Gene expression divergence between zebra finch (*Taeniopygia guttata*) subspecies and gene regulation in hybrids

By John H. Davidson

July, 2015

Director of Thesis: Christopher Balakrishnan

Major Department: Biology

Abstract: Since the publication of Mayr's Systematics and the Origin of Species, birds have been important model organisms for studies regarding evolutionary processes such as sexual selection and speciation. In this study, I will examine patterns of expression divergence in two subspecies of the zebra finch *Taeniopygia guttata*. These two zebra finch populations have been allopatrically isolated for about one million years, thus allowing for divergence in aspects of morphology, genetics and behavior. Based on previous research, we have a detailed portrait of the timing and demographic components of divergence in these birds. However, little is known about how gene expression may contribute to genomic divergence. Among sexually reproducing organisms, the conclusion of the speciation process is thought to be evidenced by the buildup of genomic incompatibilities. The two zebra finch subspecies can mate and generate viable hybrids, allowing the use of expression data from hybrids to test for such genomic incompatibilities. Here, I will use RNA sequencing technology to quantify genome-scale patterns of expression divergence, highlighting patterns of functional divergence in the zebra finch genome.

Gene expression divergence between zebra finch (*Taeniopygia guttata*) subspecies and gene regulation in hybrids

A Thesis

Provided to the Faculty of the Department of Biology

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Biology

By John H. Davidson

July, 2015

© John H. Davidson, July, 2015

Gene expression divergence between zebra finch (*Taeniopygia guttata*) subspecies and gene regulation in hybrids

by

John H. Davidson

APPROVED BY:

DIRECTOR OF
THESIS: ______________________________

Christopher Balakrishnan, PhD

COMMITTEE MEMBER: ______________________________

Keith L. Keene, PhD

COMMITTEE MEMBER: ______________________________

Carol Goodwillie, PhD

COMMITTEE MEMBER: ______________________________

Polly Campbell, PhD

CHAIR OF THE DEPARTMENT
OF BIOLOGY: ______________________________

Jeffrey McKinnon, PhD

DEAN OF THE
GRADUATE SCHOOL: ______________________________

Paul J. Gemperline, PhD

TABLE OF CONTENTS

LIST OF TABLES v

LIST OF FIGURES vi

INTRODUCTION 1

METHODS 10

RESULTS 16

DISCUSSION 20

REFERENCES 26

LIST OF TABLES

1. Table 1. List of Mapping Strategies.. 37

2. Table 2. Gene Ontology results for Australian vs Timor.................................... 38

3. Table 3. KEGG Pathway results for Australian vs Timor 39

4. Table 4. Gene Ontology results for Australian vs Hybrid 40

5. Table 5. Gene Ontology results for Timor vs Hybrid... 41

LIST OF FIGURES

1. Figure 1. Difference in mapping rates for alternative mapping strategies.......... 42

2. Figure 2. MA plot showing the number of differentially expressed genes 43

3. Figure 3. Heatmap and PCA showing similarity based on gene expression 44

4. Figure 4. Expression differences between Zebra Finch subspecies……………. 45

5. Figure 5. Overlap differentially expressed genes between pairwise comparisons… 47

6. Figure 6. Expression of POMC in F1 versus Parental Zebra Finches ….. 49

INTRODUCTION:

The Evolution of Gene Expression

Since Darwin first published *On the Origin of Species* (1859), biologists have been fascinated by ideas surrounding the mechanisms of species formation. Many evolutionary studies have focused on how changes in the genome may contribute to new phenotypes and how natural selection may act on these genetic changes (Mitchell-Olds et al., 2007 and Caporale 2003). Although changes in gene expression have long been recognized as biologically important to the process of evolution (Britten and Davidson 1969, Britten and Davidson 1971, King and Wilson 1975), there is no consensus as to how important gene expression changes may be for speciation. Furthermore, little is known about the relative importance of natural selection and genetic drift in driving expression divergence.

As with nucleotide evolution, one school of though is that gene expression divergence is predominantly neutral (King and Jukes, 1969, Khaitovich et al., 2004, Yanai et al. 2005). Under neutral theory (Kimura, 1968), most selection primarily functions to eliminate deleterious alleles. As a result, most of the changes that are free to accumulate in a population are neutral. According to this idea, the vast majority of changes in gene expression do not affect fitness and occur in a regular clock-like manner.

However, recent studies have added evidence that gene expression divergence may not be neutral. It has been shown that gene expression level is strongly negatively correlated with gene sequence divergence (Pal et al., 2001, Bloom and Adami, 2004, Rocha and Danchin 2004). These studies have shown that highly expressed genes tend to evolve more slowly and that more highly expressed genes are on average more constrained by purifying selection (Pal et al., 2001).

It has been shown that genes that are more broadly expressed tend to be more conserved in expression and that interspecific change in gene expression is positively correlated with the level of sequence divergence (Jordan et al., 2005). Taken together, these studies provide evidence for the adaptive evolution of gene expression.

Compensatory evolution of regulatory networks has also been demonstrated to be an important component of regulatory evolution. For example, it has been shown that there can be similar overall expression output despite considerable diversification at the regulatory sequence level (Tirosh et al., 2009). This has been shown to be especially true in instances involving *cis*-regulatory factors (Hinman et al., 2003, Ihmels et al., 2004). The fact that gene expression output remains similar despite regulatory sequence level diversification is strong evidence suggesting that natural selection is acting on gene expression as it suggests that compensatory changes are favored. Co-adaptation between *cis* and *trans* regulatory networks within lineages acts to maintain optimal gene expression levels (Porter and Johnson, 2002). This has been shown in diverse clades from sea urchins (Wray et al., 2003) to mammals (Cuadrado et al., 2001). Also in mammals, comparative analyses of gene sequence divergence and gene expression divergence have been performed in order to examine whether or not gene expression evolution is independent of gene sequence evolution (Jordan et al., 2005). These studies have shown that gene expression divergence is subject to the effects of purifying selective constraints and suggests that it might also be influenced by positive Darwinian selection. Altogether, these recent studies demonstrate the potential for disparity between sequence divergence and gene expression divergence.

Speciation

The transition from a unified initial species into two distinct daughter species is a gradual process that generally occurs over long periods of time. In allopatry, the initial population is separated into two subgroups when gene flow between them is prevented (Mayr, 1942). Over time, due to random changes in allele frequencies through genetic drift, the two subpopulations will begin to diverge from one another genetically. In some cases, divergent selection may further expedite divergence between the two populations (Rice and Hostert, 1993). One byproduct of this genetic divergence is the development of genetic incompatibilities (Orr, 1995). Genetic incompatibilities are the result of combinations of alleles that decrease the fitness of hybrid offspring either by decreasing survival rate or by decreasing reproductive success. Development of such incompatibilities limits gene flow between the two populations. Over time, the severity of the consequences of these incompatibilities may increase, eventually leading to reproductive isolation between the two populations. This step is seen as the final step in the speciation process.

Models for Genomic Incompatibilities

How genetic incompatibilities develop was a long-standing conundrum. The existence of incompatibilities between species implies that the genetic basis of these incompatibilities arose in the past. By definition, when these incompatibilities arose, they should have reduced the fitness of the ancestral individuals. How could incompatibilities arise and increase in frequency when the evolution of such traits would in fact reduce the fitness of the organisms in question? This problem was neatly solved in a now classic model of the evolution of genetic incompatibilities, the Dobzhansky- Muller model (Dobzhansky, 1937 and Muller, 1942). According to the Dobzhansky-Muller model, hybrid incompatibility results as a consequence of changes at more than one locus. Different allelic substitutions may occur in each of the two geographically

isolated populations. These new mutations may stochastically increase to fixation in the populations due to genetic drift or they may increase due to influence by selection. Additionally, given enough time, successive substitutions may occur in a population after the first allele has fixed, which also increase in frequency until they become fixed. No ancestral individuals are at a distinct fitness disadvantage since these new alleles are not deleterious in the background in which they arose. When offspring are formed between the two populations, certain alleles would interact within individuals that had never previously interacted. Due to epistasis, these new genetic combinations may result in previously unseen phenotypes.

Dobzhansky-Muller incompatibilities are alleles at different loci that interact with each other in a deleterious epistatic manner. While the alleles may be completely harmless in their respective parental genetic background, the combination of these alleles interacting with one another in one individual may reduce its fitness. As genetic incompatibilities continue to increase, the fitness of hybrid offspring decrease. Eventually, Dobzhansky-Muller incompatibilities may result in post-zygotic isolation (Orr and Turelli, 2001). Given this theoretical framework, a key unanswered question is how gene expression contributes to the origin of species and maintenance of species boundaries. One way in which we may look for evidence of Dobzhansky-Muller incompatibilities is by examining gene expression patterns of hybrid offspring of closely related species (Brideau, et al., 2006, Fishman and Willis, 2001).

Evidence of Incompatibilities from Gene Expression Studies

Recent technological advancements have granted us the ability to examine and compare gene expression levels at the whole-genome scale (Seehausen et al., 2014). In addition to providing information on the ways by which species are diverging, gene expression analysis may also provide insight for the evolutionary process of speciation. Components of a regulatory

pathway (e.g., a transcription factor and its binding site), may represent exactly the type of interaction envisaged in the Dobzhansky-Muller model (Johnson and Porter, 2000). Thus, aberrant gene expression may be an important contributor to hybrid dysfunctions. Accordingly, several recent studies have linked gene expression anomalies with hybrid dysfunctions (Renaut et al., 2009, Nowick et al., 2013, Barrieto et al., 2014). Gene misexpression may occur at disproportionately great levels in regulatory pathways containing rapidly evolving or male-biased genes as these genes have been shown to evolve faster and have highly divergent patterns of expression (Ortiz-Barrientos et al., 2007). Gene loss in one population of a species may allow the accumulation of mutations in downstream targets making that population genetically incompatible with other populations of that species, resulting in reproductive isolation (Zufall and Rausher, 2004). By examining these expression levels in hybrid offspring of two closely related species and comparing them to the expression level of the parents, we may shed light on the question of what genes are likely to contribute to hybrid incompatibility and thus contribute to speciation.

