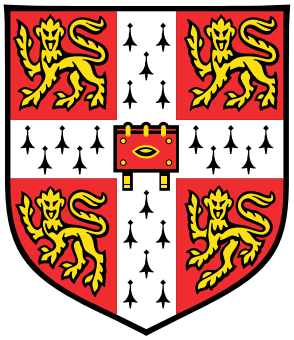
Genomic diversity and speciation in East African cichlid fish



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

List of tables

Chapter 1

Introduction

> *"How have all those exquisite adaptations of one part of the organisation to another part, and to the conditions of life, and of one distinct organic being to another being, been perfected?"*
>
> — Charles Darwin, *On the origin of species*, 1859, p.49

1.1 Genetic and phenotypic evolution

1.1.1 Introduction

First prompted by observations made as a naturalist on board of HMS *Beagle* in South America and following more than two decades of accumulating and reflecting on facts that could shed light on the origin of species, Charles Darwin proposed natural selection on phenotypic variation as the main mechanism of evolutionary change [1]. The essence of Darwin's argument [1, 2] can be summarised as follows:

1. There is a tendency of all organisms for geometric increase in numbers. For example, "if an annual plant produced only two seeds... and their seedlings next year produced two, and so on, then in twenty years there would be a million plants".

2. Despite this tendency, the numbers of organisms on earth cannot increase geometrically - the world could not hold them; in fact, the numbers remain more or less constant. Therefore, there must be a struggle for existence: since more eggs

or seeds and young are produced than can survive and reproduce, it follows that there must be competition for survival and reproduction[1].

3. There is variation between individuals, even within single species. Individuals with variations that confer an advantage in the struggle for existence have a greater probability to survive and procreate, passing on any heritable element of such variations to their offspring. On the other hand, disadvantageous variations will tend to be eliminated.

The action of natural selection critically depends on heritability - the extent to which variation among individuals in a population is predictably transmitted to their offspring [3]. Even after the scientific community rediscovered Gregor Mendel's work on patterns of inheritance [4], much critique of Darwinism, especially in the first half of 20th century, centred around the degree to which traits are heritable [2]. Darwin strongly believed the majority of inter-individual variation to be heritable, but his evidence was anecdotal, based on observation of plant and animal breeding and on examples of characters passed on in human families [1, pp.13-14]. Today, thanks to the statistical methods of quantitative genetics, we have measured the relative importance of genes and environment for over 17,000 human traits by studying over 2 million twin pairs [5] (Figure 1.1), and also for many traits in important agronomic and natural species [3]. Artificial selection studies have shown that "almost any species will respond to selection for virtually any trait" [3, p.682], even leading to claims that human twin studies "provide compelling evidence that all human traits are heritable: not one trait had a weighted heritability estimate of zero" [5].

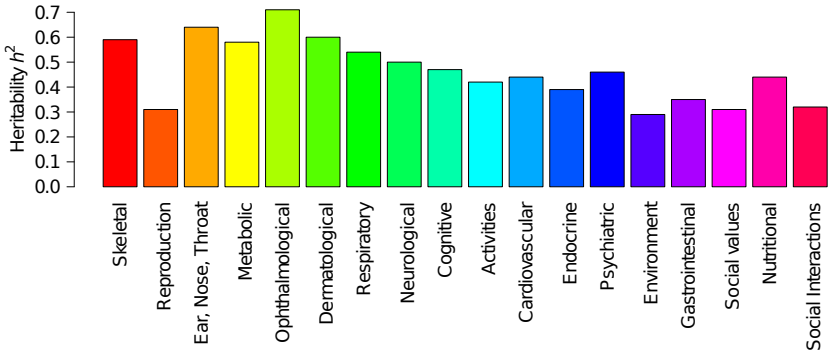


Fig. 1.1 **Heritability (h^2) estimates for over 17,000 human traits classified into 18 functional domains.** Data from [5].

[1]The concept of *struggle for existence* was key to Darwin's development of the theory of evolution by means of natural selection. However, some authors in do not mention it in modern presentations of natural selection, because (for example) selection can be effective even in situations when all individuals survive and reproduce.

In the early 20th century, chromosomes became recognised as the physical carriers of genes between generations [7], and the fact that genes are composed of DNA first demonstrated in 1944 by Avery *et al.*, showing the ability of DNA to transform bacterial cells [8]. The role of DNA as the genetic material was confirmed in 1952 by the Hershey-Chase experiment, showing that during a bacteriophage infection only DNA is injected into the bacterium, while protein is discarded and has no further function [9]. In animals, a division is made early in development between cells that will produce eggs or sperm (the germ line) and the rest of the body (somatic cells). The German biologist Albert Weismann was the first to propose in 1893 that germ cells are the only cells that can pass genetic information between generations, and no hereditary information can pass from somatic cells to the germ line and on to the next generation (Figure 1.2) [10]. This postulate is known as the Weismann barrier and is widely accepted to apply to all vertebrates, although recent studies show that animal germ cells can in some circumstances be heritably modified by signals from somatic cells or the environment (e.g. in nematodes [11, 12, 13]).

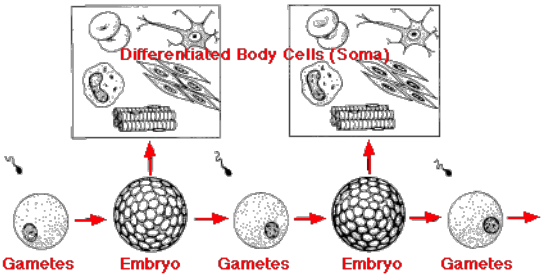


Fig. 1.2 **Separation of germline from soma.** Gametes (eggs or sperm) pass hereditary information between generations. Figure from [6].

Biological information is passed down the generations as the DNA of germ line cell lineage. Therefore, changes (mutations) in germ line DNA sequence are the ultimate source of the heritable component of phenotypic variation. The majority of germ line mutations occur as errors in DNA replication during cell division, arise at random, and can affect any part of the DNA sequence. Typically, a uniform mutation rate is considered to apply along the DNA sequence [14], although recent studies have shown that there is some minor variation in the mutation rate along the genome [15]. The genome-wide variation in germ line mutation rate is related to base composition (e.g. methylated sites with cytosine located next to guanine - CpG sites - in vertebrates have a least 10 times higher than average mutation rate), to timing of DNA replication (regions of DNA replicating late in the cell cycle tend to have a higher mutation rate), and to transcription [15].

1.1.2 Vertebrate genome architecture

Known vertebrate genome sizes vary over almost three orders of magnitude: between ~320 million base-pairs (Mb) in Green pufferfish *Tetraodon fluviatilis* and 120,000Mb in Marbled lungfish *Protopterus aethiopicus* [16]. The size of the human genome is ~3,000Mb. There appears to be with little correspondence between genome size and the external characteristics of an organism. Similarly, there is little correspondence between the number of protein-coding genes and the size, cognitive capabilities, or the number of distinct tissue types in an organism throughout eukaryotes: the number of genes in the nematode worm *Caenorhabditis elegans*, pufferfish, cichlids, and human is comparable (20,447 *C. elegans*, 18,523 in *Takifugu rubripes*, 21,437 in the Nile tilapia cichlid, and 20,300 in human; according to Ensembl annotation [17]).

Until the late 1990s, vertebrate genomes were thought to contain mainly protein coding genes but this view has changed dramatically over the last 15 years [19]. The Human Genome Project and subsequent DNA sequencing of dozens more vertebrate species revealed that protein coding sequences comprise only a small fraction of the genome, while a large proportion is taken up by mostly defunct transposable elements (~45% of the human genome; Figure 1.3). Initially viewed purely as 'junk DNA', transposon derived sequences have been shown to play a constructive role in evolution, for example the evolution of functional non-coding sequences in mammals [20]. It is now also clear that vertebrate genomes contain tens of thousands of non-coding genes. This includes small RNA genes: microRNAs, post-transcriptional regulators of gene expression [21, 22], and piRNAs, involved in repressing transposable element activity and regulating gene expression specifically in germ-line cells [23]. Thousands of long intergenic non-coding RNA genes (lincRNAs) were also identified, with diverse roles in multiple biological processes [19]. In fact the majority of the non-repetitive genome is transcribed at least in some cell types and/or has been assigned a biochemical function (such as protein binding), leading to claims by the ENCODE consortium that 80% of the genome is functional [24]. However, it is necessary to distinguish between 'selected

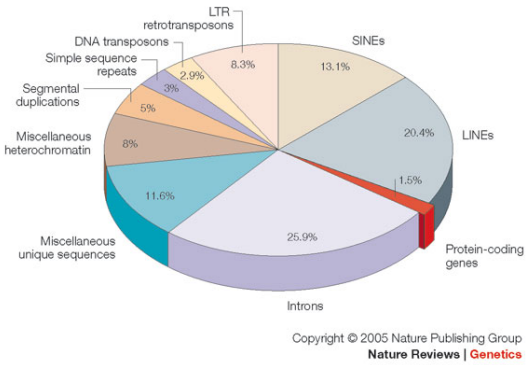


Fig. 1.3 **Components of vertebrate genomes.** Data for the human genome, from [18].

effect' function of a sequence (i.e. the effect for which it was selected) and 'causal function' (i.e. any function it performs)[2] [25]. Selected effect function can be inferred from evolutionary conservation: comparisons of multiple vertebrate genomes revealed that approximately 6% (3-8% depending on the estimation procedure) of the sequence is biologically functional in this sense [26], and all comparative studies reviewed by Ponting and Hardison in 2011 put the proportion of functional bases in the human genome at below 15%. In summary, the genome encodes a wide variety of genic and regulatory elements, most of vertebrate functional sequence is non-coding, and at least 85% of the genome sequence appears to be evolving without perceptible evolutionary constraint.

1.1.3 Molecular evolution

The view that most DNA nucleotide substitutions may be selectively neutral or nearly neutral became increasingly influential during the 1960s. Following the discovery that amino acid substitutions in protein sequences accumulate at a constant rate, Pauling and Zuckerkandl coined the term 'molecular clock' and proposed using DNA and protein sequences to infer phylogenetic relationships between taxa [27]. Deciphering the genetic code [28] revealed that the DNA sequence of a protein-coding gene can change in a way that does not affect its amino acid sequence (synonymous DNA substitutions), and Motoo Kimura suggested that "as functional constraint diminishes, the rate of evolution converges to that of synonymous substitutions" [29]. The 'neutral theory' of molecular evolution was developed and provided a set of baseline expectations for DNA variation and change in the absence of natural selection.

For an *ideal population* satisfying assumptions of the Wright-Fisher model (including, importantly, that "all parents have an equal expectation of being the parents of any progeny" [30, p.205], and that the population size does not vary over time), the neutral substitution rate can be derived as follows: let μ be the average mutation rate per base-pair (bp) per generation and let there be N_e breeding diploid individuals. The population generates $2N_e\mu$ mutations per bp per generation. The frequency of a new neutral allele arising due to mutation is initially $\frac{1}{2N_e}$. The allele frequency can change between generations due to random sampling ('genetic drift') and the probability of

[2]For example, the selected function of a heart is to pump blood, whereas causal functions may include "adding 300g to body weight, producing sounds, and preventing the pericardium from deflating onto itself" [25]

fixation of a neutral allele is equal to the allele frequency [30]. Therefore the:

$$\text{neutral substitution rate} = 2N_e\mu \times \frac{1}{2N_e} = \mu, \tag{1.1}$$

so independent of demography, and the mean time to fixation of a new mutation is $4N_e$ [30].

The *effective population size* N_e is an important factor influencing the strength of genetic drift, and also its effect on selection. The main forms of natural selection on the molecular level are listed in Table 1.1. For an allele under directional selection, the relative influence of selection and genetic drift is determined by N_e and by s the *selection coefficient*. If $N_e s \ll 1$ then the fate of the allele will be determined primarily by genetic drift. On the other hand, if $N_e s \gg 1$, it will be determined primarily by selection. For example, an allele with a selective advantage $s = 0.001$ and current frequency of 0.1 has 86.5% probability to reach fixation in a population where $N_e = 10,000$ ($N_e s = 10$), but only 10.9% probability to reach fixation in a population where $N_e = 100$ ($N_e s = 0.1$). It is clear that alleles under weak selection behave as nearly neutral in small populations, and demographic changes that affect N_e have a profound effect on genetic variation in a population.

