated SNP selection to a manual approach. Therefore, we included an additional 24 SNPs identified by manual inspection of BLAST matches between single JGI traces and *Tetraodon* chromosome 11; we have previously shown *Tetraodon* 11 to share orthologs with cichlid chromosome 5 (Streelman and Albertson 2006). Note that these SNPs were most often not discovered by our automated procedure because they (i) originated in single traces that did not meet percentage quality cutoffs and/or they (ii) did not align into comparative contigs because of overlap cutoffs.

Our validation strategy sought to document the general use and segregation of these markers among Lake Malawi cichlids. Given recent divergence times among species (some as recent as 1000 years; Won *et al.* 2005), we expected that SNPs might segregate throughout the assemblage. Therefore, Malawi samples comprised about ten individuals from each of ten populations of MZ and LF, as well as one to five individuals of 77 additional species (25 of which were rock-dwelling mbuna). Taxa were included to

represent the morphological, functional and behavioral diversity of the Malawi lineage, which may contain more than 800 species (Turner *et al.* 2001).

Ten out of 13 (about 77%) positive controls gave reliable genotypes and were variable across the dataset. For the 59 SNPs predicted by our automated procedure, 11 were fixed (i.e., no variation) in all samples, indicating an error in sequencing (or genotyping), an error in prediction or the presence of a low frequency allele in the sequenced samples. Six predicted SNPs did not produce data reliable enough for genotype calls. The remaining 42 loci from automated predictions (about 71%) were polymorphic across the data set. For 24 SNPs predicted using manual similarity searches, four were fixed and four failed reliability for genotype calls, with the remaining 16 loci (about 67%) showing polymorphism (Table 2.2). Twelve of 20 (60%) predicted SNPs with PQS of 3 or less were successful while 30 of 39 (76%) predictions with PQS of at least 4 yielded polymorphisms (Table 2.3). There is evidence of ascertainment bias in our genotypic data (see Appendix A Table A5). For example, three SNP loci (Aln100674, Aln114498 and Aln102321) exhibit alleles unique to *Rhamphochromis*. Similarly, SNPs predicted from comparisons of RE and mbuna (LF, MA, MZ) are sometimes fixed in mbuna. Polymorphisms predicted from comparisons of mbuna taxa are more likely to vary within LF and MZ populations and across mbuna species.

Table 2.2. SNP genotyping success categorized by detection method.

SNP Detection Method	Control Genes	Automated	Manual Blast
Number of genotyped loci	13	59	24
Number of polymorphic loci	10	42	16
Number of fixed loci	3	11	4
Number of failed loci	0	6	4
Successful SNP detection (%)	76.9	71.2	66.7

BLAST, Basic Local Alignment Search Tool; SNP, single nucleotide polyorphism.

Table 2.3. SNP genotyping success categorized by polymorphic quality score.

Polymorphic Quality Score	2	3	4	5
Number of genotyped loci	5	15	28	11
Number of polymorphic loci	2	10	24	6
Number of fixed/failed loci	3	5	4	5
Successful SNP detection (%)	40	66.7	85.7	54.5

SNP, single nucleotide polyorphism.

2.3.6 Genetic polymorphism and divergence at multiple scales

Strikingly, among all 68 loci showing polymorphism, no SNP locus was alternately fixed between LF and MZ, nor between rock-dwelling mbuna and non-mbuna. We thus sought to investigate the degree of polymorphism versus divergence at multiple evolutionary scales. The data (Appendix A Table A5) support previously reported population structure in MZ (Danley *et al.* 2000, Streelman *et al.* 2007) and LF (Arnegard *et al.* 1999), as well as the genetic distinction between these species (MC Mims, unpublished). For example, mean genetic differentiation (F_{ST}) in MZ is 0.148 and in LF is 0.271. Mean F_{ST} between LF and MZ was 0.215 and between mbuna (25 species) and non-mbuna (52 species) was 0.224, demonstrating that most genetic variation segregates within and not between lineages, regardless of evolutionary scale. Nevertheless, these distributions of F_{ST} yielded statistical outliers, which are indicative of genetic differentiation (Figure 2.2). Four loci were found to be statistical outliers for F_{ST} among MZ and LF populations. In MZ, opsin loci *lws* (F_{ST} = 0.514), *sws1* (0.572) and *rh1* (0.733) and in LF, opsin locus *rh1* (0.853) exhibit differentiation between populations. Between LF and MZ, three loci were identified as outliers: a non-synonymous polymorphism in *csrp1* (F_{ST} = 0.893), a synonymous polymorphism in *β-catenin* (Aln101106_1089, F_{ST} = 0.904), and an intronic polymorphism in *ptc2* (Aln100281_1741, F_{ST} = 0.863). Two statistical outliers were identified for F_{ST} between rock-dwelling mbuna and non-mbuna groups: a non-synonymous polymorphism in *irx1* (Aln102504_1609, F_{ST}

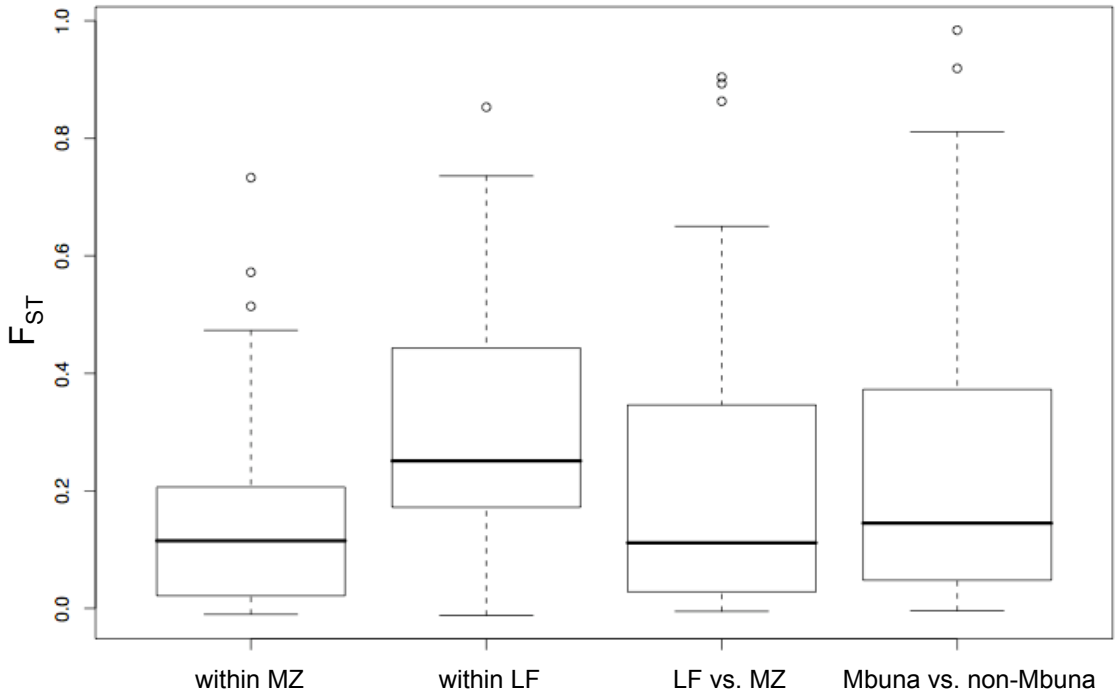


Figure 2.2. Box-and-whisker plots of F_{ST} values calculated for: within MZ, within LF, LF versus MZ and Mbuna versus non-Mbuna. Upper and lower box bounds represent 75th and 25th percentiles, respectively. The solid lines within boxes represent the median value. Whiskers mark the furthest points from the median that are not classified as outliers. Unfilled circles represent outliers that are more than 1.5 times the interquartile range higher than the upper box bound. F_{ST}, genetic differentiation; LF, *Labeotropheus fuelleborni*; MA, *Melanochromis auratus*; Mb, megabases; MC, *Mchenga conophorus*; MZ, *Maylandia zebra*.

= 0.984), and a non-genic polymorphism (Aln103534_280, F_{ST} = 0.919) in sequence with similarity to pufferfish and stickleback genomes between *contactin 3* and *ncam L1*.

2.3.7 Genetic clustering and ancestry

To further visualize the segregation of SNPs across the Malawi cichlid flock, we utilized a Bayesian approach that assigns individuals to a predefined number of genetic clusters (Pritchard *et al.* 2000). Specifically, we were interested in how species would be assigned to major Malawi cichlid lineages identified in previous studies (Won *et al.* 2006, Hulsey *et al.* 2007, Kocher *et al.* 1995). There are three such groups supported by the majority of molecular data: (i) the rock-dwelling mbuna, (ii) pelagic and sand-dwelling species, and (iii) a group comprised of *Rhamphochromis*, *Diplotaxodon* and other deep-water taxa. Analysis of 68 SNP loci accurately classifies species to respective lineages (Figure 2.3). For instance, all species considered mbuna (blue) cluster with other mbuna, to the exclusion of other groups; species thought to represent the earliest divergence within the species flock (*Rhamphochromis*) clustered together as a separate group (green); all remaining non-mbuna species formed the third group (red). Notably, deepwater genera *Diplotaxodon* and *Pallidochromis* contain individuals with mosaic genomes (red and green) and *Astatotilapia calliptera*, a non-endemic species and possible Malawi ancestor (Seehausen *et al.* 2003) combines mbuna and non-mbuna genomes.

For comparison, additional analyses were performed setting the predefined number of genetic clusters to from two to five. When set to two genetic clusters, species were accurately classified as mbuna or non-mbuna. At settings of four or five, the program was unable to yield stable classification results between replicate runs. Thus these latter three sets of analyses (data not shown) did not provide any further insights into the genetic lineages of Malawi cichlids.

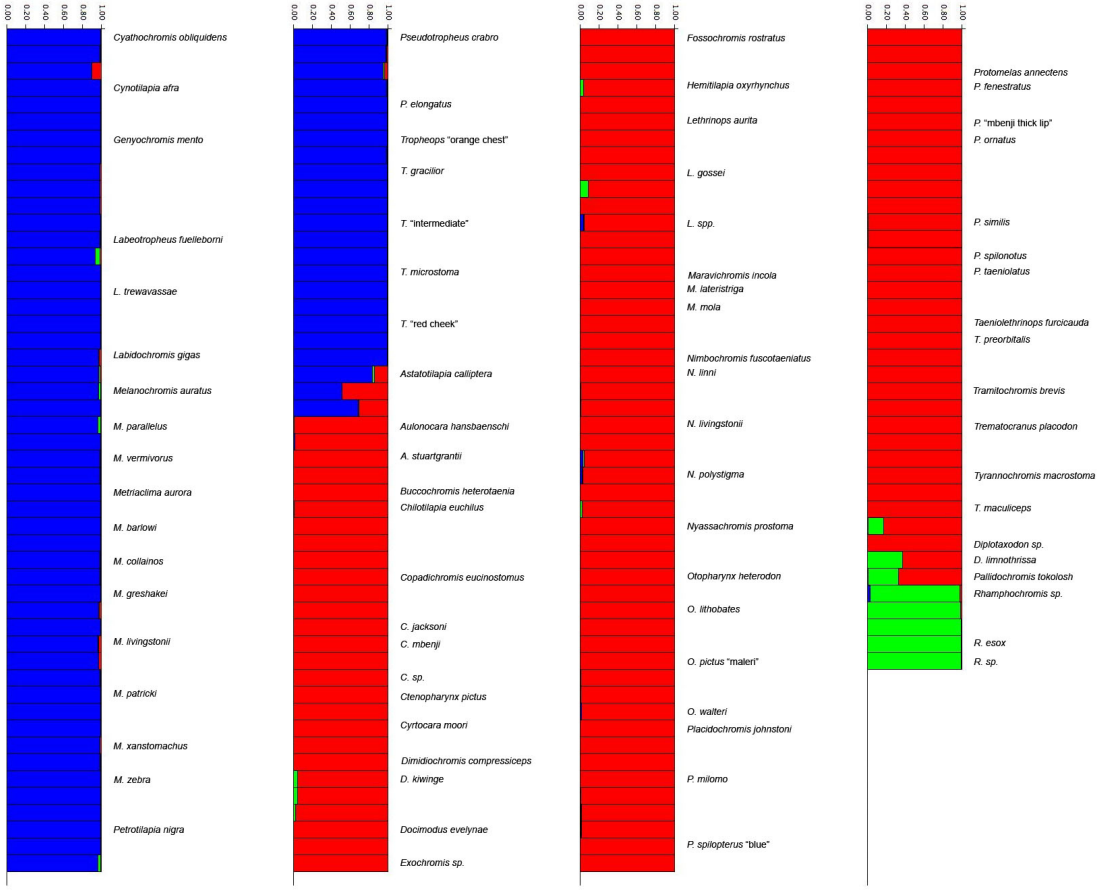


Figure 2.3. Bayesian assignment of Lake Malawi cichlids to different evolutionary lineages. We show the contribution to each individual genome (q, which ranges from 0 to 100%) from each of K = 3 predefined genetic clusters (blue, red, green), for data derived from single nucleotide polymorhisms (SNPs) in Tables 2.2 and 2.3. Note that this method predefines the number, but not the identity of genetic clusters. Species names are written once; multiple individuals from species are grouped together (for example, four individuals of *Pseudotropheus crabro*). Species considered mbuna (blue) cluster with other mbuna, to the exclusion of other groups; species thought to represent the earliest divergence within the species flock (*Rhamphochromis*) clustered together as a separate group (green); and all remaining non-mbuna species formed the third group (red).

2.4 Discussion

African cichlid fishes are important models of evolutionary diversification in form and function (Streelman *et al.* 2007). They are singularly remarkable for the extent of phenotypic and behavioral diversity on a backdrop of genomic similarity. Lake Malawi is home to the most species rich assemblage of African cichlids; as many as 800 – 1000 species are thought to have evolved from a common ancestor in the last 500K to 1MY (Turner *et al.* 2001). These recently formed species segregate ancestral polymorphism and exchange genes by hybridization (Moran and Kornfield 1993, Smith *et al.* 2003, Streelman *et al.* 2004). Such circumstances present both opportunities and challenges for understanding evolutionary history and biological diversity. Opportunistically, researchers have used molecular markers across studies to interrogate the genetic basis of phenotypic differentiation (Streelman *et al.* 2003, Lee *et al.* 2005, Albertson *et al.* 2005, Streelman and Albertson 2006). This approach views Malawi cichlid species as natural mutants screened for function by natural selection; with essentially identical ancestral genomes honed by contrasting historical processes. By contrast, the task of reconstructing a phylogeny of species has been hindered by the very same phenomena of genomic similarity and mosaicism (Won *et al.* 2005, Won *et al.* 2006); even the promising approach of Amplified Fragment Length Polymorphism (AFLP) does not provide strong resolution of the relationships among genera (Albertson *et al.* 1999, Allender *et al.* 2003, Seehausen *et al.* 2003, Kidd *et al.* 2006). The data we present here should provide new resources and perspectives for cichlid evolutionary genomics.

2.4.1 Cichlid species exhibit genomic polymorphism

Lake Malawi cichlid species sequenced by the JGI embody the phylogenetic, morphological and behavioral diversity found within the assemblage. *Rhamphochromis esox* is a large (about 0.5m) pelagic predator representing one of the basal lineages of

the species flock (Kocher *et al.* 1995, Won *et al.* 2006, Hulsey *et al.* 2007). *Mchenga conophorus* is a sand-dwelling species that breeds on leks where males construct 'bowers' to attract females. *Melanochromis auratus*, *Maylandia zebra* and *Labeotropheus fuelleborni* are rock-dwelling (mbuna) species that differ in color pattern, trophic ecology, body shape and craniofacial morphology (for pictures of these and others, see malawicichlids.com).

Our data confirm the conclusions from previous genetic analyses on a smaller scale: Lake Malawi species are genetically similar. Nucleotide diversity observed among the 5 cichlid species (Watterson's θ_w = 0.26%) is less than that found among laboratory strains of the zebrafish, *Danio rerio* (Watterson's θ_w = 0.48%; Guryev *et al.* 2006). Although overall nucleotide diversity is less than that observed in *Danio*, the ratio of replacement to silent change is nearly fivefold higher in the Lake Malawi genomes. Such a result might suggest that East African cichlid evolution is characterized by adaptive molecular evolution, as has been indicated in a few instances (Terai *et al.* 2002, Spady *et al.* 2005), or a relaxation of purifying selection attributable to small effective population size. However, we should view this estimate of K_a/K_s with caution, because of one of the remarkable features of these data (below). Variable sites identified from cross-species alignments are not substitutions fixed between species. The K_a/K_s approach to identifying selection may be largely inappropriate for such young species where ancestral alleles segregate as polymorphisms.

The pattern of variation observed across the approximately 75 species genotyped in this study demonstrates that biallelic polymorphisms segregate widely throughout the Malawi species flock. SNPs segregate within and between MZ and LF populations, as well as within and among mbuna species and other lineages. No SNP locus surveyed is alternately fixed in LF versus MZ, nor between mbuna and non-mbuna. Remarkably, the degree of genetic differentiation (F_{ST}) within species is roughly equivalent to that

between species and to that between major lineages. Lake Malawi cichlid species are mosaics of ancestrally polymorphic genomes. Add to this a propensity of recently diverged species to exchange genes (Won *et al.* 2005), and Malawi cichlids present a case of complex and dynamic evolutionary diversification, where recombination and the sorting of ancestral polymorphism may be more important than new mutation as sources of genetic variation. Despite allele sharing, SNP frequencies contain a clear signal of ancestry for the entire flock. Rock-dwelling mbuna comprise a genetic cluster, as do pelagic and sand-dwelling species, in addition to *Rhamphochromis*. Notably, *Astatotilapia calliptera*, one of a few non-endemic haplochromines in Lake Malawi, appears to retain a reservoir of ancestral polymorphisms from which mbuna and non-mbuna genomes have emerged.

2.4.2 Genomic polymorphism and the divergence of Malawi cichlids

Our hierarchical sampling design allows us to ask if there are loci exhibiting extreme genetic differentiation against the background of shared polymorphism (i) within species, (ii) between species and (iii) between major lineages. Strikingly, regardless of the evolutionary scale, statistical outliers comprise approximately 3 to 5% of loci surveyed. Opsin loci *lws*, *rh1* and *sws1* are differentiated among populations of LF and MZ, adding to reports that opsin polymorphisms are associated with population-specific color patterns or visual environments (Carleton *et al.* 2005).

Single nucleotide polymorphisms in *csrp1*, *β-catenin*, and *ptc2* exhibit greater than expected differentiation between LF and MZ. *Csrp1* (cysteine-rich protein) is a vertebrate LIM-domain family member acting in the non-canonical WNT pathway, expressed in gut, intestine and cardiac mesoderm (Miyasaka *et al.* 2007). *β-catenin* acts to transduce signals in the canonical WNT pathway (Chenn and Walsh 2002) and is expressed in developing cichlid fins, dentitions, brains and lateral lines (GJ Fraser and JT Streelman,

unpublished). Patched is a receptor for sonic hedgehog (Koudijs *et al.* 2008); *shh* is expressed in developing cichlid dentitions, jaws and brains (GJ Fraser, JB Sylvester and JT Streelman, unpublished). A SNP in *irx1* nearly perfectly differentiates rock-dwelling mbuna from the remainder of the Malawi species flock. *Irx1* acts to position the boundary between the telencephalon and the posterior forebrain (Scholpp *et al.* 2007). Finally, a SNP located between *contactin 3* and *ncam L1* exhibits differentiation between mbuna and non-mbuna lineages; these genes are linked in other genomes and functionally interact to pattern dendritic branching in the neocortex (Ye *et al.* 2008). Taken together, these genes are interesting in the context of cichlid diversification because they affect the phenotypes that vary among lineages: color and vision (Spady *et al.* 2005, Parry *et al.* 2005), guts (Reinthal 1990), dentitions (Streelman and Albertson 2006, Fraser *et al.* 2008), jaws (Albertson *et al.* 2003, Albertson *et al.* 2005) and brains (Huber *et al.* 1997).

2.4.3 Discovery for evolutionary biology

There are obvious challenges when attempting to extract information from low coverage genomic sequence, and also obvious payoffs (Kirkness *et al.* 2003, Margulies *et al.* 2005, Venkatesh *et al.* 2007, Pontius *et al.* 2007). Most previous studies have used this information for species-specific discovery (e.g., dog breeds) or broad evolutionary comparisons with respect to a reference genome (e.g., dog-human, shark-human, cat-mammal). Our goals in the present analysis stem from the unique characteristics of Lake Malawi cichlids; these are biological species that behave genetically like a single subdivided population. Therefore, our biggest challenge was to devise a strategy that retains information from these low coverage survey sequences (75% genomic covereage spread over five closely related species), but minimizes error and bias in assembly and cross-species alignment for SNP identification. For example, we excluded many contigs because they appeared to be over-assembled, and we excluded multi-

species alignments if they exceeded a polymorphism threshold. The over-assembly problem limits the coverage of these genomes in relation to expectation; this phenomenon, observed in the cat genome and in simulation, has complex and varying causes and has yet to be fully resolved (Greep 2007). It is likely to be mitigated to some degree by comparison to a higher-coverage reference sequence. The power of the data we present comes from the broad utility of the genic sequences and SNPs we have identified for many questions in genomic evolutionary biology.

Our analyses identified about 12,000 Lake Malawi cichlid sequences with similarity to human and fish proteins. This is a significant advance in our understanding of cichlid genomic content. To put this in context, approximately 13,500 unique ESTs, from three different East African cichlids, represent the sum total of such publicly released sequences (The Gene Index Project; compbio.dfci.harvard.edu/tgi). Our contribution roughly doubles the available data.

The approximately 32,000 (2,700 genic) SNPs we identified should provide a wealth of molecular markers for studies of population genetics and molecular ecology, linkage and QTL mapping, association mapping and phylogeny. We convert about 70% of predicted SNPs to polymorphic markers; this percentage is comparable to other studies from white spruce (74 to 85% depending on quality cutoffs; Pavy *et al.* 2006), zebrafish (65%; Guryev *et al.* 2006) and cow (43%; Moon *et al.* 2007). We have shown these biallelic markers to be of general use, many segregating across the major cichlid lineages of Lake Malawi. We used the SNPs to assign Malawi species to ancestral genetic clusters, and this approach should hold promise for similar questions of genetic structure that span the population *vs*. species continuum. It is important to note that early runs of this analysis, with fewer SNP loci, resulted in stable results with more individuals showing mosaic genomes. This suggests that careful consideration should be paid to the number of polymorphic loci necessary to yield confidence in evolutionary interpretation.

As more SNP loci (with known genome coordinates) are assayed, it will be possible to compute and compare ancestry proportions across scales (e.g., genome *vs*. chromosome *vs*. gene cluster).

Notably, we have used the background level of genomic similarity and polymorphism to identify loci that may have experienced a history of selection within species, between species and between major lineages. Because SNP markers are (i) co-dominant, (2) easy to genotype, (3) reliable and reproducible from lab to lab and (4) readily mapped *in silico* (NHGRI will sequence a related cichlid, the tilapia, to 7-fold draft assembly coverage in 2008) they are likely to complement microsatellites and AFLP for most applications in cichlid evolutionary genomics. Given the unique mosaic structure of Lake Malawl cichlid genomes, it is exciting to envision experiments employing SNPs to identity genotype-phenotype associations, using the entire species flock as a mapping panel. Finally, as sequencing costs continue to drop, the approach we outline here should prove applicable to those studying evolutionary and phenotypic diversity among closely related species (Streelman *et al.* 2007).

2.5 Materials and methods

2.5.1 Samples

Individuals of *Mchenga conophorus* (MC), *Labeotropheus fuelleborni* (LF), *Melanochromis auratus* (MA), *Maylandia zebra* (MZ) and *Rhamphochromis esox* (RE), were sampled from the wild during an expedition to Malawi in 2005. Specimens prepared for survey sequencing by the JGI were collected from Mazinzi Reef (MZ), Domwe Island (LF, MA) and Otter Point (MC, RE), all locales in the southeastern portion of the lake. High-quality DNA was extracted and prepared in the laboratory of TDK.

2.5.2 Trace sequences

Trace sequences generated by the Joint Genome Institute (JGI) for MC, LF, MA, MZ and RE, together with their sequence quality scores, were downloaded (6 May 2007) from the NCBI Trace Archive. The dataset for each species consisted of an average of about 152,000 individual trace reads, generated by the Sanger sequencing method, with total read lengths ranging from 137 to 185 million bases. Detailed sequence statistics for each species are provided in Appendix A Table A1.

2.5.3 Sequence pre-processing and assembly

The trace and quality sequences were first pre-processed for assembly by masking out all possible vector sequences available from the NCBI UniVec vector sequence database (downloaded 6 May 2007). The vector masking was performed using the cross_match.pl perl script provided by the Phred-Phrap package (Ewing *et al.* 1998). In order to reduce the computational complexity and time required for the final assembly, repeat sequences were masked prior to assembly using RepeatMasker version 3.1.8 (Smit AFA, Hubley R and Green P, unpublished) in conjunction with the latest repeatmasker libraries from RepBase Update (Jurka *et al.* 2005). Bases with sequencing quality score of less than 20 were also masked. The actual assembly of each species' trace sequences into contiguous sequences (contigs) was then performed using the Phrap version 0.990329 assembly program from the Phred-Phrap package. Contigs with more than 80% low quality bases (defined as <20 assembly quality score) were removed from the assembly. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the project accessions ABPJ00000000 (MC), ABPK00000000 (LF), ABPL00000000 (MA), ABPM00000000 (MZ) and ABPN00000000 (RE). The versions described in this paper are the first versions, ABPJ01000000, ABPK01000000, ABPL01000000, ABPM01000000 and ABPN01000000.

2.5.4 Similarity search and alignment

Orthologous genomic contig pairs were first identified using reciprocal BLASTN similarity searches with a strict E-value cutoff of 10^{-100}, performed across the sequence contigs of all possible species pairs. To reduce spurious ortholog assignments, putative ortholog contig pairs were only retained if their regions of high sequence similarity (1) formed good end-to-end overlaps (defined as within 100 bases of the 5' end or 30 bases from the 3' end of a sequence), or (2) overlap more than 80% of the shorter contig. Though some of the filtered regions could represent biologically relevant loci where recombination or translocations might have occurred, we decided to remove them from this analysis. Contig pair assignments were then passed to an algorithm that created clusters of contigs whereby each contig within the cluster must be related to all other contigs in the cluster through one or more putatively orthologous relations. Each cluster of contigs was then individually aligned using Phrap, resulting in a continuous alignment tiling path where each alignment position may consist of a base from any one or up to all five cichlid species (Figure 2.1). Segregating sites were then identified from alignment positions with high quality bases (>20 score) from two or more species. A polymorphism quality score (PQS) was defined, corresponding to the first digit of the lowest Phrap quality score among the nucleotides of the different species present at the polymorphic site (e.g., a polymorphic site between 4 species with base quality scores of 34, 45, 46 and 50 would be assigned a PQS of three). To compare the extent of nucleotide diversity among the five cichlid species, we calculated Watterson's theta (θ_w; Watterson 1975). This measure takes into account the number of variable positions and the sample size analyzed. Our data violate the assumption of an infinite, interbreeding population, but we chose this metric to in order to make direct comparisons to similar measures from study of other genomes (e.g., zebrafish).

2.5.5 Protein-coding sequence identification

Cichlid protein coding sequences were inferred based on similarity searches to known protein databases of fishes and humans. BLASTX searches with E-value cutoff of 10^{-10} were performed for the each cichlid genomic assembly as well as the overall consensus sequence of the cluster alignments, against a protein database made up of all GenBank *Actinopterygii* (ray-finned fishes) sequences (downloaded 02 June 2007; 163,471 entries) and all human RefSeq proteins (downloaded 25 June 2007; 34,180 sequences). The alignment with the highest scoring hit for each genomic locus was then used as a reference to determine the coding strand and phase of the protein-coding cichlid locus.

2.5.6 Evolutionary sequence divergence among JGI species

All cluster alignment segments with contributing bases from two or more species were split into pairwise alignments (each two, three, four or five species alignment position can be split into one, three, six or ten pairwise alignments respectively). Pairwise alignments within each of the ten possible species pair combinations (MC-LF, MC-MA, MC-MZ, MC-RE, LF-MA, LF-MZ, LF-RE, MA-MZ, MA-RE, MZ-RE) were then concatenated and the number of substitutions counted. Jukes-Cantor correction for multiple substitutions was applied to these direct distance measurements (Jukes and Cantor 1969). Pairwise alignments consisting of only genic sequences were obtained from multi-species cluster alignment segments in a manner similar to that described above. The DNAStatistics package of Bioperl (www.bioperl.org) was then used to calculate the K_a/K_s values of pairwise alignments.