Gene misexpression may be due to direct incompatibilities between *cis* and *trans* regulatory elements (Wittkopp et al., 2004, Gracer et al., 2009, McManus et al., 2010). A lack of expression divergence between parental taxa is a poor indicator of regulatory incompatibility in hybrid offspring (Haerty and Singh, 2006). Regulatory dysfunctions are only revealed when differentiated elements are forced to interact in hybrid genomes (Barrieto et al., 2014). As such, hybrids play an important role in painting the speciation picture. *Cis* versus *trans* effects can be distinguished by using interspecific hybrids (Wittkopp et al., 2004). Expression patterns for alleles in hybrids reveal *trans* regulatory differences between parental species. Like any trait, expression phenotypes can be inherited in a codominant, dominant or overdominant fashion. If

hybrids are intermediate in expression between both parents, this reflects codominance whereas if hybrid offspring resemble one parent, this reflects dominance (Tirosh et al., 2009). Overdominance is observed when expression in hybrids is outside the range of expression variation seen in the parental species.

Several studies have attempted to examine hybrid incompatibility through gene expression (Ortiz-Barrietos et al., 2007, Renaut et al., 2009). In *Drosophila*, it has been shown that regulatory factors and metabolic regulator genes are often overdominant, or misexpressed, in hybrid offspring (Ortiz-Barrietos et al., 2007). In whitefish, *Coregonus clupeaformis*, gene expression studies, it has been shown that hybrids have low fitness when natural selection has independently altered the binding affinity between transcription factors and DNA binding sites (Renaut et al., 2009). This same study found evidence for increased gene misexpression in offspring produced with a hybrid and parental backcross. Thus, comparing gene expression levels in backcrossed offspring may provide further information regarding genetic incompatibilities.

A Second Rule of Speciation: Haldane's Rule and the Evolution of Sex Chromosomes

Haldane's rule is the observation that the heterogametic sex is the first to evolve hybrid sterility (Haldane, 1922). The observation is particularly interesting in that it has been shown in many taxa, including Lepidoptera and mammals (Haldane, 1922), other insects (Orr, 1993), amphibians (Hillis and Green, 1990), and birds (Haldane, also Price & Bouvier, 2002). It is thought that hybrid sterility evolves faster in the heterogametic sex due to deleterious recessive alleles on the sex chromosome, which are fully expressed due to the lack of a second sex chromosome (Turelli and Orr, 1995). Due to this faster evolution, Haldane's rule may help

explain the development of hybrid incompatibilities and the maintenance of post-zygotic reproductive isolation.

Historical Significance of Speciation in Birds

Since the Systematics and the Origin of Species was published (Mayr, 1942), birds have served as an historically important taxa when testing for speciation. The role played by differences in song, plumage, and behavior for sexual selection in birds have been extensively studied and documented (Edwards et al., 2005). Additionally, birds display a diversity of mating systems and natural history of female preference driving trait differences that have been studied more in birds than in any other clade.

Unlike mammals, in birds, females are the heterogametic sex. As such, studying gene expression in birds will allow us a rare opportunity to look for evidence of Haldane's rule and other sex chromosome related hypotheses in a female heterogametic system. According to Haldane's rule, hybrids of the heterogametic sex are more likely to be sterile or inviable. Studies in male heterogametic systems have hypothesized an "X-effect" where the X chromosome has evolved at a faster rate due to effects of dominance. We can similarly test for a Z effect. The Z chromosome has a higher neutral mutation rate than autosomes since it passes through the male germ line twice and the male germ line undergoes a greater number of cell divisions (Edwards et al., 2005).

Sexual conflict might also increase the rate at which mutations that could contribute to hybrid incompatibility arise. Certain mutations on the sex chromosomes may arise that increase the rate at which those chromosomes are transmitted (Hurst et al., 1996). A mutation that increases male fitness by stimulating females to invest more in reproduction would provide an

advantage that might drive it towards fixation. Similarly a mutation that allows for differential transmission of one sex chromosome over the other might arise. Such mutations would result in a biased sex ratio, so mutations that counter these biases would be favored. These types of rapidly evolving sex chromosome situations may allow for an easier proliferation of mutations that contribute to hybrid incompatibility as well. These types of mutations have been well documented in many species of birds as well (Price, 2008).

Project Objectives

In order to examine patterns of expression divergence and to look for genetic contributions to hybrid incompatibility, I have examined populations that are geographically isolated and are on their way to becoming distinct species. Roughly one million years ago, zebra finches, *Taeniopygia guttata*, from mainland Australia colonized several of the surrounding Lesser Sunda islands (Mayr, 1944, Balakrishnan and Edwards, 2009). This allowed for the formation of a distinct subspecies on the Lesser Sunda Islands (Zann, 1996). Over the past million years, an initial founder event followed by genetic drift and divergent selection has led to the development of distinct Australian (*T. guttata castanotis*) and island or Timor (*T. guttata guttata*) subspecies (Balakrishnan and Edwards, 2009). While the two subspecies are genetically differentiated (Balakrishnan and Edwards 2009), they are still able to interbreed and produce viable hybrid offspring.

Most previous hybrid gene expression studies have looked at taxa with high levels of reported misexpression throughout the genome (Roraz et al., 2004). However, much of this difference in expression may be influenced by genetic divergence after reproductive isolation. In order to understand the role of disruption of gene expression in speciation, I examined a case where hybrid breakdown might be less severe (Barrieto et al., 2014) as in the case with zebra

finches. In order to understand evolutionary processes such as speciation, it is important to study organisms at different points in the speciation process (Seehausen et al., 2014). By examining organisms at diverse time points on this continuum, we are able to answer questions such as: What types of genes are more prone to early gene expression divergence? Is there a particular location such as the sex chromosomes where gene expression diverges earlier? Does genomic divergence tend to follow a common trajectory during the speciation process? Are these findings consistent for diverse types of taxa? The zebra finch can be used as a model organism in examining how gene expression patterns diverge between two geographically isolated subspecies.

METHODS:

RNA Preparation & Sequencing

Gene expression data in the form of RNA-Seq were compared between *T. guttata castanotis* (Australian), *T. guttata guttata* (Timor), and hybrid zebra finch populations. Birds were housed in captivity at the Institute for Genomic Biology at the University of Illinois at Urbana-Champaign. Three male birds were sampled from each of the three populations (Australian, Timor and hybrid). All of the hybrid birds studied were the result of crosses between female Australian zebra finches and Timor males. This crossing directionality was chosen because female Australian zebra finches breed more readily in captivity than do female Timor finches.

In order to control for environmental on gene expression, each individual bird was isolated in an acoustic isolation chamber the night before they were to be sacrificed. Birds were then euthanized by decapitation the next morning to control for circadian effects on gene expression. Tissues were isolated and then snap-frozen on dry ice.

For this study, whole brain tissue was homogenized in Tri-Reagent (Molecular Research Company) for RNA purification and total RNA was extracted following manufacturer's instructions. Total RNA was then DNase treated (Qiagen, Valencia CA) to remove any genomic DNA contamination and the resulting RNA was further purified using Qiagen RNeasy columns. Purified total RNA was assessed for quality using an Agilent Bioanalyzer. Library preparation and sequencing were done at the University of Illinois Roy J. Carver Biotechnology Center. Library preparation used Illumina TruSeq RNA Sample Prep Kit and manufacturer's protocols (Illumina, San Diego, CA). Sequencing was performed in a single lane of an Illumina HiSeq

2000 using a TruSeq SBS sequencing kit version 3 producing single end 100 base pair reads which were analyzed with Casava 1.8.2.

Reads were trimmed using the program ConDeTri (Smeds and Kunster, 2011) with the minimum length of the reads set to 25. Across all samples, between 72.47% and 73.40% of the reads were kept with at least 23 million reads for each sample. TopHat 2 (Kim et al., 2013) was used to map the reads to the zebra finch reference genome (taeGut3.2.4, Warren et al., 2010). For mapping, the read gap length and the read edit distance were both set to 3. Otherwise, default settings in Tophat2 were used. Samtools (Li et al., 2009) were used to convert the output files into SAM files, which were then converted to BAM files and sorted also using Samtools. Ht-seq (Anders et al., 2014) was used to count the total number of reads per gene.

Comparative Analysis of Read Mapping Software and Parameters

Preliminary analyses revealed that *T. guttata guttata* reads mapped to the reference genome at lower rate than *T. guttata castanotis*, likely due to sequence divergence. To eliminate this bias in comparisons of Timor and Australian zebra finches, I tested whether different read mapping algorithms and stringency settings could correct for mapping bias (Table 1). Three different mappers – Bowtie 2, BWA and STAMPY – were used (Langmead and Salzberg, 2012, Li and Durbin, 2009, Lunter and Goodson, 2011). The previously trimmed genomic DNA reads from both the Australian population and the Timor population were mapped to an Australian reference genome. Reads were mapped using all three mappers at default settings. An additional mapping for each program was performed using reads from each bird sample where the default settings were altered in order to allow less stringent mapping. For Bowtie 2, the alternative mapping strategy involved altering the "score-min" parameter from -0.6 to -0.5. For BWA, the "mismatch penalty" parameter was altered from 4 to 5. For STAMPY, the substitution rate

parameter was increased to 0.05. Differences between the rate at which the Australian population and the Timor population reads mapped to the reference Australian genome at each of the default and altered mapping settings were then compared.

Creation of a Timor Zebra Finch Reference Genome

As an additional strategy, I generated a Timor zebra finch genome using a reference-guided approach. Genomic DNA was extracted from muscle tissue samples of the same three Australian zebra finch and the same three Timor zebra finch individuals used in the RNA-Seq experiment. DNA was extracted using Qiagen DNA extracting kits. Purified total DNA was assessed for quality using an Agilent Bioanalyzer. Library preparation and sequencing were done at the University of Illinois Roy J. Carver Biotechnology Center. Library preparation was done using Illumina TruSeq DNA Sample Prep Kit and manufacturer's protocols (Illumina, San Diego, CA). Sequencing was done in a single lane of an Illumina HiSeq 2000 using a TruSeq SBS sequencing kit version 3 producing paired end 100 base pair reads, which were analyzed with Casava 1.8.2.

Genomic DNA reads from all samples were trimmed using ConDeTri with minimum read length set to 25. The mapping program STAMPY was used to map the Timor genomic DNA reads to the reference zebra finch genome. Samtools were used to convert the output files into SAM files, which were then converted to BAM files and sorted also using Samtools. Samtools-merge was used to merge the Timor DNA BAM files, the Timor RNA BAM files for the same three individuals and DNA BAM files generated from seven other Timor zebra finch. Samtools mpileup was used to generate pileup in a fastq file. The fastq file was converted to a fasta file using the conversion program seqtk (Li, 2015). Bowtie2-build was used to create a reference index.