Table 1.1 **Main forms of natural selection.** Based on [3].

Category	Selection	Description
Directional	Positive selection	Exerts force to increase the frequency of alleles that confer selective advantage
	Purifying selection	Acts to reduce the frequency of disadvantageous alleles, thus preserving functional sequences from being degraded by new mutations
Non-directional	Balancing selection	Selection acts to maintain two or more alleles at one locus, for example because heterozygous individuals have a higher fitness than homozygous ones (heterozygote advantage) or because the fitness of an allele depends on its frequency in the population

Figure 1.4 summarises the forces that influence genetic composition of natural populations. New mutations and inward migration create new *segregating sites* (polymorphic loci with two or more alleles), while directional selection and genetic drift remove variation. The balance between these forces is influenced by demographic events which change the effective population size N_e.

It is important to note that alleles on the same chromosome are not inherited independently but are physically linked to each other. The combination of alleles located together on the same physical chromosome is called a *haplotype*[3]. Linkage is

[3]The word *haplotype* may refer to all alleles on a chromosome or just alleles that are physically close in a particular region: e.g. a 5Mb haplotype on chromosome 1.

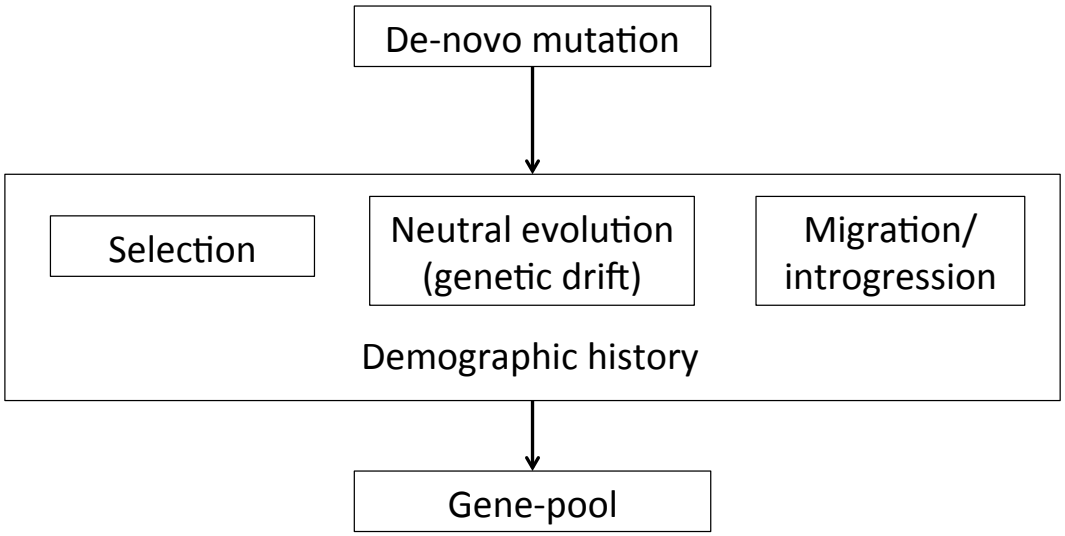


Fig. 1.4 **Forces influencing genetic composition of natural populations.**

eroded by recombination during meiosis and the probability of recombination between an allele and one of the other linked alleles on the chromosome is a monotonic function of the physical distance between them (i.e. the further the other allele, the greater the probability of recombination). Therefore, it is possible to use recombinant frequency in offspring between alleles heterozygous in the parents to detect linkage (i.e. determine if two alleles are on the same chromosome) and to generate a *genetic map* - a map describing the order of alleles along a chromosome and the distances between them in centimorgans (cM), units corresponding to 0.01 probability of recombination [3].

Consider two biallelic loci: {A,a} and {B,b}. If the relationship between the A and B alleles was random, we would expect the frequency of the haplotype AB (f_{AB}) to be the product of the frequencies of the A and B alleles (f_A and f_B):

$$f_{AB} = f_A \times f_B \tag{1.2}$$

and the two loci would be said to be at *linkage equilibrium*. Against this baseline, non-random association between alleles, referred to as *linkage disequilibrium* (LD), can be assessed. LD provides a measure of the effect of linkage on the genetic variation in a natural population. Two common measures of LD are D and r^2 [31]:

$$D = f_{AB} - f_A f_B \tag{1.3}$$

$$r^2 = \frac{D^2}{f_A(1-f_A)f_B(1-f_B)} \tag{1.4}$$

Linkage gives rise to the *hitchhiking effect* whereby an allele can get a lift in frequency from positive selection acting on a nearby allele on the same chromosome [32], and also to *background selection* whereby purifying selection against deleterious alleles reduces genetic variability at linked neutral sites [33]. Marked reduction of genetic variation in regions of low recombination has been observed in many organisms, including human, but the relative contributions of hitchhiking and background selection to this phenomenon are hard to ascertain [34]. However, in regions of normal recombination, the hallmark of fixation of a new strongly beneficial allele is a marked reduction in genetic diversity due to hitchhiking - a pattern referred to as *selective sweep*. Such a pattern cannot be caused by background selection in regions of normal recombination and therefore is considered a signature of recent positive selection [34].

1.1.4 Speciation

Evolution generates genetically and phenotypically distinct groups of organisms [35]. In sexually reproducing organisms, the discontinuities between the groups arise from reproductive isolation. Biological species can be considered to be units of evolution [36] and arise when reproductive isolation is complete, although limited gene exchange that does not affect the essential integrity of the species is possible [35, 36][4]. The process of speciation, i.e. the origin of species, is then a process of building up of reproductive barriers, and increasing genetic and phenotypic differentiation (Figure 1.5).

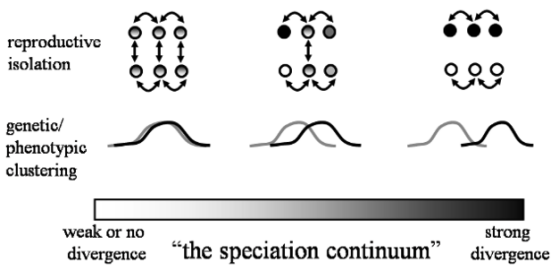


Fig. 1.5 **The speciation continuum.** Figure from [37].

Divergence during speciation is continuous. The strength of reproductive isolation and the degree of genetic and phenotypic clustering have been shown to vary quantitatively [38]. Charles Darwin wrote: *"I look at individual differences...as of high importance to us as being the first step towards such slight varieties as are barely thought worth recording... And I look at varieties which are in any degree more permanent, as steps leading to more strongly marked and more permanent varieties; and these latter as leading to sub-species, and to species."* [39, p.42].

[4] I adopt the widely accepted 'biological species concept'. For alternative species definitions and a discussion of different species concepts see [35].

Speciation in animals typically progresses too slowly for humans to be able to directly observe it happening. We are able to measure the isolating barriers between current pairs of species and study their genetic and ecological causes. However, isolating barriers continue to evolve and accumulate after speciation is complete and the causes that initially led to speciation may be obscured by evolution post speciation. Therefore, speciation studies often focus on populations that are only partially reproductively isolated, obtaining 'snapshot' views of different stages along the speciation continuum. Not all partially isolated populations will go on to eventually form species, but investigation of many different taxa at many different stages all along the speciation continuum, from newly forming varieties to pairs of well defined species, will allow us to draw general conclusions about the process of speciation [35, 38].

It is useful to understand that the word *species* does not describe any intrinsic characteristics: a species can only be defined relative to other species [35]. Coyne and Orr note two common features of speciation: *epistasis*, whereby "genes that evolve in one group produce reproductive isolation by interacting with genes evolving in another group" and *pleiotropy*, whereby "characters that evolve within a group have the side effect of producing isolating barriers" [35, p.56].

In the literature, there has been a lot of focus on the geographical context of speciation [35]. A range of definitions, based either on biogeography or on population genetics, have been proposed for different geographical modes of speciation [40]. For example, speciation may be deemed *sympatric* when isolating mechanisms develop in the absence of geographic barriers to free interbreeding (i.e. individuals are within the average dispersal distance of one another [41]); *allopatric* when geographic separation causes initial split between populations, gene exchange ceases, and reproductive barriers evolve in isolation[5]; and *parapatric* when gene exchange at the onset of speciation is partially limited by geography.

1.1.5 Genome sequencing and bioinformatics

Technological advances in DNA sequencing [42, 43] since the introduction of the Sanger method in 1977 [44, 45] have led to there now being over 10^{15}bp of sequence in the ENA data depository [46], and the cost of sequencing is continuing to fall exponentially [47]. Nevertheless, significant challenges remain.

All of the currently available sequencing methods fall far short of being able to obtain the end-to-end sequence of a whole vertebrate chromosome. Instead, sequencers

[5]Reproductive barriers evolved in allopatry can only be observed in cases where the species come into secondary contact, or in laboratory experiments

output large numbers of short reads that contain errors and need to be analysed computationally in order to provide useful information. Illumina, currently the most cost effective sequencing platform [47], produces ~100bp reads. To obtain a new genome sequence for a species, the reads need to be joined up through overlaps into a continuous sequence. Such *de novo* genome assembly is a task analogous building a jigsaw puzzle consisting of hundreds of millions or even billions of pieces, and is one of "the most complex computations in all of biology" [48]. To obtain sufficient overlaps between reads and also to be able to recognise errors in individual reads, every site in the genome needs to be sequenced multiple times (typical average *genome coverage* required for *de novo* assembly of 100bp reads is 40-100×). On the other hand, when an assembled *reference genome* is already available, it is possible to align to this reference reads from other individuals of the same species, or of a closely related species, and to infer genetic variation in the form of differences between the mapped reads and the reference [49]. The alignment approach is substantially cheaper and faster because it typically requires lower genome coverage (~5-20×) and is much less computationally intensive than *de novo* assembly [43].

1.2 Evolutionary and speciation genomics

The search for the molecular basis of evolutionary processes leading to adaptation and organismal diversity predates DNA sequencing. In one of the earliest and best known studies, V. M. Ingram, working at the Cavendish Laboratory of the University of Cambridge, discovered in 1957 the single amino acid difference between normal and sickle cell haemoglobin [50]. The sickle cell allele confers a large survival advantage to heterozygous individuals in regions with a high incidence of malaria, but is highly deleterious and often lethal when homozygous [3]. Therefore, the allele is under balancing selection in regions with high incidence of malaria, but is under purifying selection in other parts of the world. Other pre-genomic studies focussed on differences between the rates of synonymous and non-synonymous nucleotide substitutions in protein-coding genes. McDonald and Kreitman in 1991 compared within-species to between-species polymorphisms in the alcohol dehydrogenase *Adh* gene of three *Drosophila* fruit fly species. They found an excess of non-synonymous changes in between-species comparisons - evidence for repeated fixation of advantageous alleles and long-term adaptive evolution in this gene [51]. This is consistent with the role of *Adh* in alcohol tolerance and utilisation and adaptation of *Drosophila* to new feeding niches involving fermenting fruit [52].