2.5.7 Genotyping and validation of SNPs

We genotyped 96 SNPs in 364 diverse Lake Malawi cichlid samples. These SNPs included 13 positive controls, 59 loci from the automated procedure described in this report, and an additional 24 loci chosen manually by BLAST of individual traces to the *Tetraodon* genome (see main text for further description). The GenomeLab SNPstream Genotyping System Software Suite v2.3 (Beckman Coulter, Inc., Fullerton, CA) was used for experimental setup, data uploading, image analysis, genotype calling and QC review, at Emory University's Center for Medical Genomics. In brief, marker panel data (i.e., multiplexed SNP panel designed by SNPstream's Primer Design Engine website; www.autoprimer.com) were first uploaded to the SNPstream database using the PlateExplorer application software. Also uploaded was the Process Group Data containing all test sample information generated through a Laboratory Information Management System (Nautilus 2002, Thermo Fisher Scientific, Waltham, MA). An on-board CCD camera of the SNPstream Imager took two snapshot images of each well of the 384-well tag array, one under a blue excitation laser, the other under a green excitation laser. Image application software was used to analyze the captured images to detect spots, overlay an alignment grid, and determine spot intensity. The fluorescent pixel intensity data for each SNP under the two channels, representing the relative abundance of the two alleles, were uploaded to the database. The GetGenos application software was used to calculate and generate a Log(B+G) *vs*. B/(B+G) plot, where B and G were the pixel intensities under the blue and green channels, respectively, for each sample and each SNP. Next, automated genotype calling was accomplished using the QCReview application software based on a number of criteria (e.g., signal baseline, clustering pattern of the three genotypes, Hardy-Weinberg score). A genotype summary was generated using the Report application software.

2.5.8 Genetic differentiation within and among lineages

Locus specific F_{ST} (Weir and Cockerham 1984) was calculated using FSTAT version 2.9.3.2 (Goudet 1995) for three evolutionary scales: (i) within LF and MZ, (ii) between LF and MZ and (iii) between mbuna and non-mbuna. We determined that a SNP locus was a statistical outlier using the empirical distribution of F_{ST} values. F_{ST} outliers exceed the sum of the upper quartile value and 1.5 times the inter-quartile range.

2.5.9 Genomic assignment

We used a Bayesian method (STRUCTURE v.2.2; Pritchard *et al.* 2000) to ask how well our SNP genotypes assigned individuals to evolutionary lineages. We chose to define the number of K genetic clusters in accord with previous research showing about three major evolutionary groups of Lake Malawi cichlids (Moran and Kornfield 1993, Kocher *et al.* 1995, Won *et al.* 2006, Hulsey *et al.* 2007). Note that we do not intend this to mean that 3 is the best supported estimate of K in these data; our rationale is rather to demonstrate how individual genomes are composites (or not) of the major evolutionary lineages found in the lake. Thus, we used the admixture model to estimate q, the proportion of each genome derived from each of K genetic clusters. For comparison, we also ran analyses with K set to two, four or five (not shown). Each run of the program included 50,000 cycles of burn-in and run length of 50,000 steps. Multiple runs were conducted to ensure reliability and consistency of results.

2.6 Acknowledgements

We thank members of the Streelman lab, Karen Carleton and two anonymous reviewers for comments on previous drafts of this manuscript. The research is supported by grants from the NSF (IOS 0546423), NIH (R21 DE017182) and Alfred P. Sloan Foundation (BR-4499) to JTS. Drs. Karen Carleton and Federica DiPalma extracted high

quality DNA from the five species of Malawi cichlid. Library construction and sequencing was performed by the Joint Genome Institute under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.

2.7 References

Albertson RC, Markert JA, Danley PD, Kocher TD. 1999. Phylogeny of a rapidly evolving clade: the cichlid fishes of Lake Malawi, East Africa. *Proc Natl Acad Sci USA*. 96:5107-5110.

Albertson RC, Streelman JT, Kocher TD. 2003. Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. *Proc Natl Acad Sci USA*. 100:5252-5257.

Albertson RC, Streelman JT, Kocher TD, Yelick PC. 2005. Integration and evolution of the cichlid mandible: the molecular basis of alternative feeding strategies. *Proc Natl Acad Sci USA*. 102:16287-16292.

Allender CJ, Seehausen O, Knight ME, Turner GF, Maclean N. 2003. Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptual coloration. *Proc Natl Acad Sci USA*. 100:14074-14079.

Arnegard ME, Markert JA, Danley PD, Stauffer JR, Ambali AJ, Kocher TD. 1999. Population structure and colour variation of the cichlid fish *Labeotropheus fuelleborni* along a recently formed archipelago of rocky habitat patches in southern Lake Malawi. *Proc Biol Sci*. 266:119-130.

Aubin-Horth N, Desjardins JK, Martei YM, Balshine S, Hofmann HA. 2007. Masculinized dominant females in a cooperatively breeding species. *Mol Ecol*. 16:1349-1358.

Blais J, Rico C, van Oosterhout C, Cable J, Turner GF, Bernatchez L. 2007. MHC adaptive divergence between closely related and sympatric African cichlids. *PloS ONE*. 2:e734.

Carleton KL, Streelman JT, Lee B-Y, Garnhart N, Kidd MR, Kocher TD. 2002. Rapid isolation of CA microsatellites from the cichlid genome. *Anim Genet*. 33:140-144.

Carleton KL, Parry JW, Bowmaker JK, Hunt DM, Seehausen O. 2005. Color vision and speciation in Lake Victoria cichlids of the genus *Pundamilia*. *Mol Ecol*. 14(14):4341-4353.

Chenn A, Walsh CA. 2002. Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science*. 297(5580):365-369.

Danley PD, Markert JA, Arnegard ME, Kocher TD. 2000. Divergence with gene flow in the rock-dwelling cichlids of Lake Malawi. *Evolution*. 54:1725-1737.

Ewing B, Hiller L, Wendl M, Green P. 1998. Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res*. 8:175-185.

Fraser GJ, Bloomquist RF, Streelman JT. 2008. A periodic pattern generator for dental diversity. *BMC Biol*. 6:32.

Genner MJ, Turner GF. 2005. The mbuna cichlids of Lake Malawi: a model for rapid speciation and adaptive radiation. *Fish and Fisheries*. 6:1-34.

Goudet J. 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. *J Hered*. 86:485-486.

Green P. 2007. 2x genomes – does depth matter? *Genome Res*. 17:1547-1549.

Gregory TR, Nicol JA, Tamm H, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size database. *Nucleic Acids Res*. 35:D332-D338.

Guryev V, Koudijs MJ, Berezikov E, Johnson SL, Plasterk RHA, van Eeden FJM, Cuppen E. 2006. Genetic variation in the zebrafish. *Genome Res*. 16:491-497.

Huber R, van Staaden MJ, Kaufman LS, Liem KF. 1997. Microhabitat use, trophic patterns, and the evolution of brain structure in African cichlids. *Brain Behav Evol*. 50:167-182.

Hulsey CD, Mims MC, Streelman JT. 2007. Do constructional constraints influence cichlid craniofacial diversification? *Proc Biol Sci*. 274:1867-1875.

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*. 431(7011):946-957.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet*. 19:68-72.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism.* Edited by Munro HN. New York: Academic Press. 21-132.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. 110:462-467.

Katigiri T, Kidd CE, Tomasino E, Davis JT, Wishon C, Stern JE, Carleton KL, Howe AE, Kocher TD. 2005. A BAC-based physical map of the Nile tilapia genome. *BMC Genomics*. 6:89.

Kidd MR, Kidd CE, Kocher TD. 2006. Axes of differentiation in the bower-building cichlids of Lake Malawi. *Mol Ecol*. 15:459-478.

Kijimoto T, Watanabe M, Fujimura K, Nakazawa M, Murakami Y, Kuratani S, Kohara Y, Gojobori T, Okada N. 2005. *cimp*, a novel astacin family metalloproteinase gene from East African cichlids, is differentially expressed between species during growth. *Mol Biol Evol*. 22:1649-1660.

Kirkness EF, Bafna V, Halpern AL, Levy S, Remington K, Rusch DB, Decher AL, Pop M, Wang W, Fraser CM, Venter JC. 2003. The dog genome: survey sequencing and comparative analysis. *Science*. 310:1898-1903.

Kocher TD, Conroy JA, McKaye KR, Stauffer JR, Lockwood SF. 1995. Evolution of NADH dehydrogenase subunit 2 in East African cichlid fish. *Mol Phylogenet Evol*. 4:420-432.

Kocher TD, Lee W-J, Sobolewska H, Penman D, McAndrew B. 1998. A genetic linkage map of the cichlid fish, the tilapia (*Oreochromis niloticus*). *Genetics*. 148:1225-1232.

Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat Rev Genet*. 5:288-298.

Kornfield I, Smith PF. 2000. African cichlid fishes: model systems for evolutionary biology. *Ann Rev Ecol Evol Syst*. 31:163-196

Kotarski MA, Leonard DA, Bennett SA, Bishop CP, Wahn SD, Sedore SA, Shrader M. 1998. The *Drosophila* gene *asteroid* encodes a novel protein and displays dosage-sensitive interactions with *Star* and *Egfr*. *Genome*. 41:295-302.

Koudijs MJ, den Broeder MJ, Groot E, van Eeden GF. 2008. Genetic analysis of the two zebrafish patched homologues identifies novel roles for the hedgehog signaling pathway. *BMC Dev Biol*. 8:15.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*. 2:231-239.

Lee B-Y, Hulata G, Kocher TD. 2004. Two unlinked loci controlling the sex of blue tilapia (*Oreochromis aureus*). *Heredity*. 92:543-549.

Lee B-Y, Lee W-J, Streelman JT, Carleton KL, Howe AE, Hulata G, Slettan A, Stern JE, Terai Y, Kocher TD. 2005. A second-generation genetic linkage map of tilapia (*Oreochromis* spp). *Genetics*. 170:237-244.

Margulies EH, Vinson JP, NISC Comparative Sequencing Program, Miller W, Jaffe DB, Lindblad-Toh K, Chang JL, Green ED, Lander ES, Mullikin JC, Clamp M. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci USA*. 102(13):4795-4800.

Margulies EH, Birney E. 2008. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet*. 9:303-313.

Miyasaka KY, Kida YS, Sato T, Minami M, Ogura T. 2007. Csrp1 regulates dynamic cell movements of the mesendoderm and cardiac mesoderm through interactions with Dishevelled and Diversin. *Proc Natl Acad Sci USA*. 104(27):11274-11279.

Moon S, Shin HD, Cheong HS, Cho HY, Namgoong S, Kim EM, Han CS, Kim H. 2007. BcSNPdb: Bovine coding region single nucleotide polymorphisms located proximal to quantitative trait loci. *J Biochem Mol Biol*. 40(1):95-99.

Moran P, Kornfield I. 1993. Retention of ancestral polymorphism in the Mbuna species flock of Lake Malawi. *Mol Biol Evol*. 10:1015-1029.

Nagl S, Tichy H, Mayer WE, Takahata N, Klein J. 1998. Persistence of neutral polymorphisms in Lake Victoria cichlid fish. *Proc Natl Acad Sci USA*. 24:14238-14243.

Parry JW, Carleton KL, Spady T, Carboo A, Hunt DM, Bowmaker JK. 2005. Mix and match color vision: tuning spectral sensitivity by differential gene expression in Lake Malawi cichlids. *Curr Biol*. 15:1734-1739.

Pavy N, Parsons LS, Paule C, MacKay J, Bousquet J. 2006. Automated SNP detection from a large collection of white spruce expressed sequences: contributing factors and approaches for the categorization of SNPs. *BMC Genomics*. 7:174.

Pontius JU, Mullikin JC, Smith DR, Agencourt Sequencing Team, Lindblad-Toh K, Gnerre S, Clamp M, Chang J, Stephens R, Neelam B *et al*. 2007. Initial sequence and comparative analysis of the cat. *Genome Res*. 17:1675-1689.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945-959.

Reinthal PN. 1990. The feeding habits of a group of herbivorous rock-dwelling fishes from Lake Malawi, Africa. *Env Biol Fishes*. 27:215-233.

Renn SC, Aubin-Horth N, Hofmann HA. 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics*. 6:42.

Scholpp S, Foucher I, Staudt N, Peukert D, Lumsden A, Houart C. 2007. Otx1l, Otx2 and Irx1b establish and position the ZLI in the diencephalon. *Development*. 134:3167-3176.

Seehausen O, Koetsier E, Schneider MV, Chapman LJ, Chapman CA, Knight ME, Turner GF, van Alphen JJM, Bills R. 2003. Nuclear markers reveal unexpected genetic variation and a Congolese-Nilotic origin of the Lake Victoria cichlid species flock. *Proc Biol Sci*. 270(1533):129-137.

Seehausen O. 2004. Hybridization and adaptive radiation. *Trends Ecol Evol*. 19:198-207.

Smith PF, Konings A, Kornfield I. 2003. Hybrid origin of a cichlid population in Lake Malawi: implications for genetic variation and species diversity. *Mol Ecol*. 12:2497-2504.

Spady TC, Seehausen O, Loew ER, Jordan RC, Kocher TD, Carleton KL. 2005. Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid fishes. *Mol Biol Evol*. 22:1412-1422.

Streelman JT, Albertson RC, Kocher TD. 2003. Genome mapping of the orange blotch colour pattern in cichlid fishes. *Mol Ecol*. 12:2465-2471.

Streelman JT, Gmyrek SL, Kidd MR, Kidd CE, Robinson RL, Hert E, Ambali AJ, Kocher TD. 2004. Hybridization and contemporary evolution in an introduced cichlid fish from Lake Malawi National Park. *Mol Ecol*. 13:2471-2479.

Streelman JT, Albertson RC. 2006. Evolution of novelty in the cichlid dentition. *J Exp Zoolog B Mol Dev Evol*. 306:216-226.

Streelman JT, Peichel CL, Parichy DM. 2007. Developmental genetics of adaptation in fishes: the case for novelty. *Ann Rev Ecol Evol Syst*. 38:655-681.

Takahashi K, Okada N. 2002. Mosaic structure and retropositional dynamics during evolution of subfamilies of short interspersed elements in African cichlids. *Mol Biol Evol*. 19:1303-1312.

Terai Y, Morikawa N, Okada N. 2002. The evolution of the pro-domain of bone morphogenetic protein 4 (*Bmp4*) in an explosively speciated lineage of East African cichlid fishes. *Mol Biol Evol*. 19:1628-1632.

Turner GF, Seehausen O, Knight ME, Allender CF, Robinson RL. 2001. How many species of cichlid fishes are there in African lakes? *Mol Ecol*. 10:793-806.

Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, Strausberg RL, Brenner S. 2007. Survey sequencing and comparative analysis of the elephant shark (*Callorhinchus milii*) genome. *PloS Biology*. 5:e101.

Watterson GA. 1975. On the number of segregating sites in genetic models without recombination. *Theor Pop Biol*. 7:256-276.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution*. 38:1358-1370.

Won Y-J, Sivasundar A, Wang Y, Hey J. 2005. On the origin of Lake Malawi cichlid species. *Proc Natl Acad Sci USA*. 102:6581-6586.

Won Y-J, Wang Y, Sivasundar A, Raincrow J, Hey J. 2006. Nuclear gene variation and molecular dating of the cichlid species flock of Lake Malawi. *Mol Biol Evol*. 23:828-837.

Ye H, Tan YL, Ponniah S, Takeda Y, Wang SQ, Schachner M, Watanabe K, Pallen CJ, Xiao ZC. 2008. Neural recognition molecules CHL1 and NB-3 regulate apical dentrite orientation in the neocortex via PTP alpha. *EMBO J*. 27(1):188-200.

CHAPTER 3

EARLY ORIGINS OF GENETIC VARIATION IN LAKE MALAWI CICHLIDS

3.1 Abstract

Cichlid fishes have evolved tremendous morphological and behavioral diversity in the lakes and rivers of East Africa. Within each of the Great Lakes Tanganyika, Malawi and Victoria, the dual processes of hybridization and the retention of ancestral polymorphism explain allele sharing across species. Here, we investigate the sharing of single nucleotide polymorphism (SNP) between the major East African cichlid assemblages. A set of about 200 genic and non-genic SNPs was ascertained in five Lake Malawi species and successfully genotyped in a diverse collection of around 160 species from across the East African basin. We observed segregating polymorphism outside of the Malawi lineage for more than 40% of loci; this holds similarly for genic versus non-genic SNPs, as well as for SNPs at putative CpG sites vs. non-CpG sites. Bayesian analysis of genetic structure in the data supports the hypothesis that Lake Malawi cichlids are not monophyletic and that riverine species have contributed significantly to their genomes. We observed strong genetic differentiation between major Malawi groups for about 8% of loci, with contribution from both genic and non-genic SNPs. Notably, more than half of these outlier loci for genetic differentiation among Malawi cichlids likely originated prior to the radiation of the Malawi endemic species flock. Our data suggest that cichlid fishes have evolved diversity in Lake Malawi as new mutations combined with standing genetic variation shared across East Africa.

3.2 Introduction

The understanding of how organismal diversity is achieved lies at the heart of evolutionary biology. From a molecular perspective, genetic variation provides the substrate on which selection may act, allowing the adaptation to new ecological niches that may have been unfavorable to the parental species, which may then lead to organismal diversification and eventual speciation (Gavrilets and Losos 2009, Cristescu *et al*. 2010). Genetic variation may arise in the form of new random mutations, or it may already be present as standing variation, via processes such as recurrent mutations, ancestral inheritance of polymorphisms, or inter-specific hybridization and introgression (Barrett and Schluter 2008). The presence and distribution of genetic polymorphism provides us with the opportunity to study and better understand the underlying evolutionary processes of organismal diversification. One powerful system on which we can conduct such studies is the diverse but closely related species flock of East African cichlid fishes.

The cichlid fishes of the East Africa's Great Lakes, made up of an estimated 2000 species, is well acknowledged as one of the most spectacular example of rapid evolutionary radiation in vertebrates. Lake Tanganyika, the oldest lake at 9-12 million years, contains about 250 cichlid species. Lake Malawi (2-5 million years old) cichlids, with up to 1000 species, represents the richest cichlid species flock that had evolved over a relatively young evolutionary age of 1 million years. The Lake Victoria superflock, made up of 500-700 species of cichlids, mostly from Lake Victoria itself (250,000-750,000 years old), but also includes cichlids from its neighboring lakes Albert, Edward, George, Kyoga and Kivu, is evolutionarily the youngest at about 100,000 years old. In addition, some 200 cichlid species also inhabit the rivers and smaller lakes throughout Africa. Remarkably, almost all of the species found in the East African cichlid assemblage are endemics, with no single species found to be common among any of

the three East African Great Lakes. (species estimates, lake and cichlid evolutionary ages referenced in recent reviews; Kornfield and Smith 2000, Kocher 2004, Turner 2007, Kuraku and Meyer 2008, Salzburger 2009).

Knowledge on the evolutionary history of East African cichlid radiation has advanced tremendously over the past decade. Phylogenetic analyses on mitochondrial sequences of the East African cichlids have revealed that Lake Tanganyika contains at least 12 eco-morphologically distinct cichlid tribes, and that one of the tribes, the haplochromines, expanded out of Lake Tanganyika to colonize and explosively radiate into almost all of the cichlid species that can be found in the entire East Africa outside of Lake Tanganyika, that is, Lake Malawi, Lake Victoria and neighboring lakes, as well as the river and drainage systems (Salzburger *et al.* 2002, 2004, 2005). While these studies were able to resolve the broad relationships between cichlid tribes and major assemblages with high confidence, they were unable to unambiguously resolve the relationships between smaller lineage groups or species (Salzburger *et al.* 2004, 2005). This is possibly due to the maintenance of ancestral polymorphisms that is known to exist in cichlids, and previously reported independently in Lake Malawi (Moran and Kornfield 1993), Victoria (Nagl *et al.* 1998), and Tanganyika (Koblmuller *et al*. 2010).

Beyond their evolutionary histories, the rapid cichlid diversifications brought about a tremendous array of behavoiral and phenotypic variations that makes the cichlid system a good model for evolutionary genomic and developmental research. Cichlid evolution has been described as being analagous to a ‘mutagenic screen’ (Kocher 2004), except that it had occurred naturally under adaptive selection regimes. Additionally, homoplasies from convergent evolution of numerous traits have been frequently observed in independent cichlid radiations (Kocher *et al.* 1993, Kuraku and Meyer 2008, Salzburger 2009), suggesting that independent radiations of cichlids are not always totally random, but that similar adaptations, possibly under constraints, have re-evolved

repeatedly (Kuraku and Meyer 2008). These evolutionary diversifications have allowed scientists to study the evolutionary and genetic basis of many traits, including behavior (Aubin-Horth *et al.* 2007), olfaction (Blais *et al.* 2007), pigmentation (Streelman *et al.* 2003, Allender *et al.* 2003, Lee *et al.* 2005), vision (Spady *et al.* 2005, Parry *et al.* 2005, Seehausen *et al.* 2008), acoustic projection and perception (Simoes *et al.* 2008, Verzijden *et al.* 2010), sex determination (Lee *et al.* 2004, 2005, Ser *et al.* 2010), the brain (Huber *et al.* 1997, Sylvester *et al.* 2010), and craniofacial development (Albertson *et al.* 2003, 2005, Streelman *et al.* 2006, Fraser *et al.* 2008).

Nonetheless, as we progress into the genomics era, much more awaits to be discovered with regards to the evolution of cichlids, and the evolution of species in general. We want to find out where cichlids obtain the genetic diversity for radiation. Ancestral polymorphisms and allele sharing has been shown in small-scale studies within each lake, but to what extent are interlucastrine polymorphisms being maintained? And what can we infer about consequences these might have on cichlid diversifications in the different lakes? Phylogenetic studies are only able to reveal the bi- and multi-furcating relationships between species and lineages, but there is much more to learn about the genomic content, structure and relationships between cichlid species and lineages. On the molecular level, the specific positions of the polymorphisms on the genome and their allele segregation patterns would provide a clue to the selective forces that are active and their functional consequences. Would we be able to discover differentiated alleles and use them to aid functional studies? Ultimately, how would the knowledged gained about cichlid evolutionary diversification be applicable also to the adaptive evolution of species in general?

In this study, we conducted an expanded genotyping analysis of 280 SNPs, mostly sourced from Lake Malawi cichlid comparisons but also including other African cichlid comparisons, in a diverse set of 576 cichlid samples from throughout Africa. We

observed widespread sharing of about 40% of polymorphisms between lake assemblages, representing divergences of up to 12 million years. We found from a bayesian analysis of genetic structure that East African cichlids generally clustered into 6 major goups, with additional groups showing interesting admixture patterns of genomic contributions from multiple lineages, and evidence that riverine species have contributed significantly to the genomes of Malawi cichlids. The data also supports the hypothesis that Lake Malawi cichlids are not monophyletic. We found strong genetic differentiation between major Malawi groups for about 8% of loci, which may be indicative of the functional divergences that had occurred.

3.3 Materials and methods

3.3.1 Fish samples and genotyping

576 wild-caught fish samples, encompassing 78 genera and more than 161 species and strains, were collected from the major East African Rift Lakes Malawi, Victoria and Tanganyika, as well as numerous other smaller lakes and rivers throughout the African continent (Figure 3.1). High quality DNA was extracted from fin clippings using standard molecular biology protocols in the laboratories of Kocher TD, Streelman JT, Seehausen O and Salzburger W.

280 SNP positions were used for genotyping, including 214 (147 non-coding, 67 coding) that were previously identified from comparisons among Lake Malawi species (hereby termed “Malawi SNPs”; Loh *et al.* 2008), 28 “Victoria SNPs” identified from Lake Victoria species, 21 “Tanganyika SNPs” identified among Lake Tanganyika species, and 17 “Riverine SNPs” identified in *Astatotilapia burtoni*, a riverine species that is also found in Lake Tanganyika. SNP genotyping was carried out by the Broad Institute on the Sequenom(®) MassArray™ iPLEX Gold platform, which uses MALDI-TOF mass spectrometry to determine genotypes based on the mass of allele-specific extension

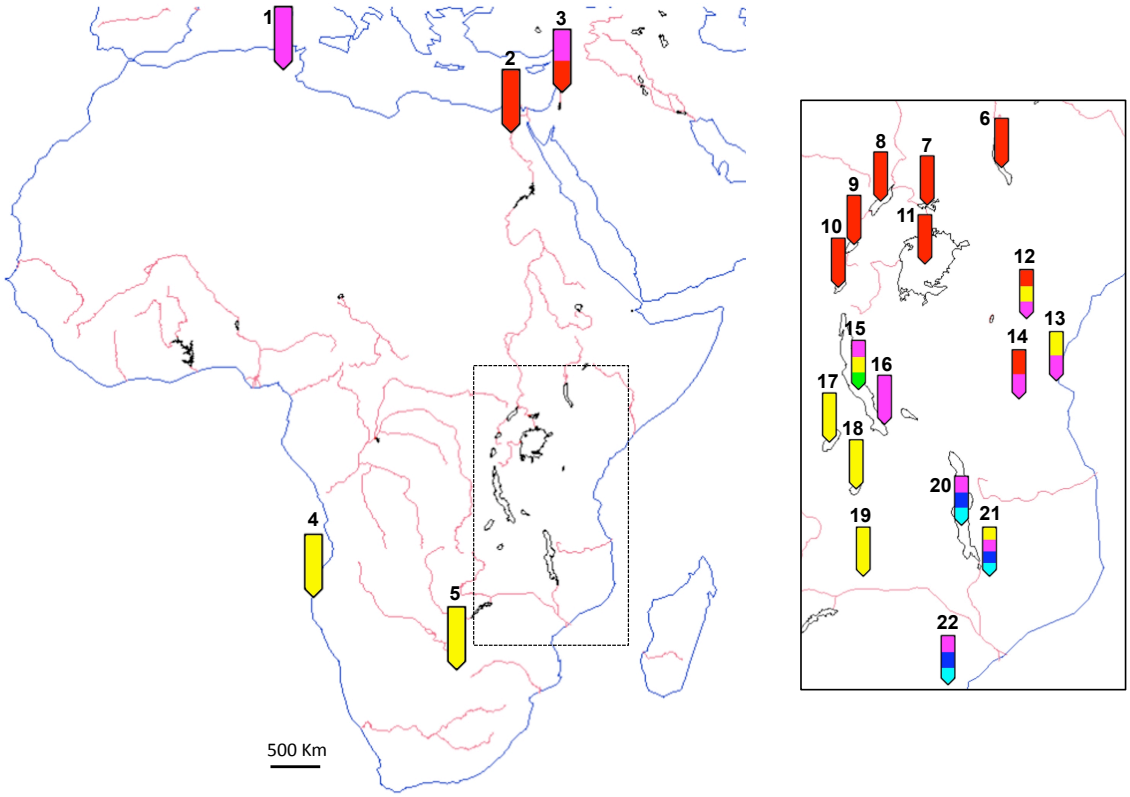


Figure 3.1. Map of Africa showing cichlid sampling locations. Section within dotted box expanded and displayed in right solid box. Numbered arrows indicate location where cichlid samples were collected. Colors on labels (not to scale) correspond to the genetic clustering colors of Figure 3.4. 1, Tunisia; 2, Egypt; 3, Kinneret; 4, Cunene; 5, Lisikili; 6, Lake Turkana; 7, Lake Kyoga; 8, Lake Albert; 9, Lake Edward; 10, Lake Kivu; 11, Lake Victoria; 12, Nyumba; 13, Bagamoyo; 14, Ilonga; 15, Lake Tanganyika; 16, Kalambo; 17, Lake Mweru; 18, Lake Bangweulu; 19, Kafue; 20, Lake Malawi; 21, Lake Chilwa; 22, Mozambique; Light blue, Malawi mbuna; Dark blue, Malawi non-mbuna; Red, Victoria superflock; Yellow, Tanganyika and riverine Haplochrominii and Tropeinii; Green, older Tanganyika tribes.

products. The assays were designed using Sequenom's MassARRAY® Design Software.