To test whether mapping bias persisted when using this newly created genome, Genomic DNA reads were mapped to a zebra finch reference genome using Bowtie2 for both the Australian and Timor populations. Australian DNA reads were mapped to an Australian reference genome while the Timor DNA reads were mapped to the newly created Timor reference genome. Samtools were used to convert the output files into SAM files, which were then converted to BAM files and sorted also using Samtools. Ht-seq was used to count the total number of reads per gene. DESeq 2 was used to calculate variance in read frequency between head to head comparisons of sample groups. Genes that were significantly different ($P < 0.01$) were then uploaded into the biomart tool at the Ensembl website.

Differential Expression Testing

After optimizing mapping settings, RNA-Seq reads were re-mapped using STAMPY with the parameter substitutionrate = 5%. DESeq 2 (Anders and Huber, 2010), was then used to test for differential expression between populations. DESeq 2 models variance in gene expression as a negative binomial distribution. This approach has demonstrated the ability to fit well with observed patterns of variation in RNA-Seq data (Love et al., 2014). Genes were considered to be differentially expressed at FDR (False Discovery Rate) adjusted P-value < 0.01.

Functional Annotation of Differentially Expressed Genes

Gene Ontology (GO) was used in functional annotation. All genes were uploaded into the biomart tool at the Ensembl website in order to retrieve standardized gene IDs and location of all of the genes in the dataset. The lists of significantly different genes ($P < 0.01$) output from the differential expression analysis were used as test genes, which were uploaded to the zebra finch gene ontology program available at http://www.ark-genomics.org/tools/GOfinch and compared to a complete reference of 18,276 genes. This tested for enrichment of GO categories within the

list of differentially expressed genes relative to what you would expect based on the whole genome. The adjusted Fisher *P*-value was used to determine significance ($P < 0.05$) for the difference in expected versus observed values for comparisons between these large data sets.

Gene Expression Divergence Between Subspecies

The RNA-Seq data from the preliminary experiment was used once again. STAMPY with a substitution rate of 0.05 was used to map the reads to a reference zebra finch genome. Samtools were used to convert the output files into SAM files, which were then converted to BAM files and sorted also using Samtools. Ht-seq was used to count the total number of reads per gene.

DESeq 2 was used to calculate variance in read frequency between head to head comparisons of sample groups. Genes that had significantly different ($P < 0.01$) expression values were uploaded to the same zebra finch gene ontology program. The adjusted Fisher *P*-value was used to determine significance for the difference in expected versus observed values for comparisons between these large data sets. Significantly different genes were also uploaded to the zebra finch KEGG pathway program available at http://www.ark-genomics.org/tools/KEGG in order to identify divergent KEGG pathways.

RNA-Seq Hybrid Comparison

The RNA-Seq data for the hybrid zebra finches were used in comparisons of hybrid finches with both of the parental populations. Optimized settings were used to map the read of the hybrid zebra finch to the same Australian reference zebra finch genome. Samtools were used to convert the output files into SAM files, which were then converted to BAM files and sorted also using Samtools. Ht-seq was used to count the total number of reads per gene.

Head to head comparisons between Australian (n=3), Timor (n=3), and hybrid (n=3) zebra finch were made for a total of 3 comparisons. Genes that were significantly different in expression ($P < 0.01$) were then uploaded to the zebra finch gene ontology program. Significance was determined using Fisher *P*-values.

RESULTS:

Read Mapping to the Timor Zebra Finch Reference Genome

Despite building a Timor zebra finch reference genome, I still observed a substantial bias in mapping rate. Over 54 million (75.7%) of the generated Australian genomic DNA reads mapped to the Australian reference genome using Bowtie 2. More than 51 million (70.5%) of the Timor DNA reads mapped to the newly assembled Timor reference genome. Of the genes that mapped, 42.4% of the Australian reads mapped to Ensembl genes and 34.8% of the Timor reads mapped to Ensembl genes. Using DE-Seq 2, 2,407 genes were found to be differentially expressed between populations, but many of these may simply reflect poor mapping to the Timor reference genome. Due to the large disparity in mapping rates, I did not use this genome in further analyses.

Consequences of Alternate Read Mapping Strategies

In order to determine the effectiveness of the different mapper programs, the differences in mapping rates between the subspecies for the three mapper programs at default and modified settings were compared. These differences were plotted in Figure 1. Mappers were considered to have more effectively corrected for mapping bias if the rate at which Australian reads mapped to the genome and the rate at which the Timor reads mapped were more similar (the difference between mapping rates was decreased). Mappers with less stringent parameters performed better than mappers at default settings. Differences in mapping rates ranges from 0.73% (Bowtie2 default) to 0.01% (STAMPY with modified parameters). Overall, STAMPY performed the best, with Timor zebra finch reads mapping at 0.01% higher proportion than Australian zebra finch reads. Since STAMPY effectively eliminated read mapping bias, I used STAMPY for downstream gene expression analyses.

Gene Expression Divergence Between Subspecies

In the analysis between populations using STAMPY, 632 genes were found to have differed in gene expression between the two subspecies (Figure 2). Of the 632 genes, 23 were located on the Z chromosome (3.6%) despite the fact that the Z chromosome comprises 4.3% of the zebra finch genome. As such, you would expect 27 of the differentially expressed genes to have been located on the Z chromosome. A chi squared test showed that there was no significant difference ($\chi^2 = 0.67$, *P*-value = 0.41) between the number of differentially expressed genes on the Z chromosome, and that which would be expected based on the size of the chromosome.

Euclidean distance-based clustering of expression profiles generated in this analysis did not group samples by population of origin (Figure 2). Similarly, principal components analysis revealed that PC2, but not PC1, grouped samples by population of origin. PC1 explained 47% of the variance in expression profile. Using Gene Ontology and PCA loadings, categories relating to PC1 were based on differences in protein domain specific binding (GO:0019904), calcium ion binding (GO:0005509), protein binding (GO:0005515), nucleotide binding (GO:0000166) and transition metal ion binding (GO:0046914) (Figure 2). Gene ontology categories relating to glycolysis (GO:0006096), beta-amyloid binding (GO:0001540), L-lactate dehydrogenase activity (GO:0004459), and regulation of long-term neuronal synaptic plasticity (GO:0048169) contributed to PC2.

Gene Ontology analysis of the differentially expressed genes showed three functional categories to differ significantly in comparisons between the Australian and Timor populations (adj. $P < 0.05$). GO categories relating to oxidation-reduction processes and oxidoreductase activity were divergent between the two populations with more genes differentially expressed than expected (Table 2). On the other hand, genes involved in protein binding processes were

shown to be more conserved in expression than expected. Thirty-five genes contributed to the oxidation-reduction processes and oxidoreductase activity and were found to be significantly different in expression between the two populations (Figure 3). Of these 35 genes, 8 genes were under-expressed in the Timor population when compared to the Australian population and 27 genes were relatively over-expressed in the Timor population indicating these genes were more likely (χ^2 = 10.31, *P*-value=0.00132) to be overexpressed in the Timor population than the Australian population. KEGG pathway analysis showed pathways relating to homologous recombination (5 genes) and Oxidative phosphorylation (8 genes) involved genes that were differentially expressed (Table 3).

Gene Expression in Hybrids Versus Parental Species

Two hundred and five genes were found to be differentially expressed between Australian and hybrid zebra finches. Of these, 142 were also differentially expressed in the comparison between the Australian and Timor zebra finches indicating broad overlap in these analyses (Figure 5). One GO category was significantly enriched among these differentially expressed genes (Table 4). This category related to cytochrome-c (mitochondrial protein involved in electron transport chain and oxidation-reduction) activity. No KEGG pathways were over or under-represented among the differentially expressed genes. The comparison between the hybrid and the Timor zebra finches showed 55 genes that were found to be differentially expressed. Of these, three genes were also differentially expressed in the Australian vs Timor comparison. No GO categories or KEGG pathways were over- or under-represented.

In order to test for overdominant gene expression, I compared all six parental zebra finches with the three hybrid finches. In this analysis, only one gene (POMC) was differentially expressed

(Figure 6). POMC was expressed significantly more highly in hybrids than in either parental population with on average 93.9 reads in hybrids, and 32.3 reads in parentals (*P*-value= 9.6E-8).

DISCUSSION:

Gene Expression Divergence

In my study I have broadly characterized patterns of gene expression divergence between two allopatric subspecies. It has previously been shown that the two zebra finch subspecies have diverged genetically at the sequence level (Balakrishnan and Edwards, 2009). Despite their relatively recent divergence and similar ecology, we found evidence for expression divergence of many genes between subpecies. Comparison between the Australian and Timor expression levels showed 632 genes that were differentially expressed. This provides evidence that, in addition to neutral sequence-level divergence (Balakrishnan & Edwards 2009), gene expression levels between the two populations have diverged after 1 million years in allopatry.

We also found differences in expression between parental species and hybrid birds. Two hundred and five genes were found to be differentially expressed between the Australian subspecies and hybrids, and 55 genes were found to be differentially expressed between the Timor subspecies and hybrids. As such, there were many fewer genes differentially expressed in the hybrid comparisons than between subspecies, suggesting that hybrid expression was largely intermediate, and not overdominant. Only the gene POMC was found to be overdominant in the hybrid population.

Lack of Evidence for Reproductive Gene Expression Divergence

Genes found on the sex chromosome or reproductive genes that are involved in male-biased pathways or sexual conflict have been shown to evolve relatively rapidly (Hurst et al., 1996, Price 2008). These types of genes have also been shown to have highly divergent patterns of gene expression (Ortiz-Barrientos et al., 2007). More rapid divergence in gene expression in

male-biased genes parallels observations that male morphological features evolve more rapidly than in females. In *Drosophila*, substantial differences sex-dependent regulation have been demonstrated (Ranz et al., 2003). Contrary to our initial predictions, neither genes on the sex chromosomes nor reproduction-related genes (as defined by GO categories) were found to be enriched among the set of genes with divergent expression profiles between subspecies. Given the strength of the pattern in other taxa, we speculate that the lack of evidence for accelerated expression divergence in sex linked genes may have to do with my choice of focal tissue. It is possible that the sex linked and reproductive genes that evolve quickly, tend to be those that expressed in reproductive tissues.