While the above and other similar studies provided fascinating insights, they were limited to single genes. The availability of data from many genes and, later, whole genomes enabled scientists to draw more general conclusions about adaptive evolution. For example, a comparison of 43 genes between *Drosophila yakuba* and *D. simulans* suggested that "45% of all amino-acid substitutions have been fixed by natural selection, and that on average one adaptive substitution occurs every 45 years in these species" [53]. The comparison of the human and mouse genomes revealed that ~80% of mouse genes have clear 1:1 *orthologs* in human, originating from common ancestral sequence [54], but for example the *V1r* family of pheromone receptors has ~160 functional genes in mice and only 5 in human [3], consistent with the key role of pheromones in social and reproductive behaviour in mice [55]. Finally, comparisons of human and chimpanzee genomes provided a virtually complete catalog of genetic differences between our species and our closest relatives, revealing ~35 million single nucleotide substitutions, ~5 million small insertions and deletions, differential signatures of transposable element activity, and a number of larger chromosomal rearrangements [56]. In 2006, Katherine Pollard and her colleagues compared 17 available reference genomes and found 49 'human accelerated regions' (HARs) with significantly accelerated rate of substitution in the human lineage, but strong sequence conservation across reptiles, birds, and other mammals [57]. Many HARs are located near genes involved in neurodevelopment, and the top HAR contains a long non-coding RNA expressed in the developing neocortex [57]. 96% of HARs are not in protein-coding genes, and broader comparisons of vertebrate genomes also suggest that functional non-coding sequences are more abundant and tend to evolve faster than protein coding sequences [19]. This evidence points to changes in gene regulatory regions, rather than differences in genes themselves, being primary genetic drivers of adaptive differences between closely related species. Overall, all the above examples illustrate the power of comparative genomics to shed light on the molecular basis of organismal diversity.

Population genetic data (i.e. patterns of genetic variation within populations) are informative about more recent selection. A variety of methods have been developed to infer positive selection by comparing observed population genetics data with expectations under neutrality [58]. As noted in section 1.1.3, a hallmark of a recent selective sweep is a marked reduction in genetic diversity in the genomic region surrounding the beneficial allele (Figure 1.6a). The process is accompanied by changes in the population frequencies of segregating alleles in the region (Figure 1.6b). As the beneficial allele rises in frequency, derived (i.e. arising via recent mutations) alleles on the same haplotype also rise to high frequencies and the excess of high-frequency

derived alleles can be detected using Fay & Wu's H statistic [59]. Then, after fixation of the beneficial allele, diversity starts returning in the form of new mutations, all of which are initially at low frequencies and the surplus of rare alleles is detected using the Tajima's D statistic [60].

As the haplotype with a beneficial allele rises in frequency it creates a local distortion in the patterns of LD. A high-frequency long haplotype may be a sign of its rapid rise in prevalence, as recombination has not had enough time to break the LD and shorten the haplotype. The extended haplotype homozygosity (EHH) statistic has been designed to capture such pattern of locally elevated LD (Figure 1.6c).

Finally, if positive selection acts on a locus in one population but not in another related population, it may result in differences in allele frequencies between the two populations (Figure 1.6c). Measures of population divergence, such as F_{ST}, d_f, and d_{XY} will be discussed further in section 1.2.1.

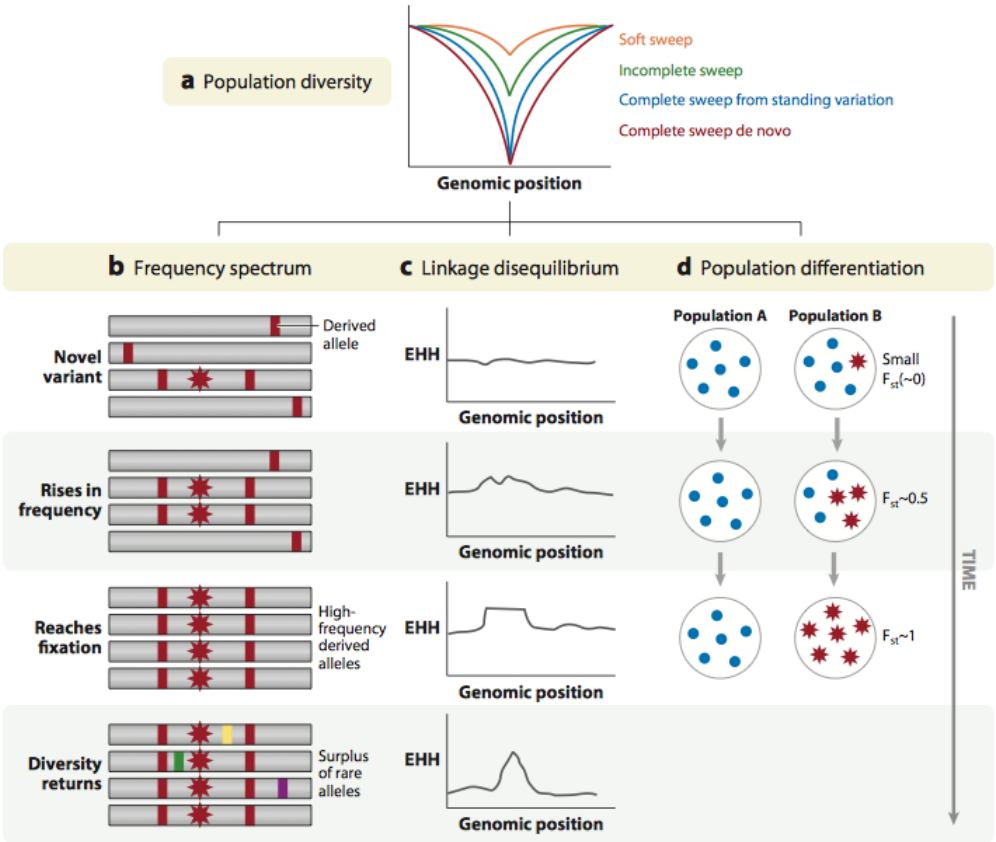


Fig. 1.6 **Signatures of selective sweeps in population genetics data. (a)** A local reduction in sequence diversity is the hallmark of all types of selective sweeps. **(b-d)** Changes in allele frequencies, local patterns of LD, and divergence between populations under different selective pressures can be used to detect positive selection in population genetic data. Figure from [58].

Population genetics is typically thought of as being focussed only on variation within a group of individuals of the same species [3]. This view is based on the assumption that genealogies of sequences sampled from a particular locus in multiple individuals *coalesce* (have a common ancestor) within the species and within-species variation has little impact on inferences concerning between-species divergence. This model (Figure 1.7A) relies on "all branches in the species tree being very long compared to within-species coalescence times" [61, p.871], and on complete isolation between species. When assumptions of hard split with long intraspecific branches are severely violated, as in Figure 1.7B, by *incomplete lineage sorting*, i.e. lineages not coalescing within the duration of their species, and/or by *introgression*, i.e. gene flow between species, the population genetics framework and methods are also appropriate to use for multi-species datasets.

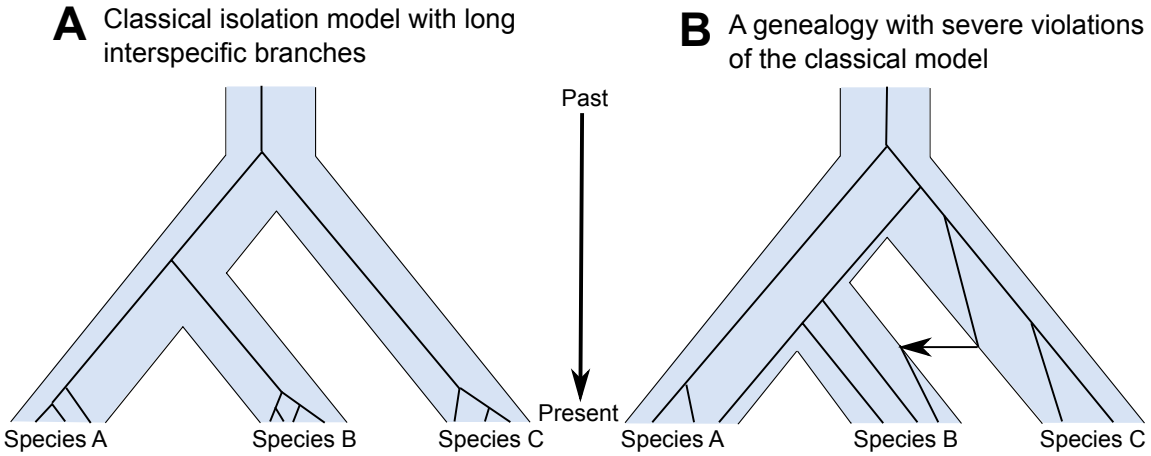


Fig. 1.7 **Sequence genealogies and speciation. (A)** A classical model of speciation with hard split between species, full isolation, and within-species coalescence times much shorter than between-species branches. **(B)** An example of sequence genealogy severely violating assumptions of the classical model due to pervasive incomplete lineage sorting and gene flow from species C to species B.

One of the central goals in the emergent field of speciation genomics is describing genome-wide patterns of divergence at different stages along the speciation continuum and understanding: 1) how are these patterns affected by the geographic mode of speciation (allopatric $\leftrightarrow$ sympatric); 2) how much of the divergence at different stages is driven by selection and how much by neutral processes; and 3) the degree to which genomic signatures are common between different organisms (e.g. are the patterns of divergence along the speciation continuum in wild mice similar to the patterns in cichlid fishes?) [62].

Theory predicts a fundamental difference in genomic signatures of allopatric divergence, where reproductive isolation can in principle be simply a result of a prolonged period of genetic drift, and signatures of sympatric or parapatric speciation, where geography permits at least some continuing gene-flow. For speciation to occur in the sympatric or parapatric case, the homogenising effect of gene-flow needs to be counteracted by divergent selection. Moreover, the accumulation of loci contributing to reproductive isolation in the face of gene-flow is counteracted by recombination. Linkage between genomic loci underlying different isolating barriers mitigates the effects of recombination, and so is thought to be conductive to speciation in these scenarios. This has been demonstrated for example in flycatchers [63] and in stickleback [64]. Felsenstein showed that linkage between loci underlying prezygotic (e.g. mate choice) isolation barriers with loci underlying postzygotic (e.g. low survival of hybrids) barriers can be especially important in facilitating speciation with gene-flow [65].

1.2.1 Measures of sequence divergence

When sampling the genomes of multiple individuals from each population or species, there are multiple ways to measure their level of divergence. The currently most commonly used measures have been reviewed by Cruickshank and Hahn [66]. It is useful to distinguish relative measures (e.g. F_{ST}, d_f) and absolute measures (d_{XY}). A brief description of these statistics is provided in Table 1.2. The distinguishing feature of relative measures of differentiation is that they are influenced by within-population levels of variation. For example, a decrease in variation in population `A` due to a selective sweep would cause a rise in F_{ST} between populations `A` and `B`, but the absolute divergence between `A` and `B` would remain constant.

Table 1.2 **Commonly used measures of population divergence.** Adapted from [66].

	Measure	Description
Relative measures	F_{ST}	Normalised measure of allele frequency differences between populations
	d_f	Number of fixed differences between populations or species
Absolute measure	d_{XY}	Average number of pairwise differences between sequences from two populations, excluding all comparisons between sequences within populations.

1.2.2 'Islands of speciation' or 'incidental islands'?

A pattern with well defined genomic regions of markedly elevated F_{ST} divergence between incipient species has been observed in a large number of studies of speciation

with gene-flow [e.g. 67, 68, 69, 70]. In these studies, which include cases where gene-flow occurred due to divergence in sympatry or due to secondary contact after a period of allopatry, 'islands' of high differentiation have been interpreted as loci resistant to gene-flow. These empirical observations have been accompanied by theoretical models of speciation with gene-flow, with divergent selection generating genomic 'islands of speciation', while the rest of the genome is homogenised by gene-flow [e.g. 71, 72]. Under these models, highly diverged regions (HDRs) have lower effective migration rates and harbour alleles underlying isolating traits between the incipient species - i.e. reproductive isolating genes generating 'islands of speciation'.

Others have offered alternative explanations [66, 73, 74], highlighting that 'islands' of high relative divergence, as measured for example by F_{ST} and d_f (described in Section 1.2.1) can also be caused by other forces in the absence of gene-flow. One alternative model that has received much attention suggests that post-split selection (e.g. a selective sweep) generates regions of reduced genetic diversity in one or both of the newly formed species, thus increasing F_{ST} in the vicinity of the selected locus [66, 73, 75]. Under this model, 'islands' of high relative divergence may be involved in post-split adaptation, or specialisation, or may reflect background selection present in any organism [66]. They do not directly cause speciation, and, therefore, have been referred to as 'incidental islands' [74].