3.3.2 Coincident polymorphism

To first determine a broad based pattern of allele sharing between cichlid lineages of the different lakes, we grouped the cichlid samples into 4 main catchment groups, namely, the cichlids of (i) Lake Malawi, (ii) Lake Victoria superflock, (iii) Lake Tanganyika, and (iv) Other African rivers and regions. In each group, observed polymorphism at each SNP position was established when the minor allele was present in at least 2 cichlid samples. This criterion was defined to conservatively reduce polymorphism calls that may be due to possible genotyping errors. Coincident polymorphism sharing between the catchment groups was then determined. For a finer scale study of coincident polymorphism in 180 Malawi SNPs, the cichlid fish samples were grouped based on previously determined phylogenetic lineages (Salzburger and Meyer 2004), and polymorphism was determine by any occurrence of the minor allele within each lineage.

3.3.3 Genetic clustering

We utilized a Bayesian approach implemented in the STRUCTURE v.2.2 analysis package (Pritchard *et al.* 2000) to assign individuals (with admixture allowed) to a predetermined number (K) of genetic clusters based on their SNP genotypes. Each Markov-Chain Monte Carlo (MCMC) run performs 10,000 burn-in cycles followed by 10,000 cycles of data collection. Eleven replicate runs were performed for each value of K ranging from two to eight, following which the optimal number of genetic clusters best representing the data was then determined. This was based on the ad-hoc statistic ΔK suggested by Evanno *et al.* 2005, which selects the K value that had the largest second

order rate of change of the log probablility of data with respect to the number of clusters. The clustering pattern that was most often obtained among the eleven runs was then selected. We observed that for runs at K=7 and higher, even though MCMC stability was achieved well before the 10,000 runs were completed, there was considerable variability in the results between runs, which prevented the determination of any consistent genetic clustering results.

3.3.4 Genetic differentiation

To investigate the levels of genetic differentiation among Lake Malawi cichlid populations, F_{ST} (Weir and Cockerham 1984) for each SNP was calculated using FSTAT version 2.9.3.2 (Goudet 1995). Several F_{ST} comparisons were performed: among mbuna (M), non-mbuna (N) and other deep water and pelagic (D) populations; among pairs of M, N and D lineages; among populations (with >5 samples) grouped by their genus; and between the *Labeotropheus* and *Metriaclima* genus. The empirical distribution of F_{ST} values at each SNP was used to determine statistical outliers, defined as values exceeding the sum of the upper quartile value and 1.5 times the interquartile range.

3.4 Results and discussion

3.4.1 *Genotype data*

A wide selection of 576 fish samples, representative of the diversity of East African cichlids and encompassing 78 genera and more than 161 species and strains, were genotyped at 280 SNP positions. More than 161,000 genotypes were collected, with 86.3% successful reads. We performed an initial quality analysis of the SNP and cichlid sample results, which led to 61 SNP results being discarded due to high genotyping failure rates of more than 25% of samples, allele monomorphism, or had widespread heterozygosity suggestive of non-specificity of the genotyping probes. Thirteen cichlid

samples were also removed as they failed genotyping or had data indicating probable DNA contamination. The remaining 123,297 genotypes (563 samples x 219 SNPs) had a successful genotyping yield of 95.3% and were used for subsequent analyses.

The resultant 219 polymorphic and informative SNPs used for analyses consisted of 180 Malawi SNPs (119 coding, 61 non-coding), 21 Victoria SNPs, 9 Tanganyika SNPs and 9 Riverine SNPs (see Methods and Table 3.1). As these SNPs were identified from

Table 3.1. Source and genotyping success of sampled SNPs.

SNP Source	Total Number Genotyped	Failed, Low Quality, Monomorphic or Excessive Heterozygosity	Informative SNPs
Malawi SNPs; non-coding	147	28	119
Malawi SNPs; coding	67	6	61
Victoria SNPs	28	7	21
Tanganyika SNPs	21	12	9
Riverine SNPs	17	8	9
Total	**280**	**61**	**219**

cichlids belonging to allopatric lakes and river systems, we expected to our data to show some ascertainment bias. Indeed, when we calculated the average heterozygosity of the different cichlid assemblages for the different classification of SNPs (Appendix B Figure B1), we observed that the ascertained lineage often had a higher, though not statistically significant, average heterozygosity value. The disproportionate distribution of SNPs, with a majority being identified from Lake Malawi cichlids, also produced ascertainment bias in the information content obtained from the genotyping results, as evidenced by our observation of longer branch lengths calculated for the evolutionarily younger Malawi lineages compared to the older Tanganyika lineages, when we attempted to build a phylogeny (not shown) from the data obtained. However, as the current study is mostly

focused on Lake Malawi cichlids, the ascertainment bias is not expected to adversely affect the types of analyses we conduct and the conclusions made.

3.4.2 Origins of Lake Malawi polymorphism

We wanted to investigate how much polymorphism sharing occurs among East African Cichlids. Using the subset of 180 Malawi SNPs, we tabulated the extent of polymorphism sharing between cichlids that were categorized into four groups based on their catchments: (i) the Lake Malawi assemblage, (ii) the Lake Victoria superflock, (iii) the Lake Tanganyika assemblage, (iv) all other cichlids. Initially using the widest definition (i.e. any occurrence of the minor allele) to define polymorphism within a catchment, we found that a surprisingly high 61.7 % (111 out of 180) of all Malawi SNPs were polymorphic both inside and outside of Lake Malawi. We recognized that there might be low levels of genotyping error inherent in the data, and therefore sought to reduce the possibility of errorneous results by redefining polymorphism to be present only when the minor allele occurred in at least 2 fish samples within the catchment. This conservative definition reduced the percentage of shared polymorphism to 48.9% (88 SNPs), which still represents a relatively large proportion of Malawi SNPs (Figure 3.2A).

This trend of high levels of polymorphism sharing is similar for both the subsets of coding and non-coding SNPs, demonstrating that polymorphism sharing is pervasive phenomena irrespective of general selective constraints. We repeated this analysis for the much smaller set of Victoria (18) and Tanganyika (9) SNPs cichlids, and found similarly high proportions of polymorphism sharing (Appendix B Figure B2 and B3). The Riverine SNPs (9), originally identified from a single species (*A. burtoni*) that was present both in Lake Tanganyika and the nearby rivers, was not found to be polymorphic in Lake Malawi cichlids or the Lake Victoria superflock.

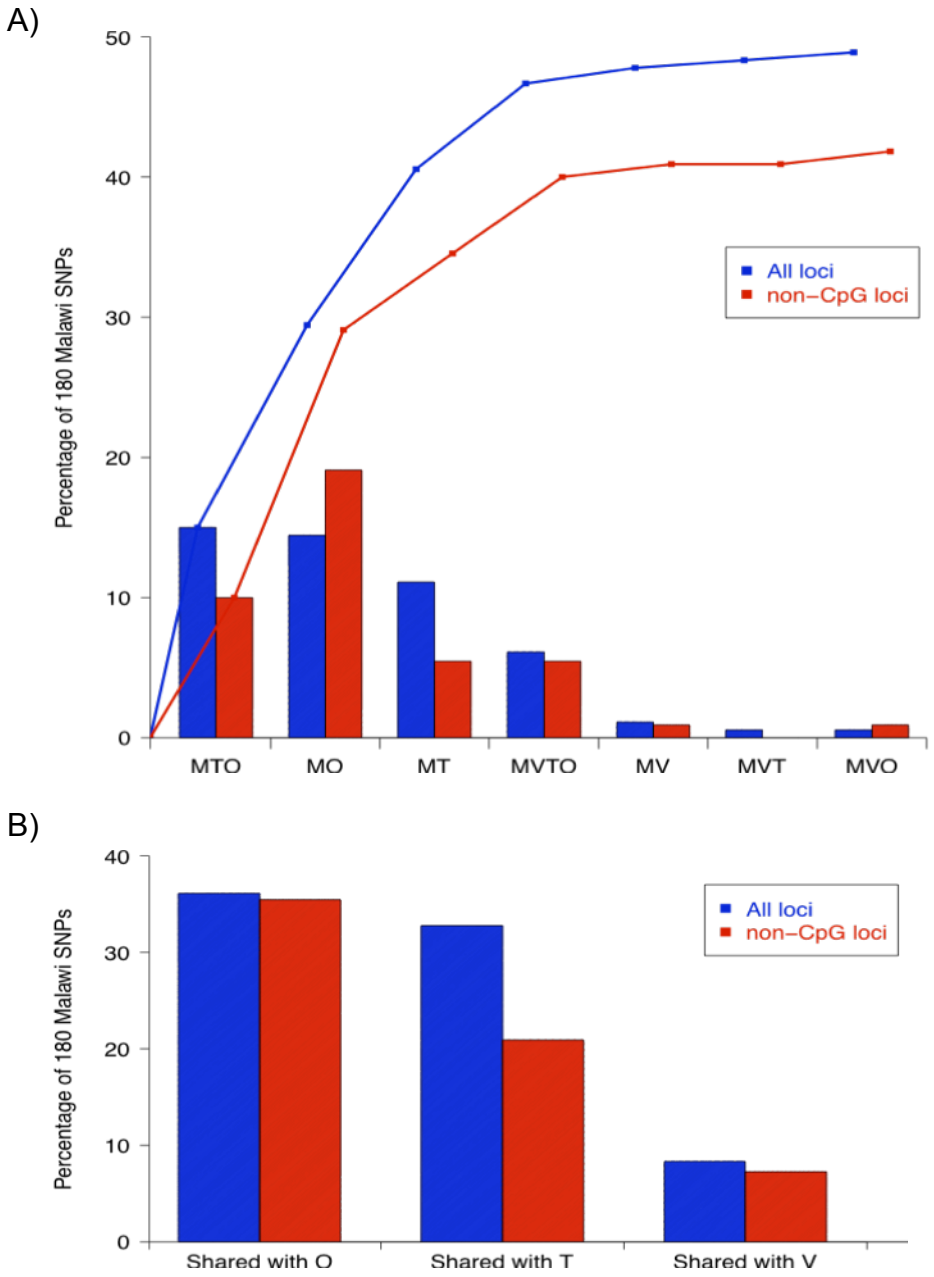


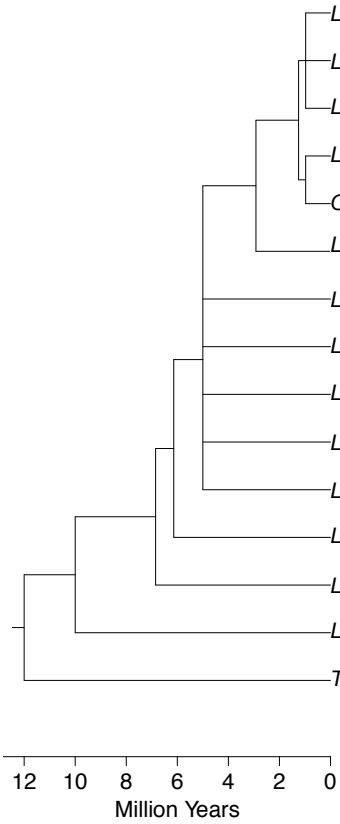
Figure 3.2. Percentage of shared polymorphism of 180 Malawi SNPs (108 non-CpG) with cichlids in other catchments. A) Strict polymorphism sharing with each catchment combination indicated by the category labels. B) Total polymorphism sharing with one other catchment. Bar graphs show percentage polymorphism sharing for each category while line graphs tally cumulative percentages. M, Malawi assemblage; V, Victoria superflock; T, Tanganyika assemblage; O, other rivers and drainages.

Such high levels of coincident polymorphism is unexpected, given that the average nucleotide diversity of cichlids was found to be a low 0.26% (or 1 variable site every 385 nucleotides; Loh *et al.* 2008), and that these cichlid lineages have diverged up to 12 million years ago (Figure 3.3). However, there could be several possible biological phenomena that could explain high levels of coincident polymorphism.

There could be variations in mutation rate along the genome that is context dependent, such as those sites consisting of a cytosine immediately followed by guanine (CpG). Methylation of the cytosines at CpG sites is widespread in vertebrate genomes (Suzuki and Bird 2008), forming unstable methyl-cytosines that are capable of spontaneous deamination. which leads to a high rate of C-to-T and G-to-A transitions. We removed all SNPs that could be produced by CpG mutations, but continued to observe similarly high polymorphism sharing rates of 41.8% among non-CpG Malawi SNPs (Figure 3.2; Appendix B Figures B2 and B3).

Recent reports described cryptic variation in the human mutation rate that could be responsible for elevated levels of coincident SNPs between human and chimpanzees (Hodgkinson *et al.* 2009, Hodgkinson and Eyre-Walker 2010). The authors in these studies were unable to define the specific context effects (hence 'cryptic') to explain the coincident SNPs, but they did observe a 15-fold excess of A-T coincident SNPs when compared to expected transition and transversion SNP rates, and concluded that some other mechanism beyond ancestral polymorphism was responsible for the the elevated coincident SNP. In our current analysis, we did not observe the transition and transversion distribution of coincident SNPs to be significantly different from the average distribution over all SNPs (chi-square test; $P = 0.481$), and therefore have no evidence of similar cryptic variation occurring in cichlids.

Coincident SNPs in divergent lineages could also be due to ancestral polymorphism. Ancestral (or trans-specific) polymorphism, the inheritence of polymorphisms from a



	A	B	C	D
LM mbuna	71	0.24	0.49	0
LM nonmbuna	68	0.07	0.06	0.83
LM "deep"	31	0.14	0.14	0.50
LV superflock	19	0.23	0	0.004
Other Haplochromini	67	0.17	0.02	0
LT Haplochromini and Tropheini	36	0.37	0	0
LT Ectodini	5	0.06	0	1
LT Cyphotilapiini	6	0	0	0
LT Cyprichromini	4	1	0	0
LT Perissodini	1	0	0	0
LT Limnochromini	12	0.33	0	1
LT Lamprologini	16	0.47	0	1
LT Eretmodini	2	1	0	0
LT Bathybatini	28	0.09	0	1
Tilapiini	19	0.67	0	0

Figure 3.3. Chronogram and polymorphism information of East African Cichlid lineages. A, Number of SNPs out of 88 coincident Malawi SNPs that are polymorphic; B-D, lineage minor allele frequency patterns of several SNP examples; B, SNP Aln112626_241 shows widespread polymorphism in eight out of twelve lineages outside of Lake Malawi; C, SNP Aln116141_779 shares polymorphism with riverine haplochromines which belong to a sister clade; D, SNP Aln104822_926 is technically not polymorphic in each of the Lake Tanganyika lineages but frequent fixation of alternate alleles indicates early ancestral origins of the polymorphism.

common ancester and their subsequent maintenance in extant species, has been found to be prevalent in intra-lucastrine cichlids. (Moran and Kornfield 1993, Nagl *et al.* 1998, Koblmuller *et al*. 2010). Using the set of 180 Lake Malawi SNPs, we conducted a finer resolution study of polymorphism sharing by dividing the cichlids outside of Lake Malawi into 12 previously known lineages (see Methods and Figure 3.3). Table 3.2 shows the distribution of the 88 coincident Malawi SNPs based on the number of lineages outside of Lake Malawi that is also polymorphic.

Table 3.2. Distribution of the 88 coincident SNPs based on the number of lineages outside of Lake Malawi that is also polymorphic.

Number of lineages (outside malawi) that are also polymorphic	8	7	6	5	4	3	2	1
Number of Malawi SNPs	1	2	4	4	7	16	19	35
Cumulative number of Malawi SNPs	1	3	7	11	18	34	53	88
Cumulative percentage over 180 Malawi SNPs	0.6	1.7	3.9	6.1	10.0	18.9	29.4	48.9

Fifty-three of these coincident SNPs had polymorphisms in at least two non-Malawi lineages (example in Figure 3.3, column B). This could mean that at least three independent mutations (including within Lake Malawi) had occurred at exactly the same nucleotide position to produce the coincident SNP, but this is very unlikely. It is thus likely that the coincident SNPs were the result of ancestral polymorphisms that had been maintained since the lineage splits. Even from among the 35 Malawi SNPs that were found to be polymorpic in only one other lineage outside of Lake Malawi, 3 and 24 SNPs were polymorphic within the sister clade of Lake Victoria superflock and riverine (which includes many species of the *Astatotilapia* genus) cichlids respectivey (example in Figure 3.3, column C). Given that the polymorphism is mostly shared between sister

clades, and having found a close relationship between *Astatotilapia* and Lake Malawi cichlids (see genetic admixture section below), it is therefore reasonable to expect that these coincident SNPs could be the result of ancestral polymorphisms. Also, there were several SNPs whereby fixation of alternate alleles was frequently observed among lineages (example in Figure 3.3, column D). These lineages had to have been polymorphic at some earlier time along the lineage branch, thus "adding" to the total number of polymorphic lineages and making multiple independent coincident mutations even more unlikely. We thus believe that a significant proportion of the coincident SNPs would have been inherited ancestrally, initiated either by a mutation event in a common ancestor, or from a very early hybridization event that introduced the polymorphism to the ancestors of currently polymorphic lineages. Recent hybridization between species across different lakes is unlikely, as the lakes are geographically distinct and hundreds of miles apart.

We also found that the level of Malawi-Tanganyika polymorphism sharing (32.3%) was higher than Tanganyika-Victoria sharing (23.3%), which was in turn higher than Malawi-Victoria polymorphism sharing (8.5%). This was not expected, given that well established phylogenies show the Lake Victoria superflock being a sister clade to the Lake Malawi assemblage, to the exclusion of the Lake Tanganyika assemblage (Meyer 1993). However, it has been suggested that the cichlids of Lake Victoria experienced a severe population bottleneck when the lake was thought have dried out and refilled about 14,000 years ago (Johnson *et al.* 1996, Seehausen 2002, Verheyen *et al.* 2003), and this bottleneck could possibly explain the reduced polymorphism sharing of Lake Victoria polymorphisms.

Our finding of extensive ancestral polymorphism sharing across lakes sheds new light on the often observed evolution of similar traits in cichlids from different lakes, such as physical morphologies (fusiform bodies, fleshly lips, nuchal humps, horizontal striping

etc.; Kocher *et al.* 1993) , behaviour (brood-care; Goodwin *et al.* 1998), or even molecular changes (rhodopsin genes; Suguwara *et al.* 2005). While many of these examples have been often been drawn from comparisons of species in Lake Malawi and Tanganyika, the evolution of similar traits are also present in Lake Victoria cichlids (Salzburger *et al.* 2007, Ole Seehausen, personal communication). One of these earlier reports by Kocher and colleagues (1993) tested the genetic divergence of a group of Lake Malawi cichlids from their "twins" in Lake Tanganyika, and concluded that this phenomenon of similar trait evolution was caused by morphological convergence and not migration of ancestral species across lakes. Our data suggests that such textbook examples of 'convergent' evolution could in fact be the result of deeply rooted molecular parallelisms.

3.4.3 Genetic clustering of East African cichlids

We first investigated how our cichlid samples would be genetically clustered based on their genotypes, blind from any prior knowledge of species lineages or phylogeny. We applied a bayesian analysis using the STRUCTURE package (Pritchard *et al.* 2000), which found that our samples were best described by six genetic clusters (see Methods; mean ln probability of data = -28,353.7). The inferred ancestry of each of the 563 cichlid samples was calculated and reported as the fraction assigned to each of the 6 clusters (Figure 3.4). We observed two general patterns of inferred ancestries. A majority of the cichlids displayed a pattern of single ancestry, where they were assigned to a single genetic cluster. The remaining cichlid samples had admixed ancestry patterns, with genetic contributions from two or more of the six genetic clusters (discussed in the next section).

The cichlids found with single ancestry were divided into six groups based on the genetic clusters that they had been assigned to. Matching up the cichlids assigned to

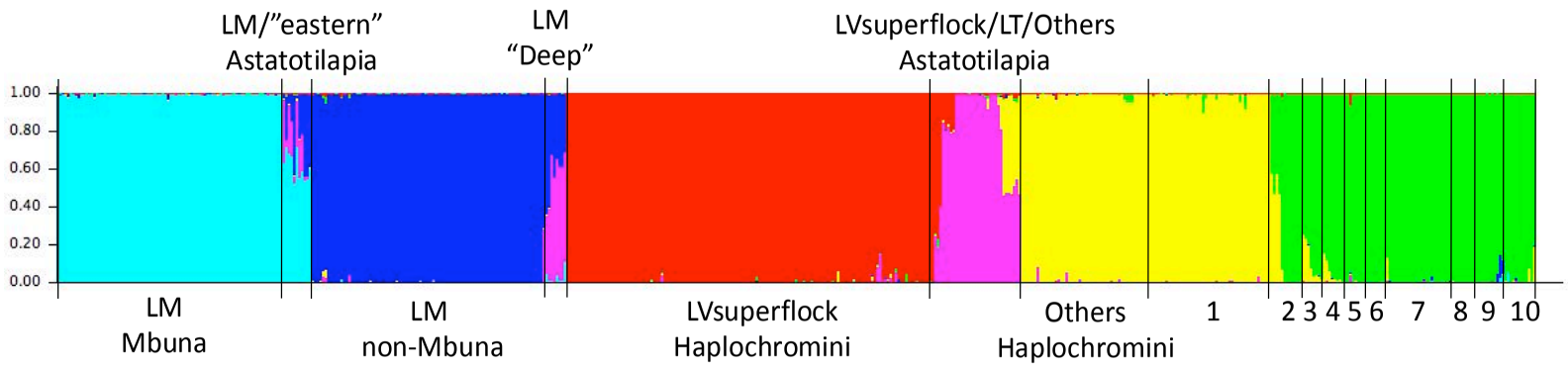


Figure 3.4. Bayesian assignment of individual cichlid samples into six genetic clusters. The color chart is made up of 563 individual vertical bars, each representing a single cichlid sample and proportionally colored based on their fraction assignment to the six clusters. Black vertical bars split the chart into segments, with each segment label describing the catchment (LM, LVsuperflock, LT, Others etc.) followed by lineage (Mbuna, Haplochromini, Ectodini etc). LM, Lake Malawi; LV, Lake Victoria; LT, Lake Tanganyika; 1, LT Haplochromini/Tropheini; 2, LT Limnochromini; 3, LT Ectodini; 4, LT Cyprichromini; 5, LT Cyphotilapiini; 6, LT Perissodini; 7, LT Lamprologini; 8, LT Eretmodini; 9, LT Bathybatini; 10, LT Tilapiini.

these groups with their actual species identities, we found that these six groupings corresponded very well to known cichlid lineages. For example, the cichlids belonging to the first group, represented by the light blue color in Figure 3.4, was found to contain all of the samples of the mbuna (rock-dwelling) lineage of Lake Malawi that was used in this study. Two other groups showed similar exact correspondence to known lineages: the non-mbuna lineage of Lake Malawi; the Lake Victoria superflock of cichlids. The three remaining groups generally corresponded well to known lineages, the Lake Tanganyika and other African Haplochromini and Tropheini tribes, other evolutionarily older cichlid tribes of Lake Tanganyika, and cichlids from the Astatotilapia genus, though a small number of species within these latter three groups displayed admixed ancestries (discussed in next section). Overall, the results obtained here show that these lineages, known to be separated due to allotropy or very early divergences within their respective catchments, are also well diverged genetically and enough to be distinct and distinguishable from one another.

In addition, this current study genotyped SNPs identified, and therefore predominantly polymorphic, in Lake Malawi (180), Lake Victoria (21), Lake Tanganyika (9) and other rivers and drainages (9). The Lake Malawi SNPs represented a more than two-and-a-half fold increase from our earlier study (Loh *et al.* 2008), but did not further resolve beyond the three main Lake Malawi lineages previously observed (mbuna, non-mbuna and deep water species). This strongly suggests that the species within each lineage had not yet sufficiently diverged to be further separated into smaller cluster groupings. The same may not be concluded for the Lake Victoria and Tanganyika lineages though, as the ascertainment bias caused by the low number of Victoria and Tanganyika SNPs used yields less predictive power. However, the two groupings obtained in Lake Tanganyika cichlids, compared to the single group for the Lake Victoria

superflock, despite the smaller number of Tanganyika-specific SNPs, can be attributed to the large number of Lake Malawi SNPs that also share polymorphism with Tanganyika cichlids (see above). Future genotyping studies increasing the number of SNPs identified from Lake Victoria and Taganyika cichlids may yield further cluster separation within these groups.

3.4.4 Genetic admixture in cichlid species

The STRUCTURE analysis revealed appreciable levels of admixture in certain cichlid species, with generally consistent admixture patterns among multiple samples within a species. Our most interesting result was several different admixture patterns belonging to different species and populations of the *Astatotilapia* genus (Figure 3.4), which belongs to one of the few genera that can be found distributed throughout Africa and not endemic to any one location. Previous studies had also postulated that members of the Astatotilapia genus had contributed genetically to the genomes of Lake Malawi cichlids (Seehausen *et al*. 2003, Loh *et al*. 2008, Joyce *et al*. 2011). As a basis for comparison, *Astatotilapia burtoni* from Lake Tanganyika and the connected Kalambo river, as well as *Astatotilapia desfontainii* from Tunisia in North Africa, had displayed single ancestry (discussed above) genetic patterns unique to Astatotilapias (i.e. pink color in Figure 3.4).

Astatotilapia calliptera from Lake Malawi displayed an admixture of mainly Lake Malawi mbuna and lower levels of non-mbuna (14%) and *Astatotilapia* (18%) contribution. However, this admixture pattern, with low levels of *Astatotilapia* contribution, need not be taken to necessarily imply high divergence of these species from other *Astatotilapia* species found elsewhere. Rather, it serves to emphasize *A. calliptera's* extremely close relationship with Lake Malawi cichlids, possibly due to its genetic contribution to Lake Malawi cichlids. In addition, we now also observe that almost half the contribution to the *Rhamphochromis*, *Diplotaxodon* and *Pallidochromis*

genera, which represent the deep water and pelagic lineages of Lake Malawi and thought to be evolutionarily basal to the mbuna and non-mbuna lineages, are actually of *Astatotilapia* origins, where previously the contribution was thought to be specific to the deep water lineage when the sample set then contained only Lake Malawi cichlids (Loh *et al.* 2008). Our current findings further support the hypothesis that Lake Malawi was possibly founded by one or more *Astatotilapia* ancestors from which the mbuna, non-mbuna and deepwater genomes have emerged.

Interestingly, several other species of the *Astatotilapia* genus (*A. swynnertoni*, and other undescribed *Astatotilapia*), sampled from other locations of the "eastern" Indian Ocean drainage systems (Lake Chilwa and Buzi river), also displayed the same admixture pattern as Lake Malawi *A. calliptera*. The clustering and sharing of admixture patterns by these allopatric lineages suggests that the Lake Malawi flock is not monophyletic. Lake Malawi non-monophyly has recently been demonstrated in a mitochondrial study using these same samples (Joyce *et al.* 2011). Our SNP genotyping adds further nuclear DNA support to the evidence from mitochondrial data. Yet other *Astatotilapia* species (*A. bloyeti*, *A. flavijospehi*, *A. tweddlei* and some *A. burtoni* populations), collected from around Africa (outside of Lakes Malawi, Tanganyika, Victoria superflock), displayed admixture with either Lake Victoria superflock or Lake Tanganyika and Riverine *Haplochromini* genomes.

Finally, several species of Lake Tanganyika *Limnochromini*, *Ectodini* and *Cyprichromini* tribes show genomic contributions from the evolutionarily younger *Haplochromini*/*Tropheini* tribes (Salzburger and Meyer 2004). This repeated but similar genomic admixture pattern over several different tribes suggests that cross species hybridization might have occurred. Together, these tribes are also the youngest within Lake Tanganyika, which is in line with the observation that genetic admixture is not

prevalent among older Lake Tanganyika tribes, as species hybridization would be less likely given that more genetic incompatibiities would have been accumulated.

Combining both genetic clustering and admixture analyses, this study revealed a logical continuum of cluster and admixture patterns, from the Lake Tanganyika haplochromines, to the non-endemic *Astatotilapia* genus, and onward to the Lake Malawi assemblage and the Lake Victoria superflock. It strongly suggests an extensive role played by the *Astatotilapia* genus in expanding the East African cichlid radiation. This continuum is also visible in the context of cichlid phylogeographic distribution (Figure 3.1), where we observed spatial concentrations of the predominant genetic clusters at the major lakes of Malawi, Victoria and Tanganyika, with directionally influenced admixture patterns in the intervening rivers and lakes. These results are in agreement with earlier findings that the haplochromines expanded out of Lake Tanganyika to populate the all the major lakes, rivers and drainage systems of East Africa (Salzburger *et al.* 2002, 2004, 2005).