Metabolic Gene Expression has Diverged

I found that genes involved in metabolism processes are enriched among the differentially expressed genes. Protein-binding genes were shown to be relatively conserved in their expression. The gene categories that most significantly differed between the two zebra finch subspecies dealt with the metabolism processes of oxidation-reduction and oxidoreductase activity. Metabolic processes diverging in expression between closely related species has previously been seen (Ortiz-Barrietos et al., 2007). It is not known why these are often some of the first genes to diverge in expression level between separated populations. This trend may be impacted by differences in ecological divergence in populations that have been separated in allopatry (Arnegard et al., 2014). Metabolic adaptation may be expected in different habitats, such as those experienced by birds in arid island Australia versus the Lesser Sunda islands.

Possible Implications and Future Directions

Our findings are limited by the fact that the tissue samples used in this experiment were collected from the brains of the birds since gene expression levels for certain genes may vary in

different tissue types. In order to test for differences in expression of genes relating to the sex chromosome, reproductive regulatory genes, or male-biased genes it may be necessary to perform a similar experiment examining differences in gene expression of gonadal tissue samples. Furthermore, differences in gene expression between the two subspecies may occur at higher rates in gonadal tissue than in biologically conserved areas such as brain tissue. Future comparisons similar to this one using gene expression data generated by gonadal tissue may provide evidence for this process.

While it is now known that the zebra finch subspecies have diverged in gene expression in allopatry an open question regarding this progression remains: what is the evolutionary cause of this difference in gene expression? The genes that were shown to have diverged in expression levels related to metabolic processes. However, the population bottleneck of the Timor subspecies (Balakrishnan and Edwards, 2009) may have contributed to divergence in gene expression between the two populations due to lower effective population size (Phifer-Rixey et al., 2014) indicating genetic drift may have played a role in expression divergence. All birds used in this experiment were domesticated birds. However, the Australian lineage has been domesticated for much longer than the Timor lineage. As such, differences in domestication histories rather than differences in environmental factors and divergence due to genetic drift may have impacted the divergence in gene expression patterns between the two subspecies that were observed. Differences in expression of metabolism-related genes may reflect adaptation to captivity in domesticated birds, rather than ecological adaptation.

My study examined only F1 hybrid birds. However, in some cases hybrid compatibilities are not observed until F2 or backcross generations (Barrieto et al., 2014, Phifer-Rixey et al., 2014). For example in *Drosophila* males have been used to track infertility in order to find what

chromosomes and genes contribute to hybrid sterility (Wu &Ting, 2004). These studies showed that gene expression patterns contributing to hybrid incompatibility may not be expressed unless recombinant chromosomes are generated, altering regulatory interactions in *cis*. As such, some traits that contribute to hybrid incompatibility may not arise until the F2 generation or backcrossed hybrids. Future studies should incorporate F2 generations, which may result in novel gene expression interactions and potential decreases in fitness not seen in the F1 generation hybrids.

In addition to looking at F2 generation hybrids, future studies may find increased evidence for misexpression in hybrid zebra finch by performing the opposite experimental cross (male Australian with female Timor) than the one performed for this experiment (male Timor with female Australian). In Drosophila, genes having the greatest effect on hybrid sterility and inviability are found on the X chromosome (Coyne & Orr, 1988, Charlesworth et al., 1987). In birds, females are heterogametic with ZW sex chromosomes while males receive two copies of the Z chromosome. Future studies using F1 females from both hybrid cross directions would potentially allow for expression of sex linked genes that contribute to speciation, which are not present in the current set of crosses.

Accommodating Sequence Divergence in RNA-Seq

Our study also suggests several strategies that may be used to deal with mapping bias, which may be observed when attempting to map divergent genes to a reference genome. In order to quantify gene expression, RNA reads were mapped to a reference genome. It has been observed that it is sometimes difficult for bioinformatics programs to distinguish between reads that are derived from alternative alleles for the same gene (Stevenson et al., 2013). For instance, in a study where reads of a human heterozygous genotype were mapped to a reference genome,

reads that were an exact match to the reference genome mapped at higher rates than those that were not despite the fact that reads were simulated at equal numbers for each allele (Degner et al., 2009). This difference in ability to map reads to a reference genome results in mapping bias. As such, reads that are exact matches to reference alleles map at higher rates than alternative alleles.

I observed that using standard pipelines, the Australian zebra finch reads mapped at a higher rate to the zebra finch reference genome than the Timor zebra finch reads did. I hypothesized that this was due to sequence divergence between Australian and Timor zebra finches. Given a million years of divergence, even neutral divergence in DNA sequence would result in substantial mapping bias. In order to eliminate mapping bias, two methods were used: creation and use of a Timor zebra finch reference genome for the Timor mapping and the use of alternative mapping strategies using less stringent mapping parameters.

When a comparison was made between the Timor read mapping rate to the Timor reference genome and the Australian mapping rate to the Australian reference genome, it was found that the Australian reads still mapped at a higher rate (75.7% compared to 70.5%). This persistence in disparity between mapping rates shows us that the assembled Timor zebra finch reference genome was not able to effectively eliminate the mapping bias. This may have been due to low quality of the assembled reference genome.

The alternate mapper settings showed a decrease in the disparity between rates at which Australian reads and Timor reads mapped to the same Australian zebra finch reference genome. The program that most successfully eliminated this bias was the mapper STAMPY with a substitution rate of 0.05. The substitutionrate parameter allows for easier comparisons to be made between divergent populations by factoring in changes in sequence due to substitutions

over time. When this setting was set to 0.05, there was only a 0.01% difference between the mapping rates of the two populations. Interestingly, under this setting the Timor reads were the data that mapped at a higher rate to the Australian zebra finch reference genome. Since this setting most effectively corrected for the previously observed mapping bias, this mapper and mapper setting was used in the comparisons between the three zebra finch populations. I suggest that the STAMPY aligner, designed exactly for this purpose, will be useful in future studies in taxa lacking a reference genome.

REFERENCES

Anders, S., Huber, W. “Differential expression analysis for sequence count data.” *Genome Biology* 11.10 (2010): R106. doi:10.1186/gb-2010-11-10-r106.

Anders, S., Pyl, PT., Huber, W. “HTSeq – A Python framework to work with high-throughput sequencing data.” *Bioinformatics* 31.2 (2014): 166-69.

Arnegard, ME., McGee, MD., Matthews, B., Marchinko, KB., Conte, GL., Kabir, S., Bedford, N., Bergek, S., Chan, YF., Jones, FC., Kingsley, DM., Peichel, CL., Schluter, D. “Genetics of ecological divergence during speciation.” *Nature* 511 (2014): 307-11.

Balakrishnan, CN., Edwards, SV. “Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*).” *Genetics* 181.2 (2009): 645-60.

Balakrishnan, CN., Lin, YC., London, SE., Clayton, DF. “RNA-Seq transcriptome analysis of male and female zebra finch cell lines.” *Genomics* 100.6 (2012): 363-69.

Barreto, FS., Pereira, RJ., Burton, RS. “Hybrid dysfunction and physiological compensation in gene expression.” *Molecular Biology and Evolution* 32.3 (2014): 613-22.

Bloom, JD., Adami, C. “Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response.” *BMC Evolutionary Biology* 4:14 (2004): *PMC*.

Brideau, NJ., Flores, HA., Wang, J., Maheshawari, S., Wang, X., Barbash, A. “Two Dobzhansky-Muller genes interact to cause hybrid lethality in *Drosophila*.” *Science* 314.5803 (2006): 1291-95.

Britten, RJ., Davidson, EH. “Gene regulation for higher cells: a theory.” *Science* 165 (1969): 349-57.

Britten, RJ., Davidson, EH. “Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty.” *The Quarterly Review of Biology* 46.2 (1971): 111-38.

Belle, GDM., Kane, NC., Rieseberg, LH., Adams, KL. “RNA-Seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations.” *Genome Biology and Evolution* 5.7 (2013): 1309-23.

Cabot, EL, “Genetics of reproductive isolation: complex epistasis underlying hybrid sterility in the *Drosophila simulans* clade.” *Genetics* 137 (1994): 175-89.

Caporale, LH. “Natural selection and the emergence of a mutation phenotype: an update of the evolutionary synthesis considering mechanisms that affect genome variation.” *Annual Review of Microbiology*. 57 (2003): 467-85.

Charlesworth, B., Coyne, JA., Barton, NH. “The relative rates of evolution of sex chromosomes and autosomes.” *The American Naturalist* 130 (1987): 113-46.

Coyne, JA., Charlesworth, B. “Location of an x-linked factor causing male sterility in hybrids of *Drosophila simulans* and *D. mauritiana.” Heredity* 57 (1986): 243-46.

Coyne, JA., Charlesworth, B. “Genetic analysis of X-linked sterility in hybrids between three sibling species of *Drosophila*.” *Heredity* 62 (1989): 97-106.

Coyne, JA., Orr, HA. “Two rules of speciation.” *Speciation and its Consequences* Sunderland, Mass. Sinauer Associates, 1989. pp. 180-207.

Cuadrado, M., Sacristan, M., Antequera, F. “Species-specific organization of cpg island promoters at mammalian homologous genes.” *EMBO Reports* 2 (2001): 586–92.

Darwin C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London. Murray, 1859. Print.

Degner, JF., Marioni, JC., Pai, AA., Pickrell, JK., Nkadori, E., Gilad, Y., Pritchard, JK. “Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.” *Bioinformatics* 25.24 (2009): 3207-12.

deQueiroz, K. “Ernst Mayr and the modern concept of species.” *Proceedings of the National Academy of Sciences* 102 (2005): 6600-07.

Dobzhansky, T. “Studies on hybrid sterility II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids.” *Genetics* 21 (1936): 113-35.

Dobzhansky, T. *Genetics and the Origin of Species*. New York, Columbia University. 1937.

Edwards, SV., Kingan, SB., Calkins, JD., Balakrishnan, CN., Jennings, WB., Swanson, WJ., Sorenson, MD. “Speciation in birds: genes, geography and sexual selection.” *Proceedings of the National Academy of Sciences* 102 (2005): 6550-57.

Ekblom, R., Galindo, J. “Applications of next generation sequencing in molecular ecology of non-model organisms.” *Heredity* 107 (2011): 1–15.