The two contrasting hypotheses with regards to the origin of 'islands' of high differentiation are summarised in Table 1.3. A crucial prediction of the 'islands of speciation' model is that that both relative and absolute measures of divergence should be high in regions resistant to gene-flow, whereas the 'incidental islands' model predicts that absolute divergence will not be affected [66, 73]. Mindful of this prediction, Cruickshank and Hahn [66] last year re-analysed published sequence data for eight recently diverged taxa with putative 'islands of speciation' and showed that, in all cases, absolute divergence (d_{XY}) in the islands was lower than outside - not consistent with the 'islands of speciation' hypothesis. It is therefore likely that many previously reported genomic islands do not represent the first steps in the formation of new species. The question is not settled, however, with prominent evolutionary biologists defending the role of the divergent islands in speciation [76], citing for example a recent study of genomic divergence in the face of gene-flow between the European hooded and carrion crows. This study revealed a single prominent island of high F_{ST} and d_f that exhibits low absolute divergence but nevertheless is very likely involved in the maintenance of reproductive isolation [77].

Table 1.3 **Interpreting genomic 'islands' of high differentiation.** Two alternative hypotheses are introduced, together with their predictions for levels of relative (F_{ST}, d_f) and absolute (d_{XY}) differentiation in the islands. Based on [66, 74].

Model	Cause	Interpretation	Predictions
Islands of speciation	Lower effective migration	Loci underlying reproductive isolating barriers/traits thereby causing speciation	High F_{ST}, d_f High d_{XY}
Incidental islands	Lower effective population size	Loci involved in post-split adaptation or simply background selection; alleles within islands do not cause speciation	High F_{ST}, d_f Low d_{XY}

To better understand why a selective sweep in a newly formed species does not affect d_{XY}, consider the sequence genealogies shown in Figure 1.8. Each genealogy connects haplotypes from the same individuals belonging from two species (haplotypes H1, H2 from species 1 and haplotypes H3, H4 from species 2); the (A) panel shows the genealogy at a neutral locus, whereas the (B) panel shows a locus under recent (post-speciation) positive selection in species 1, reflected in the more recent common ancestor of haplotypes H1 and H2. Despite the differences between the genealogies, the average distance (H1,H2 $\longleftrightarrow$ H3,H4) is exactly the same between the two panels. Thus, d_{XY} remains unchanged.

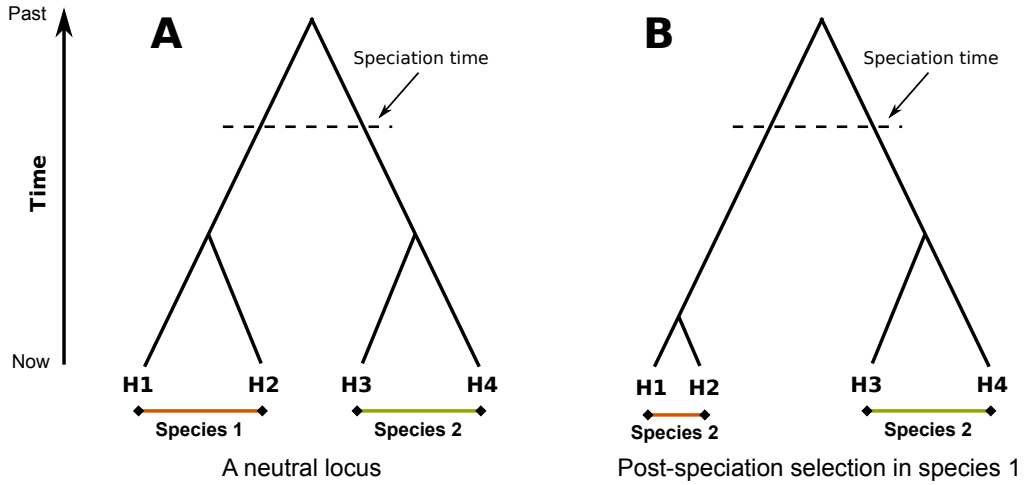


Fig. 1.8 **Effect of different sequence genealogies on F_{ST} and d_XY.** The figure shows the genealogies of four haplotypes from two species (two haplotypes from species 1 and two haplotypes from species 2) at **(A)** a neutral locus and **(B)** a locus under recent positive selection (selective sweep) in species 1. Figure adapted from [66].

1.3 East African cichlid radiation

1.3.1 Cichlid fish

Cichlids (*Cichlidae*) are one of 448 families of modern bony fish (*Teleostei*) and belong to the *Labroidei* suborder [78]. The geographical distribution of cichlids and phylogenetic relationships suggest that the cichlid family originated on the Gondwana supercontinent about 150 million years ago [79]. Subsequent fractioning of Gondwana means that, today, cichlids are found in Southern India, on Madagascar, across Sub-Saharan Africa and along the Nile, and in the tropics of the Americas. Although cichlid morphology is incredibly diverse, all have a common body plan in terms of the position of the fins and the arrangement of parts of the jaws (Figure 1.9).

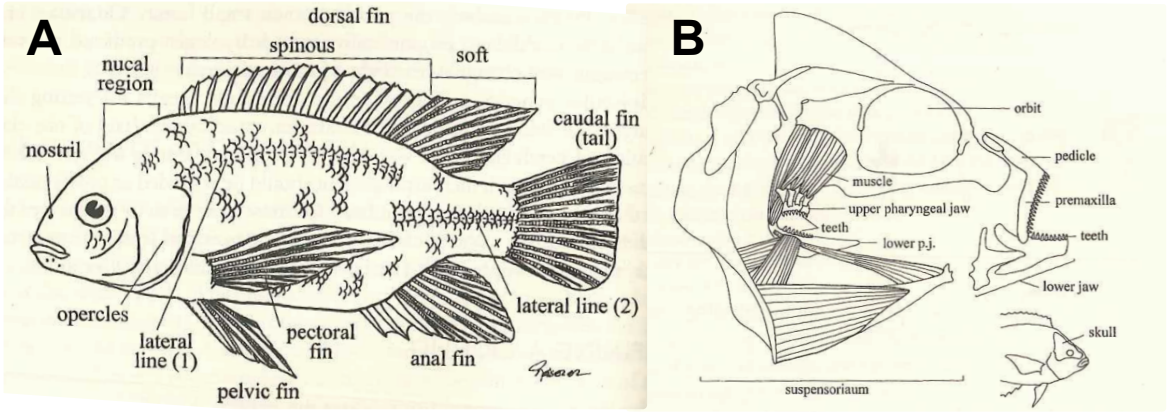


Fig. 1.9 **Major physical features of cichlids. (A)** The basic body plan showing the position of fins and two other distinctive features: a single nostril and an interrupted lateral line. **(B)** A cross section through the head of a cichlid, showing the unique arrangement of pharyngeal jaws. Both figures from [78].

Characteristic features of cichlids include having only one nostril on each side of the snout (fish typically have two), having a lateral line that is interrupted about two thirds of the way to the tail (fish typically have a single uninterrupted lateral line on each side), and, crucially, a unique arrangement of pharyngeal jaws.In most fishes, the upper pharyngeal jaw floats freely and the lower pharyngeal bones are split and limited to simple rocking action. Therefore, their utility for chewing food is limited. On the other hand, the upper pharyngeal jaw in cichlids has developed a protruding connection with the base of the skull and no longer floats freely. The lower pharyngeal bones in cichlids are fused into a single jaw (in common with other fishes of the suborder *Labrodei*) and developed bony outgrowths for attachment of more muscles, enabling it to produce a much greater force. Having two sets of jaws capable of chewing is

believed to facilitate rapid evolution of new feeding specialisations and to contribute to the great evolutionary success of cichlids and other Labrodei [78, 80].

1.3.2 East African cichlids

Cichlids are the most species rich and diverse family of vertebrates [81]. They have undergone repeated adaptive radiations in more than 30 African lakes, although 120 more lakes across the continent have been colonised by cichlids without speciating [82]. Lake depth and sexual dichromatism (colour differences between males and females) are significantly associated with African cichlid radiations [82]. There are approximately 2,000 distinct cichlid species in East Africa, with great diversity in habitat, behaviour, feeding apparatus, craniofacial morphology, pigmentation, and body shapes [83]. The diversity of tooth and jaw morphologies and their association with habitats and ecological niches is evidence for the importance of these traits in the East African radiations [83] (Figure 1.10).

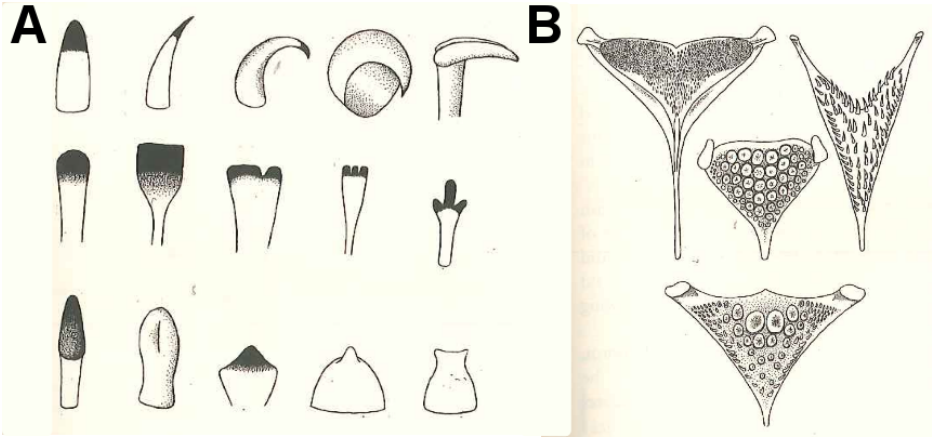


Fig. 1.10 **Evolution of cichlid feeding apparatus in East Africa. (A)** Outer jaw teeth of several species of African cichlids. Teeth with cusps and a flat top are common in algae scrapers, while pointy conical teeth are common in predators. **(B)** Lower pharyngeal jaws from an algae feeder (top left), piscivore (top right), a specialised molluscivore (middle) and a more generalist molluscivore (bottom). Figures from [78].

Nowhere have cichlids radiated faster and more spectacularly than in the Great Lakes of East Africa: Lakes Malawi, Tanganyika, and Victoria. The lakes contain hundreds of species, virtually all of which are endemic to each. The ~200 species rich Lake Tanganyika radiation is believed to have been seeded by eight independent cichlid lineages and, being 9-12 million years old, it hosts the oldest and most diverse assemblage of cichlid fishes [84]. Both the smallest and the largest known cichlid species are endemic to Tanganyika: *Neolamprologus brevis* at 3 cm and *Boulengerochromis*

microlepis at 80 cm. Tanganyikan cichlids are grouped into 12 tribes, one of which, the tribe *Haplochromini*, contains two endemic species and several lake-river generalists found throughout the Tanganyikan catchment area.

The cichlid radiations in Lakes Malawi and Victoria have been seeded by Haplochromini riverine lineages that shared a common ancestor with the Tanganyikan haplochromines approximately two million years ago. It is thought that ancestral haplochromines evolved in Lake Tanganyika, colonised river systems of East Africa and through the rivers also entered the newly emerging Lakes Malawi and Victoria [85] (Figure 1.11). Therefore, while there are only a handful of Haplochromini cichlids in Lake Tanganyika, almost all of the over 1,000 species of Lakes Malawi and Victoria are haplochromines [85].