3.4.5 Genetic divergence in Lake Malawi cichlids

For each SNP genotyped, we calculated the F_{ST} value (Weir & Cockerham 1984) which measures the levels of genetic differentiation among Lake Malawi cichlid populations (Figure 3.5). This was performed at different evolutionary levels among (i) the major lineages of mbunas (M), non-mbunas (N) and deep-water species (D); (ii) all combinations of pairwise populations of M, N and D. (iii) all genus level populations (with at least five cichlid samples); and (iv) populations of the genus *Labeotropheus* and *Metriaclima*, which have often been used in previous Lake Malawi cichlid evolutionary studies (eg. Albertson *et al.* 2003, Streelman and Albertson 2006, Loh *et al.* 2008). The median genetic differentiation found in these comparisons ranged from 0.020 to 0.209 (mean range: 0.167 to 0.302), indicating that genetic variation mostly segregates within

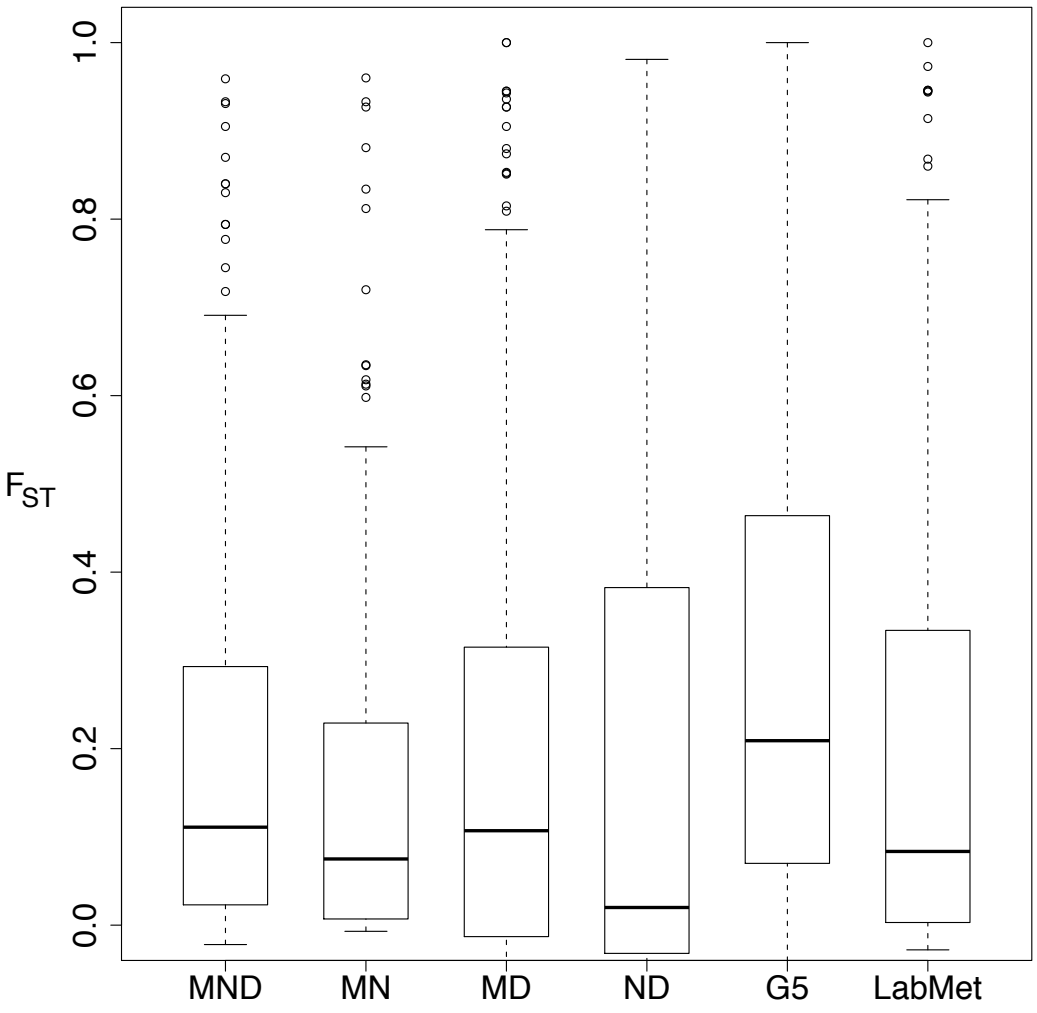


Figure 3.5. F_{ST} distribution and outliers with significant genetic differentiation. A) Box-and-whisker plots of F_{ST} distribution with upper and lower box bounds representing 75th and 25th percentiles, respectively. The solid lines within boxes represent the median value. Whiskers mark the furthest points from the median that are not classified as outliers. Unfilled circles represent outliers that are more than 1.5 times the interquartile range higher than the upper box bound. Category labels describe the populations used in the F_{ST} calculation: MND, Mbuna versus Non-Mbuna versus Deep; MN, Mbuna versus Non-Mbuna; MD, Mbuna versus Deep; ND, Non-Mbuna versus Deep; G5, genus populations with more than 5 samples; LabMet, *Labeotropheus* versus *Metriaclima*.

and not between lineages. This finding is also reflected by our observation that only 5 out of 180 Malawi-identified SNPs were differentially fixed at the species level, while the remaining SNPs showed widespread polymorphism still being maintained in many species.

We were interested to discover SNP loci that displayed high F_{ST} values that were outliers to their own empirical distribution, which would then be indicative of high genetic differentiation. A simple strategy of assigning the upper tail ends of F_{ST} histograms as outliers had been used previously (Luikart *et al.* 2003), and was found to fare no worse (Narum and Hess 2011) than more sophisticated methods which incorporate different evolutionary models and/or heterozygosity correlations (e.g. FDIST2, Beaumont and Nichols 1996; LOSITAN, Antao *et al.* 2008; Arlequin, Excoffier and Lischer 2010; BayeScan, Foll and Gaggiotti 2008). We applied boxplot statistics to the empirical distribution in order to determine outliers, an additional statistical filter to the histogram strategy. We had used this same F_{ST} outlier approach in an earlier study to detect genetic divergence (Loh *et al.* 2008), and it has proven to produce significant results. Two out of eight F_{ST} outlier loci detected in that study, in the *irx1* and *ptc2* genes, have been further studied in the time since publication and shown to be associated with developmental brain patterning (Slyvester *et al.* 2010) and craniofacial development (Roberts R and Kocher TD, unpublished) respectively.

We found that in the MND, MN, MD and LabMet analyses, an average of 7.9% of SNPs were statistical outliers with high F_{ST} values (Figure 3.5). We note that results of the MND analysis would be correlated to the subsequent three pairwise analyses, and expected to see that a MND F_{ST} outlier would necessarily produce two high (but not necessarily an outlier) and one low F_{ST} calculation among the three pairwise analyses. However, performing these three additional analyses remained valuable as they may also reveal additiona F_{ST} outliers that are biologically relevant only to the pair of

populations being tested and not the third. The ND and G5 analyses did not yield any significant outliers, as the F_{ST} distribution had a wider spread of intermediate values (compare box bounds in Figure 3.5). Nonetheless, we do observe high F_{ST} values of 1 (alternately fixed in populations) or slightly below, which could still be biologically relevant.

In total, we identified 33 SNP loci as F_{ST} outliers. This included a mix of both genic and non-genic loci. Thirty-six percent of the outliers could be inferred as recent SNPs, as their polymorphism is only present within Lake Malawi, while the remaining 64% share ancestral polymorphism outside the lake and could be inferred as old. The outlier SNPs included some of the loci that were picked up in our previous study (*rh1, csrp, irx1, ptc2*; Loh *et al.* 2008), plus several other interesting genes. One of them is the *transforming growth factor beta 2* (*tgfb2*) gene, which showed strong genetic differentiation between mbuna and other Lake Malawi cichlids (non-mbunas and "deep" lineages). *tgfb2* belongs to a superfamily of multifunctional cytokines with important regulatory roles during development, including neuromuscular (McLennen and Koishi 2002), eye (Saika 2006), cranofacial (Behnan et al. 2005) and tooth (Huang and Chai 200) development – topics that are frequently studied in cichlids (see Introduction).

It was recently reported that divergent selection on miRNA target sites may have contributed to the diversification of cichlids (Loh *et al.* 2011). The same *hoxa10* SNP, found in that study to be a well differentiated miRNA target site and predicted to influence muscle development and regeneration, was also found here to be significantly differentiated between mbuna and non-mbuna. *dicer 1*, found here to be well differentiated in mbuna from other Malawi cichlids, is a key processing enzyme which cleaves double-stranded RNAs and pre-microRNAs in the RNA interference (RNAi) and microRNA (miRNA) pathways (Jaskiewicz and Filipowicz 2008). These links to miRNA regulation makes the differentiation found here in *dicer 1* and *hoxa10* very interesting

leads for further study. The full list of outlier SNPs and their associated genes are provided in Table 3.3.

Table 3.3. List of outlier SNPs and calculated F_{ST} values. Significant outlier F_{ST} values highlighted in red. Associated gene names printed in grey represent the closest gene within 100 kilobases from SNP position. Dashes indicate no F_{ST} values calculated due to monomorphism among populations. NA, not applicable.

SNP Name	SNP origin*	Associated Gene#	MND	MN	MD	ND	G5	LabMet
Aln101510_393	recent	transforming growth factor, beta 2	0.959	0.96	0.936	-0.032	0.948	0.072
Aln102749_378	old	glutamate receptor, ionotrophic, AMPA 4	0.933	-0.001	1	0.981	0.879	-
Aln102504_1609	old	iroquois homeobox protein 1, b	0.931	0.933	1	-0.033	1	-
Aln113666_686	old	dicer 1, ribonuclease type III	0.905	0.927	0.565	0.768	0.986	0.001
Aln110417_383	recent	neuroligin 1	0.87	0.881	0.262	0.77	0.909	-
Aln105577_385	recent	TOX high mobility group box family member 3	0.84	-	0.945	0.95	1	-
Aln103506_276	recent	pre-B-cell leukemia homeobox 3	0.84	-	0.945	0.95	1	-
Aln103131_1413	old	NA	0.83	0.834	-0.025	0.77	0.769	0.072
Aln102321_608	old	Zic family member 1 (odd-paired homolog, Drosophila)	0.794	-	0.927	0.933	0.917	-
Aln118947_983	recent	tubulin folding cofactor D	0.794	-	0.927	0.933	0.917	-
Aln104822_926	old	solute carrier family 4, anion exchanger, member 1	0.777	0.812	0.852	0.21	0.782	-
Aln101222_933	recent	serine palmitoyltransferase, long chain base subunit 3	0.745	-	0.905	0.914	0.835	-
Aln112709_570	old	CUB and Sushi multiple domains 2	0.718	0.72	0.809	0.033	0.723	0.275
Aln100532_2174	old	potassium channel, subfamily K, member 9	0.691	-	0.88	0.891	0.741	-
Aln109969_676	recent	homeobox A10	0.626	0.635	-	0.484	0.566	-
Aln105584_365	old	cathepsin A	0.622	0.634	0.459	-0.032	0.668	0.817
Aln106343_852	recent	homeobox B9	0.599	0.618	0.26	0.403	0.707	-
Aln103262_483	old	chromodomain helicase DNA binding protein 4	0.649	0.613	0.24	0.979	0.542	0.064
Aln112165_601	old	NA	0.6	0.611	0.44	-0.032	0.586	0.316
Aln100281_1741	old	patched 1	0.592	0.598	0.588	-0.004	0.728	0.914
Aln102003_434	old	thrombospondin, type I, domain containing 7A	0.562	0.034	0.943	0.809	0.956	-
Aln104744_1075	old	POU class 3 homeobox 3	0.559	0.542	0.874	0.175	0.57	-0.019
Aln110178_952	old	ATPase, Na+/K+ transporting, alpha 2 polypeptide	0.273	0.087	0.853	0.481	0.517	-
Aln102499_612	recent	PRKC, apoptosis, WT1, regulator	0.636		0.851	0.864	0.659	-
Aln113582_375	old	membrane frizzled-related protein	0.271	0.042	0.815	0.514	0.454	-
Aln102027_539	old	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	0.236	0.249	0.097	-	1	1
Aln105956_1118	recent	carbonyl reductase 1	0.498	0.511	0.324	-	0.699	0.973
Aln101293_1168	old	membrane protein, palmitoylated 2	0.08	0.021	0.257	0.408	0.425	0.946
csrp1	recent	cysteine and glycine-rich protein 1	0.348	0.361	0.188	-	0.783	0.946
Aln107567_398	old	NA	0.376	0.39	0.213	-	0.878	0.945
rhodopsin	old	rhodopsin	0.42	0.376	0.204	0.848	0.666	0.944
Aln103439_528	recent	NA	0.378	0.392	0.217	-	0.633	0.868
Aln122064_679	old	aquaporin 3 (Gill blood group)	0.451	0.463	0.285	-0.037	0.717	0.86

* SNP origin defined as recent if polymorphism is present only in Lake Malawi, or old if polymorphism is shared with lineages outside Lake Malawi
# Due to lack of cichlid genome annotation, the gene associated with a SNP is determined via comparative analyses with other fish genomes.

3.5 Conclusion

The high species richness and rapid evolutionary radiation of East African cichlids continue to remain an intriguing question studied by evolutionary biologists. The rapid technological advances in genome sequencing and other molecular techniques over the last decade have allowed us to obtain a closer peek into the genetic variation of cichlids. Our study traced the evolution of cichlid genetic structure, and showed the close relationship between the riverine *Astatotilapia* genus and the Malawi assemblage, and that the Malawi assemblage is non-monophyletic. High genetic differentiation was found in a small subset of loci with interesting gene associations, which will allow us to initiate

future investigations into the functional underpinnings of adaptive evolution. More significantly, knowing that the high levels of cichlid morphological and behavioral diversity had arisen from relatively low levels of genetic variation (Loh *et al*. 2008), we have found here (focusing on Lake Malawi cichlids but with evidence pointing to the same trends in other East African cichlids) that in addition to more recently-arisen mutations within the flock, a significant portion of genetic variation had been inherited ancestrally prior to the diversification of the species flocks. Together with repeated hybridization and introgressions that are known to occur within the lakes (Salzburger *et al*. 2002b, Bell and Travis 2005, Joyce et al. 2011), these mechanisms together serve to maintain the high levels of allele sharing and polymorphisms (i.e. standing variation) among cichlids. Adaptive diversifications from standing variation, for multiple reasons, is likely to occur much faster: beneficial alleles are immediately available; alleles usually start at higher frequencies with higher fixation probabilities; the allele is "older", and might have been pre-tested by selection in other environments, thus increasing the likelihood of large beneficial effects (Barrett and Schluter 2008). Conversely, mathematical modelling of the speciation process involving new mutations generally found waiting times for speciation to occur to be extremely long (Gavrilets 2003). In addition, parallel evolution of similar traits, as is often observed in cichlids, is much more probably from selection on standing variation, as was the case demonstrated by parallel evolution of freshwater stickleback adaptations from their marine ancestors (Schluter and Conte 2009). Overall, this study suggests that the rapid radiation of cichlid diversity in Lake Malawi was probably greatly influenced by high standing genetic variation shared across East Africa, though diversity arising from new mutations was also involved. This is a phenomenon that might be shared by other rapidly radiating model systems.

3.6 Acknowledgements

We thank members of the Streelman lab for comments and discussions relating to this study, and the National Human Genome Research Institute for funding the genotyping experiments under the auspices of the International Cichlid Genome Consortium.

3.7 References

Albertson RC, Streelman JT, Kocher TD. 2003. Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. *Proc Natl Acad Sci USA*. 100:5252-5257.

Albertson RC, Streelman JT, Kocher TD, Yelick PC. 2005. Integration and evolution of the cichlid mandible: the molecular basis of alternative feeding strategies. *Proc Natl Acad Sci USA*. 102:16287-16292.

Allender CJ, Seehausen O, Knight ME, Turner GF, Maclean N. 2003. Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptual coloration. *Proc Natl Acad Sci USA*. 100:14074-14079.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G. 2008. LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*. 9:323.

Aubin-Horth N, Desjardins JK, Martei YM, Balshine S, Hofmann HA. 2007. Masculinized dominant females in a cooperatively breeding species. *Mol Ecol*. 16:1349-1358.

Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. *Trends Ecol Evol*. 23(1):38-44.

Beaumont MA, Nichols RA. 1996. Evaluating loci for use in the genetic analysis of population structure. (1996). *Proc R Soc Lond B*. 263:1619-1626.

Behnan SM, Guo C, Gong TW, Shum L, Gong SG. 2005. Gene and protein expression of transforming growth factor beta 2 gene during murine primary palatogenesis. *Differentiation*. 73(5):233-239.

Bell MA, Travis MP. 2005. Hybridization, transgressive segregation, genetic covariation, and adaptive radiation. Trends Ecol Evol. 20(7):358-361.

Blais J, Rico C, van Oosterhout C, Cable J, Turner GF, Bernatchez L. 2007. MHC adaptive divergence between closely related and sympatric African cichlids. *PloS ONE*. 2:e734.

Cristescu ME, Adamowicz SJ, Vaillant JJ, Haffner DG. 2010. Ancient lakes revisited: from the ecology to the genetics of speciation. *Mol Ecol*. 19(22):4837-4851.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol*. 14(8):2611-2620.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour*. 10:564–567

Foll M, Gaggiotti O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 180(2):977-993.

Fraser GJ, Bloomquist RF, Streelman JT. 2008. A periodic pattern generator for dental diversity. *BMC Biol*. 6:32.

Gavrilets S. 2003. Perspective: models of speciation: what have we learned in 40 years? *Evolution*. 57(10):2197-2215.

Gavrilets S, Losos JB. 2009. Adaptive radiation: contrasting theory with data. *Science*. 323(5915):732-737.

Goodwin NB, Balshine-Earn S, Reynolds JD. 1998. Evolutionary transitions in parental care in cichlid fish. *Proc R Soc Lond B*. 265(1412):2265-2272.

Goudet J. 1995. FSTAT (Version 1.2): A computer program to calculate F-statistics. *J Hered*. 86:485-486.

Hodgkinson A, Ladoukakis E, Eyre-Walker A. 2009. Cryptic variation in the human mutation rate. *PLoS Biol*. 7(2):e1000027.

Hodgkinson A, Eyre-Walker A. 2010. The genomic distribution and local context of coincident SNPs in human and chimpanzee. *Genome Biol Evol*. 2:547-557.

Huang XF, Chai Y. 2010. TGF-beta signalling and tooth development. *Chin J Dent Res*. 13(1):7-15.

Huber R, van Staaden MJ, Kaufman LS, Liem KF. 1997. Microhabitat use, trophic patterns, and the evolution of brain structure in African cichlids. *Brain Behav Evol*. 50:167-182.

Jaskiewicz L, Filipowicz W. 2008. Role of Dicer in posttranscriptional RNA silencing. *Curr Top Microbiol Immunol*. 320:77-97.

Johnson TC, Scholz CA, Talbot MR, Kelts K, Ricketts RD, Ngobi G, Beuning K, Ssemmanda II, McGill JW. 1996. Late Pleistocene Desiccation of Lake Victoria and Rapid Evolution of Cichlid Fishes. *Science*. 273(5278):1091-1093.

Joyce DA, Lunt DH, Genner MJ, Turner GF, Bills R, Seehausen O. 2011. Repeated colonization and hybridization in Lake Malawi cichlids. *Curr Biol*. 21(6):526.

Koblmüller S, Egger B, Sturmbauer C, Sefc KM. 2010. Rapid radiation, ancient incomplete lineage sorting and ancient hybridization in the endemic Lake Tanganyika cichlid tribe Tropheini. *Mol Phylogenet Evol*. 55(1):318-334.

Kocher TD, Conroy JA, McKaye KR, Stauffer JR. 1993. Similar morphologies of cichlid fish in Lakes Tanganyika and Malawi are due to convergence. *Mol Phylogenet Evol*. 2(2):158-165.

Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat Rev Genet*. 5:288-298.

Kornfield I, Smith PF. 2000. African cichlid fishes: model systems for evolutionary biology. *Ann Rev Ecol Evol Syst*. 31:163-196.

Kuraku S, Meyer A. 2008. Genomic analysis of cichlid fish 'natural mutants'. *Curr Opin Genet Dev*. 18(6):551-558.

Lee B-Y, Hulata G, Kocher TD. 2004. Two unlinked loci controlling the sex of blue tilapia (*Oreochromis aureus*). *Heredity*. 92:543-549.

Lee B-Y, Lee W-J, Streelman JT, Carleton KL, Howe AE, Hulata G, Slettan A, Stern JE, Terai Y, Kocher TD. 2005. A second-generation genetic linkage map of tilapia (*Oreochromis* spp). *Genetics*. 170:237-244.

Loh YH, Katz LS, Mims MC, Kocher TD, Yi SV, Streelman JT. 2008. Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biol.* 9(7):R113.

Loh YH, Yi SV, Streelman JT. 2011. Evolution of microRNAs and the diversification of species. *Genome Biol Evol*. 3:55-65.

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P. 2003. The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet. 4(12):981-994.

McLennan IS, Koishi K. 2002. The transforming growth factor-betas: multifaceted regulators of the development and maintenance of skeletal muscles, motoneurons and Schwann cells. *Int J Dev Biol*. 46(4):559-67.

Meyer A. 1993. Phylogenetic relationships and evolutionary processes in East African cichlid fishes. *Trends Ecol Evol*. 8(8):279-284.

Moran P, Kornfield I. 1993. Retention of ancestral polymorphism in the Mbuna species flock of Lake Malawi. *Mol Biol Evol*. 10:1015-1029.

Nagl S, Tichy H, Mayer WE, Takahata N, Klein J. 1998. Persistence of neutral polymorphisms in Lake Victoria cichlid fish. *Proc Natl Acad Sci USA*. 24:14238-14243.

Narum SR, Hess JE. 2011. Comparison of F(ST) outlier tests for SNP loci under selection. Mol Ecol Resour. 11(1):184-194.

Parry JW, Carleton KL, Spady T, Carboo A, Hunt DM, Bowmaker JK. 2005. Mix and match color vision: tuning spectral sensitivity by differential gene expression in Lake Malawi cichlids. *Curr Biol*. 15:1734-1739.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945-959.

Saika S. 2006. TGFbeta pathobiology in the eye. *Lab Invest*. 86(2):106-115.

Salzburger W, Meyer A, Baric S, Verheyen E, Sturmbauer C. 2002. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Syst Biol*. 51(1):113-135.

Salzburger W, Baric S, Sturmbauer C. 2002b. Speciation via introgressive hybridization in East African cichlids. *Mol Ecol*. 11:619-625.

Salzburger W, Meyer A. 2004. The species flocks of East African cichlid fishes: recent advances in molecular phylogenetics and population genetics. *Naturwissenschaften*. 91(6):277-290.

Salzburger W, Mack T, Verheyen E, Meyer A. 2005. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. *BMC Evol Biol*. 5:17.

Salzburger W, Braasch I, Meyer A. 2007. Adaptive sequence evolution in a color gene involved in the formation of the characteristic egg-dummies of male haplochromine cichlid fishes. *BMC Biol*. 5:51.

Salzburger W. 2009. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. *Mol Ecol*. 18(2):169-185.

Schluter D, Conte GL. 2009. Genetics and ecological speciation. *Proc Natl Acad Sci U S A*. 106:9955-62.

Seehausen O. 2002. Patterns in fish radiation are compatible with Pleistocene desiccation of Lake Victoria and 14,600 year history for its cichlid species flock. *Proc Biol Sci*. 269(1490):491-497.

Seehausen O, Terai Y, Magalhaes IS, Carleton KL, Mrosso HD, Miyagi R, van der Sluijs I, Schneider MV, Maan ME, Tachida H, Imai H, Okada N. 2008. Speciation through sensory drive in cichlid fish. *Nature*. 455(7213):620-626.

Ser JR, Roberts RB, Kocher TD. 2010. Multiple interacting loci control sex determination in lake Malawi cichlid fish. *Evolution*. 64(2):486-501.

Simões JM, Duarte IS, Fonseca PJ, Turner GF, Amorim MC. 2008. Courtship and agonistic sounds by the cichlid fish *Pseudotropheus zebra*. *J Acoust Soc Am.* 124(2):1332-8.

Spady TC, Seehausen O, Loew ER, Jordan RC, Kocher TD, Carleton KL. 2005. Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid fishes. *Mol Biol Evol*. 22:1412-1422.

Streelman JT, Albertson RC, Kocher TD. 2003. Genome mapping of the orange blotch colour pattern in cichlid fishes. *Mol Ecol*. 12:2465-2471.

Streelman JT, Albertson RC. 2006. Evolution of novelty in the cichlid dentition. *J Exp Zoolog B Mol Dev Evol*. 306:216-226.

Sugawara T, Terai Y, Imai H, Turner GF, Koblmüller S, Sturmbauer C, Shichida Y, Okada N. 2005. Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi. *Proc Natl Acad Sci U S A*. 102(15):5448-5453.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 9(6):465-476.

Sylvester JB, Rich CA, Loh YH, van Staaden MJ, Graser GJ, Streelman JT. 2010. Brain diversity evolves via differences in patterning. *Proc Natl Acad Sci USA.* 107(21):9718-9723.

Turner GF. 2007. Adaptive radiation of cichlid fish. *Curr Biol*. 17(19):R827-831.

Verheyen E, Salzburger W, Snoeks J, Meyer A. 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. *Science*. 2003 300(5617):325-329.

Verzijden MN, van Heusden J, Bouton N, Witte F, ten Cate C, Slabbekoorn H. 2010. Sounds of male Lake Victoria cichlids vary within and between species and affect female mate preferences. *Behav Ecol.* 21(3):548-555.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution*. 38:1358-1370.

CHAPTER 4

EVOLUTION OF MICRORNAS AND THE DIVERSIFICATION OF SPECIES[2]

4.1 Abstract

MicroRNAs (miRNAs) are ancient, short, non-coding RNA molecules that regulate the transcriptome through post-transcriptional mechanisms. miRNA riboregulation is involved in a diverse range of biological processes and mis-regulation is implicated in disease. It is generally thought that miRNAs function to canalize cellular outputs, for instance as 'fail-safe' repressors of gene mis-expression. Genomic surveys in humans have revealed reduced genetic polymorphism and the signature of negative selection for both miRNAs themselves and the target sequences to which they are predicted to bind. We investigated the evolution of miRNAs and their binding sites across cichlid fishes from Lake Malawi (East Africa), where hundreds of diverse species have evolved in the last million years. Using low-coverage genome sequence data, we identified 100 cichlid miRNA genes with mature regions that are highly conserved in other animal species. We computationally predicted target sites on the 3'-UTRs of cichlid genes to which miRNAs may bind, and found that these sites possessed elevated single nucleotide polymorphism (SNP) densities. Furthermore, polymorphic sites in predicted miRNA targets showed higher minor allele frequencies on average and greater genetic differentiation between Malawi lineages when compared to a neutral expectation and non-target 3' UTR SNPs. Our data suggest that divergent selection on miRNA riboregulation may have contributed to the diversification of cichlid species, and may similarly play a role in rapid phenotypic evolution of other natural systems.

---

[2] Loh YH, Yi SV, Streelman JT. 2011. Evolution of microRNAs and the diversification of species. *Genome Biol Evol*. 3:55-65.

4.2 Introduction

Ever since King and Wilson compared protein sequence between chimpanzee and human and concluded that there was insufficient coding divergence to explain phenotypic differences (King and Wilson 1975), biologists have highlighted regulatory change in gene expression as a source for adaptive evolution (Wray 2007, Carroll 2008). There is now ample direct evidence that cis-acting mutations cause phenotypic variation among closely related organisms by modulating gene expression (Sucena *et al*. 2003, Miller *et al*. 2007). These data, coupled with the signature of divergent and positive selection at putative gene regulatory elements (Haygood *et al*. 2007, Sethupathy *et al*. 2008), have established the general consensus that 5' promoters act as evolutionary engines of transcriptional change (e.g., "tinker where the tinkering's good" [Rockman and Stern 2008]).

Plausible scenarios for the evolution of animal diversity hinge on the ever-growing complexity of 5' promoters and the modification of transcriptional regulatory networks (Levine and Tjian 2003). Notably, evolutionary 'tinkering' with transcription at 5' promoters may have evolved in concert with post-transcriptional safeguards encoded at the 3' end of cistrons. Reports suggest that microRNAs (miRNAs), potent agents of riboregulation, are as old as metazoan 5' cis-regulatory logic (Grimson *et al*. 2008, Wheeler *et al*. 2009). miRNAs are short (~22 nucleotides), endogenous, non-coding RNA molecules that regulate gene expression after transcription. Generally, animal miRNA targeting is achieved by complementary base pairing between the miRNA and specific sequences in the 3' untranslated region (3'-UTR) of messenger RNAs (mRNAs). Target recognition is thought to be determined by perfect Watson-Crick base pairing at a miRNA 'seed' region (base positions 2-7 counting from the 5' end; [Lewis *et al*. 2005]), although this is not a necessary condition and targeting may include other determinants

(Grimson *et al*. 2007, Barbato *et al*. 2009). Transcript silencing then occurs through inhibition of translation, or via mRNA degradation (Bartel 2004). Individual miRNAs may regulate hundreds of loci and it has been estimated that a majority of human genes are potential miRNA targets (Lewis *et al*. 2005, Friedman *et al*. 2009).