Fishman, L., Willis, JH. “Evidence for Dobzhansky-Muller incompatibilities contributing to the sterility of hybrids between *Mimulus guttatus* and *M. nasutus*.” *Evolution* 55.10 (2001): 1932-42.

Gagnaire, PA., Normandeau, E., Bernatchez, L. “Comparative genomics reveals adaptive protein evolution and a possible cytonuclear incompatibility between European and American eels.” *Molecular Biology and Evolution* 29.10 (2012): 2909-19.

Gilad, Y., Pritchard, JK., Thornton, K. “Characterizing natural variation using next-generation sequencing technologies.” *Trends in Genetics* 25 (2009): 463–71.

Gould, SJ. “Is a new and general theory of evolution emerging?” *Paleobiology* 6 (1980): 119-30.

Graze, RM., McIntyre, LM., Main, BJ., Wayne, ML., Nuzhdin, SV. “Regulatory divergence in *Drosophila melanogaster* and *D. simulans*, a genomewide analysis of allele-specific expression.” *Genetics* 183 (2009): 547-61.

Haerty, W., Singh, RS. “Gene regulation divergence is a major contributor to the evolution of Dobzhansky–Muller incompatibilities between species of *Drosophila*.” *Molecular Biology Evolution* 23 (2006): 1707-14.

Haldane, JBS. “Sex-ratio and unisexual sterility in hybrid animals.” *Journal of Genetics* 12 (1922): 101— 109.

Hillis, DM., Green, DM. “Evolutionary changes of heterogametic sex in the phylogenetic history of amphibians.” *Journal of Evolutionary Biology* 3.1-2. (1990): 49-64.

Hinman, VF., Nguyen, AT., Cameron, RA., Davidson, EH. “Developmental gene regulatory network architecture across 500 million years of echinoderm evolution.” *Proceedings of the National Academy of Sciences* 100 (2003): 13356–61.

Hurst, LD., Atlan, A., Bengtsson, BO. “Genetic conflicts.” *Quarterly Review of Biology* 71 (1996): 317-64.

Ihmels, J., Bergmann, S., Barkai, N. "Defining transcription modules using large-scale gene expression data." *Bioinformatics* 166 (2004)

Johnson, N. "Hybrid incompatibility and speciation." *Nature Education* 1.1 (2008): 20.

Johnson, NA. "Gene interaction and the origin of species." *Epistasis and the Evolutionary Process* Oxford. University Press, New York. 2000. 197-212.

Johnson, NA., Porter, AH. "Rapid speciation via parallel, directional selection on regulatory genetic pathways." *Journal of Theoretical Biology* 205 (2000): 527-42.

Jordan, IK., Marino-Ramirez, L., Koonin, EV. "Evolutionary significance of gene expression divergence." *Gene* 345.1 (2005): 119-26.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Wolfgang, E., Muetzel, B., Wirkner, U., Ansorge, W., Paabo, S. "A neutral model of transcriptome evolution." *PLoS Biol* 2.5 (2004): 132 doi:10.1371/journal.pbio.0020132.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, SL. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." *Genome Biology* 14.4 (2013).

Kimura, M. "Evolutionary rate at the molecular level." *Nature* 217 (1968): 624-26.

King, JL., Jukes, TH. "Non-Darwinian evolution." *Science* (1969): 788-798.

King, MC., Wilson, AC. "Evolution at two levels in humans and chimpanzees." *Science* 188.4184 (1975): 107-16.

Langmead, B., Salzberg, S. "Fast gapped-read alignment with Bowtie 2." *Nature Methods* 9 (2012): 357-59.

Li, H. "seqtk Toolkit for processing sequences in FASTA/Q formats." *GitHub* GitHub, Inc. 2015. February, 10th 2015. https://github.com/lh3/seqtk.

Li, H., Durbin, R. "Fast and accurate short read alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (2009): 1754-69.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis G., 1000 Genome Project Data Processing Subgroup. "The Sequence alignment/map (SAM) format and SAMtools." *Bioinformatics* 25 (2009): 2078-9.

Love, MI., Huber, W., Anders, S. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq 2." *Genome Biology* 15.550. (2014): doi: 10.1186/s13059-014-0550-8.

Lunter, Goodson. "Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads." *Genome Research* 21 (2011): 936-39.

Mayr, E. *Systematics and the Origin of Species* (Columbia Univ. Press, New York). 1942.

Mayr, E. "Timor and the colonization of Australia by birds." *Emu* 44 (1944): 113-30.

Mayr E. Animal Species and Evolution. Harvard University Press, Cambridge. MA. 1963.

Maheshwari, S., Barbash, DA. "The genetics of hybrid incompatibilities." *Annual Review Genetics* 45 (2011): 331–55.

Masly, JP., Presgraves, DC. "High-resolution genome-wide dissection of the two rules of speciation in *Drosophila." PLoS Biology* 5 (2007): 243.

McManus, CJ., "Regulatory divergence in Drosophila revealed by mRNA-seq." *Genome Results* 20 (2010): 816–25.

Mitchell-Olds, T., Willis, JH., Goldstein, DB. "Which evolutionary processes influence natural genetic variation for phenotypic traits?" *Nature Reviews Genetics* 8 (2007): 845-56.

Muller, HJ. "Isolating mechanisms, evolution, and temperature." *Biology Symposium* 6 (1942): 71-125.

Nowick, K., Carneiro, M., Faria, RA. "A prominent role of KRAB-ZNF transcription factors in mammalian speciation?" *Trends in Genetics* 29 (2013): 130–139.

Orr, HA. "Haldane's rule has multiple genetic causes." *Nature* 361 (1993): 532-33.

Orr, HA. "The population genetics of speciation: the evolution of hybrid incompatibilities." *Genetics* 139.4 (1995): 1805-13.

Ortiz-Barrientos, D., Counterman, BA., Noor, MAF. "Gene expression divergence and the origin of hybrid dysfunctions." *Genetica* 129 (2007): 71-81.

Pal, C., Papp, B., Hurst, LD. "Highly expressed genes in yeast evolve slowly" *Genetics* 158.2 (2001): 927-31.

Phifer-Rixey, M., Bomhoff, M., Nachman, MW. Genome-wide patterns of differentiation among house mouse subspecies. *Genetics* 198.1 (2014). 283-97.

Porter, AH., Johnson, NA. "Speciation despite gene flow when developmental pathways evolve." *Evolution* 56 (2002): 2103–11.

Presgraves, DC. "The molecular evolutionary basis of species formation." *Nature Review Genetics* 11 (2010): 175–180.

Price, TD., Bouvier, MM. "The evolution of F1 postzygotic incompatibilities in birds." *Evolution* 56.10 (2002): 2083-89.

Price T. *Speciation in Birds*. Ed. by G Hefta and J Besch. *Roberts and Company Publishers.* Greenwood Village, CO. 2008.

Ranz, JM., Castillo-Davis, CI., Meiklejohn, CD., Hartl, DL. “Sex-dependent gene expression and evolution of the *Drosophila* transcriptome.” *Science* 300 (2003): 1742-45.

Ranz, JM., Namgyal, K., Gibson, G., Hartl, DL. “Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*.” *Genome Results* 14 (2004): 373-79.

Rice, WR., and Hostert, EE. “Laboratory experiments in speciation: What have we learned in 40 years?” *Evolution* 47.6 (1993): 1637-53.

Renaut, S., Nolte, AW., Bernatchez, L. “Gene expression divergence and misexpression between lake whitefish species pairs (*Coregonus spp. Salmonidae*).” *Molecular Biology Evolution* 4 (2009): 925-36.

Rocha, EP., Danchin, A. “An analysis of determinants of amino acids substitution rates in bacterial proteins.” *Molecular Biology Evolution* 21 (2004): 108–16.

Seehausen, O., Butlin, RK., Keller, I., Wagner, CE., Boughman, JW., Hohenlohe, PA., Peichel, CL., Saetre, GP., Bank, C., Brannstrom, A., Brelsford, A., Clarkson, CS., Eroukhmanoff, F., Feder, JL., Fischer, MC., Foote, AD., Franchini, P., Jiggins, CD., Jones, FC., Lindholm, AK., Kucek, K., Maan, ME., Marques, DA., Martin, SH., Matthews, B., Meier, JI., Most, M., Nachman, MW., Nonaka, E., Rennison, DJ., Schwarzer, J., Watson, ET., Westram, AM., Widmer, A. “Genomics and the origin of species.” *Nature Review Genetics* 15 (2014): 176-92.

Smeds, L., Kunstner, A. "ConDeTri – A content dependent read trimmer for Illumina data." *PLoS ONE* 6.10 (2011): e26314. doi:10.1371/journal.pone.0026314.

Song, R., Messing, J. "Gene expression of a gene family in maize based on noncollinear haplotypes." *Proceedings of the National Academy of Science* 100 (2003): 9055–60.

Springer, NM., Stupar, RM. "Allele-specific expression patterns reveal biases and embryo-specific parent-of-origin effects in hybrid maize." *Plant Cell* 19 (2007): 2391–2402.

Stevenson, KR., Coolon, JD., Wittkopp, PJ. "Sources of bias in measures of allele-specific expression derived from RNA-Seq data aligned to a single reference genome." *BMC Genomics* 14.536 (2013): doi:10.1186/1471-2164-14-536.

Templeton, AR. "Mechanisms of speciation—a population genetic approach." *Annual Review Ecological Systems* 12 (1981): 23-48.

Tirosh, I., Reikhav, S., Levy, AA., Barkai, N. "A yeast hybrid provides insight into the evolution of gene expression regulation." *Science* 324.5927 (2009): 659-62.

Trapnell, C., Patcher, L., Salzber, SL. "TopHat: discovering splice junctions with RNA-Seq." *Bioinformatics* 25.9 (2009) 1105-11.

Turelli, M., Orr, HA. "The Dominance Theory of Haldane's Rule." *Genetics* 140 (1999): 389-402.

Turelli, M., Orr, HA. "The evolution of postzygotic isolation: accumulating Dobzhansky-Muller incompatibilities." *Evolution* 55.6 (2001): 1085-94.