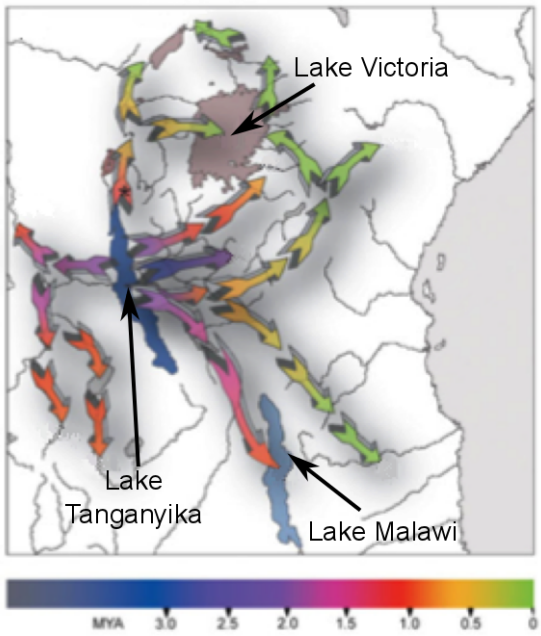


Fig. 1.11 **The 'out of Tanganyika' model of East African cichlid radiations.** The ancestors of today's haplochromines used in Lake Tanganyika and used existing and ancient river systems to spread across East Africa, also seeding the spectacular radiations of Lake Malawi and Lake Victoria. Figure adapted from [85].

The ~1-2 million years old cichlid flock of Lake Malawi has most likely been seeded by fish similar and ancestral to the lake-river generalist *Astatotilapia calliptera*, possibly in multiple repeated colonisation events over hundreds of thousands of years [86]. Most Lake Malawi cichlids can be assigned to two major lineages (each with ~250 species): the rock-dwelling 'mbuna', and the 'sand-dwellers'. There are also two genera of open-water predators, each with ~10-20 species: *Diplotaxodon* and *Rhampochromis* [87].

Lake Victoria, the youngest of the three Great Lakes, dried up some 15,000 years ago. This event created a bottleneck in the cichlid population of the Lake Victoria basin. One study suggests that a 30-50 fold decline in cichlid populations ensued, with surviving cichlids occupying refugia in rivers and other lakes in the area [88]. These haplochromine cichlids began to re-enter the re-emerging Lake Victoria from multiple directions (Figure 1.11) some 12,000 years ago and formed a genetically diverse founding population. Therefore, even though the Lake Victoria species flock is only

thousands of years old, the most common ancestor of these fish can only be traced back to the time Haplochromini left Lake Tanganyika [88]. The ancestral genetic variants present in the fish re-entering the lake 12,000 years ago have since been 'mixing-up' through hybridisation and new species are emerging at an astonishing rate, adapting to conditions in different parts of the lake (e.g. ref [89]).

Frequent occurrence of independent evolution of similar phenotypes in different lakes and lineages suggests an important role of natural selection in generating these radiations [90, 91]. Examples of ecologically driven adaptations include pelagic zooplanktivores with a large number of gill rakers, rock-dwelling algae scrapers with many rows of fine brush-like teeth, snail crushers with hypertrophied pharyngeal jaws, several groups of dark-adapted species with greatly enlarged lateral-line sensory apparatus, large pelagic piscivores with a single row of unicuspid teeth, and insect-eaters with 'fat' lips used to seal crevices in the rocks and prevent prey from escaping. These populations of cichlid 'variants' present an unparalleled opportunity to study the genetic basis of how complex traits evolve in vertebrates. This includes traits occurring in the later stages of the life cycle, not included in many traditional induced mutagenesis screens.

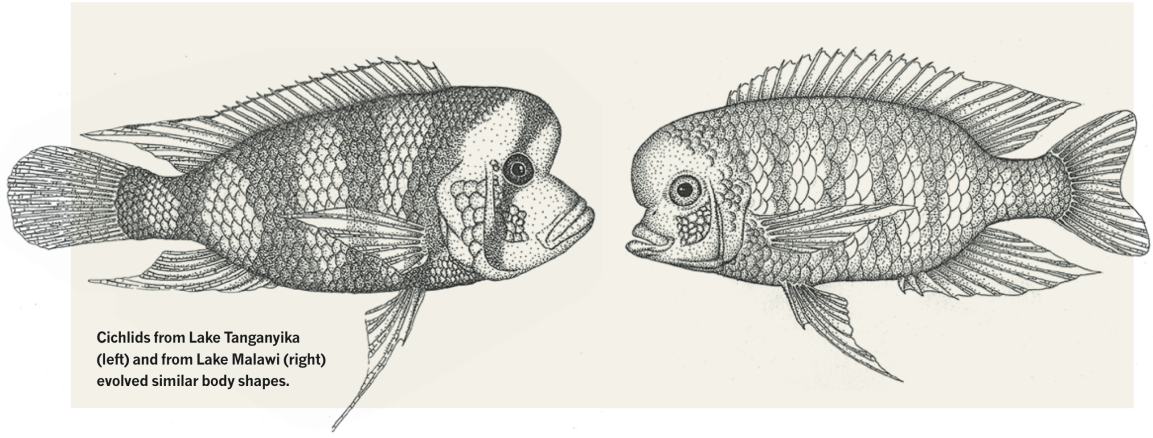


Fig. 1.12 **Parallel evolution of body shapes in Lakes Malawi and Tanganyika.** Figure from [92].

1.4 Overview of the remainder of this thesis

This thesis describes some of the first steps in the application of whole genome sequence data to study the exceptional diversity of East African cichlids. The rest of the thesis is organised in six chapters:

- Chapter 2 introduces reference genomes and annotations generated by the Cichlid Genome Consortium, and describes my contribution: microRNA gene annotation. This work has been published in Brawand, Wagner, Li, Malinsky *et al.*, 2014 [93]. In addition, the chapter contains unpublished work on evolution of target genes regulated by microRNAs.
- Chapter 3 provides an overview of the new whole genome data generated during my PhD and describes the bioinformatics groundwork I have done to facilitate the use of this data for scientific research on East African cichlid evolution.
- Chapter 4 takes a short detour into computational biology and describes a new method for reducing problems caused by heterozygosity during genome assembly called `trio-sga`. The method is applied to generate improved *de novo* genome assemblies from cichlid data, and its benefits are demonstrated more dramatically using the highly heterozygous *Heliconius* butterfly data. A manuscript on the `trio-sga` method is in preparation.
- Chapter 5 contains the initial analysis of Lake Malawi cichlid samples, focussing on characterisation of genomic diversity, relationships between species, and the extent of introgressive hybridisation. This will be extended by adding further data from Summer 2015 and additional analysis to form the basis of a future publication.
- Chapter 6 presents a detailed characterisation of early-stage adaptive divergence of two cichlid fish ecomorphs in Lake Massoko, a small (700m diameter) isolated crater lake in Tanzania. The work described in this chapter has been submitted as a 'Report' to the journal *Science* in July 2015. The manuscript received favourable reviews and has been upgraded to a full 'Research Article'. The revised article has been submitted as Malinsky, Challis, Tyers *et al.*, 'Genomic Islands of Speciation Separate Cichlid Ecomorphs in an East African Crater Lake' on 11th September 2015.
- In chapter 7, I first summarise the main contributions and findings of the research presented in this thesis of and then discuss ongoing work and future directions.

Chapter 2

The cichlid genome project

2.1 Introduction

2.1.1 Publication note

The work described in section 2.1.2 and sections 2.3.1 to 2.3.3 was previously published in Brawand, Wagner, Li, Malinsky *et al.*, 2014 [93]. This publication was the result of over five years of work by the 'Cichlid Genome Consortium' and its 76 members. As always, results obtained by others are indicated in the text; everything else is my own work.

2.1.2 Five reference genomes

Recognising the potential of the East African cichlid radiations to generate many important insights into the genetic basis of speciation and functional diversification, a Cichlid Genome Consortium led by the Broad Institute generated draft reference genome assemblies for five species. The genomes were selected to provide one reference for each of five major lineages of East African cichlids, focussing on Lakes Tanganyika, Malawi, and Victoria (Figure 2.1) The reference genomes, listed in Table 2.1, are used throughout the rest of this thesis. *Oreochromis niloticus*, commonly known as Nile tilapia, is a riverine cichlid that shared a common ancestor with the highly radiating lake cichlids approximately 25-50 million years ago and so provides an outgroup for the study of their evolution. *Neolamprologus brichardi*, a reef-dwelling planktivore species, is representative of Lamprologini, the most numerous tribe of Lake Tanganyika comprising 79 endemic species. *Astatotilapia burtoni* is a Tanganyikan representative of the tribe Haplochromini. It is one of the few species able to cross the lake-river

boundaries and therefore is also found in the rivers throughout the Lake Tanganyika catchment. *Metriaclima zebra* is a representative of the rock-dwelling 'mbuna' lineage of Lake Malawi, an exemplar of a rapidly speciating genus, and is a highly specialised lake species, an algae scraper. Finally, *Pundamilia nyererei*, a widely distributed reef-dwelling planktivore, is a representative of the young cichlid radiation of Lake Victoria.

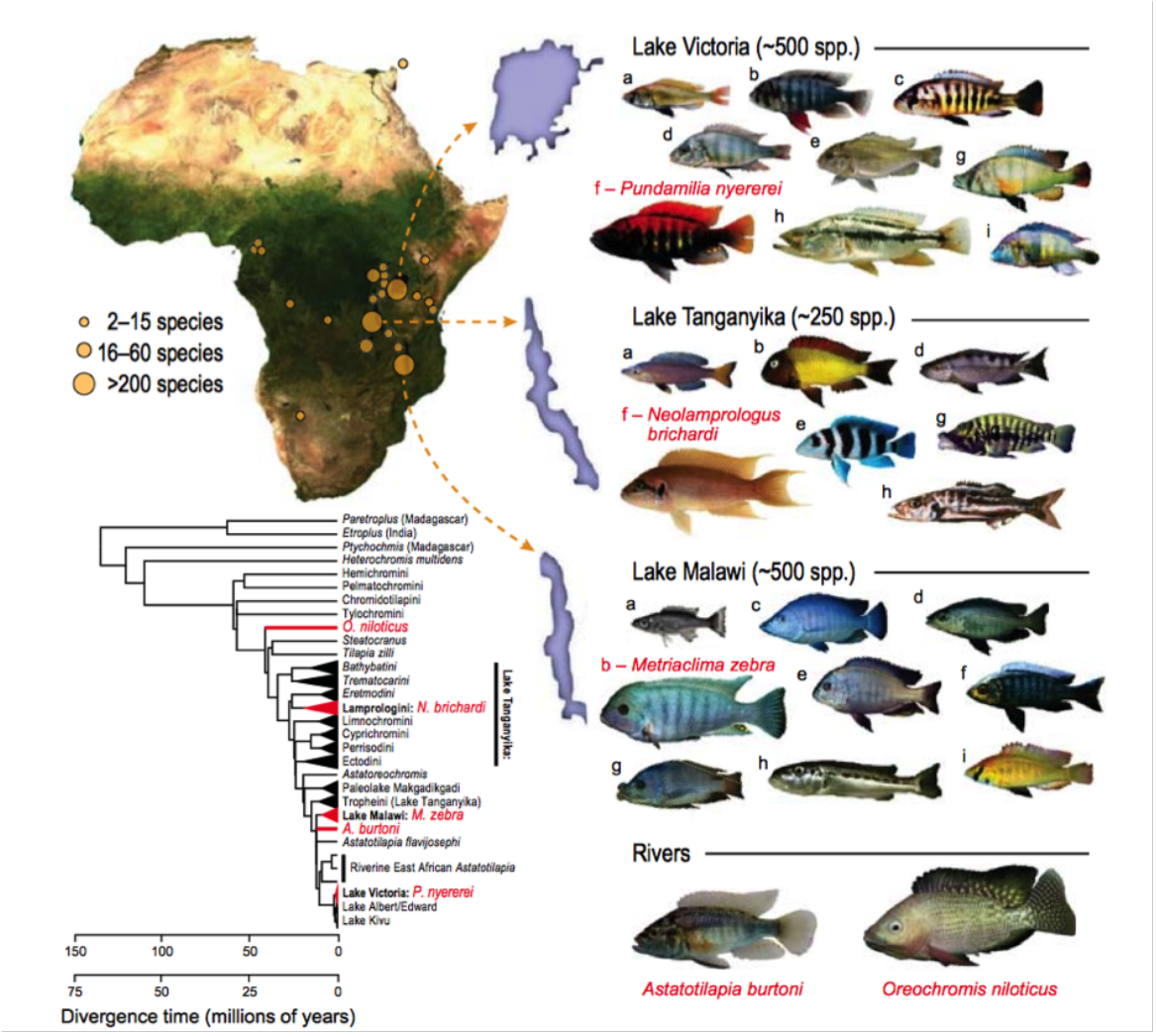


Fig. 2.1 **East African radiation of cichlid fish. Top left:** A map showing the location of African lakes in which cichlids have radiated. **Bottom left:** A phylogenetic tree showing the relationship between the sequenced species (red) and other cichlid lineages, with timescales reflecting two different estimates. **Right:** The sequenced species (red) and examples of major ecotypes in each lake: a, pelagic zooplanktivore; b, rock-dwelling algae scraper; c, paedophage (absent from Lake Tanganyika); d, scale eater; e, snail crusher; f, reef-dwelling planktivore; g, lobe-lipped insect eater; h, pelagic piscivore; i, ancestral river-dweller also found in lakes. Figure from [93].