MicroRNAs generally act as 'fail-safe' buffers against gene mis-expression in time and/or space, in effect canalizing the transcriptome (Carrington and Ambrose 2003, Stark *et al*. 2005). Consistent with this notion, miRNA mis-expression and/or genetic polymorphism in target sequences can cause abnormality and disease (Clop *et al.* 2006, Sethupathy and Collins 2008, Eberhart *et al*. 2008, Mencía *et al*. 2009). Likewise, and in contrast to predicted transcription factor binding sites in 5' promoters, human miRNAs and their 3' UTR target sequences evolve under purifying selection (Sethupathy *et al*. 2008, Chen and Rajewsky 2006, Saunders *et al*. 2007).

As humans and chimps diverged from a common ancestor during the last 5-7 million years, the East African Rift lakes Tanganyika, Malawi and Victoria spawned three of the most spectacular evolutionary radiations known to biology (Kornfield and Smith 2000, Salzburger *et al*. 2005). In Lake Malawi alone, hundreds of cichlid fish species have evolved from a common ancestor over the last million years (Won *et al*. 2005). These species are remarkably diverse in size, shape, color and behavior (Streelman *et al*. 2003, Albertson *et al*. 2005, Fraser *et al*. 2008, Carleton *et al*. 2008, Sylvester *et al*. 2010), yet their genomes are highly similar and share ancestral polymorphism (Moran and Kornfield 1993, Loh *et al*. 2008). We have shown recently that most of the genome is not genetically differentiated among Malawi species and major lineages; only 2-4% of single nucleotide polymorphism (SNP) loci exhibit the statistical signature of strong evolutionary divergence (Loh *et al*. 2008). Cichlids are models of the mapping of phenotype to genotype; the problem of so many biological species in so little time

(Kocher 2004) is equally matched by the problem of rapid diversification and evolutionary novelty (Streelman *et al*. 2007).

We hypothesized that divergence of miRNAs or their target sequences might be one of the genomic mechanisms contributing to the rapid phenotypic evolution observed in Lake Malawi cichlids. To this end, we analyzed available low-coverage genome sequence and SNP data (Loh *et al*. 2008) and computationally identified (i) putative cichlid miRNAs and (ii) the target sequences in 3' UTRs to which miRNAs may bind. Most studies of miRNA focus on evolutionary conservation of the molecules and their target sites (Barbato *et al*. 2009, Bartel 2004, Alexiou *et al*. 2009). Our goal of evaluating the link(s) between miRNAs, polymorphism in putative miRNA targets and diversity among Lake Malawi cichlid species predicates that we not only consider target sequences conserved for hundreds of millions of years, but also those that may have evolved more recently. Such 'non-conserved' targets are known to be functional and may be generated by single mutations to standing sequence (Clop *et al*. 2006, Farh *et al*. 2005).

We observed that predicted cichlid mature miRNAs are strongly conserved in sequence. On the other hand, miRNA targets exhibited greater SNP densities than flanking sequences and the overall 3' UTR average. Moreover, polymorphic sites in target sequences showed higher minor allele frequencies and divergence among Malawi evolutionary lineages when compared against a neutral expectation and non-target SNPs in the same set of 3'-UTRs. Our data reveal a signature of divergent selection on cichlid miRNA binding sites and suggest an evolutionary role for miRNA riboregulation in the diversification of species.

4.3 Materials and methods

4.3.1 Lake Malawi Genomes

We obtained Lake Malawi cichlid genomic data, consisting of 304,310 sequences from 5 species, 25,458 multi-species alignments and 32,417 SNPs, from a previous study (Loh *et al*. 2008), which applied various criteria to ensure that alignments are allelic and not products of paralogous loci. Sequence data were generated by the Sanger method, allowing the detection of variable sites with an even distribution across the dataset and with high confidence (Loh *et al*. 2008). Examination of these data and subsequent genotyping revealed very low genetic variation, and the persistence of ancestral polymorphism across the Malawi cichlid flock. Molecular genetic analyses across multiple cichlid species are thus highly analogous to within-species polymorphism studies conducted in other organisms (e.g., humans; [Chen and Rajewsky 2006, Saunders *et al*. 2007]). Our use of the term “SNP” in this context therefore extends to include variable sites across multiple cichlid species (see Loh *et al*. 2008 for more details).

4.3.2 miRNA Gene Detection

A database of 623 known teleost precursor miRNA (pre-miRNA) sequences was downloaded from miRBase Release 14.0 (Griffiths-Jones *et al*. 2008). To detect miRNA genes in cichlids, we conducted a BLASTN similarity search of these pre-miRNAs against the cichlid genomic sequences described above, with an E-value cutoff of 0.001. The BLASTN hits were then manually inspected and compared to their query sequences in order to extract adjacent nucleotides that might form part of the pre-miRNA. RNA secondary structure of the cichlid putative miRNA sequences was predicted using Mfold (Zuker 2003) to ensure proper stem-loop folding, and excess bases were trimmed. A reciprocal BLASTN of the putative cichlid miRNAs against known teleost miRNAs was performed to identify the cichlid miRNA and to assign orthology. Multiple sequence

alignments of the putative cichlid miRNAs and their orthologs were then generated using ClustalW (Larkin *et al*. 2007). Mutations in the alignments were marked and counted based on the region (mature miRNA, stem or loop) where they reside.

4.3.3 3'-UTR Annotation

Cichlid genomes have yet to be fully sequenced and annotated; therefore we first annotated cichlid 3'-UTRs from partial genomic sequence. We chose to work with genomic and not transcript sequences because our ultimate goal was to map SNPs to putative miRNA targets found within 3'-UTRs (below); SNP data exist for genome survey sequences (Loh *et al*. 2008), but not for the small number of publicly available cichlid ESTs. Sequences used to annotate cichlid 3'-UTRs include *Fugu rubripes*, *Tetraodon nigroviridis*, *Oryzias latipes*, *Gasterosteus aculeatus* and *Danio rerio* proteins (98,037 entries) downloaded from Ensembl Version 56, all *Actinopterygii* proteins (41,746 entries) from Refseq Release 39, and all *Eukaryota* proteins (158,696 entries) from UniProtKB/Swiss-Prot Release 2010_02 databases.

We applied the TBLASTN algorithm with an E-value cutoff of $1e^{-10}$, to identify similarity between the protein sequences above and cichlid multi-species alignments (Loh *et al*. 2008). High-scoring Segment Pairs (HSPs) of the TBLASTN output with lengths of at least 30 amino acids were parsed and retained, and in cases where the end position of a HSP query was found to be within 3 amino acids from the known 3'-end of the full-length query protein, it was deemed that a corresponding cichlid coding region might also have ended in this region. We further looked within the ±9 nucleotide region of the HSP subject (i.e. cichlid) end to confirm that a stop codon was indeed present and in frame with codon phase of the HSP. Cichlid 3'-UTRs were thus annotated to begin at the next nucleotide beyond the stop codon and presumed to continue for 500 nucleotides in length. This approximation of 3'-UTR length was based on a calculation of

the mean 3'-UTR length in zebrafish (513 nucleotides), as annotated by Ensembl. During our work on this project, an additional ~56,000 unique ESTs were released for the tilapia cichlid, roughly 10-15 million years divergent from the Malawi assemblage (Lee *et al*. 2010). Comparing our annotations to these data, we observed that 66% of our predicted 3'-UTRs showed significant similarity (E-value < $1e^{-5}$) to ESTs.

4.3.4 miRNA Target Prediction

A total of 249 unique mature miRNA sequences, consolidated from the 623 known pre-miRNAs from *Fugu*, *Tetraodon* and *Danio* (miRBase), and the 100 derived from miRNA loci in cichlids (this study), was used for the prediction of target sites on annotated cichlid 3'-UTRs. The target prediction algorithm (hereby termed the SeedMatch algorithm) was written in Perl programming language, implementing the seed-matching requirements similar to that of TargetScanS (Lewis *et al*. 2005): namely, (i) a six nucleotide Watson-Crick complementary match between miRNA and mRNA at position 2-7 of the miRNA, plus (ii) an anchor of either an adenosine at the mRNA target aligned to miRNA position 1, and/or a Watson-Crick match at position 8 of the miRNA.

Conservation of predicted cichlid miRNA target sites in other fish species was determined by (i) generating multiple sequence alignments (MLAGAN; [Brudno *et al*. 2003) of cichlid 3'-UTRs and their orthologs (when determined) in pufferfishes, medaka, stickleback and zebrafish, (ii) applying the SeedMatch algorithm separately to each sequence in the multiple alignment to identify target sites, and (iii) calling a cichlid target site conserved when an identical target site was found in at least one other fish at a location within 50 nucleotide positions along the alignment. We defined conservation as such, in contrast to other target prediction strategies requiring strict conservation across multiple species (Barbato *et al*. 2009, Alexiou *et al*. 2009) for two reasons. First, the fishes with complete genome sequences noted above are all at least 100 million years

divergent from Malawi cichlids. Second, fish genomes are generally more divergent with greater neutral nucleotide substitution rates compared to mammals (Brunet *et al*. 2006). The latter consideration influences the degree of target conservation observed between species, and also our initial task of generating robust multiple sequence alignments.

4.3.5 Target SNP Density Calculations

Subsequent to predicting miRNA targets sites on 3'-UTRs, we mapped SNPs to these same data (Loh *et al*. 2008). For statistical analysis of observed SNP densities in predicted miRNA targets, we obtained a distribution of randomized target SNP densities by running 1000 simulations that permute the occurrence of SNPs along the 3'-UTRs. In each simulated run, every empirical SNP in the 3'-UTRs was shuffled to a random position maintaining the same trinucleotide sequence (i.e., the SNP position itself and the nucleotides immediately before and after). For example, a G[A/T]C trinucleotide where [A/T] represents the SNP would be shifted to a random GAC or GTC position. The 'randomized' target SNP density was then calculated for each run. This simulation strategy controls for neighbor-dependant mutation rates and has been used previously to investigate SNP densities in miRNA target sites (Hiard *et al*. 2010).

4.3.6 3'-UTR Re-sequencing, Alignment and Target Prediction

The analyses described above using data from Loh *et al*. (2008) allow us to identify cichlid miRNAs, their putative targets, and to calculate SNP densities in target sequence. However, because those data do not represent full genomes from the 5 species sequenced, alignments of orthologous sequence rarely contain more than 3 species (Loh *et al*. 2008). To better understand evolutionary processes acting on putative cichlid miRNA target sequences, we re-sequenced annotated 3'-UTRs in a diverse and standardized collection of species. Polymerase Chain Reaction (PCR) primers were

designed (Appendix C Table C2) and used for amplification and sequencing of a subset of annotated 3'-UTRs from the genomic DNA of eight individuals: *Labeotropheus fuelleborni* (LF)*, Melanochromis auratus* (MA) and *Maylandia zebra* (MZ) are members of the rock-dwelling mbuna lineage; *Tyrranochromis maculiceps* (TM)*, Docimodus evelynae* (DE)*, Nimbochromis polystigma* (NP) and *Mchenga conophorus* (MC) belong to a sister lineage of pelagic and sand-dwelling species (henceforth termed non-mbuna); *Rhamphochromis esox* (RE) represents an early-diverging, deepwater group within the radiation (pictures at http://www.malawicichlids.com). The individuals of LF, MA, MZ, MC and RE were those survey-sequenced by the JGI (Loh *et al*. 2008). Sequences were aligned using ClustalW (Larkin *et al*. 2007), from which polymorphic positions were identified at locations exhibiting at least 7 species depth of coverage (Appendix C File C2). We applied the target site prediction algorithms and SNP density calculations to these data as described above. We also carried out additional analyses, described below, with these re-sequenced data.

4.3.7 Minor Allele Frequencies of SNPs in Re-Sequenced 3'-UTRs

We calculated the minor allele frequency (MAF) of each SNP (in and out of putative miRNA targets) identified in the re-sequenced data set. We then compared these MAF distributions to a neutral expectation. From a set of 70 non-genic SNPs typed across a diverse mix (183 individuals, 62 species) of Lake Malawi cichlids (Cichlid Genome Consortium, Broad Institute), we randomly sampled eight individuals to match our re-sequenced 3'-UTR data set (three mbuna, four non-mbuna and one deepwater species) and calculated the allele frequency distribution of the sample. This process was repeated 1000 times to approximate a neutral distribution of allele frequencies and the 95% confidence intervals at each allele frequency. Because we sequenced and re-sampled 8 individuals or 16 total alleles, the empirical and simulated allele frequency

data are largely discrete, with the majority of observations falling around multiples of 1/8 (0.125). Therefore, bins were set around multiples of 0.125 and bin edges fall at the midpoint of consecutive bins; for example the first bin edge (0.1675) is the midpoint between 0.125 and 0.25. Z-tests were implemented within each allele frequency bin, to detect significant shifts in the proportion of SNPs exhibiting that particular range of MAFs, between empirical and re-sampled neutral distributions.

4.3.8 Genetic Differentiation of High-MAF Target SNPs in Re-Sequenced 3'-UTRs

We observed that SNPs in predicted targets exhibited higher minor allele frequencies than expected under neutrality. To test whether these high-MAF (31.25 < MAF < 50%) miRNA target SNPs exhibited greater genetic differentiation among Malawi lineages than expected under neutrality, we generated 1000 sets of matching 'neutral' genotype data using the same non-genic SNP dataset and sampling strategy described above. For each set of genotypic data, we calculated for each SNP the (i) overall population, (ii) mbuna and (iii) non-mbuna allele frequencies, where each allele frequency value lies between 0 and 1. We defined a SNP as displaying clear lineage-specific differentiation when the difference in mbuna and non-mbuna allele frequencies was equal or greater than 0.75, and hence calculated the proportion of high-MAF SNPs that were well differentiated between lineages. Values were aggregated for the 1000 data sets to obtain a distribution from which a Z-test was used to determine the statistical significance of our observed data.

4.4 Results

4.4.1 miRNA Prediction

We used a reference set of 623 known teleost pre-miRNA sequences from *Fugu, Tetraodon and Danio*, obtained from miRBase Release 14.0 (Griffiths-Jones *et al*. 2008),

in a similarity search (see Methods) against a database of 304,310 cichlid genomic sequences (Loh *et al*. 2008). We manually curated the similarity hits to extract putative cichlid pre-miRNAs, and confirmed that they were able to fold into the secondary stem-loop structure necessary for miRNA biogenesis (Bartel 2004). This resulted in the identification of 100 distinct cichlid pre-miRNA genes (Appendix C Table C1) that produce 87 unique mature miRNAs.

We compared cichlid pre-miRNA loci to their orthologues in other fish species and found a total of 1002 out of 6422 nucleotide positions where substitutions had occurred. This results in an overall nucleotide divergence of 0.156 (variable sites/nucleotide positions). When the pre-miRNAs were divided into mature miRNA, stem and loop regions (Figure 4.1A), we observed nucleotide divergences of 0.015, 0.172 and 0.485 respectively (Figure 4.1B), with no mutations found in the miRNA 'seeds.' A similar trend of region-specific variation holds for the subset of substitutions where cichlids exhibit a different nucleotide than all other species; a divergence of 0.008, 0.060 and 0.185 at the mature miRNA, stem and loop regions respectively (Figure 4.1B).

4.4.2 Polymorphism in Cichlid miRNA Targets

To study genetic variation in putative cichlid miRNA targets, we mapped SNPs (Loh *et al*. 2008) to target sequences predicted to fall within 3'-UTRs. We first annotated 731 cichlid 3'-UTRs (Appendix C File C1) that contained 367 SNPs (0.28% SNP density). To direct our computational prediction of targets, we used 249 unique mature miRNAs, derived from miRNA loci in cichlids (above) as well as known miRNAs from other fish species *Fugu*, *Tetraodon* and *Danio*. These miRNAs are highly conserved among vertebrates; 86% are in miRNA families that extend outside of fishes. Note that the 100 cichlid miRNAs we identified here (above) possess identical 'seed' sequences to their

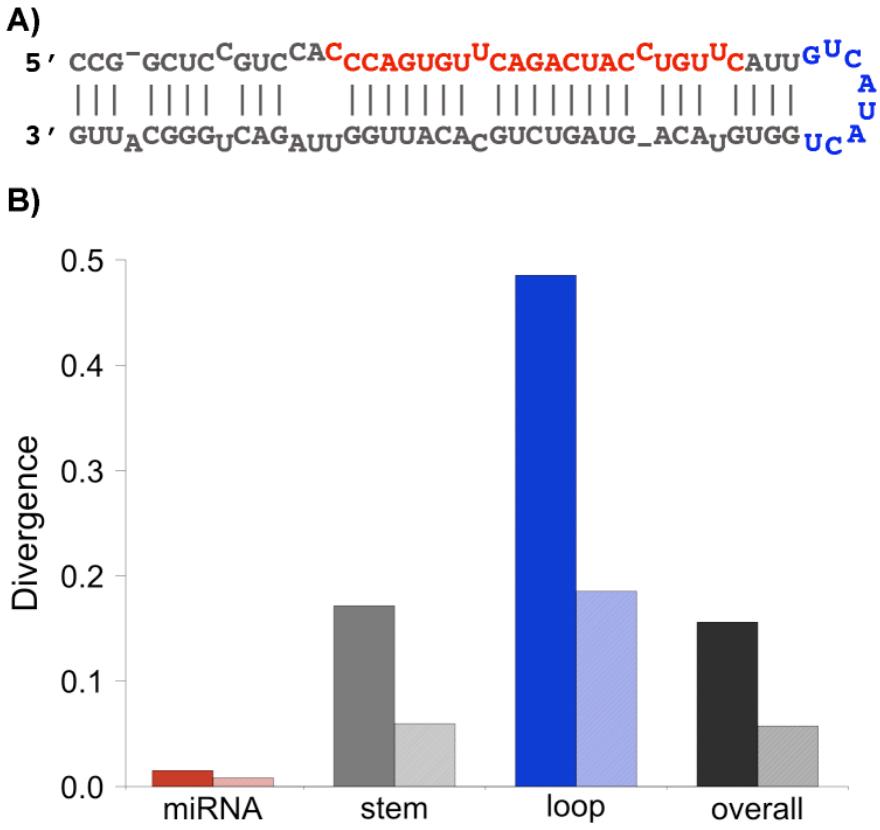


Figure 4.1. Evolutionary divergence in pre-miRNA sequences. A) An example of predicted stem-loop secondary structure for a cichlid miRNA (lfu-mir-199-1 shown here), classified into separate regions for analysis. Nucleotide symbols are colored red for the mature miRNA region, blue for the loop region, and grey for the stem region excluding the mature miRNA. Vertical bars represent Watson-Crick or G:U base-pairing matches. B) Distribution of divergence across different regions of the pre-miRNA. Bar colors correspond to the regions defined in A., with black representing the divergence over the entire molecule. Solid-colored bars are calculated from all observed variable sites. Shaded bars are calculated from variable sites where cichlids displayed a different nucleotide than all other species.

fish orthologues; this justifies our use of additional fish miRNAs, conserved among vertebrates but not yet identified in cichlids (see below), to facilitate target prediction.

Putative miRNA binding sites in 3'-UTR sequences were predicted using a Perl script written to implement a 'SeedMatch' algorithm incorporating rules similar to those of TargetScanS (Lewis *et al*. 2005). Briefly, 7- and 8-mer target sites were identified that had exact Watson-Crick base-pair matches at 'seed' sequences (positions 2-7 counting from the miRNA 5' end), plus a corresponding base anchor at position 1 and/or 8 (see Methods).

Considering all putative 3'UTRs identified from the Loh *et al*. (2008) data, we detected 6,299 miRNA target sites on 719 of 731 3'-UTR sequences (an average of 8.62 miRNA target sites per 3'-UTR; Table 4.1). As expected, we observed overlaps among predicted target sites for multiple miRNAs; 13.0% of the total 3'-UTR length (39,660 nucleotides) was predicted to be bound by one or more miRNA(s), similar to results reported in human and mouse (Hiard *et al*. 2010). Seventy-eight SNPs mapped within 17,607 informative bases of miRNA target sites. Thus, the SNP density for miRNA target sites is 0.44%, higher than (i) the average 3'-UTR SNP density (0.28%), (ii) the SNP densities of target flanking sequence (0.21-0.28%) and (iii) the average 'randomized'

Table 4.1. miRNA target prediction results on all putative and select re-sequenced 3'-UTRs.

	All 731 putative 3'-UTRs			130 re-sequenced 3'-UTRs		
	All (731 UTRs)	Conserved Targets (481 UTRs)	Non-Conserved Targets (481 UTRs)	All (130 UTRs)	Conserved Targets (124 UTRs)	Non-Conserved Targets (124 UTRs)
Number of targets predicted	6,299	875	3307	1,296	360	639
Number of targets (per 3'-UTR)	8.62	1.82	6.88	9.97	2.90	5.15
Total coverage of targets (nt)	39,660	5,505	21,157	6,602	2,159	4,089
3'UTR coverage by targets (%)	13.0	2.7	10.5	13.7	4.76	9.01
Informative sites within targets* (nt)	17,607	2,761	9,355	6,602	2,159	4,089
Number of SNPs in targets	78	8	40	40	7	29
SNP density in targets (%)	0.443	0.290	0.428	0.606	0.324	0.709

* only a subset of 3'-UTR positions had multi-species sequence data to determine polymorphism.

target SNP density of 0.28% (Z-test, $P=2.41\times10^{-6}$; Figure 4.2A). For reference, the SNP densities of synonymous and replacement coding sites in the same set of data is 0.42% and 0.20%, respectively (Loh *et al*. 2008).

Enforcing a criterion of target site conservation reduced the size of our data set considerably (see Methods and below; Table 4.1). We assigned orthology to single genes in other fish genomes for 481 out of 731 predicted cichlid 3'-UTRs. Other predicted 3'-UTRs showed similarity to members of gene families, or to specific pairs of duplicated loci, but we could not specify reciprocal orthology with confidence. Conserved sites accounted for 21% of cichlid miRNA targets (875 of 4182), similar to previous study (Friedman *et al*. 2009, Hiard *et al*. 2010), and covered only 2.7% of nucleotides in these 481 3'-UTRs. The SNP density in conserved target sites was 0.29%, similar to the average SNP density for flanking and overall 3'-UTRs and within the 95% confidence interval of randomized target SNP densities (Appendix C Figure C1).

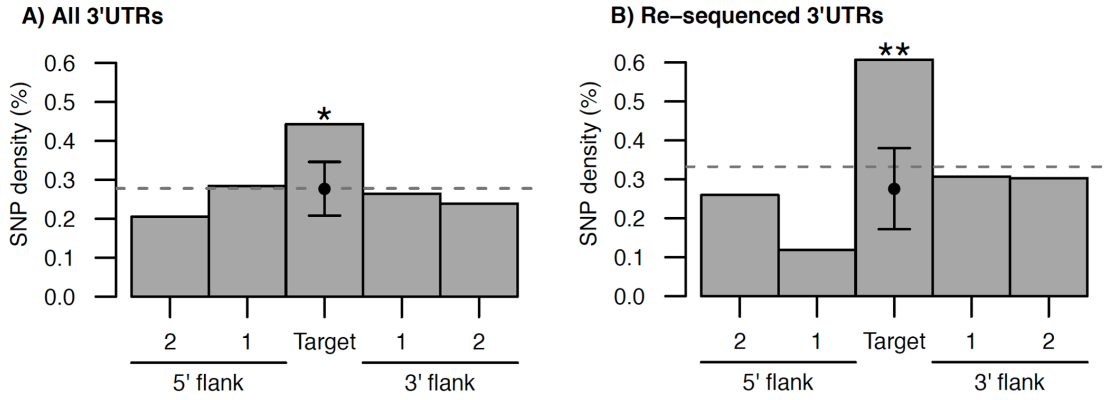


Figure 4.2. SNP densities within computationally predicted miRNA target sites and their flanking regions. Data from A) all predicted 3'-UTRs and B) select re-sequenced 3'-UTRs. Flanking regions 1-2 on both 5' and 3' ends of 'target' represent successive, non-overlapping windows of sizes equal to that of the target sites. Dotted lines show the average 3'-UTR SNP density. Filled circle with error bars represent the mean and 95% confidence intervals of SNP densities calculated from 1000 simulated replicates of randomized SNP shuffling. Asterisk symbols indicate significant deviation from simulated distributions (Z-test, * $P<10^{-5}$; ** $P<10^{-9}$).

4.4.3 MAFs and Genetic Differentiation of 'Target' SNPs in Re-Sequenced 3'-UTRs

We re-sequenced a set of 130 3'-UTRs in eight individuals of Malawi cichlid species spanning a range of morphologies and behaviors, representing the three major evolutionary lineages in the lake (Loh *et al*. 2008, Won *et al*. 2006). Our rationale here was twofold. First, we reasoned that 3'-UTR sequence variation across samples, in and out of putative miRNA target sites, could be examined for the evolutionary signature of natural selection (Chen and Rajewsky 2006, Saunders *et al*. 2007). Second, in order to better validate predicted miRNA-mRNA interactions against previous literature, we chose certain gene subsets whose molecular functions have been well characterized for interactions with miRNAs (e.g., development [Plasterk 2006], immunity [Xiao and Rajewsky 2009]).

From 48,114 base positions of multiple sequence alignments (Appendix C File C2), we identified 160 SNPs, an overall SNP density of 0.33%. We then applied the SeedMatch algorithm to these data. SeedMatch targets covered 6,602 total bases, within which we mapped 40 SNPs (Table 4.1). This resulted in a SNP density in predicted targets of 0.606%, higher than the overall average in re-sequenced data (0.33%), target flanking sequence (0.12-0.31%), and randomized target SNP densities (0.28%; Z-test, $P=4.88\times10^{-10}$; Figure 4.2B). Similar to the analysis of all putative 3'-UTRs (above), enforcing a strong conservation criterion for target sites reduced the size of the data set (only 4.8% of 3'-UTR bases are covered by conserved target sites). Conserved sites accounted for 36% of all targets on 124 cichlid 3'-UTRs; the empirical SNP density in conserved targets was 0.32%, elevated from flanking sequence but similar to the overall 3'-UTR and randomized target SNP densities (Appendix C Figure C1).

Next, we examined the allele frequency distribution of SNPs in predicted miRNA target sites in relation to 3'-UTR non-target sites, compared against a neutral expectation. We approximated a 'neutral' distribution by sub-sampling from a data set of

70 randomly chosen, non-genic SNPs typed in a diverse mix of Lake Malawi cichlids. Significant departure from a neutral distribution of allele frequencies might be indicative of natural selection (Sethupathy *et al*. 2008, Chen and Rajewsky 2006, Drake *et al*. 2006). Notably, allele frequencies at non-target 3'-UTR SNPs did not depart from the neutral distribution (nearly 80% of polymorphisms exhibit minor alleles that are relatively rare), but predicted 'target' SNPs differed significantly, with a bias towards high minor allele frequencies (MAFs, Figure 4.3, Appendix C Figure C2).

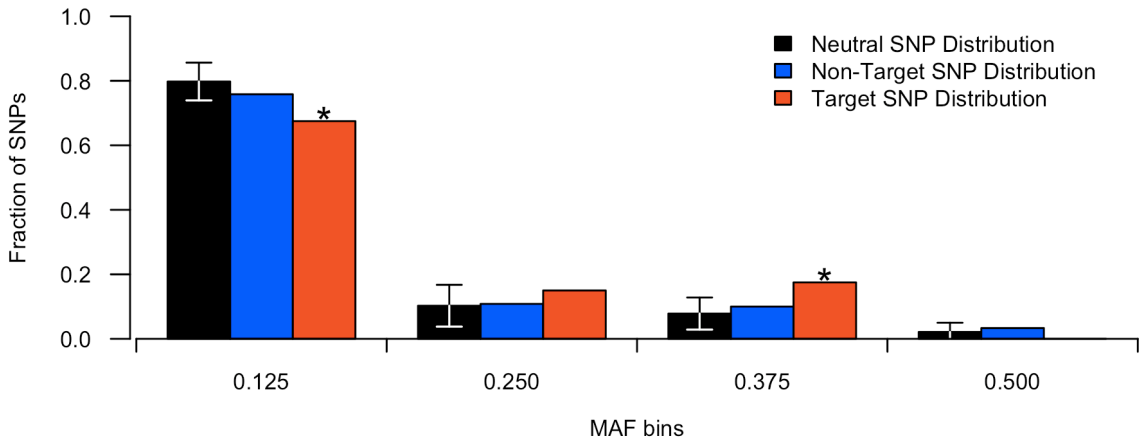


Figure 4.3. Comparison of minor allele frequency distributions. 3'-UTR miRNA target SNPs are colored in red, non-target SNPs in blue and non-genic (neutral) SNPs in black. Error bars represent the 95% confidence interval of the neutral expectation. Asterisk symbols indicate significant deviation from neutral expectation within each bin (Z-test, * $P<10^{-4}$).

