Genomic diversity and speciation in East African cichlid fish

Milan Malinsky

Wellcome Trust Sanger Institute

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Girton College September 2015

To Alena and Sasha . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This thesis does not exceed the length limit of 60,000 words specified by the Biology Degree Committee.

Milan Malinsky
September 2015

Acknowledgements

I would like to acknowledge the people who have helped along the way to becoming a scientist. First my parents and especially my grandparents who encouraged my natural curiosity from an early age. They did not burden me by making me believe in given wisdoms, but instead led me to follow my observations and analysis and make my own conclusions about the world we live in. For this, I am and always will be grateful.

I would like to thank my teachers in the Czech Republic, the most important among them Martina Aberlová whose influence in the final three years of secondary school helped me to get a well-rounded basic education, and at high school Alena Šteklová who encouraged me to continue learning the English language. This enabled me to enter the English speaking world and thus the global scientific enterprise.

Many specific skills I used during my PhD, especially in computer programming and algorithms, were developed during an undergraduate degree in Computer Science at the University of Birmingham, UK. Here, my thanks go to William Edmondson, the Admission Tutor for his confidence in me, Peter Coxhead, my personal tutor, for his help and guidance, and Hamid Dehghani, my final year project supervisor, for introducing me to the world of cutting edge science at an international conference.

Next, my thanks go to Stephen Eglen and Simon Tavaré, for running the Cambridge MPhil course in Computational Biology and for helping me to make the most of it. It was while I was a student on this course that Simon introduced me to Eric Miska who, fascinated by the extraordinary evolution and diversity of East African cichlids, provided the initial spark for this PhD project.

Soon, I too became excited by cichlid evolution and could see that computational analysis of genomic data would provide important clues about the amazing history of this fish family. My PhD supervisor Richard Durbin is a research leader in analysing population genomic data and his help has been invaluable to this project. Access to the Sanger Institute facilities and Richard's expert strategic guidance are what made this project truly possible. I am also very appreciative of the specific insights that Richard provided during the project, often helping to push forward when working on my own felt like I was reaching a dead-end. In the first year in Richard's group I especially enjoyed working with Jared Simpson, who introduced me to algorithms for genome assembly and to C++ programming. In the next three years, I enjoyed collaborating with Stephan Schiffels with whom I had numerous discussions of population genetics topics, each of which increased my understanding of this scientific field.

George Turner and Martin Genner have been my principal collaborators and advisors on evolution and ecology of cichlid fish. They helped me to develop from being a 'Computer Scientist' at the beginning of the PhD to the 'computer-savvy' evolutionary biologist I am today. Without their input this work and my personal development would be greatly diminished.

I owe thanks to the Wellcome Trust for financially supporting my PhD.

Abstract

Unravelling the genetic basis of functional diversification is fundamental for our understanding of vertebrate evolution and can also have significant implications for animal and human health. Speciation leads to phenotypic diversity by producing new units of evolution - species. In less than five million years, East African cichlids have radiated into thousands of species that differ in craniofacial morphology, pigmentation, behaviour and many other traits. In this thesis, I take advantage of recent advances in DNA sequencing technologies to study the genetic basis of this exceptional diversity. First, as a member of the Cichlid Genome Consortium (CGC), I identified and characterised over 1,000 loci generating microRNAs, non-coding RNA genes that regulate expression and may play a role in the evolution of cichlid traits. Next, at the Sanger Institute, we obtained whole genome sequences of 271 individuals from over 70 species from in and around Lake Malawi. I aligned the data to a reference genome generated by CGC, and used the results to: 1) ascertain the overall levels of genetic variation and allele sharing within and between species; 2) reconstruct relationships between the species; 3) study in detail the genetic causes and consequences of early stages of speciation in Lake Massoko, a small isolated crater lake in southern Tanzania. I found that that the genetic distance between the most diverged Lake Malawi species is surprisingly low, comparable to the distance between two strains of zebrafish, that there are discrepancies between relationships inferred from molecular phylogeny and from traditional taxonomy, and that measurable introgression between species occurs but does not seem to be common. In Lake Massoko, I identified clearly demarcated genomic regions of differentiation between incipient species in sympatry. Interestingly, there are no fixed differences; instead I found a genome-wide pattern with dozens of loci of moderate divergence. With collaborators, we found that alleles in the regions are associated to mate preferences in the laboratory, and genes in the regions are enriched for molecular functions consistent with morphological and sensory system adaptation. To facilitate this work, I constructed whole genome alignments between CGC genome assemblies, assigned ancestral alleles to genetic variants in Lake Malawi, and built a genome browser that can be used to visualise datasets produced by us and the CGC. The browser website has been visited over 650 times since March 2014. In addition, I developed a new method for genome assembly to reduce problems caused by heterozygosity, taking advantage of mother-father-offspring trio data. I applied this method to obtain *de novo* genome assemblies of three cichlid species, and also three highly heterozygous *Heliconius* butterfly species. These datasets, tools, and findings make significant contributions to evolutionary genetics and will provide a foundation for future research on processes underlying the evolution of phenotypic diversity, especially in cichlids.