Table 2.1 **Versions of cichlid genome assemblies used in this thesis**.

Species	Broad Institute assembly	URL used to download
M. zebra	MetZeb1.1_prescreen	http://www.broadinstitute.org/ftp/pub/assemblies/fish/M_zebra/MetZeb1.1_prescreen/M_zebra_v0.assembly.fasta
P. nyererei	PunNye1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/P_nyererei/PunNye1.0/P_nyererei_v1.assembly.fasta.gz
A. burtoni	HapBur1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/H_burtoni/HapBur1.0/H_burtoni_v1.assembly.fasta.gz
N. brichardi	NeoBri1.0	http://www.broadinstitute.org/ftp/pub/assemblies/fish/N_brichardi/NeoBri1.0/N_brichardi_v1.assembly.fasta.gz
O. niloticus	Orenil1.1	http://www.broadinstitute.org/ftp/pub/assemblies/fish/tilapia/Orenil1.1/20120125_MapAssembly.anchored.assembly.fasta

In addition to generating the reference genomes, the project team obtained RNA sequence data from multiple tissues in each species and used it to generate high quality annotation of protein coding genes. Comparisons between the annotated cichlid genomes and also utilising the existing genomes of other teleost fish (medaka, stickleback, tetraodon, and zebrafish) then provided interesting insights into the genomic changes underlying evolution of cichlids in East Africa. Details of these findings are described in the Cichlid genome consortium publication [93]. Here I briefly describe three highlights. First, the consortium members discovered 4.5- to 6-fold increase in the rate of gene duplications in the ancestors of the rapidly radiating lake cichlids and of the haplochromines. Duplicate genes can evolve new functions through neo- or subfunctionalisation [3] and thus facilitate adaptation and speciation. Second, the rate of evolution of genes associated with changes in jaw morphology is accelerated in haplochromines, providing evidence of repeated positive selection on these genes. And third, 625 noncoding regions were found to have a significantly accelerated rate of substitution in one or more of the East African lake cichlids, but strong sequence conservation across teleosts. Laboratory experiments indicated the ability of these non-coding sequences to alter the strength and patterns of expression of nearby protein coding genes, further strengthening the evidence for their function in cichlid evolution.

Another type of gene regulation, regulation by microRNAs (miRNAs), could also play an important role in cichlid evolution. I worked under the supervision of Eric Miska at the University of Cambridge to generate a miRNA annotation for each of the five reference genomes, and predictions of genes whose expression may be regulated by the miRNAs (target genes). I also categorised miRNAs in terms of sequence novelty, and assessed the hypothesis that accelerated rate of change in pairing between miRNAs

and target genes could have contributed to the divergence of the five cichlid species sequenced by the consortium.

2.2 Background

2.2.1 The nature and functions of microRNAs

Like protein coding genes, animal microRNAs are transcribed by RNA polymerase II. After transcription, a ∼70 to ∼100bp long section of the primary transcript (pri-miRNA) must fold into a so-called 'hairpin' structure (Figure 2.2), the hallmark of microRNAs. The hairpin is then excised from the flanking sequence and exported out of the nucleus. This so-called hairpin precursor (pre-miRNA) is further processed in the cytoplasm. First, its terminal loop is cleaved off, resulting in a ∼22bp double-stranded RNA molecule (Figure 2.2). The two strands separate and one associates with the microRNA-induced silencing complex (miRISC). The other strand is degraded [21, 94]. The selection of the strand to be loaded into miRISC is based on the thermodynamics of the double-stranded RNA molecule. The strand more often detected with miRISC is called mature miRNA, and the more often degraded strand in called miRNA* [94].

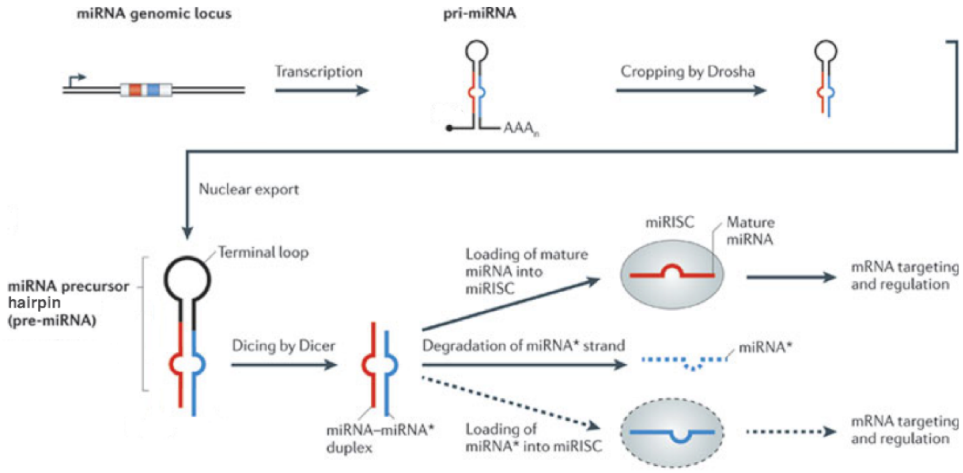


Fig. 2.2 **Biogenesis of microRNAs.** After transcription, a part of the RNA molecule folds into a hairpin structure. Still in the nucleus, regions flanking the hairpin are cut off by the RNAase type III enzyme Drosha. After being exported out of the nucleus, the miRNA precursor hairpin (pre-miRNA) itself is cleaved in another reaction catalysed by the RNAase type III enzyme Dicer. The resulting miRNA-miRNA* duplex then separates. The strand more often loaded into the microRNA-induced silencing complex (miRISC) is referred to as mature miRNA or simply miRNA and the other, more often degraded, strand is miRNA*. Figure adapted from [21].

The mature miRNA acts as an adaptor. Through complementary base pairing with sequences in 3′ untranslated regions (3′UTRs) of particular protein coding transcripts in the cytoplasm, it guides the miRISC towards them [94, 95]. The binding of miRNA-miRISC to a protein coding transcript leads to degradation of the transcript and/or to inhibition of protein synthesis [95]. Therefore, the mechanistic effect of individual microRNAs is to downregulate their targets.

Regardless of which protein coding gene is targeted, the miRNA-miRISC binding to mRNA normally requires continuous base pairing between the 3′UTR and bases 2 to 7 of the mature miRNA (Figure 2.3). Additional match at nucleotide 8 is required if the mature miRNA sequence does not start with an A [96]. The 7 nucleotides at bases 2 to 8 are known as the seed region. Extensive complementarity in the rest of the sequence is unusual in animal miRNAs [96].

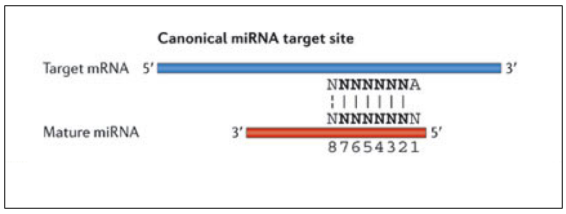


Fig. 2.3 **The canonical miRNA target site**. The mature miRNA guides miRISC to its target by complementary base pairing in the seed region. Figure adapted from [21].

2.2.2 MicroRNAs in vertebrate development and evolution - a hypothesis

On the whole, protein coding genes tend to evolve very slowly. We share almost all of our protein coding genes with the chimpanzee, and approximately 88% with rodents [97]. Many aspects of development are governed by transcription factors shared across the animal kingdom, from the fruit fly to human,and genes in evolutionarily distant organisms are remarkably similar, especially in functional terms, and often can be substituted for one another [98]. These and other observations reviewed by Caroll [98] provide a compelling argument that changes in gene regulation are likely to be important genetic drivers of adaptive differences between closely related animal species.

Regulatory changes affect development and can facilitate adaptive evolution by inducing differences in the timing or strength of gene expression. Many microRNAs play precisely these roles in development. Two first two microRNA genes that were identified, *lin-4* and *let-7*, were discovered by screening for genes that control developmental timing in *C. elegans* larvae [99, 100]. mir-10, a microRNA conserved from *C. elegans* to human,

originates from genomic regions known to harbour the Hox genes, crucial regulators of animal development, and it is known to target mRNAs of several Hox genes to inhibit protein synthesis [101]. Finally, the extensiveness of microRNA involvement in development is exemplified for example by the haematopoiesis pathway, where "virtually every step seems to be fine-tuned by specific microRNAs" [102].

Many microRNAs also impact evolution by fine-tuning gene expression. There is growing evidence that this leads to added robustness in regulatory networks and increased phenotypic reproducibility in the face of environmental perturbations [103, 104, 105]. Such 'canalisation' of development increases heritability, making traits more responsive to natural selection [21]. For example, a comparative study of the localisation of ancient microRNAs in a sea anemone, marine worms, and a sea urchin revealed close connection between establishment of new tissues and new microRNAs in early bilaterian evolution [106].

Therefore,we hypothesise that a) differential miRNA regulation of protein coding genes may contribute to some of the phenotypic differences between cichlids; b) miRNAs, through increased canalisation, may contribute to the increased speed of evolution in some cichlid lineages.

2.3 MicroRNA annotation

2.3.1 Small RNA sequencing

To generate experimental evidence for microRNA gene annotation, I prepared small RNA sequencing libraries from late stage embryos of all five species whose genomes were assembled by the Broad Institute. All samples were obtained from the same strains of fish used for genome sequencing by the Broad Institute, and were staged according to developmental milestones set out in ref [107]. Stage ∼22 embryos (corresponding to length of approximately 7.2mm and age of 8 days post fertilisation in *O. niloticus*) were homogenised in TRIsure (Bioline) reagent and total RNA extracted using the manufacturer's protocol. Using 5μg of total RNA as input, I used the Illumina TruSeq Small RNA Sample Preparation Kit to generate small RNA sequencing libraries, again following the manufacturer's protocol. The libraries were sequenced on the Illumina MiSeq platform, yielding between 2.4 and 4.4 million 36bp single-end reads per sample.

Because the mature miRNA themselves are ∼22bp long, the ends of 36bp reads contained sequences from 3′ adaptors (specific sequences attached to the ends of the small RNA molecules to facilitate sequencing). The adaptor sequences were

removed from reads using the `cutadapt v1.0` tool [108]. Figure 2.4 shows the length distribution of sequences after adaptor removal, and thus provides an indication about the proportion of reads in the dataset that are likely to correspond to sequences of mature miRNA molecules. The vast majority of reads came from molecules within the usual miRNA size range (20 to 24bp).