We asked if high-MAF SNPs in predicted miRNA targets were differentiated among lineages (i.e., mbuna *vs*. non-mbuna) to a degree beyond expectation under neutrality. We found that a significantly elevated proportion (86%) of high-MAF (31.25 < MAF < 50%) target SNPs exhibit genetic differentiation between Malawi evolutionary groups (Z-

test, $P=9.32\times10^{-7}$). Predicted miRNA-gene interactions, highlighting evolutionarily differentiated SNPs, are shown in Figure 4.4 and discussed below.

4.5 Discussion

Lake Malawi cichlids have evolved in a brief evolutionary window. Their genomes are highly similar and segregate ancestral polymorphism. For comparison, nucleotide diversity across the flock (0.26%, [Loh *et al*. 2008]) is less than that observed among laboratory strains of the zebrafish (0.48%, [Guryev *et al*. 2006]), comparable to that of chimpanzees (0.24%, [Fischer *et al*. 2004]) and humans (0.11%, [The International Hapmap Consortium 2007]), and contrasts against the ~1.2% divergence between chimps and humans (King and Wilson 1975, Chen and Li 2001). It is notable then that the range of variation across Malawi species for many phenotypes (body size, tooth and taste bud number) spans an order of magnitude and that the diversity of other traits (color pattern, feeding and breeding biology, brain organization) is comparable to that observed in other vertebrate taxonomic orders. The cichlid system is thus a model of the genotype to phenotype mapping function (Streelman *et al*. 2007), with speculation revolving around the rapid evolution of novelty. Here, we test the hypothesis that evolutionary divergence of microRNAs and/or their binding sites may have contributed to the diversification of species (Plasterk 2006).

4.5.1 Cichlid miRNA Target Sites Exhibit Elevated SNP Densities

We identified 100 distinct miRNA loci in the genomes of cichlid fishes. The mature miRNAs encoded by these loci are highly conserved among fishes (Figure 4.1B). The trend of higher divergence in stems and loops (*vs*. the mature miRNA) has been observed in other species (Hertel *et al*. 2006), and may be indicative of purifying selection against change to the functional component of the miRNA molecule (and/or a

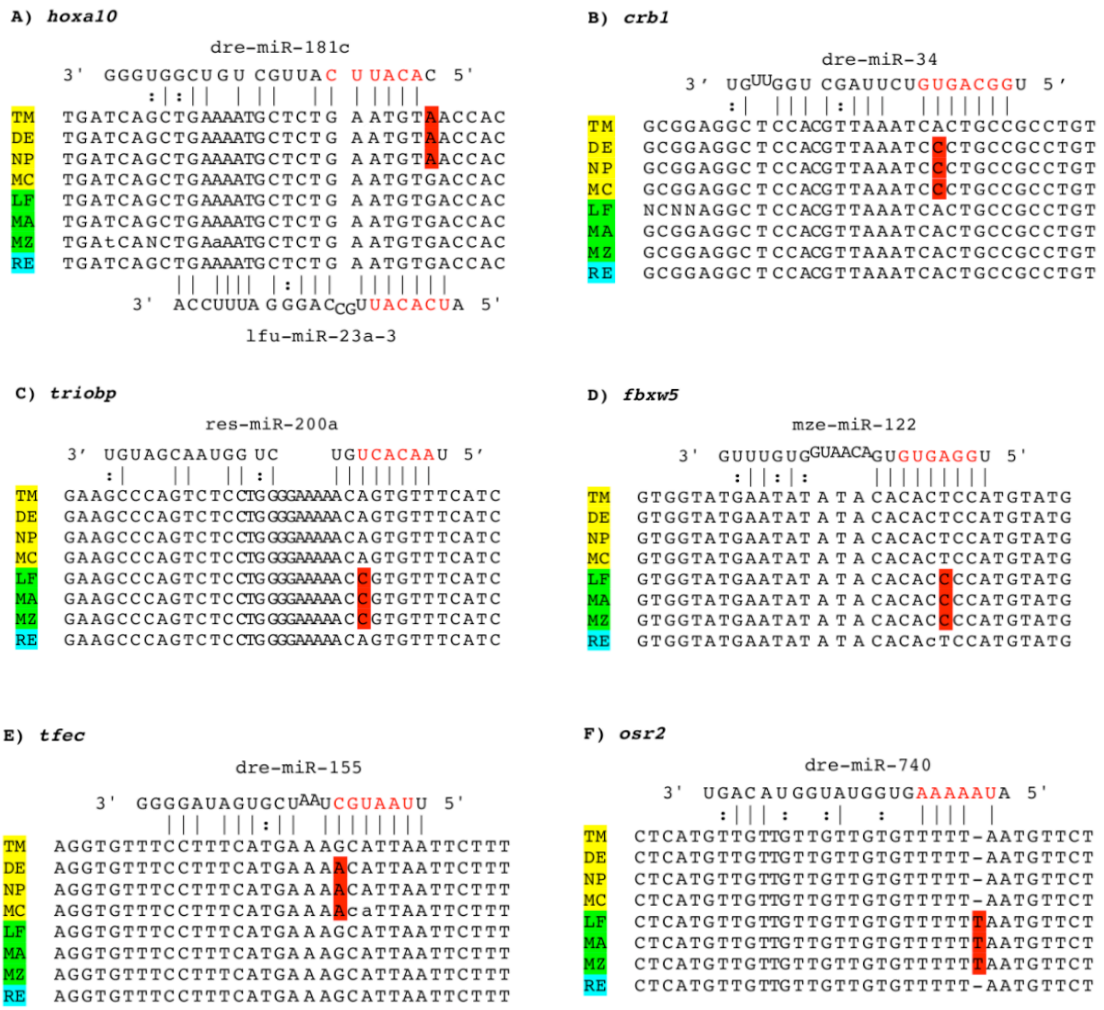


Figure 4.4. Multiple sequence alignments of several miRNA targets containing differentiated SNPs. Red blocks indicate SNP minor alleles. Dashes represent gaps in sequence (indel in *osr2*). miRNAs predicted to bind to the target are shown, with the seed region in red font. Vertical bars represent Watson-Crick base-pairing and colons represent G:U base-pairing. Raised and lowered nucleotides illustrate bulges in the predicted miRNA binding. TM, *Tyrranochromis maculiceps*; DE, *Docimodus evelynae*; NP, *Nimbochromis polystigma*; MC, *Mchenga conophorus*; LF, *Labeotropheus fuelleborni*; MA, *Melanochromis auratus*; MZ, *Maylandia zebra*; RE, *Rhamphochromis esox*. Yellow, green and blue boxes over abbreviated species names represent non-mbuna, mbuna and deepwater lineages respectively.

relaxation of constraint at stems and loops). The number of miRNAs we identified is likely to be an incomplete count, as the available cichlid genomic resources used here comprise only about 32% coverage of the cichlid genome (Loh *et al*. 2008). As a reference, there are 360 zebrafish (characterized from an assembled genome and by deep RNA sequencing; [Wienholds *et al*. 2005, Soares *et al*. 2009]) and 132 pufferfish miRNAs in miRBase.

Predicted miRNA target sites, located in the 3'-UTRs of cichlid genes, showed elevated SNP densities when compared to flanking regions, the overall 3'-UTR average and randomized simulations that account for nucleotide composition (Table 4.1; Figure 4.2). For a more restricted set of evolutionarily conserved targets, SNP densities were not distinguishable from those in flanks, the overall 3'-UTR average and simulation values. This trend held in both the genome-wide 3'-UTR data set and in the directed set of re-sequenced 3'-UTRs. Our observation of elevated or equivalent SNP densities in both conserved and non-conserved miRNA targets runs counter to results from previous study within humans, where average SNP density in predicted target sites (both conserved and non-conserved) was reduced compared to flanking regions (Chen and Rajewsky 2006, Saunders *et al*. 2007).

4.5.2 miRNA Target Sites Show the Signature of Divergent Natural Selection

The observation of increased SNP density at predicted miRNA target sites does not provide conclusive information about the evolutionary forces shaping this pattern; for instance, even though the SNP density of predicted targets is high within the context of 3'-UTR sequence, minor alleles at variable sites could be rare. We therefore re-sequenced a collection of 3'-UTRs in a standard set of species and designed a test to evaluate the allele frequency distribution of (i) SNPs predicted in miRNA binding sites and (ii) other 3'-UTR non-target SNPs, against a neutral expectation. This test is

conceptually similar to the DAF (derived allele frequency) approach (Sethupathy *et al*. 2008, Chen and Rajewsky 2006, Drake *et al*. 2006). However, because Lake Malawi cichlid fishes retain ancestral polymorphism that may pre-date the species flock (Loh *et al*. 2008) we have not attempted to designate ancestral *vs*. derived alleles.

We found that while the allele frequency distribution of non-target SNPs in 3'-UTRs was not different than the neutral expectation, the distribution of predicted miRNA target SNPs was biased towards high minor allele frequencies (MAFs, Figure 4.3). In addition, we observed that 86% of putative miRNA target SNPs with high MAFs showed a clear pattern of evolutionary divergence between major Malawi lineages (Figure 4.4 and below). To put this in greater context, we have previously observed that <5% of haphazardly chosen SNPs are outliers for genetic differentiation in a large sample of mbuna *vs*. non-mbuna (Loh *et al*. 2008). The alternative that the differentiated polymorphisms we highlight in Figure 4.4 are not in fact in miRNA targets, but are each physically linked to other, as yet unidentified nucleotide sites, is unlikely because it would require that we happened upon these unidentified sites in six independent loci through the sole discovery operation of searching for miRNA targets.

Taken together, our observations of (i) elevated SNP densities, (ii) a bias towards high MAFs and (iii) the pattern of genetic differentiation among lineages for high-MAF SNPs suggest that select miRNA binding sites have experienced divergent selection during the evolution of the Lake Malawi species flock.

4.5.3 Differentiated SNPs in miRNA Targets are Biologically Relevant

A secondary goal of our re-sequencing project was to investigate putative miRNA binding site polymorphism in gene sets whose molecular functions have been well-studied *vis-à-vis* miRNAs. We reasoned that such data would add biological plausibility to our computational predictions and population genetic analyses. Figure 4.4 displays

examples of high-MAF SNPs, genetically differentiated among Malawi cichlid lineages, mapped to miRNA target sites in 3'-UTRs. These examples represent miRNA-gene pairs supported by previous research in humans and other model organisms.

The interplay between miRNAs and Hox gene riboregulation is well known (Yekta *et al*. 2008). We predict an association between two miRNAs, miR-181c and miR-23a, which share a target site SNP in the cichlid *hoxa10* 3'-UTR (Figure 4.4A); this target site in *hoxa10* is conserved between cichlid and stickleback. The SNP differentiates non-mbuna predators (TM, DE, NP) from other species. miR-181 is known to target mouse *Hoxa11* (a Hox cluster family member of *hoxa10*) during muscle differentiation (Naguibneva *et al*. 2006); fish *hoxa10* genes are expressed in paired fins and associated musculature (Ahn and Ho 2008). Recently, it has been shown that miR-181 is up-regulated while miR-23 is down-regulated in mouse leg muscle during endurance exercise (Safdar *et al*. 2009). These data raise the possibility that a single SNP modulates the miRNA riboregulation of Hox-mediated fin muscle development and regeneration in Lake Malawi predators.

We highlight two miRNA-gene pairs that may modify sensory modalities among Lake Malawi cichlids. We predict differential binding of miR-34 to cichlid *crb1* (Figure 4.4B), a member of the Crumbs protein complex. *crb1* contributes to photoreceptor morphogenesis and sensitivity, mutations cause retinal degeneration in humans, mice and flies (Bulgakova and Knust 2009). miR-34 is expressed in neural tissue (including the optic tectum) of larval and adult zebrafish (Kapsimali *et al*. 2007), also in the retina of embryonic and adult mice (Arora *et al*. 2010). This association is of particular interest given the vast literature implicating the role of vision in Malawi cichlid ecology, mate choice and evolution (Carleton *et al*. 2008). Next, we predict that the TRIO and F-actin binding protein (*triobp*) is differentially bound by miR-200a (Figure 4.4C). *triobp* functions in the hair cell cilia of the inner ear (Kitajiri *et al*. 2010), mutations result in nonsyndromic

hearing loss (Shahin *et al*. 2006). miR-200a is expressed in sensory epithelia, including those of the inner ear of zebrafish, chicken and mouse (Soukup 2009). Recent reports have linked hearing to mate choice and communication in East African cichlids (Verzijden *et al*. 2010, Simões *et al*. 2008).

Two SNPs are predicted to affect binding of miRNAs to genes involved in immune response (Xiao and Rajewsky 2009). *fbxw5* (Figure 4.4D) is a F-box protein with a role in interleukin signaling (Minoda *et al*. 2009); a T↔C SNP differentiated among Malawi cichlids is predicted to modulate binding of miR-122, a liver-specific miRNA (Soares *et al*. 2009, Sarasin-Filipowicz *et al*. 2009). The miR-122 binding site in *fbxw5* is conserved between cichlid and medaka. Second, *tfec* (Figure 4.4E) is a macrophage-restricted BHLH transcription factor, also involved in interleukin signaling (Rehli *et al*. 2005). We predict that a differentiated G↔A SNP modifies binding of miR-155, a well-known regulator of immune function (O'Connell *et al*. 2009).

Finally, our data may be useful to identify new interactions between miRNAs and genes of interest. For example, we predict that an indel in the 3'-UTR of cichlid *osr2* should differentially regulate binding of miR-740 in mbuna cichlids (LF, MA, MZ) *vs*. others (Figure 4.4F). *Osr2* restricts the teeth of mice to a single row (Zhang *et al*. 2009), among other functions in the craniofacial skeleton. Tooth row number is highly variable among cichlid species (Fraser *et al*. 2008). miR-740 is poorly understood (Kloosterman *et al*. 2006); our data suggest it may play a role in craniofacial development.

4.6 Conclusion

Biologists recognize that 5' cis-acting mutations regulate gene expression and contribute to phenotypic evolution (King and Wilson 1975, Wray 2007, Carroll 2008). Correspondingly, studies have reported the signature of diversifying selection on population genetic variants in computationally predicted 5' promoter elements (Haygood

et al. 2007, Sethupathy *et al*. 2008). The situation is different for 3'-UTRs. miRNAs and their binding sites collaborate as post-transcriptional capacitors to canalize the transcriptome (Carrington and Ambros 2003, Stark *et al*. 2005). Evidence suggests that both miRNAs and their target sequences in 3'-UTRs evolve under purifying selection (Chen and Rajewsky 2006, Saunders *et al*. 2007). Metazoan cistrons may therefore have evolved for transcriptional exploration at 5' promoters, with post-transcriptional safeguards encoded at the back.

We provide evidence that the evolution of miRNA binding sites may play a role in evolutionary diversification. We demonstrate that (i) computationally predicted miRNA targets in cichlid 3'-UTRs harbor elevated SNP densities, that (ii) a greater frequency of polymorphic sites in predicted targets have high minor allele frequencies compared to a neutral expectation and that (iii) these sites are often genetically differentiated among Malawi lineages.

It has been argued that polymorphisms in miRNA target sites are deleterious *within species* because even single base mismatches (especially to the 'seed') can abrogate binding and disrupt riboregulation (Sethupathy *et al*. 2008, Clop *et al*. 2006, Mencía *et al.* 2009). We suggest that mutations in 3'-UTRs where miRNAs may bind, whether breaking transcriptome canalization or introducing new regulation, may contribute to phenotypic differentiation *among* rapidly evolving lineages. Further analyses, with fully annotated and assembled cichlid genomes (http://www.genome.gov/10002154), deeper genotyping, next-generation miRNA and miRNA target prediction algorithms (Barbato *et al.* 2009, Chaudhuri and Chatterjee 2007), and experimental validation of predicted miRNAs and their interactions with target genes (Sethupathy and Collins 2008, Kuhn *et al*. 2008) will reveal additional intricacies of miRNA riboregulation and evolution.

4.7 Acknowledgements

We thank Craig Albertson, Greg Gibson and two anonymous reviewers for comments on previous drafts of this manuscript; Michael Norsworthy for experimental assistance. This work was supported by the National Science Foundation [IOS 0546423]; and the National Institutes of Health [R01 DE019637].

4.8 References

Ahn D, Ho RK. 2008. Tri-phasic expression of posterior Hox genes during development of pectoral fins in zebrafish: implications for the evolution of vertebrate paired appendages. *Dev Biol.* 322(1):220-233.

Albertson RC, Streelman JT, Kocher TD, Yelick PC. 2005. Integration and evolution of the cichlid mandible: the molecular basis of alternative feeding strategies. *Proc Natl Acad Sci USA.* 102(45):16287-16292.

Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. 2009. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*. 25(23):3049-3055.

Arora A *et al.* 2010. Prediction of microRNAs affecting mRNA expression during retinal development. *BMC Dev Biol.* 10:1.

Barbato C *et al*. 2009. Computational challenges in miRNA target predictions: To be or not to be a true target? *J Biomed Biotechnol.* 2009:803069.

Bartel DP. 2004. MicroRNAs:Genomics, biogenesis, mechanism and function. *Cell.* 116(2):281-297.

Brudno M *et al*. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13(4):721-731.

Brunet FG *et al.* 2006. Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. *Mol Biol Evol.* 23(9):1808-1816.

Bulgakova NA, Knust E. 2009. The Crumbs complex: from epithelial-cell polarity to

retinal degeneration. *J Cell Sci.* 122(Pt 15):2587-2596.

Carleton KL *et al*. 2008. Visual sensitivities tuned by heterochronic shifts in opsin gene expression. *BMC Biol.* 6:22.

Carrington JC, Ambrose V. 2003. Role of microRNAs in plant and animal development. *Science.* 301(5631):336-338.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell.* 134(1):25-36.

Chaudhuri K, Chatterjee R. 2007. MicroRNA detection and target prediction: Integration of computational and experimental approaches. *DNA Cell Biol.* 26(5):321-337.

Chen FC, Li WH. 2001. Genomic divergences between humans and homonoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet.* 68(2):444-456.

Chen K, Rajewsky N. 2006. Natural selection on human microRNA binding sites inferred from SNP data. *Nat Genet.* 38(12):1452-1456.

Clop A *et al*. 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet.* 38(7):813-818.

Drake JA *et al.* 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet.* 38(2):223-227.

Eberhart JK *et al*. 2008. MicroRNA mirn140 modulates Pdgf signalling during palatogenesis. *Nat Genet.* 40(3):290-298.

Farh KK *et al.* 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science.* 310(5755):1817-1821.

Fischer A, Wiebe V, Paabo S, Przeworski M. 2004. Evidence for a complex demographic history of chimpanzees. *Mol Biol Evol.* 21(5):799-808.

Fraser GJ, Bloomquist RF, Streelman JT. 2008. A periodic pattern generator for dental diversity. *BMC Biol*. 6:32.

Friedman RC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19(1):92-105.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* 36(Database Issue):D154-158.

Grimson A et al. 2007. MicroRNA targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell.* 27(1):91-105.

Grimson A *et al*. 2008. Early origins and evolution of microRNas and Piwi-interacting RNAs in animals. *Nature.* 455(7217):1193-1197.

Guryev V *et al*. 2006. Genetic variation in the zebrafish. *Genome Res.* 16(4):491-497.

Haygood R, Fedrigo O, Hanson B, Yokoyama KD, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet.* 39(9):1140-1144.

Hertel J *et al*. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics.* 7:25.

Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. 2010. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. *Nucleic Acids Res.* 38: D640-651.

Kapsimali M *et al.* 2007. MicroRNAs show a wide diversity of expression profiles in the developing and mature central nervous system. *Genome Biol.* 8(8):R173.

King MC, Wilson AC. 1975. Evolution at two levels in Humans and Chimpanzees. *Science.* 188(4184):107-116.

Kitajiri S *et al.* 2010. Actin-bundling protein TRIOBP forms resilient rootlets of hair cell stereocilia essential for hearing. *Cell.* 141(5):786-98.

Kloosterman WP *et al*. 2006. Cloning and expression of new microRNAs from zebrafish. *Nucleic Acids Res.* 34(9):2558-2569.

Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model.

Nat Rev Genet. 5(4)**:**288-298.

Kornfield I, Smith PF. 2000. African cichlid fishes: model systems for evolutionary biology. *Ann Rev Ecol Evol Syst.* 31:163-196.

Kuhn DE *et al*. 2008. Experimental validation of miRNA targets. *Methods.* 44(1):47-54.

Larkin MA *et al*. 2007. ClustalW and ClustalX version 2. *Bioinformatics.* 23(21):2947-2948.

Lee BY *et al.* 2010. An EST resource for tilapia based on 17 normalized libraries and assembly of 116,899 sequence tags. *BMC Genomics.* 11:278.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature.* 424(6945):147-151.

Lewis BP, Burge CB, Bartel DP. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell.* 120(1):15-20.

Loh YH *et al*. 2008. Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biol.* 9(7):R113.

Mencía A *et al*. 2009. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat Genet.* 41(5):609-613.

Miller CT *el al*. 2007. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell.* 131(6):1179-1189.

Minoda Y, Sakurai H, Kobayashi T, Yoshimura A, Takaesu G. 2009. An F-box protein, FBXW5, negatively regulates TAK1 MAP3K in the IL-1beta signaling pathway. *Biochem Biophys Res Commun.* 381(3):412-417.

Moran P, Kornfield I. 1993. Retention of ancestral polymorphism in the Mbuna species flock of Lake Malawi. *Mol Biol Evol.* 10:1015-1029.

Naguibneva I *et al*. 2006. The microRNA miR-181 targets the homeobox protein Hox-A11 during mammalian myoblast differentiation. *Nat Cell Biol.* 8(3):278-284.

O'Connell RM, Chaudhuri AA, Rao DS, Baltimore D. 2009. Inositol phosphatase SHIP1 is a primary target of miR-155. *Proc Natl Acad Sci USA.* 106(17):7113-7118.

Plasterk RH. 2006. Micro RNAs in animal development. *Cell* 124(5):877-881.

Rehli M *et al*. 2005. Transcription factor Tfec contributes to the IL-4-inducible expression of a small group of genes in mouse macrophages including the granulocyte colony-stimulating factor receptor. *J Immunol.* 174(11):7711-7722.

Rockman MV, Stern DL. 2008. Tinker where the tinkering's good. *Trends Genet.* 24(7):317-319.

Safdar A, Abadi A, Akhtar M, Hettinga BP, Tarnopolsky MA. 2009. miRNA in the regulation of skeletal muscle adaptation to acute endurance exercise in C57Bl/6J male mice. *PLoS One.* 4(5):e5610.

Salzburger W, Mack T, Verheyen E, Meyer A. 2005. Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. *BMC Evol Biol.* 5(1):17.

Sarasin-Filipowicz M, Krol J, Markiewicz I, Heim MH, Filipowicz W. 2009. Decreased levels of microRNA miR-122 in individuals with hepatitis C responding poorly to interferon therapy. *Nat Med.* 15(1):31-33.

Saunders MA, Liang H, Li WH. 2007. Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci USA.* 104(9):3300-3305.

Sethupathy P, Giang H, Plotkin JB, Hannenhalli S. 2008. Genome-wide analysis of natural selection on human cis-elements. *PLoS One.* 3(9):e3137.

Sethupathy P, Collins FS. 2008. MicroRNA target site polymorphisms and human disease. *Trends Genet.* 24(10):489-497.

Shahin H *et al.* 2006. Mutations in a novel isoform of *TRIOBP* that encodes a filamentous-actin binding protein are responsible for DFNB28 recessive nonsyndromic hearing loss. *Am J Hum Genet.* 78(1):144-152.

Simões JM, Duarte IS, Fonseca PJ, Turner GF, Amorim MC. 2008. Courtship and agonistic sounds by the cichlid fish *Pseudotropheus zebra*. *J Acoust Soc Am.* 124(2):1332-8.

Soares AR *et al*. 2009. Parallel DNA pyrosequencing unviels new zebrafish microRNAs. *BMC Genomics.* 10:195.

Soukup GA. 2009. Little but loud: small RNAs have a resounding affect on ear development. *Brain Res.* 1277:104-114.

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. 2005. Animal microRNA confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell.* 123(6):1133-1146.

Streelman JT, Albertson RC, Kocher TD. 2003. Genome mapping of the orange blotch colour pattern in cichlid fishes. *Mol Ecol.* 12(9):2465-2471.

Streelman JT, Peichel CL, Parichy DM. 2007. Developmental genetics of adaptation in fishes: the case for novelty. *Ann Rev Ecol Evol Syst.* 38:655-681.

Sucena E, Delon I, Jones I, Payre F, Stern DL. 2003. Regulatory evolution of shavenbaby/ovo underlies multiple cases of morphological parallelism. *Nature.* 424(6951):935-938.

Sylvester JB *et al.* 2010. Brain diversity evolves via differences in patterning. *Proc Natl Acad Sci USA.* 107(21):9718-9723.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature.* 449(7164):851-861.

Verzijden MN *et al.* 2010. Sounds of male Lake Victoria cichlids vary within and between species and affect female mate preferences. *Behav Ecol.* 21(3):548-555.

Wheeler BM *et al*. 2009. The deep evolution of metazoan microRNAs. *Evol Dev.* 11(1):50-68.

Wienholds E *et al*. 2005. MicroRNA expression in zebrafish embryonic development. *Science.* 309(5732):310-311.

Won YJ, Sivasundar A, Wang Y, Hey J. 2005. On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence. *Proc Natl Acad Sci USA.* 102(Suppl 1):6581-6586.

Won YJ, Wang Y, Sivasundar A, Raincrow J, Hey J. 2006. Nuclear gene variation and molecular dating of the cichlid species flock of Lake Malawi. *Mol Biol Evol.* 23(4):828-837.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet.* 8(3):206-216.

Xiao C, Rajewsky K. 2009. MicroRNA control in the immune system: basic principles. *Cell.* 136(1):26-36.

Yekta S, Tabin CJ, Bartel DP. 2008. MicroRNAs in the Hox network: an apparent link to posterior prevalence. *Nat Rev Genet.* 9:789-796.

Zhang Z, Lan Y, Chai Y, Jiang R. 2009. Antagonistic actions of Msx1 and Osr2 pattern mammalian teeth into a single row. *Science.* 323(5918):1232-1234.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31(13):3406-3415.

CHAPTER 5

OVERALL CONCLUSIONS

The East African cichlid radiation remains undoubtedly one of the most spectacular radiation of vertebrates known in the natural world. This dissertation encompasses three studies that seek to decipher the underpinnings of such rapid evolutionary diversification, investigated via the genetic variations in East African cichlids in general, but focusing mainly on the cichlids of Lake Malawi.

The first study (Chapter 2) began with the generation of an informative and valuable cichlid genomic resource, from which the general properties of the cichlid genome were characterized, followed by an initial evolutionary analysis of the genetic structure and relationships between Lake Malawi cichlids. We generated five low coverage Lake Malawi cichlid genome assemblies, and were then able to comprehensively quantify the genome-wide extent of genetic variation (single nucleotide polymorphisms). Nucleotide diversity of Malawi cichlids was low at 0.26%, even less than that found among laboratory strains of the zebrafish *Danio rerio*. More significantly, we found that biallelic polymorphisms segregate widely throughout the Malawi species flock, making each species a mosaic of ancestrally polymorphic genomes. Yet these genomes continue to retain clear signals of ancestry that successfully differentiates between the clusters of rock-dwelling mbuna, the pelagic and sand-dwelling non-mbuna, as well as the deep-water *Rhamphochromis*. We also detected loci, involved in important sensory as well as developmental pathways, that exhibited extreme genetic differentiation against a backdrop of shared polymorphism, when studied at different evolutionary scales of within species, between species, and between major lineages.