Warren, WC., Clayton, DF., Ellegren, H., Arnold, AP., Hillier, LW., Künstner, A., Searle, S., White, S., Vilella, AJ., Fairley, S., Heger, A., Kong, L., Ponting, CP., Jarvis, ED., Mello, CV., Minx, P., Lovell, P., Velho, TA., Ferris, M., Balakrishnan, CN., Sinha, S., Blatti, C., London, SE., Li, Y., Lin, YC., George, J., Sweedler, J., Southey, B., Gunaratne, P.,

Watson, M., Nam, K., Backström, N., Smeds, L., Nabholz, B., Itoh, Y., Whitney, O, Pfenning, AR., Howard, J.. Völker, M., Skinner, BM., Griffin, DK., Ye, L., McLaren, WM., Flicek, P., Quesada, V., Velasco, G., Lopez-Otin, C., Puente, XS., Olender, T., Lancet, D., Smit, AF., Hubley, R., Konkel, MK., Walker, JA., Batzer, MA., Gu, W., Pollock, DD., Chen, L., Cheng, Z., Eichler, EE., Stapley, J., Slate, J., Ekblom, R., Birkhead, T., Burke, T., Burt, D., Scharff, C., Adam, I., Richard, H., Sultan, M., Soldatov, A., Lehrach, H., Edwards, SV., Yang, SP., Li, X., Graves, T., Fulton, L., Nelson, J., Chinwalla, A., Hou, S., Mardis, ER., Wilson, RK. "The genome of a songbird." *Nature* 464.7289 (2009) 757-62.

Wittkopp, PJ., Haerum, BK., Clark, AG. "Evolutionary changes in cis and trans gene regulation." *Nature* 430 (2004): 85–88.

Wolf, JBW., Lindell, J., Backstrom, N. "Speciation genetics: current status and evolving approaches." *Philosophy of Transactions of the Royal Society B Biological Science* 365 (2010): 1717–1733.

Wray, GA., Hahn, MW., Abouheif, E., Balhoff, JP., Pizer, M., Rockman, MV., Romano, LA. "The evolution of transcriptional regulation in eukaryotes." *Molecular Biology Evolution* 20 (2003): 1377– 1419.

Wu, CI., Davis, AW. "Evolution of post-mating reproductive isolation: the composite nature of Haldane's rule and its genetic basis." *The American Naturalist* 142.2 (1993): 187-212.

Wu, CI., Ting, CT., "Genes and speciation." *Nature Reviews Genetics* 5 (2004): 114-22.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., Shmueli, O. "Genome-wide

midrange transcription profiles reveal expression level relationships in human tissue specification." *Bioinformatics*. 25.5 (2005): 650-59.

Zann, RA. *The Zebra Finch: A Synthesis of Field and Laboratory Studies*. Oxford University Press. Oxford.1996.

Zufall, RA., Rausher, MD. "Genetic changes associated with floral adaptation restrict future evolutionary potential." *Nature* 428 (2004): 847–850.

Table 1. List of mappers and alternative mapping strategies used in attempts to correct for mapping bias.

Mapping Program	Altered Parameter	Default Setting	Altered Setting
Bowtie 2	score-min	-0.6	-0.5
BWA	mismatch	4	5
STAMPY	substitution rate	0.01	0.05

Table 2. Gene Ontology results for the comparison between Australian and Timor zebra finch gene expression levels. Shown are the top 10 results based on adjusted Fisher p-value. Top three results were considered significant (*P*-value < 0.01).

GO	go_description	total	expectation	observation	fisher	adj.fisher
GO:0055114	oxidation-reduction process	463	15	35	7e-06	0.0063
GO:0005515	protein binding	5548	184	140	1.3e-05	0.0063
GO:0016491	oxidoreductase activity	361	12	28	3.2e-05	0.01
GO:0008237	metallopeptidase activity	80	3	10	3e-04	0.066
GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c	5	0	3	0.00035	0.066
GO:0005622	intracellular	1674	56	33	0.00057	0.074
GO:0016884	carbon-nitrogen ligase activity, with glutamine as amido-N-donor	6	0	3	0.00067	0.074
GO:0008152	metabolic process	537	18	33	0.00075	0.074
GO:0001558	regulation of cell growth	23	1	5	0.00081	0.074
GO:0005975	carbohydrate metabolic process	125	4	12	0.00091	0.074

Table 3. KEGG pathway analysis results for comparison between Australian and Timor zebra finch gene expression levels. Top 10 results are shown.

pathway_id	pathway_title	total	expectation	observation	fisher	adj.fisher
gga03440	Homologous recombination	24	1	5	0.00078	0.043
gga00190	Oxidative phosphorylation	75	2	8	0.0022	0.061
gga03430	Mismatch repair	15	0	3	0.011	0.15
gga00980	Metabolism of xenobiotics by cytochrome P450	15	0	3	0.011	0.15
gga00511	Other glycan degradation	7	0	2	0.019	0.17
gga00280	Valine, leucine and isoleucine degradation	33	1	4	0.02	0.17
gga00860	Porphyrin and chlorophyll metabolism	19	1	3	0.022	0.17
gga00350	Tyrosine metabolism	22	1	3	0.032	0.22
gga00640	Propanoate metabolism	23	1	3	0.036	0.22
gga03030	DNA replication	26	1	3	0.049	0.27

Table 4. Gene Ontology results for the comparison between Australian and hybrid zebra finch gene expression levels. Top 10 results are shown.

GO	go_description	total	expect	observation	fisher	adj.fisher
GO:0006122	mitochondrial electron transport, ubiquinol to cytochrome c	5	0	3	1.2e-05	0.0051
GO:0008121	ubiquinol-cytochrome-c reductase activity	8	0	3	6.4e-05	0.014
GO:0008610	lipid biosynthetic process	15	0	3	0.00049	0.071
GO:0051393	alpha-actinin binding	4	0	2	0.00067	0.073
GO:0071391	cellular response to estrogen stimulus	5	0	2	0.0011	0.096
GO:0009755	hormone-mediated signaling pathway	10	0	2	0.0048	0.11
GO:0005634	nucleus	2257	24	13	0.0096	0.11
GO:0003874	6-pyruvoyltetrahydropterin synthase activity	1	0	1	0.011	0.11
GO:0009181	purine ribonucleoside diphosphate catabolic process	1	0	1	0.011	0.11
GO:0070552	BRISC complex	1	0	1	0.011	0.11

Table 5. Gene Ontology results for the comparison between Timor and hybrid zebra finch gene expression levels. Top 10 results are shown.

GO	go_description	total	expectation	observation	fisher	adj.fisher
GO:0019388	galactose catabolic process	1	0	1	0.003	0.043
GO:0003978	UDP-glucose 4-epimerase activity	1	0	1	0.003	0.043
GO:0030552	cAMP binding	1	0	1	0.003	0.043
GO:0005199	structural constituent of cell wall	1	0	1	0.003	0.043
GO:0018117	protein adenylylation	1	0	1	0.003	0.043
GO:0070733	protein adenylyltransferase activity	1	0	1	0.003	0.043
GO:0035494	SNARE complex disassembly	1	0	1	0.003	0.043
GO:0010807	regulation of synaptic vesicle priming	1	0	1	0.003	0.043
GO:0090231	regulation of spindle checkpoint	1	0	1	0.003	0.043
GO:0071207	histone pre-mRNA stem-loop binding	1	0	1	0.003	0.043

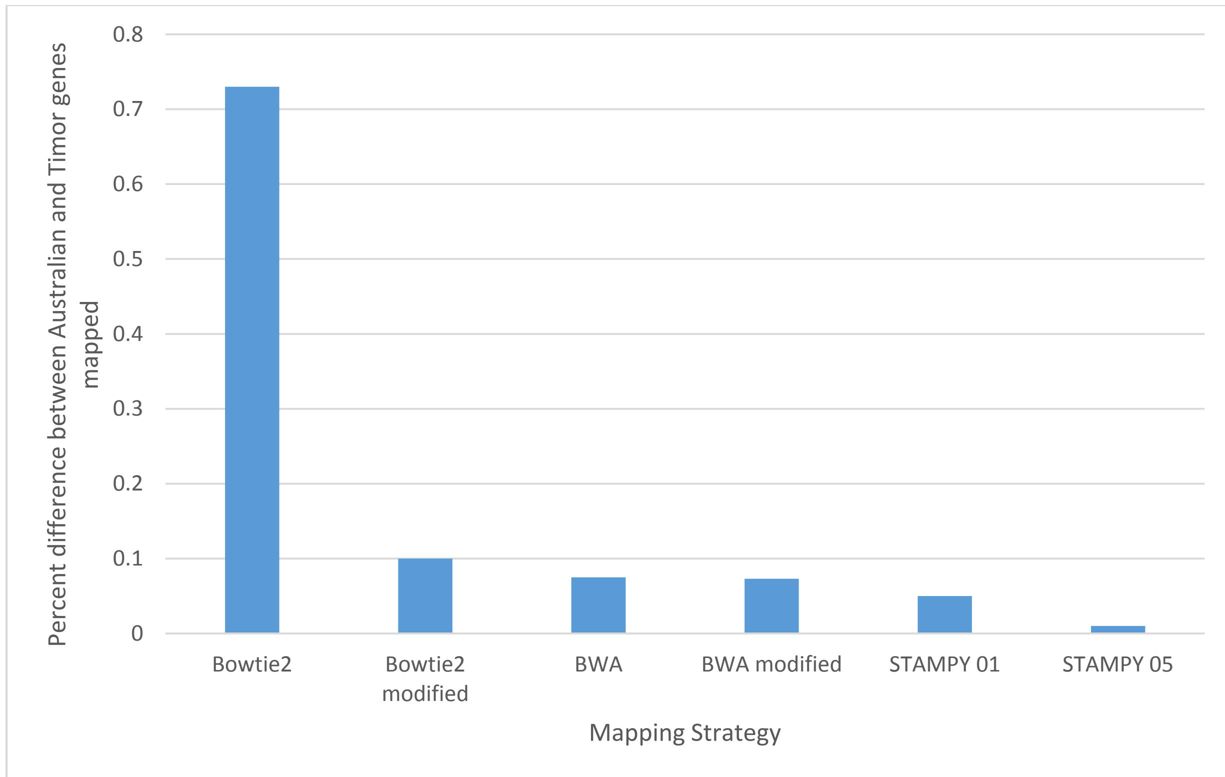


Figure 1: Bar graph showing effectiveness of mapping strategies. Default settings and alternative settings for three different mapping programs (Bowtie 2, BWA and STAMPY) are shown. The height of each bar represents the magnitude of the difference between the rates at which reads from Australian zebra finch mapped to a reference genome compared to the rates at which reads from Timor zebra finch mapped to the same reference genome. Mapping strategies with smaller bars more effectively corrected for mapping bias.