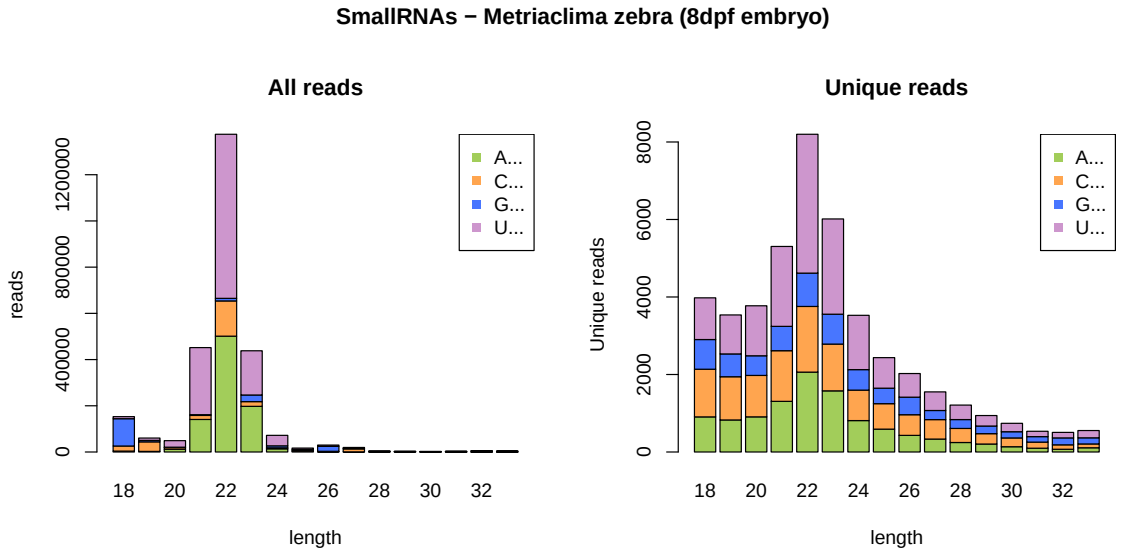


Fig. 2.4 **Length distribution of small RNAs sequenced from *M. zebra* embryos.** Colour denotes the first base of the read, as indicated in the legend. **Left:** All reads counted. The vast majority come from molecules within the usual miRNA size range. **Right:** Only unique reads counted. It is clear that the miRNA reads have low complexity - i.e. there are a lot of identical reads coming from the same miRNA.

2.3.2 Identification of miRNA loci

I used the `mirdeep2 v2.0.0.5` [109] package to detect signatures of microRNA genes from the sequencing data as follows. First, reads were 'collapsed', so that only one copy of identical sequences was kept, together with a record about how often the sequence is present in the data. Next, reads were mapped to their corresponding genome using the `bowtie v0.12.7` read aligner [110]. Reads mapping to multiple (more than five) genomic loci were discarded - these would be mostly repeat-derived sequences. Then looking at the remaining reads, the `RNAfold v2.0` program [111] was called to consider whether RNA derived from the sequence flanking any stacks of reads can plausibly fold into the hairpin structure, the signature of miRNAs. Finally, the localisation of reads within the predicted hairpin structure was evaluated. True

miRNA reads should correspond predominantly to the mature 22nt sequence and should have a well aligned 5′ start site due to consistent processing by the Dicer enzyme. For a detailed description of the process, please see the `mirdeep2` publication by Friedländer *et al.* [109].

The `mirdeep2` pipeline also takes advantage of homology with known miRNAs from related organisms. Therefore, I prepared a file containing all experimentally validated teleost miRNAs present in `miRBase v.19` (`TELEOST_MATURE_miRNA.fa`). Then the mapper.pl and miRDeep2.pl scripts from the `mirdeep2` package were executed as follows:

```
mapper.pl READS_FILE -j -l 18 -m -p GENOME_ASSEMBLY -s reads_collapsed.fa
-t reads_collapsed_vs_genome.arf -v

miRDeep2.pl reads_collapsed.fa GENOME_ASSEMBLY reads_collapsed_vs_genome.arf
none TELEOST_MATURE_miRNA.fa none 2 > report_mirbase19.log
```

Using output from the above commands, I constructed a high confidence set of cichlid miRNA loci by selecting predicted loci that received `mirdeep2` score greater than or equal to 10 in any of the five species, except where fewer than 2% of the aligned reads were a perfect match to the predicted mature sequence (± one nucleotide at the 3′ end, mismatch at the 3′ end base allowed). For example, a miRNA locus with score < 10 in *M. zebra* would be included if its predicted mature sequence were identical to a miRNA with mirdeep2 score $>= 10$ in *O. niloticus* or another cichlid species.

In this way, I identified 1,344 microRNA (miRNA) loci (259 - 286 per species). The complete annotation has been submitted to `miRBase` [112], the central miRNA repository.

2.3.3 Evolution in cichlid miRNA repertoires

The ways miRNAs evolve and change their target repertoires are illustrated in Figure 2.5. I searched for sources of novelty in cichlid miRNA repertoires by comparison with known teleost miRNAs and I discovered a) 40 cases of de-novo miRNA emergence and nine cases of apparent miRNA loss (Figure 2.6a); b) four distinct mature miRNAs with direct mutation(s) in the seed sequence; c) at least 14 cases of arm switching; d) one case of seed shifting; e) 92 distinct miRNAs with mutation(s) outside the seed sequence. For detailed methods see section 2.4.1.

I shared my results with Jeffrey Streelman at the Georgia Institute of Technology, who used RNA *in situ* hybridisation experiments to explore the spatial expression

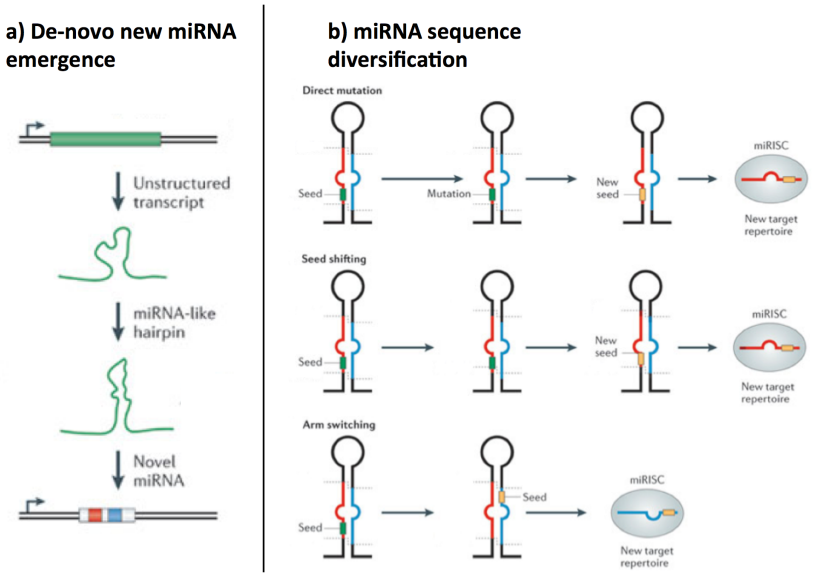


Fig. 2.5 **Evolution of miRNA novelty:** A miRNA with a new target repertoire can arise: **a)** de-novo from any transcribed sequence (e.g. intron); **b)** By diversification of an existing miRNA in the way of: 1) Direct mutation in the seed; 2) Change in hairpin processing which shifts the 5′ end of the miRNA (seed shifting); 3) Change in hairpin processing which leads to the previously degraded strand (blue) being utilised as a miRNA (arm switching); Figure from [21].

patterns in *M. zebra* for one case of arm switching (mze-miR-7132a-5p and mze-miR-7132a-3p). We found that spatial expression is clearly differentiated in the miRNA-miRNA* pair (Figure 2.6b), consistent with previous reports [113] and suggesting miRNA strand preferences may be controlled developmentally.

There is little comparable data on miRNA sequence novelty at similar evolutionary scales in other vertebrate lineages, but comparison with published data on *Drosophila* [114] suggests that the rate of de-novo miRNA emergence is not unusually elevated in cichlids. The cases of mutation in the seed sequence and of seed shifting are isolated incidents, too few to infer any trend. However, arm switching seems to be widespread (I have identified 15 cases but there are likely to be many more) and is likely the main mechanism by which cichlids generate miRNA sequence novelty. The evolution of miRNAs is intertwined with evolution of their targets, the protein coding genes they regulate. The detailed catalog of miRNA sequence novelty presented here provides a basis for exploration of the miRNA target space.

Novel cichlid miRNAs have complementary expression to predicted targets. I used the `RNAhybrid` [115] software package to predict genes that may be regulated by four de-novo cichlid-specific miRNAs: miR-10029, miR10032, miR-10044, miR-10049

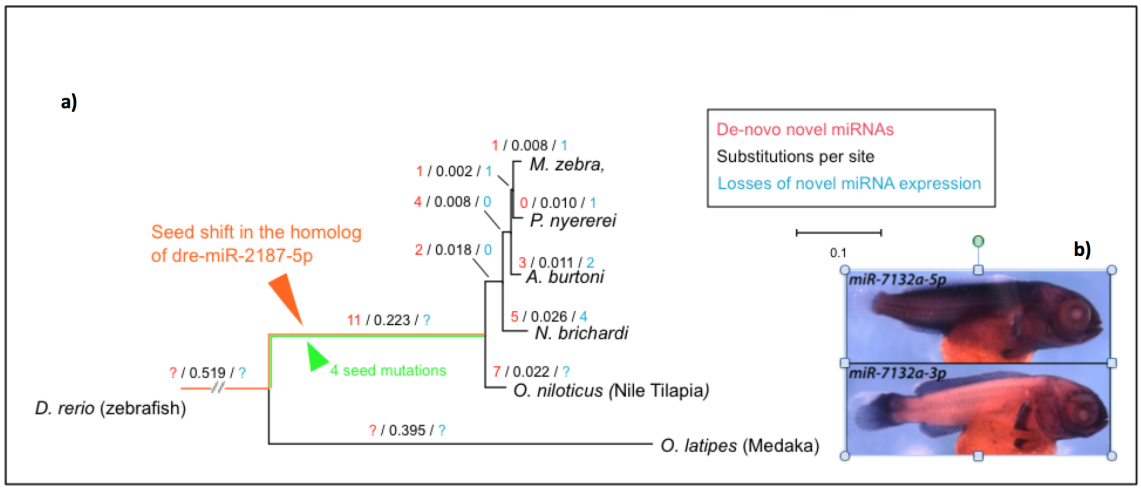


Fig. 2.6 **a)** The birth and death of de-novo novel microRNAs mapped on the phylogenetic tree of the five cichlid species. **b)** In situ hybridisation images showing the difference in spatial expression between mature miRNAs generated from 5p and 3p arms of the miR-7132a precursor in *M. zebra* (Jeffrey Streelman, Georgia Institute of Technology)

(detailed methods in section 2.4.2) and, again, shared the results with Jeffrey Streelman. His *in situ* hybridisation experiments revealed that, in *M. zebra*, the spatial expression of the four de novo miRNAs is confined to specific tissues (for example, fins, facial skeleton, brain), suggesting their expression is tightly regulated, and is strikingly complementary to genes predicted to contain target sites for these miRNAs (miR-10032 target *neurod2*, and miR-10029 target *bmpr1b*).

2.3.4 Evolution of miRNA targets

The target sequences bound by miRNAs tend to evolve much faster than miRNAs themselves [21]. To be able to observe sequence evolution at miRNA target sites across the East African cichlid radiation, I curated a set of 2359 genes with 1:1 homology in all five cichlid species and a unique 3′ untranslated region (3′UTR) sequence per gene per species. This was made possible by starting from the `V1` gene annotations generated at the Broad Institute as a part of the Cichlid genome project [93], which includes 3′UTRs and assignment of orthologs between the cichlid species. To ensure that any observed differences between species are due to substitutions or short insertions or deletions, I restricted the analyses to the portion of 3′UTRs present in all five species genome assemblies.