The second study (Chapter 3) performed an extend genotyping analysis, using SNPs that had been identified from cichlids within different lake catchments, and tested across a large representative set of cichlid samples from across Africa. This allowed us to expand our evolutionary analysis to cover the entire East African cichlid radiation. Astonishingly, more than 40% of Malawi SNPs were found to be also polymorphic in species outside of Lake Malawi, with similar trends of high allele sharing also present in SNPs identified from the other locales. We found that these coincident SNPs were most likely the result of ancestral polymorphism sharing. Bayesian analysis of genetic structure in the data supports the hypothesis that Lake Malawi cichlids are not monophyletic and that riverine species have contributed significantly to their genomes. As with the first study, we were able to further identify additional interesting loci that were well differentiated between species and lineages, and these are ideal candidate genes that should be further studied to uncover genotype-to-phenotype relationships.

The third study (Chapter 4) then investigated cichlid genetic variation in relation to the evolution of microRNA regulation. We identified 100 cichlid miRNA genes with mature regions that are highly conserved in other animal species. We found that the microRNA target sites on the 3'-untranslated regions of cichlid genes to which miRNAs may bind possessed elevated SNP densities, with polymorphic sites that showed higher minor allele frequencies on average and greater genetic differentiation between Malawi lineages when compared with a neutral expectation. These results suggest that divergent selection on miRNA riboregulation may have contributed to the diversification of cichlid species.

Overall, we noticed a common denominator that seemed to be pervasive in all these studies, which is the phenomena of extensive sharing of ancestral polymorphisms. Our studies suggest that selection on ancestral polymorphism often gives rise to evolutionary diversifications within lakes, both functionally (Chapter 2), and in terms of

gene regulation (Chapter 4). It could also possibly account for the parallel evolution of similar traits between species of different lakes (Chapter 3). We thus believe that standing genetic variation in the form of ancestrally inherited polymorphisms, as opposed to variations arising from new mutations, provides much of the genetic diversity on which selection acts, allowing for the rapid and repeated adaptive radiation of East African cichlids.

5.1 Publications

The following publications, listed in order of decreasing authorship contributions, represents the body of research conducted during my PhD candidature, arising both directly and indirectly from the studies reported in this dissertation, as well as other research not mentioned herein.

1. Loh YH, Bezault E, Muenzel FM, Roberts RB, Barluenga M, Kidd MR, Sivasundar A, Howe AE, Di Palma F, Lindbald-Toh K, Seehausen O, Salzburger W, Kocher TD, Hey J, Streelman JT. Early origins of genetic variation in Lake Malawi cichlids. *In preparation*.

2. Loh YH, Katz LS, Mims MC, Kocher TD, Yi SV, Streelman JT. 2008. Comparative analysis reveals signatures of differentiation and genomic polymorphism in Lake Malawi cichlids. *Genome Biol*. 9(7):R113.

3. Loh YH, Yi SV, Streelman JT. 2011. Evolution of microRNAs and the diversification of species. *Genome Biol Evol*. 3:55-65.

4. Sylvester JB, Rich CA, Loh YH, van Staaden MJ, Fraser GJ, Streelman JT. 2010. Brain diversity evolves via differences in patterning. *Proc Natl Acad Sci U S A*. 107(21):9718-9723.

5. O'Quin KE, Smith D, Naseer Z, Schulte J, Engel SD, Loh YH, Streelman JT, Boore JL, Carleton KL. 2011. Divergence in cis-regulatory sequences surrounding the opsin gene arrays of African cichlid fishes. *BMC Evol Biol*. 11(1):120.

6. Elango N, Lee J, Peng Z, Loh YH, Yi SV. 2009. Evolutionary rate variation in Old World monkeys. *Biol Lett*. 5(3):405-408.

APPENDIX A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Due to the large sizes of some of the tables, only the first page would be shown here to illustrate the data available. The complete set of supplementary materials for Chapter 2 are available online at http://genomebiology.com/2008/9/7/R113/additional.

Table A1. Trace sequence statistics of five Lake Malawi cichlid species. (Complete)

	C. conophorus	L. fuelleborni	M. auratus	M. zebra	R. esox
Number of trace reads	157,434	153,061	138,517	161,413	152,385
Total read length (bases)	166,071,742	167,074,220	137,257,743	184,775,275	175,769,721
Shortest read length (bases)	72	88	76	109	76
Longest read length (bases)	6,759	7,264	4,862	7,072	5,834
Mean read length (bases)	1,055	1,092	991	1,145	1,153
Q25 read length (bases)	800	893	822	844	976
Q50 (median) read length (bases)	1,040	1,092	995	1,223	1,133
Q75 read length (bases)	1,313	1,237	1,126	1,417	1,383

Table A2. Human gene homologs present in the five cichlid species. "1" and "0" indicates the presence and absence of the cichlid homolog of the human gene respectively. CC, *C. conophorus*; LF, *L. fuelleborni*; MA, *M. auratus*; MZ, *M. zebra*; RE, *R. esox*. (First page)

S/No.	Human gene description	Gene homolog identified in				
		CC	LF	MA	MZ	RE
1	11-beta-hydroxysteroid dehydrogenase 1 [NP_861420.1]	1	0	1	0	0
2	15 kDa selenoprotein isoform 1 precursor [NP_004252.2]	1	0	0	1	0
3	1-acylglycerol-3-phosphate O-acyltransferase 3 [NP_001032642.1]	0	0	1	0	0
4	1-acylglycerol-3-phosphate O-acyltransferase 4 [NP_064518.1]	0	0	1	0	0
5	1-acylglycerol-3-phosphate O-acyltransferase 5 [NP_060831.2]	1	0	1	0	1
6	1D-myo-inositol-trisphosphate 3-kinase B [NP_002212.2]	1	0	0	0	0
7	2,3-bisphosphoglycerate mutase [NP_001715.1]	0	0	1	1	0
8	2,4-dienoyl CoA reductase 1 precursor [NP_001350.1]	1	0	0	0	0
9	24-dehydrocholesterol reductase precursor [NP_055577.1]	0	0	1	0	0
10	2B28 protein [NP_056937.2]	0	0	1	0	0
11	2-hydroxyphytanoyl-CoA lyase [NP_036392.2]	0	1	1	0	0
12	2-oxoglutarate and iron-dependent oxygenase domain containing 2 [NP_078899.1]	0	0	0	0	1
13	2'-phosphodiesterase [NP_808881.2]	1	0	1	0	1
14	3'(2'), 5'-bisphosphate nucleotidase 1 [NP_006076.3]	0	0	1	1	0
15	3-hydroxy-3-methylglutaryl-Coenzyme A reductase [NP_000850.1]	0	1	1	1	1
16	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 1 (soluble) [NP_002121.3]	1	0	0	1	0
17	3-hydroxybutyrate dehydrogenase precursor [NP_004042.1]	0	0	0	1	0
18	3-hydroxybutyrate dehydrogenase, type 2 [NP_064524.3]	0	0	0	1	0
19	3-hydroxyisobutyrate dehydrogenase [NP_689953.1]	0	0	1	1	0
20	3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase (hydroxymethylglutaricaciduria) [NP_000182.2]	0	0	0	1	0
21	3-mercaptopyruvate sulfurtransferase [NP_001013458.1]	0	1	0	0	1
22	3-oxo-5 alpha-steroid 4-dehydrogenase 2 [NP_000339.2]	0	0	1	0	0
23	3-oxoacyl-ACP synthase, mitochondrial [NP_060367.1]	0	1	0	0	0
24	3'-phosphoadenosine 5'-phosphosulfate synthase 1 [NP_005434.4]	1	0	0	1	1
25	3'-phosphoadenosine 5'-phosphosulfate synthase 2 isoform a [NP_004661.2]	0	0	0	0	1
26	3'-phosphoadenosine 5'-phosphosulfate synthase 2 isoform b [NP_001015880.1]	1	0	0	1	0
27	3-phosphoinositide dependent protein kinase-1 isoform 1 [NP_002604.1]	0	0	1	0	0
28	43 kD receptor-associated protein of the synapse isoform 2 [NP_116034.2]	0	0	1	0	1
29	4-aminobutyrate aminotransferase precursor [NP_065737.2]	1	0	0	1	0
30	4-hydroxyphenylpyruvate dioxygenase [NP_002141.1]	1	0	0	0	1
31	5' nucleotidase, cytosolic IB isoform 2 [NP_150278.2]	1	0	1	0	0
32	5' nucleotidase, ecto [NP_002517.1]	1	1	0	1	0
33	5,10-methylenetetrahydrofolate reductase (NADPH) [NP_005948.3]	0	0	0	1	1
34	5',3'-nucleotidase, mitochondrial precursor [NP_064586.1]	0	1	0	0	0
35	52kD Ro/SSA autoantigen [NP_003132.2]	1	1	1	1	1
36	5'-3' exoribonuclease 1 [NP_061874.2]	0	0	1	0	1
37	5'-3' exoribonuclease 2 [NP_036387.2]	0	0	0	1	1
38	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase [NP_004035.2]	1	0	0	0	0
39	5-azacytidine induced 1 isoform a [NP_055799.1]	0	1	1	0	0
40	5-azacytidine induced 1 isoform b [NP_001009811.1]	1	1	0	0	0
41	5-azacytidine induced 2 [NP_071906.1]	1	0	0	0	0
42	5-hydroxytryptamine (serotonin) receptor 1A [NP_000515.2]	1	1	0	0	1
43	5-hydroxytryptamine (serotonin) receptor 1B [NP_000854.1]	0	0	1	0	0
44	5-hydroxytryptamine (serotonin) receptor 1D [NP_000855.1]	0	0	1	0	0
45	5-hydroxytryptamine (serotonin) receptor 2A [NP_000612.1]	0	0	0	0	1
46	5-hydroxytryptamine (serotonin) receptor 2C [NP_000859.1]	1	0	0	0	0
47	5-hydroxytryptamine (serotonin) receptor 3A isoform a precursor [NP_998786.1]	1	1	1	1	1
48	5-hydroxytryptamine receptor 3 subunit C [NP_570126.2]	0	1	0	1	1
49	5-hydroxytryptamine receptor 5A [NP_076917.1]	0	1	0	0	0
50	5-hydroxytryptamine receptor 7 isoform d [NP_062873.1]	0	0	0	1	1
51	5-methyltetrahydrofolate-homocysteine methyltransferase [NP_000245.1]	1	1	1	0	0
52	5'-nucleotidase domain containing 2 [NP_075059.1]	1	0	0	1	0
53	5'-nucleotidase domain containing 3 isoform 1 [NP_001026871.1]	1	0	0	0	0
54	5'-nucleotidase domain containing 3 isoform 2 [NP_057659.1]	0	0	0	1	0
55	5'-nucleotidase, cytosolic IA [NP_115915.1]	1	0	1	1	1
56	5'-nucleotidase, cytosolic II [NP_036361.1]	1	0	0	0	1
57	5'-nucleotidase, cytosolic III isoform 1 [NP_001002010.1]	0	0	0	0	1
58	5'-nucleotidase, cytosolic III isoform 2 [NP_001002009.1]	0	0	1	0	1
59	5'-nucleotidase, cytosolic III-like [NP_443167.2]	0	0	1	0	0
60	5'-nucleotidase, cytosolic II-like 1 protein [NP_689942.2]	0	1	0	0	0

Table A3. List of alignments and polymorphic sites. (First page)

Alignment	Number of polymorphic sites				Number of aligned positions	Fraction polymorphic	High species trace number (>= 5)
	Non-genic	Synonymous	Non-synonymous	Total			
Aln100017	2	0	0	2	2290	0	N
Aln100040	1	0	1	2	2012	0	N
Aln100041	1	0	0	1	2003	0	N
Aln100074	0	0	1	1	1734	0	N
Aln100078	1	0	0	1	1718	0	N
Aln100095	1	0	0	1	1614	0	N
Aln100102	1	0	0	1	1576	0	N
Aln100148	0	0	0	0	1478	0	N
Aln100164	1	0	0	1	1453	0	N
Aln100169	1	0	0	1	1449	0	N
Aln100170	1	0	0	1	1449	0	N
Aln100173	1	0	0	1	1445	0	N
Aln100206	1	0	0	1	1400	0	N
Aln100215	0	0	0	0	1389	0	N
Aln100230	1	0	0	1	1368	0	N
Aln100241	1	0	0	1	1362	0	N
Aln100242	0	0	0	0	1359	0	N
Aln100248	1	0	0	1	1345	0	N
Aln100252	1	0	0	1	1343	0	N
Aln100261	1	0	0	1	1333	0	N
Aln100262	1	0	0	1	1330	0	N
Aln100264	1	0	0	1	1327	0	N
Aln100268	0	0	0	0	1323	0	N
Aln100279	0	0	0	0	1314	0	N
Aln100281	1	0	0	1	1313	0	N
Aln100291	0	0	1	1	1300	0	N
Aln100292	1	0	0	1	1300	0	N
Aln100300	1	0	0	1	1294	0	N
Aln100340	1	0	0	1	1249	0	N
Aln100348	1	0	0	1	1243	0	N
Aln100349	0	0	0	0	1243	0	N
Aln100363	0	1	0	1	1234	0	N
Aln100364	1	0	0	1	1234	0	N
Aln100375	0	0	1	1	1226	0	N
Aln100380	1	0	0	1	1224	0	N
Aln100415	1	0	0	1	1194	0	N
Aln100445	1	0	0	1	1179	0	N
Aln100449	0	0	0	0	1177	0	N
Aln100452	1	0	0	1	1175	0	N
Aln100476	0	0	0	0	1157	0	N
Aln100485	0	0	0	0	1151	0	N
Aln100495	1	0	0	1	1144	0	N
Aln100502	1	0	0	1	1140	0	N
Aln100508	0	0	0	0	1135	0	N
Aln100515	1	0	0	1	1132	0	N
Aln100518	1	0	0	1	1130	0	N
Aln100541	0	0	0	0	1115	0	N
Aln100548	1	0	0	1	1111	0	N
Aln100549	0	0	0	0	1111	0	N
Aln100572	1	0	0	1	1094	0	N

Table A4. List of alignments with BLAST hits to fish and humans. (First page)

Alignment	Number of polyorphic sites		Blast hit accession no.	Blast hit description	Human homolog accession no.	Human homolog description	High fraction polymorphi	High species trace number (>=5)
	Synonymous	Non-synonymous						
Aln100017	0	0	NP_997798.1	amyloid beta precursor protein (cytoplasmic tail) binding protein 2 [Danio rerio]	NA	NA	N	N
Aln100017	0	0	CAF89015.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100017	0	0	CAG06502.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100019	0	0	ABN80445.1	ribosomal protein S6 [Poecilia reticulata]	NA	NA	N	N
Aln100019	1	0	Q90YR8	40S ribosomal protein S6	NP_001001.2	ribosomal protein S6 [Homo sapiens]	N	N
Aln100020	0	0	CAM47025.1	novel protein (zgc:100952) [Danio rerio]	NA	NA	N	N
Aln100024	1	0	XP_688755.2	PREDICTED: hypothetical protein [Danio rerio]	NP_005449.5	G protein-coupled receptor 51 [Homo sapiens]	N	N
Aln100024	0	1	XP_689136.2	PREDICTED: hypothetical protein [Danio rerio]	NA	NA	N	N
Aln100024	0	0	CAG02963.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100037	1	1	AAH95723.1	Si:dkey-90m5.4 protein [Danio rerio]	NP_443204.1	leucine-rich alpha-2-glycoprotein 1 [Homo sapiens]	N	N
Aln100040	0	1	XP_691046.2	PREDICTED: similar to FERM and PDZ domain-containing protein 3 [Danio rerio]	XP_042978.5	PREDICTED: similar to PDZ domain containing 10 [Homo sapiens]	N	N
Aln100041	0	0	CAG03826.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100042	0	0	XP_001338906.1	PREDICTED: similar to transposase [Danio rerio]	NA	NA	N	N
Aln100042	1	7	XP_001343927.1	PREDICTED: similar to pol polyprotein [Danio rerio]	NP_056196.2	family with sequence similarity 19 (chemokine (C-C motif)-like), member A5 [Homo sapiens]	N	N
Aln100046	12	5	XP_698332.2	PREDICTED: similar to pol polyprotein [Danio rerio]	NA	NA	N	N
Aln100052	0	0	CAF89574.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100052	0	0	NP_001013524.1	component of oligomeric golgi complex 8 [Danio rerio]	NA	NA	N	N
Aln100061	2	0	CAF89606.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100063	0	0	CAF92562.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100069	1	0	CAG12328.1	unnamed protein product [Tetraodon nigroviridis]	NP_065937.1	ubiquitin specific protease 28 [Homo sapiens]	N	N
Aln100074	0	1	CAF96073.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100074	0	0	CAF91401.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100074	0	0	BAE79363.1	myosin heavy chain embryonic type 3 [Cyprinus carpio]	NP_000248.1	myosin, heavy polypeptide 7, cardiac muscle, beta [Homo sapiens]	N	N
Aln100076	2	6	XP_001341052.1	PREDICTED: hypothetical protein [Danio rerio]	NP_077084.1	PHD finger protein 1 isoform b [Homo sapiens]	N	N
Aln100078	0	0	CAG07418.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100078	0	0	NP_004293.1	activin A type IB receptor isoform a precursor [Homo sapiens]	NA	NA	N	N
Aln100078	0	0	XP_687633.2	PREDICTED: similar to serine/threonine kinase receptor type1 [Danio rerio]	NA	NA	N	N
Aln100078	0	0	AAC34382.1	serine/threonine kinase receptor type1 [Takifugu rubripes]	NA	NA	N	N
Aln100084	0	0	CAG02743.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100085	0	0	XP_001340427.1	PREDICTED: hypothetical protein [Danio rerio]	NA	NA	N	N
Aln100085	0	1	CAF91654.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100099	1	1	CAG11341.1	unnamed protein product [Tetraodon nigroviridis]	NP_001002838.1	WNK lysine deficient protein kinase 3 isoform 2 [Homo sapiens]	N	N
Aln100100	0	0	NP_956930.1	hypothetical protein LOC393609 [Danio rerio]	NA	NA	N	N
Aln100100	1	0	CAF90088.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100103	1	6	AAS83204.1	reverse transcriptase [Fundulus heteroclitus]	NA	NA	N	N
Aln100110	2	0	CAG11287.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100111	0	0	CAG06117.1	unnamed protein product [Tetraodon nigroviridis]	NA	NA	N	N
Aln100112	0	0	AAW80263.1	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase [Danio rerio]	NA	NA	N	N
Aln100112	1	2	CAF93918.1	unnamed protein product [Tetraodon nigroviridis]	NP_004332.2	carbamoylphosphate synthetase 2/aspartate transcarbamylase/dihydroorotase [Homo sapiens]	N	N
Aln100113	0	0	NP_004078.1	synapse-associated protein 97 [Homo sapiens]	NA	NA	N	N

Table A5. Major allele frequency for biallelic SNPs surveyed across Lake Malawi cichlid populations and species. The first ten loci represent positive controls as explained in the text. Two SNPs were predicteded and genotyped in sws2b; genotypes were in perfect linkage so only one is shown here. (First page)

snp/pop	*aim1*	*mitf*	*ednrb*	*rhodopsin*	*sws1*	*sws2a*	*lws*	*dec1*_1	*dec1*_3	*ip3r* (EXON12)
Species	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**	**NA**
All MZ	0.73	0.989	0.8	0.48	0.95	0.82	0.65	1	0.094	0.771
All LF	0.83	0.984	0.37	0.82	1	0.99	1	0.658	0.784	0.906
F_{ST} (within MZ)	**0.17**	**0.05**	**-0.004**	**0.733**	**0.572**	**0.114**	**0.514**	**NA**	**0.02**	**0.336**
F_{ST} (within LF)	**0.172**	**0.023**	**0.736**	**0.853**	**NA**	**-0.006**	**NA**	**0.599**	**0.667**	**0.066**
F_{ST} (MZ v LF)	**0.028**	**-0.004**	**0.303**	**0.2**	**0.0444**	**0.153**	**0.358**	**0.337**	**0.65**	**0.059**
All mbuna (25 sp.)	0.7	1	0.63	0.49	0.86	0.73	0.92	0.854	0.383	0.778
All others (52 sp.)	0.992	0.73	0.042	0.08	0.71	0.044	0.792	1	0.960	0.967
F_{ST} (Mbuna v nonMbuna)	**0.179**	**0.358**	**0.485**	**0.424**	**0.227**	**0.811**	**0.005**	**0.106**	**0.502**	**0.144**

snp/pop	*sws2b*	snp14 (*csrp1*)	*sema 3c* (Exon 12)	*sema 3f* (snp32)	snp10	snp13	snp19	snp21	snp22	snp24
Species	**NA**	**LF/MZ**	**LF/MZ**	**MA/MZ**	**CC/RE**	**MZ/RE**	**LF/RE**	**CC/MZ**	**LF/RE**	**CC/LF**
All MZ	0.989	0.946	0.94	0.62	1	0.978	0.76	0.55	0.91	0.82
All LF	0.96	0.04	0.9	0.26	1	0.962	0.27	0.96	0.22	0.85
F_{ST} (within MZ)	**-0.01**	**0.033**	**0.135**	**0.195**	**NA**	**0.009**	**0.115**	**0.218**	**0.23**	**-0.002**
F_{ST} (within LF)	**0.233**	**0.076**	**0.179**	**0.557**	**NA**	**0.059**	**0.474**	**0.175**	**0.286**	**0.216**
F_{ST} (MZ v LF)	**0.009**	**0.893**	**-0.004**	**0.348**	**NA**	**-0.005**	**0.366**	**0.356**	**0.643**	**-0.002**
All mbuna (25 sp.)	0.85	0.69	0.86	0.54	1	0.9	0.62	0.91	0.92	0.43
All others (52 sp.)	0.812	0.988	0.87	0.02	0.836	1	0.55	0.992	1	0.21
F_{ST} (Mbuna v nonMbuna)	**0.097**	**0.382**	**0.009**	**0.363**	**0.249**	**0.034**	**-0.004**	**0.15**	**0.277**	**0.443**

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

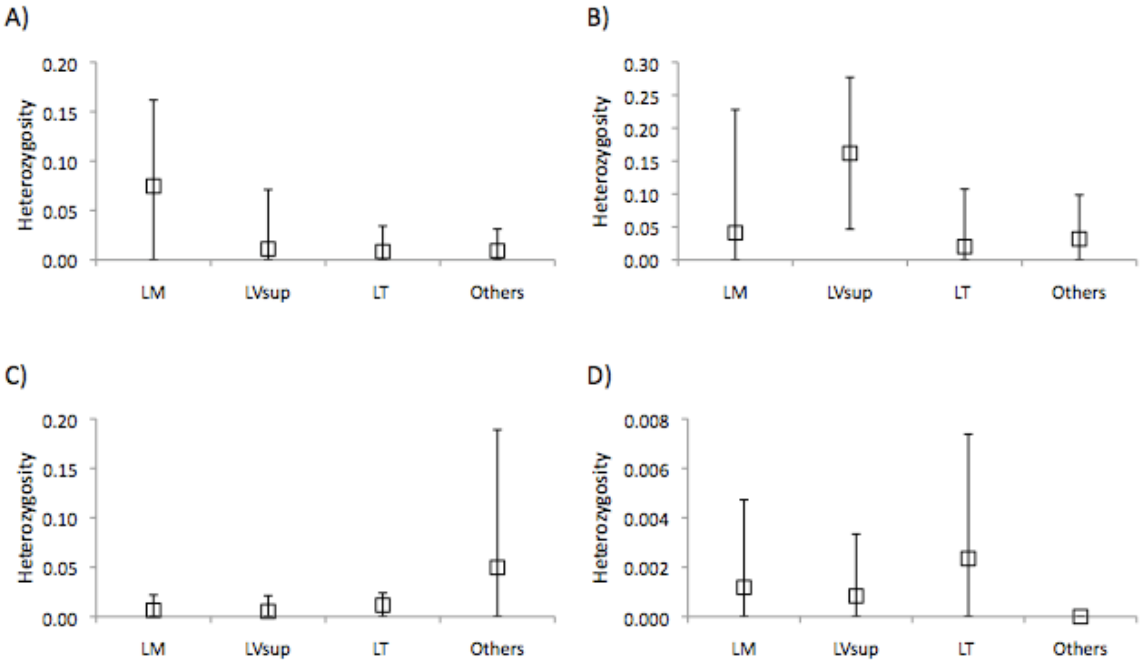


Figure B1. Observed heterozygosity of SNPs in different assemblages. SNPs were classified into A) 180 Malawi SNPs, B) 21 Victoria SNPs, C) 9 Tanganyika SNPs, and D) 9 Riverine SNPs. Boxes mark average heterozygosity, with ±1 S.D. error bars. Higher average heterozygosity generally observed for ascertained lineages (A & B). The heterozygosity values calculated for each assemblage are generally low as they contain numerous species that are not necessarily all polymorphic. LM, Lake Malawi; LVsup, Lake Victoria superflock; LT, Lake Tanganyika.

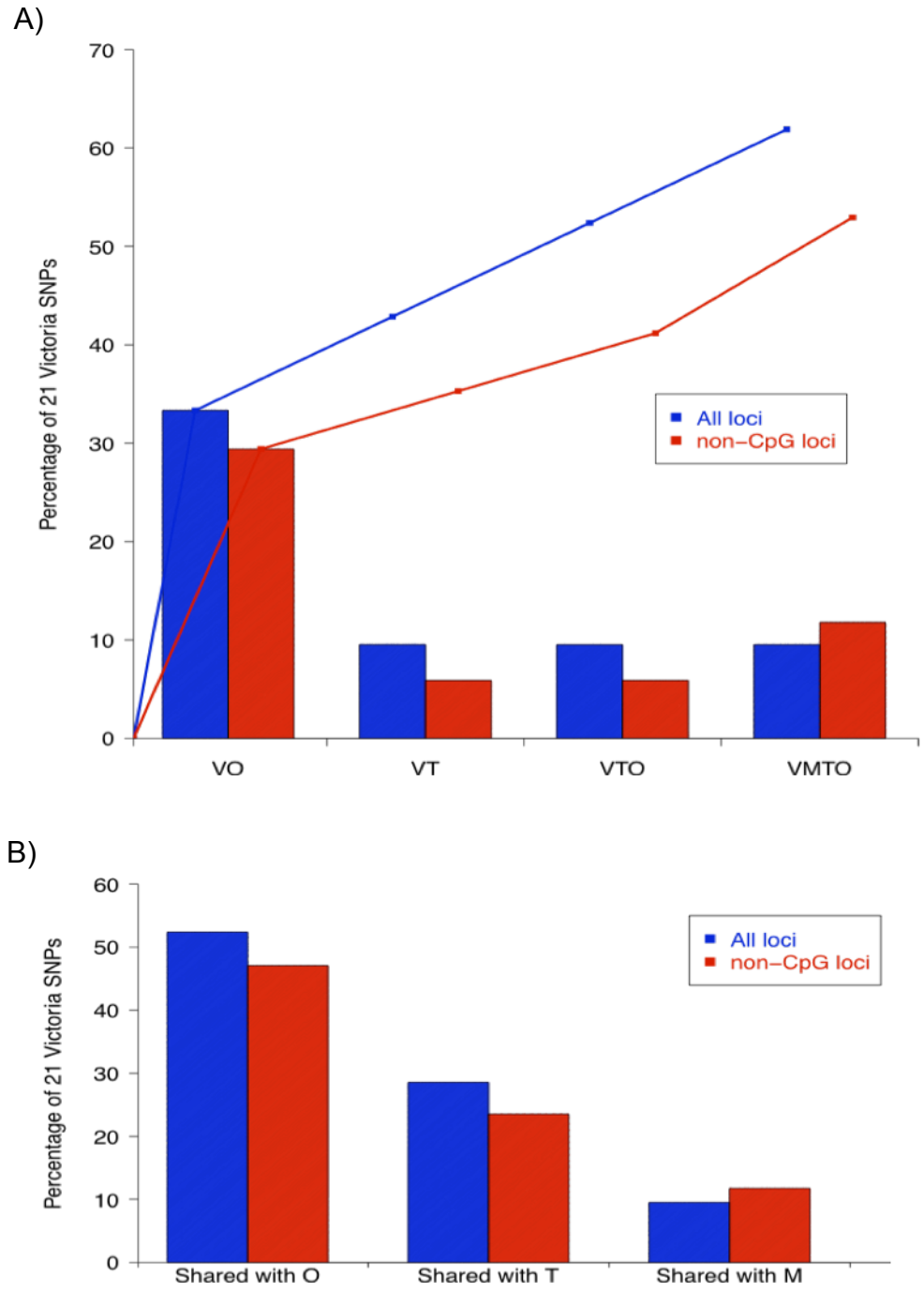


Figure B2. Percentage of shared polymorphism of 21 Victoria SNPs (17 non-CpG) with cichlids in other catchments. A) Strict polymorphism sharing with each catchment combination indicated by the category labels. B) Total polymorphism sharing with one other catchment. Bar graphs show percentage polymorphism sharing for each category while line graphs tally cumulative percentages. M, Malawi assemblage; V, Victoria superflock; T, Tanganyika assemblage; O, other rivers and drainages.