Table of contents

List of figures **xv**

List of tables **xix**

1 Introduction **1**

1.1 Genetic and phenotypic evolution 1

1.1.1 Introduction 1

1.1.2 Vertebrate genome architecture 4

1.1.3 Molecular evolution 5

1.1.4 Speciation 8

1.1.5 Genome sequencing and bioinformatics 9

1.2 Evolutionary and speciation genomics 10

1.2.1 Measures of sequence divergence 14

1.2.2 ‘Islands of speciation’ or ‘incidental islands’? 14

1.3 East African cichlid radiation 17

1.3.1 Cichlid fish 17

1.3.2 East African cichlids 18

1.4 Overview of the remainder of this thesis 21

2 The cichlid genome project **23**

2.1 Introduction 23

2.1.1 Publication note 23

2.1.2 Five reference genomes 23

2.2 Background 26

2.2.1 The nature and functions of microRNAs 26

2.2.2 MicroRNAs in vertebrate development and evolution - a hypothesis 27

2.3 MicroRNA annotation 28

2.3.1 Small RNA sequencing 28

2.3.2 Identification of miRNA loci 29
2.3.3 Evolution in cichlid miRNA repertoires 30
2.3.4 Evolution of miRNA targets 32
2.4 Detailed methods 37
2.4.1 Searching for novelty in cichlid miRNA repertoires 37
2.4.2 Target prediction by RNAhybrid 38
2.4.3 Target prediction with PACMIT 38

3 New genome sequence datasets 41
3.1 Introduction 41
3.1.1 Overview of whole-genome data 41
3.2 Alignment, variant calling, filtering, and genotype refinement 47
3.2.1 DNA extraction and sequencing 47
3.2.2 Alignment 48
3.2.3 Sample call-sets 48
3.2.4 Variant calling, filtering, genotype refinement, and haplotype phasing 48
3.3 Coverage and cross-contamination estimates from data 50
3.3.1 Lake Malawi call set 50
3.3.2 Crater lake call set 51
3.4 Whole genome alignments 52
3.5 Cichlid genome browser 54

4 *De novo* genome assembly 57
4.1 Introduction 57
4.2 trio-sga - Trio-aware genome assembly 60
4.2.1 Overview 60
4.2.2 Algorithms 61
4.3 *Heliconius* butterfly genome assemblies 65
4.4 Cichlid trio genome assemblies 70
4.5 *Andinoacara coeruleopunctatus* genome assembly 72

5 Lake Malawi genetic diversity 73
5.1 Introduction 73
5.2 Genomic diversity 74
5.3 Phylogenetic relationships 79
5.4 Exploring evidence for interspecific introgression 85

5.4.1 ABBA-BABA tests 85
5.4.2 Chromopainter and fineSTRUCTURE 87
5.5 Methods 91

6 Incipient speciation in Lake Massoko, Tanzania 95
6.1 Introduction 95
6.1.1 Publication note 95
6.1.2 Background 95
6.2 Whole-genome evidence of Massoko divergence 99
6.3 Population size estimates 104
6.4 Islands of speciation 107
6.5 Divergent SNPs associated with mate choice 117
6.6 Functions of adaptation 119
6.7 Comparisons to other systems 124
6.8 Detailed methods 125
6.8.1 Field sampling and eco-morphological analysis 125
6.8.2 RAD-seq data processing and analysis 125
6.8.3 Whole genome data processing and analysis 127
6.8.4 Mate choice experiments 132
6.8.5 Measuring rhodopsin absorption spectra 134