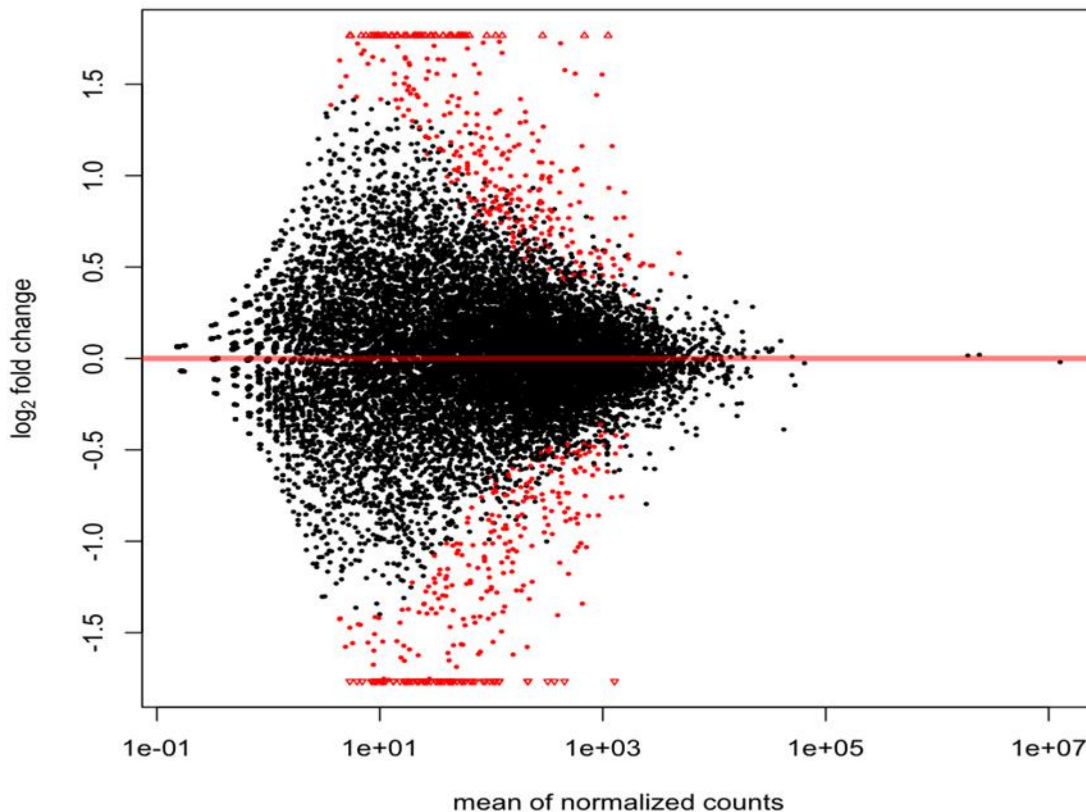
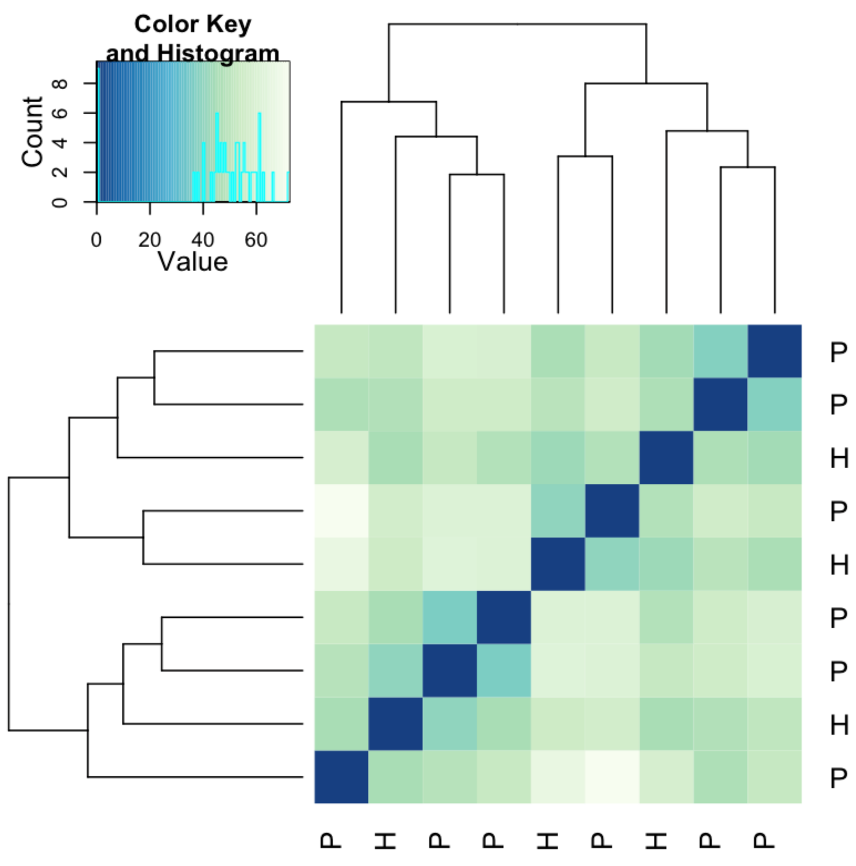


Figure 2: MA plot showing differentially expressed genes in comparison between Australian and Timor zebra finch populations. The y-axis shows log2 fold change for gene expression and the x-axis shows the mean of normalized counts. The black dots represent genes that did not significantly differ in expression between the two populations while the red dots represent genes that were significantly different in gene expression (*P*-value < 0.01). In total, 632 genes were found to significantly differ in expression values between the two populations.


Color Key
and Histogram
Count
0
2
4
6
8
0
20
40
60
Value
P
P
H
P
H
P
P
H
P
P
H
P
P
H
P
H
P
P

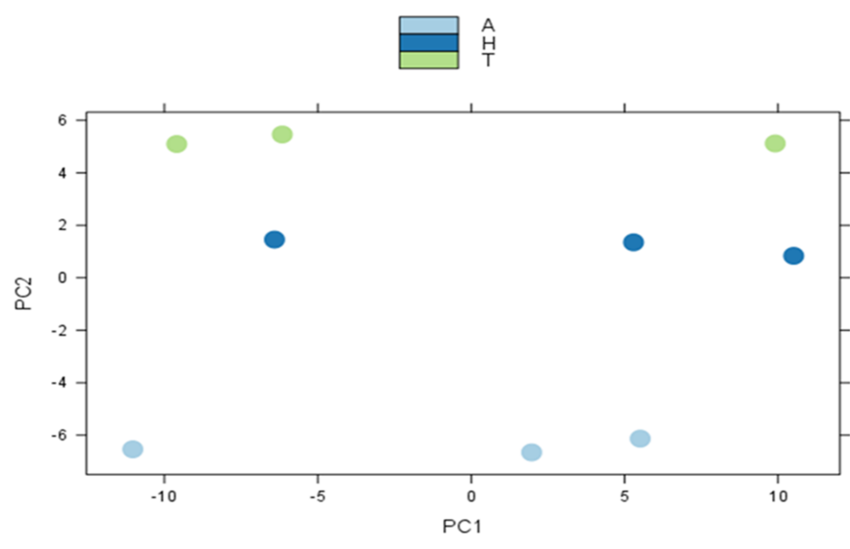
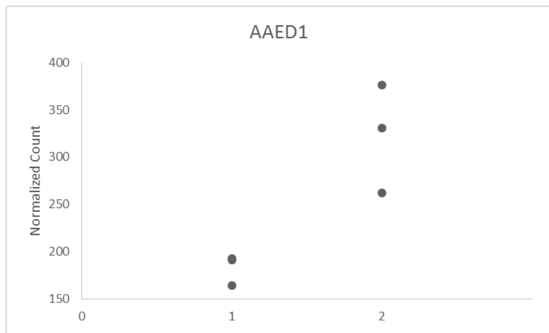
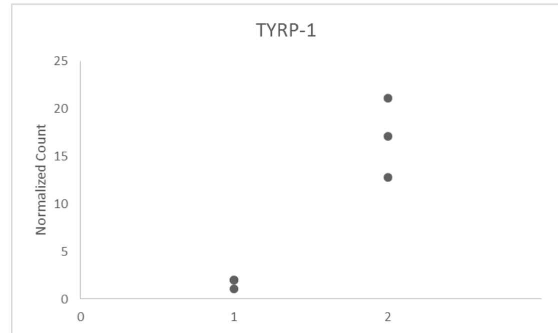

A
H
T
6
4
2
0
-2
-4
-6
PC2
-10
-5
0
5
10
PC1

Figure 3: Heatmap and principal component analysis showing similarity of individuals from all three populations based on gene expression levels. Darker squares of the heat map show closer relatedness between individuals. Hybrids are all found dispersed between the parental individuals demonstrating that the hybrids are largely intermediate in overall expression profile. Principal component analysis shows PC1 and PC2. Data points designated as H represent hybrid birds while data points designated as A represent birds from the Australian subspecies and data points designated as T represent birds from the Timor subspecies. PC1 does not segregate individuals based on population of origin. However, PC2 does. Ion and protein binding genes were shown to have contributed to the differences between individuals in PC1.