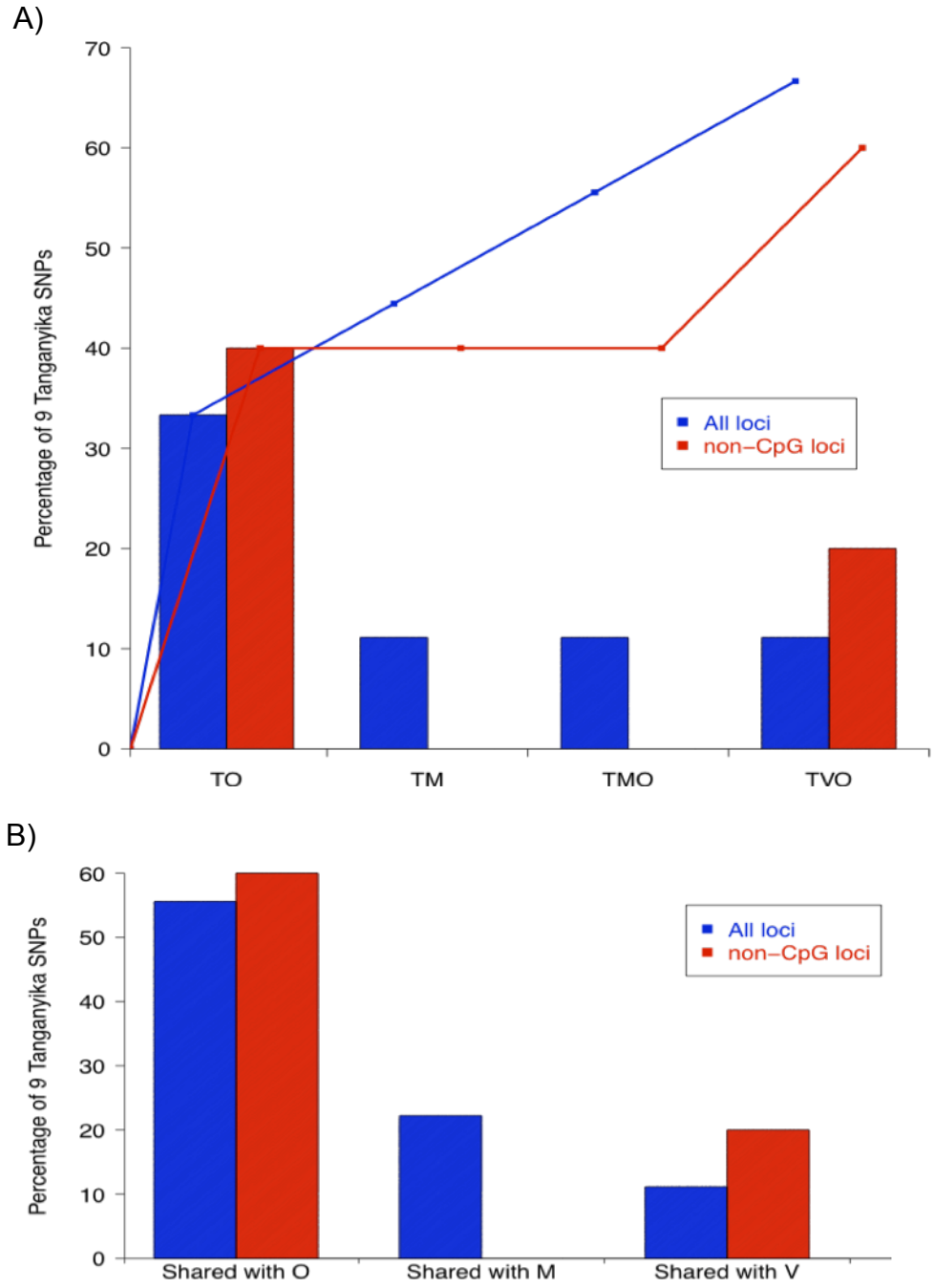


Figure B3. Percentage of shared polymorphism of 9 Tanganyika SNPs (5 non-CpG) with cichlids in other catchments. A) Strict polymorphism sharing with each catchment combination indicated by the category labels. B) Total polymorphism sharing with one other catchment. Bar graphs show percentage polymorphism sharing for each category while line graphs tally cumulative percentages. M, Malawi assemblage; V, Victoria superflock; T, Tanganyika assemblage; O, other rivers and drainages.

APPENDIX C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

Due to the large sizes of the tables and files, only the first page would be shown here to illustrate the type of data available. The complete set of supplementary materials for Chapter 4 are available online at http://gbe.oxfordjournals.org/content/3/55/suppl/DC1.

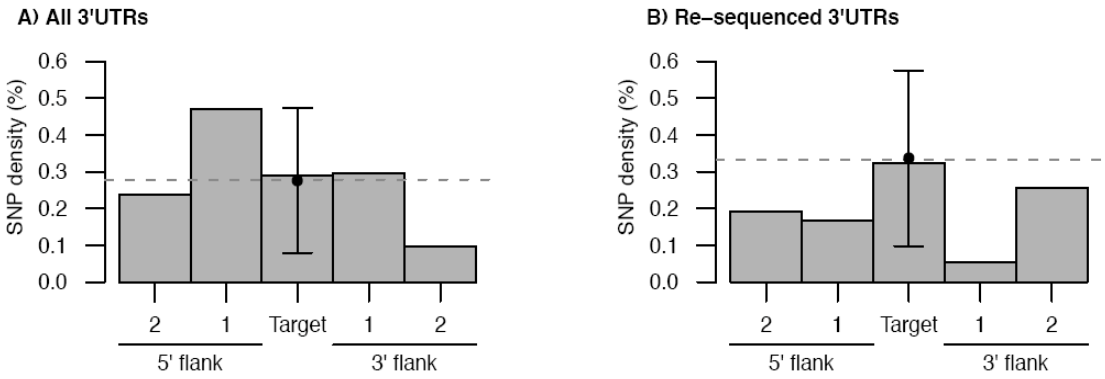


Figure C1. SNP densities within conserved miRNA target sites and their flanking regions. A) all predicted 3'-UTRs. B) select resequenced 3'-UTRs. Flanking regions 1-2 on both 5' and 3' ends of 'target' represent successive, non-overlapping windows of sizes equal to that of the target sites. Dotted lines show the average 3'-UTR SNP density. Filled circle with error bars represent the mean and 95% confidence intervals of SNP densities calculated from 1000 simulated replicates of randomized SNP shuffling.

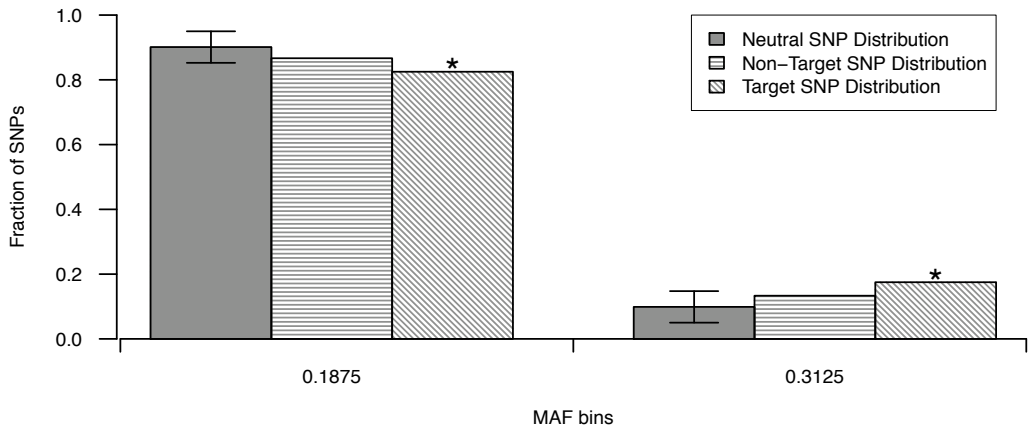


Figure C2. Comparison of minor allele frequency distributions. Minor allele frequencies grouped into 2 bins. 3'-UTR miRNA target SNPs are represented with diagonal shading, non-target SNPs with horizontal shading and non-genic (neutral) SNPs in solid grey. Error bars represent the 95% confidence interval of the neutral expectation. Asterisk symbols indicate significant deviation from neutral expectation within each bin (* P= 0.00218).

Table C1. miRNAs detected in cichlids. (First page)

S/No.	Cichlid miRNA name*	Length	Genomic Loci Accession No.	Start	End	Strand	Hairpin Sequence	Mature miRNA Sequence	Homologous to	Mfold dG (kcal/mol)
1	lfu-let-7a-1	79	ABPK01007631	1831	1909	1	UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGG UCAUACCCACACUGGGAGAUAACUAUUCAACCUA CUGUCUUUCCCA	UGAGGUAGUAGGUUGUAUAGUU	dre-let-7a-1	-34.7
2	lfu-let-7f	85	ABPK01007631	2011	2095	1	CAGUGUGAGGUAGUAGAUUGUGUAGUUUUAGGA UAGUGAUUUUGCCCCUCUUUAGGAGAUAACUAUAC AAUCUAUUGCCUUCCCUG	UGAGGUAGUAGAUUGUGUAGUU	dre-let-7f	-30
3	lfu-mir-100	63	ABPK01023571	802	864	-1	CCCAAACCCGUAGAUCCGAACUUGUGUUAAGUGA CACCACAAGCUUGUAUCUACAGGUCUGCG	AACCCGUAGAUCCGAACUUGUG	fru-mir-100	-19.6
4	lfu-mir-103-1	83	ABPK01027941	624	706	1	CUGCUCUACGCUUUUAGCCUCUUUACAGUGCUG CCUUGUCUAAUCAUGUUCAAGCAGCAUUGUACAG GGCUAUGACAGCAUAG	AGCAGCAUUGUACAGGGCUAUGA	fru-mir-103	-31.1
5	lfu-mir-135a	80	ABPK01077862	471	550	1	ACAUGCGUCUUAUGGCUUUCUAUUCCUAUGUGA GUUCUUUCUAACAUGUCAUGUAGGGUUUAAAGC CAUUGGAUACACAG	UAUGGCUUUCUAUUCCUAUGUG	fru-mir-135a	-25.3
6	lfu-mir-137	71	ABPK01025702	499	569	-1	CUCGACCACGGGUAUUCUUGGGUUGAUAAUACA GAUGUGGAUGUUAUUGCUUGAGAAUACGCGUAG UCGAG	UUAUUGCUUGAGAAUACGCGUA	tni-mir-137	-32
7	lfu-mir-138	83	ABPK01044398	1018	1088	1	UGUGUGCCGNAGCUGGUGUUGUGAAUCAGGCCG AUGACACAAGCUCUUAUAACCCGGCUAUUUCCA ACACCAGGGUGGCACC	AGCUGGUGUUGUGAAUCAGGCC	fru-mir-138	-33.1
8	lfu-mir-142a	91	ABPK01051276	493	583	-1	UGUACAGUGCAGUCAUCCAUAAAGUAGAAAGCAC UACUAAACUCCUCGCCACAGUGUAGUGUUUCCUA CUUUAUGGAUGAGUGUACUGUUG	CAUAAAGUAGAAAGCACUACU and UGUAGUGUUUCCUACUUUAUGGA	tni-mir-142a	-47.4
9	lfu-mir-15a	61	ABPK01051643	375	435	1	UGUAGCAGCACGGAAUGGUUUGUGGGUUAUACU GAGAUGCAGGCCAUACUGUGCUGCCGCA	UAGCAGCACGGAAUGGUUUGUG	fru-mir-15a	-31.1
	res-mir-15a	61	ABPN01038434	932	992	-1	UGUAGCAGCACGGAAUGGUUUGUGGGUUAUACU GAGAUGCAGGCCAUACUGUGCUGCCGCA	UAGCAGCACGGAAUGGUUUGUG	fru-mir-15a	-31.1
10	lfu-mir-16	88	ABPK01051643	648	735	1	GUCGCCUUACUGUAGCAGCACGUAAAUAUUGGA GUUAACACUCUAGCUGAAGUCUCCAGUAUUGAUC GUACUGCUGAAGCAAAGCGGG	UAGCAGCACGUAAAUAUUGGAG	fru-mir-16	-29.9
	res-mir-16	88	ABPN01038434	632	719	-1	GUCGCCUUACUGUAGCAGCACGUAAAUAUUGGA GUUAACACUCUAGCUGAAGUCUCCAGUAUUGAUC GUACUGCUGAAGCAAAGCGGG	UAGCAGCACGUAAAUAUUGGAG	fru-mir-16	-29.9
11	lfu-mir-183	89	ABPK01045721	534	622	-1	GACUCCUGUUCUGUGUAUGGCACUGGUAGAAUU CACUGUGAGAGCUCACUAUCAGUGAAUUACCAUA GGGCCAUAAACAGAGCAGAGAA	UAUGGCACUGGUAGAAUUCACUG	dre-mir-183	-39.4
	mze-mir-183	89	ABPM01011651	636	724	1	GACUCCUGUUCUGUGUAUGGCACUGGUAGAAUU CACUGUGAGAGCUCACUAUCAGUGAAUUACCAUA GGGCCAUAAACAGAGCAGAGAA	UAUGGCACUGGUAGAAUUCACUG	dre-mir-183	-39.4
12	lfu-mir-199-1	86	ABPK01006073	1042	1127	-1	CCGGCUCCGUCCACCCAGUGUUCAGACUACCUGU UCAUUGUCAUACUGGUGUACAGUAGUCUGCACAU UGGUUAGACUGGGCAUGG	CCCAGUGUUCAGACUACCUGUUC	tni-mir-199-1	-37.8
13	lfu-mir-210	88	ABPK01056774	143	230	1	UCUAAAAGCAGGUAAGCCACUGACUAACGCACAU UGUGCCAGUUUCCAGUUCCACUGUGCGUGUGAC AGCGGCUAACCUGGUUUUGGG	CUGUGCGUGUGACAGCGGCUAA	fru-mir-210	-37.1
14	lfu-mir-214	81	ABPK01006072	142	222	-1	GCAGUGUGUCUGCCUAUCUACACUUGCUGUGCA GAAUAUCCUCCAACCUGUACAGCAGGCACAGACA GGCAGAUAGACAUC	ACAGCAGGCACAGACAGGCAG	dre-mir-214	-39.4
	res-mir-214	81	ABPN01041673	234	314	1	GCAGUGUGUCUGCCUAUCUACACUUGCUGUGCA GAAUAUCCUCCAACCUGUACAGCAGGCACAGACA GGCAGAUAGACAUC	ACAGCAGGCACAGACAGGCAG	dre-mir-214	-39.4
15	lfu-mir-23a-3a	76	ABPK01027459	135	210	1	GGCCAGGGGAAUUCCUGGCAGAGUGAUUUUUUU AACUAGAGGACUGAAUCACAUUGCCAGGGAUUU CCAAUGGCU	AUCACAUUGCCAGGGAUUUCCA	tni-mir-23a-3	-38.5
16	lfu-mir-24-2	82	ABPK01046841	24	105	-1	GGGUCGGUCUCCUGUGCCUGCUGUGCUGAUAAU CAGUGUGUGACGUCGGCUGGCUCAGUUCAGCAG GAACAGGGGACUGUUC	UGGCUCAGUUCAGCAGGAACAG	fru-mir-24-2	-42.4
17	lfu-mir-25	82	ABPK01045950	493	574	1	GCUGGUGUUGAGAGGCGGAGACUUGGGCAAUUG CCGGGCAUCCCAGAGGGCAUUGCACUUGUCUCG GUCUGACAGUGCCGGC	CAUUGCACUUGUCUCGGUCUGA	fru-mir-25	-40.9
	res-mir-25	82	ABPN01027101	86	167	1	GCUGGUGUUGAGAGGCGGAGACUUGGGCAAUUG CCGGGCAUCCCAGAGGGCAUUGCACUUGUCUCG GUCUGACAGUGCCGGC	CAUUGCACUUGUCUCGGUCUGA	fru-mir-25	-40.9
18	lfu-mir-26a-1	79	ABPK01021308	683	761	1	CUGGGUCUGUUUCAAGUAAUCCACGAUACGCUU GUUACAGUGGGGAAAGCCUAUUCGGGAUGACUU GGUUCAGAAACAA	UUCAAGUAAUCCACGAUACGCU	dre-mir-26a-1	-26.9

Table C2. List of primer sequences. (First page)

S/No	Internal Identifier	Forward_primer	Reverse_Primer	3'-UTR Identifier	Description
1	P001Mir	GGTTGACCGAATGAGAAGGA	GATCTGCCAAGTGATGCTGA	Aln100017_3518_4017_1_ENSTNIP00000017408_4e-46	Amyloid protein-binding protein 2 (Amyloid beta precursor protein-binding protein 2)(APP-BP2)(Protein interacting with APP tail 1) [Source:UniProtKB/Swiss-Prot;Acc:O92624]
2	P004Mir	AACCTCTCAGCCTCAACCAG	TTCATGGAGTGCCACGTACT	Aln118712_190_689_1_ENSORLP00000004565_2e-16	RING finger protein 122 [Source:UniProtKB/Swiss-Prot;Acc:Q9H9V4]
3	P006Mir	AATCACTGGAGACGCCACAC	AAACATGACCGGGTTGTTGT	Aln117782_562_63_-1_ENSTNIP00000022738_5e-26	Transcription elongation factor SPT6 (hSPT6)(Tat-cotransactivator 2 protein)(Tat-CT2 protein) [Source:UniProtKB/Swiss-Prot;Acc:Q7KZ85]
4	P007Mir	ACAACCTGGCATGACAATGA	CACTTGTTTGCACTGCATGA	Aln112730_533_34_-1_ENSORLP00000018008_6e-22	Integrin alpha-6 Precursor (VLA-6)(CD49 antigen-like family member F)(CD49f antigen) [Contains Integrin alpha-6 heavy chain;Integrin alpha-6 light chain]
5	P008Mir	ACAGACTGGTCGGCAGAAGA	CGTTATTAACCTTTGTGCCACTT	Aln103205_816_1315_1_ENSGACP00000024930_3e-92	Prickle-like protein 1 (REST/NRSF-interacting LIM domain protein 1) [Source:UniProtKB/Swiss-Prot;Acc:Q96MT3]
6	P009Mir	ACATCTGAGATTCAGGCGCT	GTGGGTTGGTGTATGCACTG	Aln105466_576_77_-1_ENSORLP00000000542_1e-16	Add: Rho-class glutathione S-transferase
7	P010Mir	ACATGGAGCCAGCTCTGAAG	CATACCTTGGCAAATGGGAG	Aln109412_1272_1771_1_gi|157265543|ref|NP_001098071.1|_1e-100	Jxc1-B [Takifugu rubripes]
8	P011Mir	ACCAGACTGACCGACAAACC	CGTGCACGCTTATCATCAGA	Aln101940_424_923_1_ENSORLP00000004205_2e-24	Ras-related protein Rab-3B [Source:UniProtKB/Swiss-Prot;Acc:P20337]
9	P012Mir	ACCTCGCTCCACCCTCTACT	TGGCAAAGTGGTGGTCAGT	Aln109946_596_97_-1_ENSORLP00000008334_9e-35	Peroxisome proliferator activated receptor isoform b (Fragment). [Source:UniProtKB/TrEMBL;Acc:Q8UUX1]
10	P013Mir	ACCTTCTCTGACGTTCTCGC	GGCTTAGTGTTGCGCAGTCT	Aln124768_977_478_-1_ENSGACP00000026228_3e-27	Dynein light chain 4, axonemal [Source:UniProtKB/Swiss-Prot;Acc:O96015]
11	P014Mir	ACTCAGCCACATTCAGGGAC	GTTTCACAGCACAGCACGAT	Aln114663_1003_504_-1_ENSTRUP00000040309_3e-20	Nardilysin Precursor (EC 3.4.24.61)(N-arginine dibasic convertase)(NRD convertase)(NRD-C) [Source:UniProtKB/Swiss-Prot;Acc:O43847]
12	P015Mir	ACTGACCTGCTGGTCTCTCC	AGAAATGCAAATGAGCTAAATACA	Aln104526_374_1_-1_gi|82185264|sp|Q6NRP2.1|PSME4_XENLA_3e-33	RecName: Full=Proteasome activator complex subunit 4; AltName: Full=Proteasome activator PA200
13	P017Mir	AGAATGCACAAGGCTTCGAC	TCTTATCGCTTCACAGAATCAAG	Aln118553_814_1313_1_ENSGACP00000020912_1e-11	E3 ubiquitin-protein ligase NEDD4 (EC 6.3.2.-)(Neural precursor cell expressed developmentally down-regulated protein 4)(NEDD-4) [Source:UniProtKB/Swiss-Prot;Acc:P46934]
14	P018Mir	AGCACCACCAGCTAGGAAGA	TGCAAACACAAATACGCACA	Aln112306_610_111_-1_ENSTRUP00000042820_8e-29	BAH and coiled-coil domain-containing protein 1 (Bromo adjacent homology domain-containing protein 2)(BAH domain-containing protein 2) [Source:UniProtKB/Swiss-Prot;Acc:Q9P281]
15	P019Mir	AGCCTGGACCACTGAGAGAA	ATGAAGCCTGGTGACATGGT	Aln112130_330_829_1_ENSTRUP00000018069_2e-19	TRIO and F-actin-binding protein (Protein Tara)(Trio-associated repeat on actin) [Source:UniProtKB/Swiss-Prot;Acc:O9H2D6]
16	P020Mir	AGCTGAAACGCTCCAAGAAC	CTGCACGTAAACAGCCAAAC	Aln108448_920_421_-1_ENSDARP00000025183_3e-17	CUG-BP- and ETR-3-like factor 2 (CELF-2)(Bruno-like protein 3)(RNA-binding protein BRUNOL-3)(CUG triplet repeat RNA-binding protein 2)(CUG-BP2)(ELAV-type RNA-binding protein 3)(ETR-3) [Source:UniProtKB/Swiss-Prot;Acc:O6P0B1]
17	P022Mir	AGGTATGGATCAGCTGGGTG	ACTCGGCCAATCACACAATC	Aln105385_826_333_-1_ENSGACP00000009687_1e-82	Delta-type opioid receptor (DOR-1) [Source:UniProtKB/Swiss-Prot;Acc:P41143]
18	P023Mir	AGTGGCAACTGTCTCCGATT	TTGCTCTTTGGGAGTAAAGTCA	Aln106884_725_226_-1_ENSGACP00000008702_2e-17	KH domain-containing, RNA-binding, signal transduction-associated protein 1 (p21 Ras GTPase-activating protein-associated p62)(GAP-associated tyrosine phosphoprotein p62)(Src-associated in mitosis 68 kDa protein)(Sam68)(p68) [Source:UniProtKB/Swiss-Prot;Acc:O07666]

File C1. Sequences of putative cichlid 3’-UTRs. (First page)

```
################################################################################
# File Contents : 731 cichlid putative 3'-UTR sequences (with up to 500 bases    #
#                 of upstream sequences)                                         #
# Header Format   : alignmentnumber_upstreamstartpos_upstreamendpos_utrstartpos_ #
#                   utrendpos_strandorientation_proteinid description            #
#                                                                                #
# alignmentnumber   : unique identifier of cichlid alignments available from     #
#                   : http://cichlids.biology.gatech.edu                         #
# upstreamstartpos  : position of 3'UTR start with respect to cichlid alignment  #
# upstreamendpos    : position of 3'UTR end with respect to cichlid alignment    #
# utrstartpos       : position of 3'UTR start with respect to cichlid alignment  #
# utrendpos         : position of 3'UTR end with respect to cichlid alignment    #
# strandorientation : strand orientation of 3'UTR with respect to cichlid        #
#                     alignment                                                  #
# proteinid         : identifier of protein used in the prediction of the 3'UTR  #
# description       : description of protein used in the prediction of the 3'UTR #
#                                                                                #
# The provided sequences contains the putative cichlid 3’-UTRs (in upper case)   #
# as well as up to 500 bases of sequences (lower case) immediately upstream of   #
# the 3’-UTR.                                                                    #
#                                                                                #
################################################################################


>Aln100017_3018_3517_3518_4017_1_ENSTNIP00000017408_4e-46 Amyloid protein-binding
protein 2 (Amyloid beta precursor protein-binding protein 2)(APP-BP2)(Protein
interacting with APP tail 1) [Source:UniProtKB/Swiss-Prot;Acc:Q92624]
gtcctgatgttgcttcgtcacagaacatgttgtactgagatgcatcatactcaacctacttctcagttaattgccaagtttt
gctgttattgttatgtaatcatttatgtagagctgacattaatcagtccaacacctgacgttacttcgctacaatggaagag
ggtaacattatcaaaaacggggctcgagtcagagagactgaatgaaacatggtctgtgatgtttctgtcttcacaggtaaaa
agctgtttggcgagggttacagtgggctggagtacgactaccgaggcctgatcaaactctacaactcagtgggaaactacga
gaaggtgtttgaataccacaacgtactgtccaactggaaccggctgagggaccggcagtttgcagtggcggatgccctggag
gacgtcaacactacaccccagcagacccaggaagtggtacaagctttcctattggcccagagcctaggccccacccgcccct
gtctcggcTGATTGGTTGACCGAATGAGAAGGAGAAGAGAGGAGGGAAAAGAGACAAGTGGGTGGGTTTTAGCCTGGTGTGG
ACATGATTTGAGCCCAAAATGTGATGCTTCAGTTTCACACGGATTAGTAACACAGCAAATTCAAACATCACTGGCAGCTTGG
TTGACTTTGTATGCTGCCAGTACACACACACAGACATACACACGTTTTCATGGAATAAAAAAAATAAATAGAGAAGACACTT
ATCAAACCTCATCAACACACTCTCACCATCAAGATGCAGTCACTGCGTTTCTCCAACAGAAAGAGGACACATGCCCGAGACT
GGGGACTCGAAAAGAAACGTCAGACGTCCTATAAGGATGACGCACGACCTCCTCGCCTCTCATCTCACTCGCCTTTTTCTTC
CTGGACAGCAGTTTTGCTTTTGTACCGTACCAGTGTAGCTCAGATCTCCGGGAGCGAACCCGTGTTCTCAAGAGCTGCACGA
CAGCATCTGTGGAATA
>Aln100020_518_1017_1018_1514_1_ENSGACP00000017181_1e-12 Uncharacterized protein
C22orf25  [Source:UniProtKB/Swiss-Prot;Acc:Q6ICL3]
tgtctccaaatggacaggtaaaaagcagctttgaattgacagacctcaaataataatatatttcttggattcgaggggcac
ccaaactccacctacactactgtgctagacagcaagatcacagaacagtgaactcttggcacacacaaaaaatccccagtgt
naatcaatcctccacagcaatgaaacactggattattgggtacagacactctcttgtagccaaaaggtcagagggttagtcc
tttccaaagctatcaaaacatcatttagcagggcagcggcaaacagtgtctgaaaagtgacttttcataaccactgtcatac
ttattagatattaaatttacaattgttttcttctctgtattgttatccattcagaaccaatacaattatcctgatagatgca
gaagggaatgtgatcttcacagagcgcaccatgcttgactgtgacacaaccaaatggagcaccagttctttccagttcaaac
tgcaggagTGAAGACACAATGGAGGAAACCTACTCTGTGCCTTTCTGCATCATCTCTCCTTCCGTTACCTTTTCCTCACCAT
CTCCTTCAGCTGAGCCTCATGTGACCGCAAAGACTCATCAGCACTTTTTTTATTCAGCTAGTCACTAGCTGCCTCGTTAACA
TCAGTTTCATATATTTCTTACTTTTGATTTAATTTAAAGATGTATGCTAATTTTNNAAAAAAAGTAAAGATGCCAAAGGACA
TTTAGAGTTTTGTAGAGGAAAATATTTTACCTACATGTCATGAACCTTCTCAAAAATTATAGCTTGAATGTAATTTGGTAAA
```

File C2. Alignments of resequenced cichlid 3'-UTRs. (First page)

The sample alignment in this page explains how the alignments have been formatted. Actual alignments to follow in the next page