7 Conclusions 137
7.1 Conclusions 137
7.2 Future work 138

References 143

Appendix A Lake Malawi genetic diversity 161

Appendix B Lake Massoko speciation 163

List of figures

1.1 Heritability (h^2) estimates for human traits 2
1.2 Separation of germline from soma 3
1.3 Components of vertebrate genomes 4
1.4 Forces influencing genetic composition of natural populations 7
1.5 The speciation continuum 8
1.6 Signatures of selective sweeps in population genetics data 12
1.7 Sequence genealogies and speciation 13
1.8 Effect of different sequence genealogies on F_{ST} and d_{XY} 16
1.9 Major physical features of cichlids 17
1.10 Evolution of cichlid feeding apparatus in East Africa 18
1.11 The 'out of Tanganyika' model of East African cichlid radiations 19
1.12 Parallel evolution of body shapes in Lakes Malawi and Tanganyika . . 20

2.1 East African radiation of cichlid fish 24
2.2 Biogenesis of microRNAs 26
2.3 The canonical miRNA target site 27
2.4 Length distribution of small RNAs sequenced from *M. zebra* embryos . 29
2.5 Evolution of miRNA novelty 31
2.6 The birth and death of de-novo originated cichlid-specific microRNAs . 32
2.7 Sequence evolution at specific miRNA target sites 33
2.8 Sequence evolution in *tmem141* 3′UTRs 34
2.9 miR-99b downregulates luciferase with *M. zebra tmem141* 3′UTR . . . 34
2.10 Purifying selection on miRNA target sites 36
2.11 Precision versus sensitivity of a selection of miRNA target prediction algorithms 38

3.1 Map of crater lake region in Southern Tanzania 45
3.2 Collection sites of non-crater-lake *Astatotilapia calliptera* specimens . . 45

3.3 Cross-contamination and read-depth estimates for the Lake Malawi variant call-set . . . 50
3.4 Cross-contamination and read-depth estimates for the Crater lake variant call-set . . . 51
3.5 Cambridge Cichlid Browser *M. zebra* genome gateway . . . 55
3.6 Cambridge Cichlid Browser - example browser graphic . . . 55
3.7 A map showing the location of CCB users . . . 56
3.8 Ten cities with the highest contribution to CCB sessions . . . 56

4.1 An example of genome assembly graph with two bubbles due to heterozygous sites . . . 58
4.2 An illustration of the N50 measure of assembly contiguity . . . 59
4.3 Trio aware read filtering and error correction . . . 64
4.4 Estimates of genome size and heterozygosity for *Heliconius* genomes . . . 66
4.5 Read phasing reduces heterozygosity in *Heliconius* data . . . 67
4.6 The distributions of cichlid mate-pair insert sizes . . . 71
4.7 Distribution of scaffold lengths in *A. coeruleopunctatus* assembly . . . 72

5.1 Variants called against *M. zebra* genome in Lake Malawi samples . . . 75
5.2 Frequency of heterozygous sites in Lake Malawi samples . . . 75
5.3 F_{ST} divergence between Lake Malawi species . . . 77
5.4 Genome-wide F_{ST} profile between *P. subocularis* and *T. placodon* . . . 78
5.5 Principal Component Analysis of genetic variation in the Lake Malawi sample set . . . 79
5.6 Lake Malawi whole genome phylogeny . . . 81
5.7 Contrast between whole genome and mtDNA phylogenies . . . 82
5.8 Variation in phylogenies between 1,460 regions along the genome . . . 83
5.9 Difference between local phylogenies based on nuclear DNA and the mtDNA phylogeny . . . 85
5.10 The design of the ABBA-BABA test for introgression in *Lethrinops* . . . 86
5.11 The results of the ABBA-BABA test for introgression in *Lethrinops* . . . 87
5.12 Coancestry between Lake Malawi samples measured by the Chromopainter software . . . 88
5.13 *O. tetrastigma* Ilamba introgression into crater lake *Astatotilapia* . . . 89
5.14 *Placidochromis electra* introgression into *Buccochromis rhoadesii* . . . 90

6.1 Cichlid radiation in the crater lakes of southern Tanzania . . . 96
6.2 Collection sites of RAD sequenced non-crater-lake *A. calliptera* . . . 98