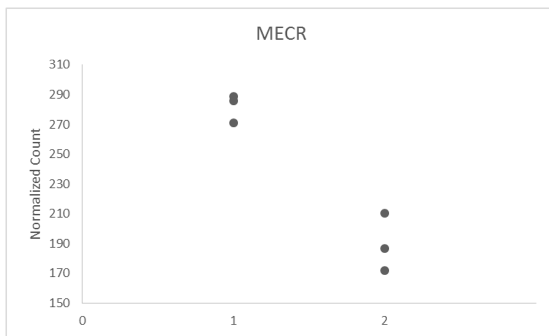
3. A)



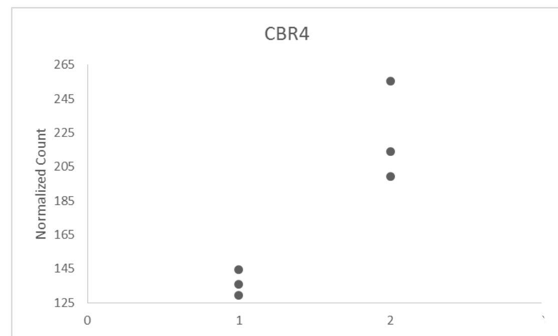
E)



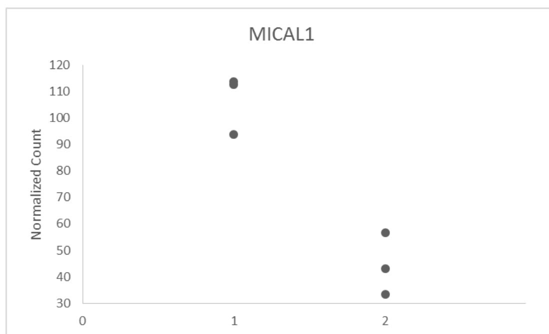
B)



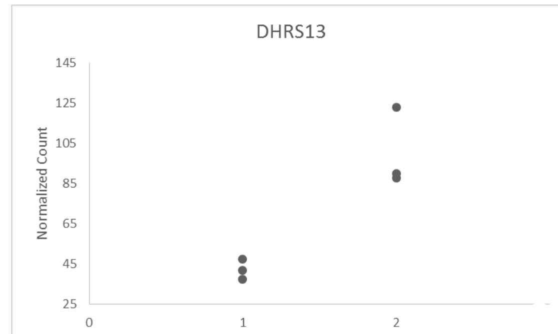
F)



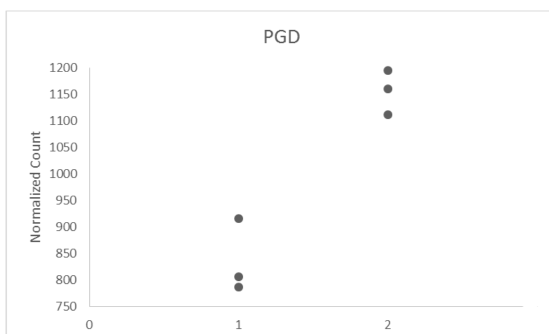
C)



G)



D)



H)

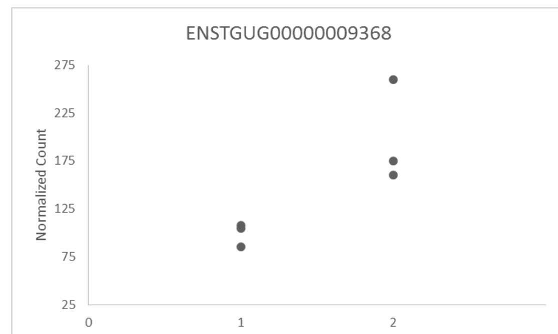


Figure 4: Plots showing differences in normalized counts of gene expression data for Australian and Timor zebra finch. X-axis one of each plot represents expression data for Australian individuals while x-axis number 2 of each plot represents expression data for Timor individuals. Plots for the top eight genes that most significantly differed in expression are shown. The y-axis shows the mean normalized count. Note the difference in scales.

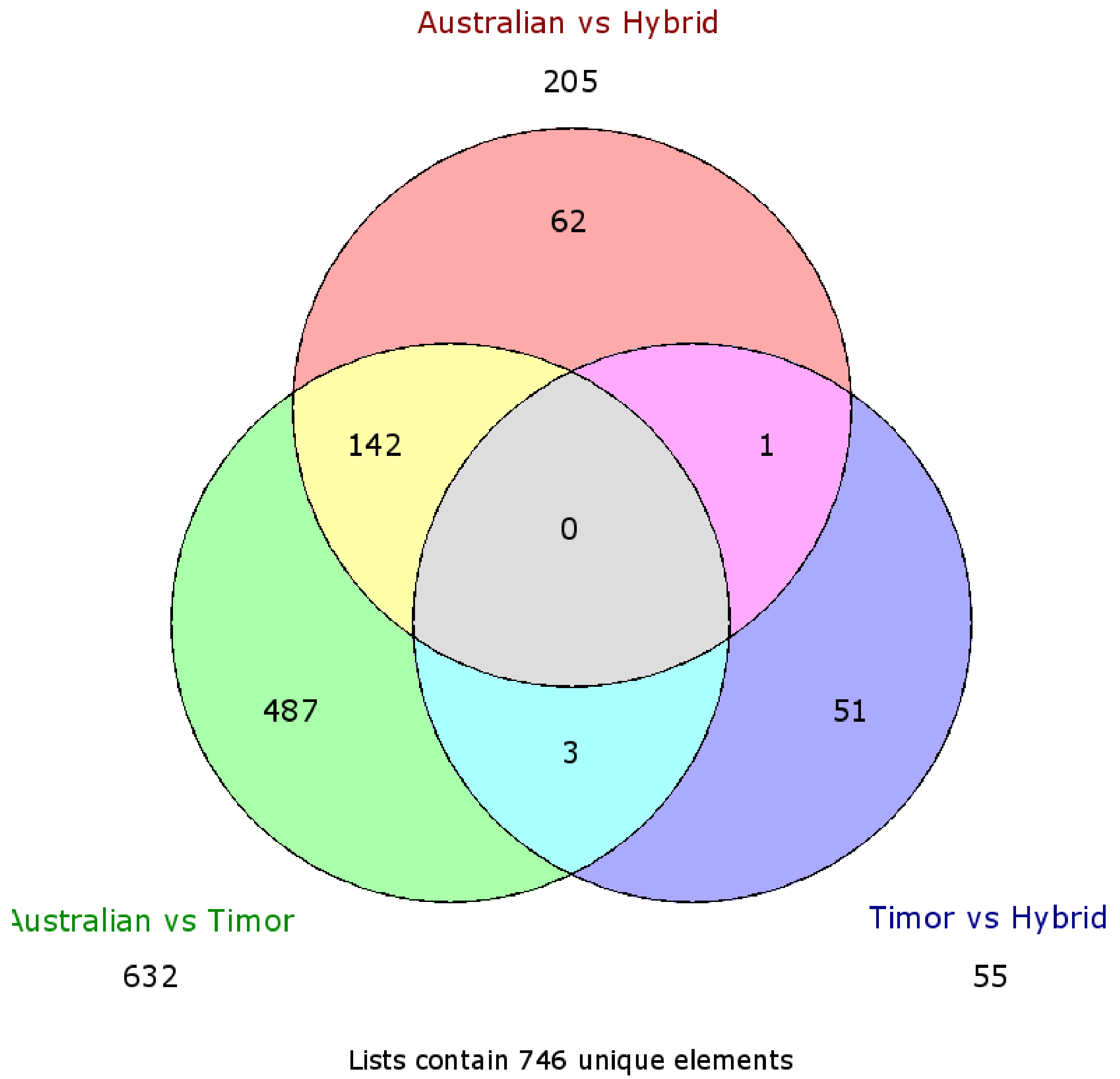


Figure 5: Venn diagram showing the overlap in the number of reads that were differentially expressed in each pairwise comparison between populations. There is a large amount of overlap in comparisons between Australian to Timor populations and the Australian to hybrid comparison indicating broad overlap in these two analyses.

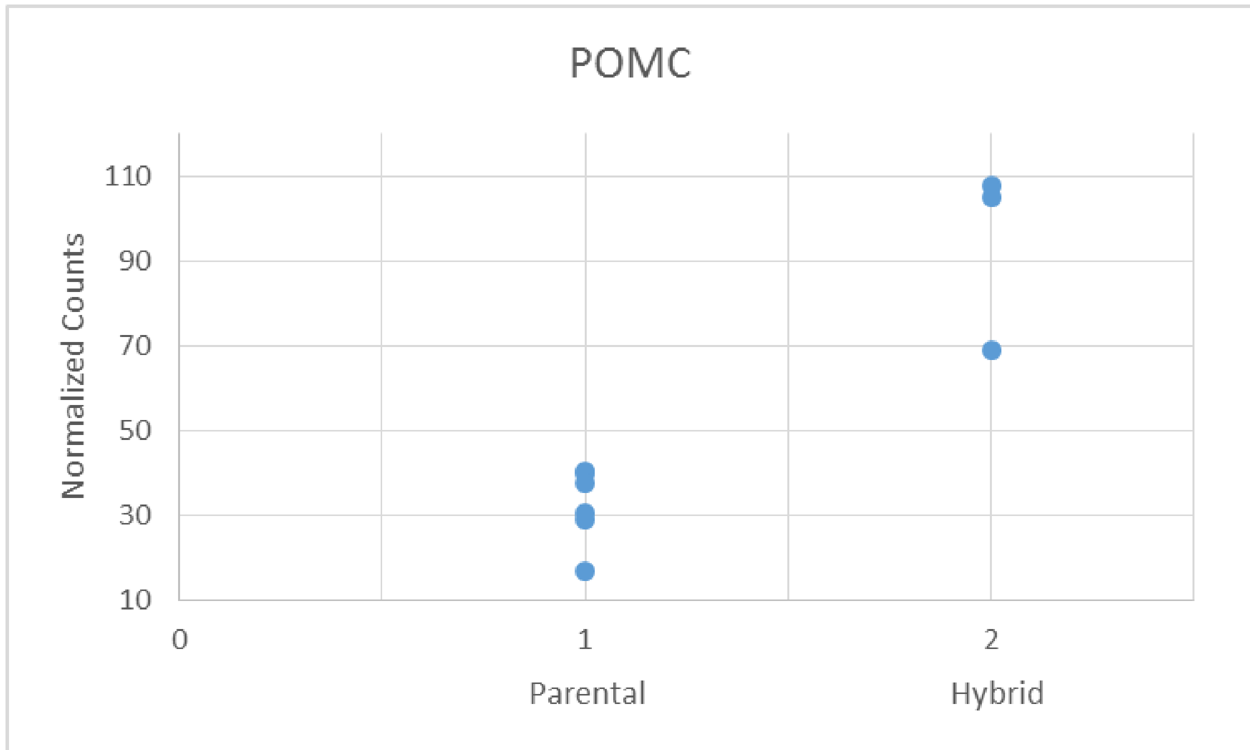


Figure 6: Plot showing the normalized counts data for the POMC gene. POMC was the gene that was found to be different in expression in the hybrid to parental comparison. X-axis number one represents birds from both of the parental populations. X-axis number two represents birds from the hybrid population. The y-axis shows the mean normalized counts. Gene expression was higher in the hybrids showing evidence of over-dominance.