6.3 Lake Massoko divergence with whole genome sequence data 100
6.4 Crater lake whole genome sequence data 101
6.5 Lake Massoko fineSTRUCTURE results 102
6.6 MSMC cross-coalescence between Massoko and Mbaka river 103
6.7 Heterozygosity in Itamba, Massoko, and additional *A. calliptera* populations 104
6.8 Decay of linkage disequilibrium in Itamba and Massoko populations . . 105
6.9 Inferred population size histories for Massoko ecomorphs, Itamba, and *A. calliptera* from Mbaka river 106
6.10 Genome-wide pattern of F_{ST} divergence 107
6.11 Models of species formation used for coalescent simulations 108
6.12 Comparing the distributions of observed and simulated F_{ST} values . . . 109
6.13 Comparing the distributions of observed and simulated F_{ST} values for additional demographic models 110
6.14 Islands of speciation between benthic and littoral ecomorphs 113
6.15 Patterns of F_{ST}, d_{XY}, and π_{diff} in a speciation cluster on scaffold 88 . 114
6.16 Difference in nucleotide diversity between benthic and littoral ecomorphs (π_{diff}) in HDRs 115
6.17 Evidence against allopatric (double-invasion) divergence for Massoko . 116
6.18 Mate-choice trials 118
6.19 Enrichment Map for significantly enriched GO terms 120
6.20 Rhodopsin and rod cells 121

7.1 Parallel evolution of the ‘thick lip’ phenotype in Lake Malawi 140
7.2 Parallel evolution between lakes Tanganyika and Malawi 141

A.1 Average bootstrap values for 1,460 phylogenies representing regions along the genome 161
A.2 An illustration of the ABBA-BABA test for introgression 162

B.1 Statistical distribution of within-Massoko F_{ST} divergence 168
B.2 Genome-wide pattern of F_{ST} divergence using sliding windows or varying sizes 169
B.3 Neutral simulations - fitting split time to match Massoko littoral-benthic F_{ST} divergence 170
B.4 The two alleles of rhodopsin found in Massoko *Astatotilapia* 170

List of tables

1.1 Main forms of natural selection 6

1.2 Commonly used measures of population divergence 14

1.3 Interpreting genomic ‘islands’ of high differentiation 16

2.1 Versions of cichlid genome assemblies used in this thesis 25

3.1 Whole genome sequencing at Sanger Institute - an overview 41

3.2 Lake Malawi sequencing 44

3.3 Cichlid samples from outside Lake Malawi 46

3.4 Versions of non-cichlid teleost assemblies used in whole-genome alignments 53

4.1 *Heliconius* contig assembly statistics 68

4.2 *Heliconius* scaffold assembly statistics 68

4.3 *Heliconius* assembly statistics - high coverage 69

4.4 Cichlid contig assembly statistics 70

4.5 Cichlid scaffold assembly statistics 71

4.6 *A. coeruleopunctatus* assembly statistics 72

6.1 Location and geographical characteristics of explored crater lakes . . . 97

6.2 Depth distribution of ecomorphs in Lake Massoko 97

6.3 Results of Massoko morphological and stable isotope analysis 98

6.4 Sliding-window based F_{ST} summary 107

6.5 Candidate ‘islands of speciation’ 112

6.6 Genotyped variants used for mate-choice trials and F_{ST} values observed in the reference sample of 18 benthic and 16 littoral males 117

6.7 Numbers of genes available for Gene Ontology enrichment analysis . . . 119

6.8 Genes contributing to GO enriched terms related to sensory perception 122

6.9 Genes contributing to GO enriched terms related to morphogenesis . . 123

6.10 Genes contributing to GO enriched terms related to (steroid) hormone signalling . . . 123

B.1 Results of a survey of fish fauna in crater lakes of Rungwe District, Tanzania . . . 163
B.2 An overview of *Astatotilapia* samples collected for RAD sequencing . . 164
B.3 The migration parameter `M` and the probability of migration 164
B.4 Location and lengths of highly diverged regions (HDRs) 165
B.5 GO enrichment terms in candidate 'islands of speciation' 166
B.6 GO enrichment terms in all HDRs (±10kb) 166
B.7 GO enrichment terms in all HDRs (±50kb) 167