Fine scale evolution of particular miRNA targets

I focussed on the evolution of sites targeted by novel miRNAs, by miRNAs involved in arm switching in cichlids, and by 29 of the 38 miRNAs whose expression profiles in the central nervous system of zebrafish have been studied in detail by Kapsimali *et al.* [116]. I used the PACMIT software package [117] for target predictions (see section 2.4.3 for detailed methods), and focussed on identifying sites where sequence variation leads to gain or loss of predicted binding sites in *M. zebra*, *P. nyererei*, and *A. burtoni*, members of the most rapidly spectating cichlid tribe Haplochromini.

Using this strategy, I was able to identify several interesting cases where sequence evolution in 3′UTRs is predicted to create differences between the haplochromines and other East African cichlids in miRNA regulation of particular genes. For example, the mature miRNA produced from the 3p arm of the arm switching miR-7132a is predicted to target the key developmental gene *bmper* in *O. niloticus* and in *N. brichardi* but a substitution destroys the predicted binding site in haplochromines, possibly removing *bmper* from the control of this miRNA. Conversely, miR-124 is predicted to regulated the gene *birc5a* in haplochromines, but the binding site is lost in *O. niloticus* and in *N. brichardi*. This is interesting because in zebrafish *birc5a* promotes neuronal differentiation [118] and expression of miR-124 is associated with the transition of neural progenitors to differentiated neurons [116]. Therefore, the substitutions in *birc5a* 3′UTRs may result in differences in neural development and function between haplochromines and other East African cichlids.

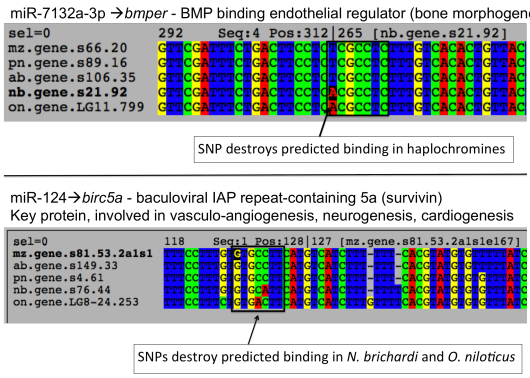


Fig. 2.7 **Sequence evolution at specific miRNA target sites**. Multiple alignments of 3′UTRs of: mz-*M. zebra*, pn-*P. nyererei*, ab-*A. burtoni*, nb-*N. brichardi*, and on-*O. niloticus*. Predicted target sites (sequences complementary to miRNA seeds) are denoted by black rectangles and substitutions (SNPs) predicted to alter binding are highlighted.

The way a miRNA targets a gene for regulation is still imperfectly understood and miRNA target prediction algorithms all suffer from large numbers of false positives (see section 2.4.3). However, it is possible to experimentally illustrate that a miRNA is interacting with a transcript, for example by showing that it can reduce the level of protein expression in an *in vitro* system using a luciferase assay [119]. My target

predictions provided a basis for such an experiment. The experiment was conducted by Joe Hanly, a research project student at the Miska Laboratory at the Gurdon Institute in spring 2013. He used the psiCHECK-2 vector[1] to assay the ability of miR-99b to downregulate the expression of the gene *tmem141*. My computational predictions suggested that the gene should be targeted in *M. zebra* and other haplochromines, but that the binding site is not present in *O. niloticus* and *N. brichardi* (Figure 2.8).

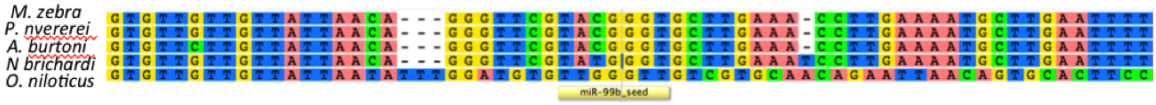


Fig. 2.8 **Sequence evolution in *tmem141* 3′UTRs** The predicted target site (sequence complementary to the seed sequence of miR-99b) is delineated by the yellow rectangle below the sequences. (Figure by Joe Hanly, research project student at Miska Lab in spring 2013)

3′UTRs of *tmem141* from *M. zebra*, *O. niloticus*, and *N. brichardi* were cloned into the psiCHECK-2 luciferase assay vector and transfected into human cell culture. Upon introduction of miR-99b, the activity of luciferase protein with the *M. zebra* 3′UTR was significantly reduced by 20% +/- 7.6% compared to control (Figure 2.9), while luciferase constructs with *O. niloticus* and *N. brichardi* 3′UTRs were not significantly downregulated (Figure 2.9). The function and expression profile of *tmem141* in fish is not known, and the gene was selected for this experiment by Joe Hanly simply for ease of 3′UTR cloning. Nevertheless, these results provide experimental evidence complementary to and in agreement with my computational predictions and constitute a first attempt at experimental validation of a miRNA target in cichlid fishes.

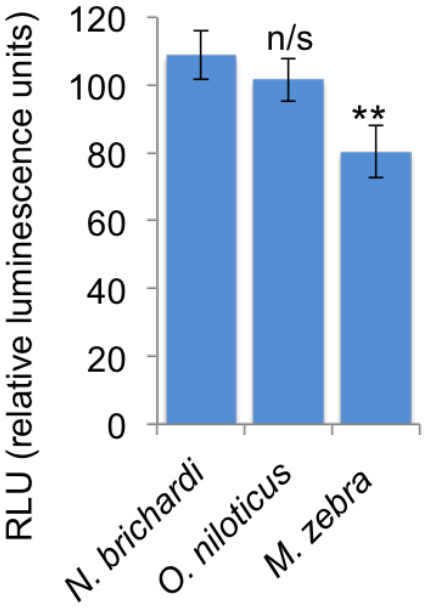


Fig. 2.9 **miR-99b downregulates luciferase with *M. zebra* *tmem141* 3′UTR**.

[1]https://www.promega.co.uk/resources/pubhub/cellnotes/microrna-biosensors-application-for-the-psicheck2-vector/

Increased conservation of target sequences across the East African radiation

In 2011, Loh, Yi, and Streelman conducted a study on low coverage genomic data from five Lake Malawi species and concluded that there is evidence of "accelerated divergence of miRNA target sites in cichlids, suggesting that the selective divergence of miRNA regulation has a role in the diversification of cichlid species." [120]. Specifically, the authors found that the density of single nucleotide variants in predicted miRNA targets was 0.44%, much higher than the average 3′UTR density of 0.28% (Figure 2.10A). While this is a very interesting finding, I note that the study relied on limited data: only 11.6Mb of multiple sequence alignment (only 1.6% of the assembled *M. zebra* genome), only computational predictions of miRNAs and protein coding genes, and an arbitrary decision that all 3′UTRs extend exactly 500bp from the end of the protein coding sequence.

With the whole genome assemblies from the Cichlid Genome Consortium and benefitting from experimental miRNA and 3′UTR annotations, I was able to confidently test whether accelerated divergence of miRNA target sites can also be observed across the broader East African radiation. I used all *M. zebra* miRNAs and all annotated 3′UTRs in my curated a set of 2359 genes to predict miRNA target sites with the PACMIT software package [117] (see section 2.4.3 for detailed methods). Then I generated multiple alignments of the 3′UTR sequences with `clustax v2.0` and found that the density of single nucleotide variants within predicted miRNA targets is 8.2% (718/8,755bp) while the density over the rest of 3′UTR sequences is 11.2% (221,452/1,962,760bp) (Figure 2.10B). Therefore, these findings contrasts with that of Loh, Yi, and Streelman [120]. I conclude that there is evidence of elevated purifying selection on miRNA targets over the evolutionary timescales that separate *O. nilotics*, *N. brichardi*, *A. burtoni*, *P. nyererei*, and *M. zebra*, consistent with functional constraint as in other functional sequence such as enhancers or protein coding sequence.

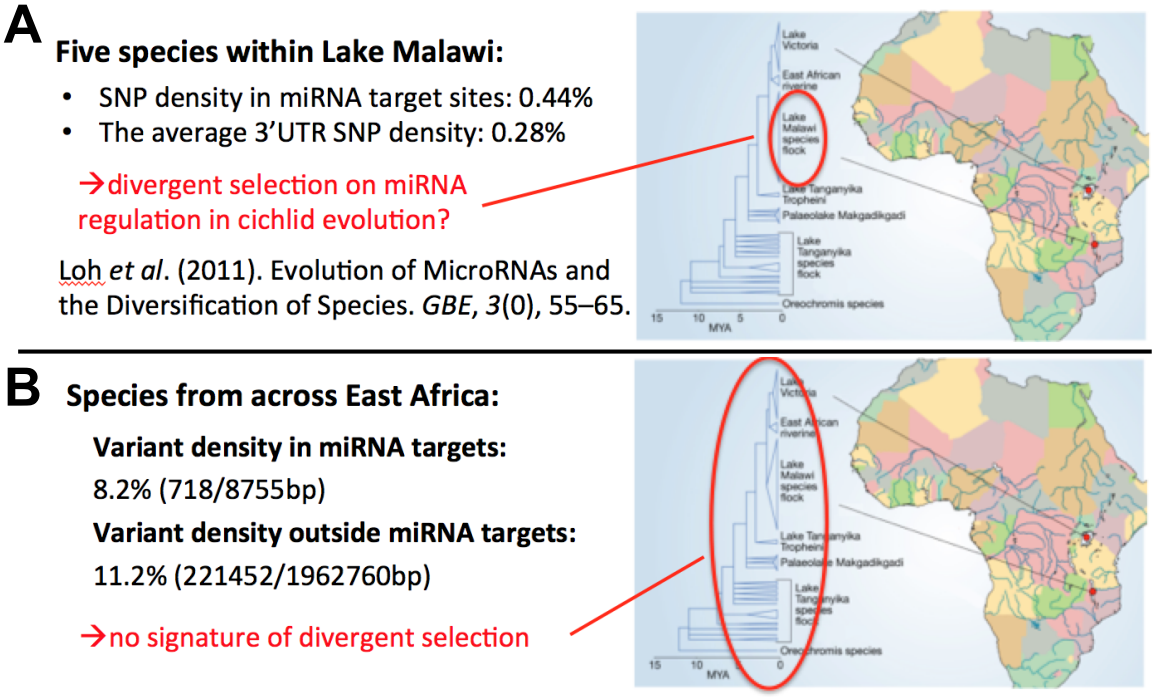


Fig. 2.10 **Purifying selection on miRNA target sites**. Contrasting results were obtained by two studies: **(A)** Loh *et al.* found accelerated target site divergence in five Lake Malawi species. **(B)** I found evidence of elevated purifying selection on miRNA targets when comparing the five species sequenced by the Cichlid genome consortium. The maps and phylogenetic trees used in this figure are from [121].

2.4 Detailed methods

2.4.1 Searching for novelty in cichlid miRNA repertoires

I used programs from the `fasta v36.3.4` software package to align the high confidence set of cichlid mature and precursor (hairpin) miRNA sequences to all experimentally validated teleost miRNAs present in `miRBase v.19`. Specifically, I used the `ssearch` (local Smith-Waterman) algorithm for hairpin→hairpin and mature→mature alignments and the `glsearch` (global-local Needleman-Wunsch) algorithm for mature→hairpin alignment as follows:

```
> ssearch36 -E 0.005 -C 25 -m 9 -3 CICHLID_HAIRPIN.fa TELEOST_HAIRPIN.fa >
HAIRPIN_TO_HAIRPIN.ssalign

> glsearch36 -E 0.1 -C 25 -m 9 -3 CICHLID_MATURE.fa TELEOST_HAIRPIN.fa >
MATURE_TO_HAIRPIN.glalign

> ssearch36 -E 0.1 -C 25 -m 9 -3 CICHLID_MATURE.fa TELEOST_MATURE.fa >
MATURE_TO_MATURE.ssalign
```

Based on analysing the alignments, I used a custom script to automatically assign all cichlid miRNAs into five categories based on their conservation/novelty. The name of the file in GREEN colour indicates presence of at least one significant alignment for a cichlid miRNA in miRBase v.19; RED colour indicates absence of any significant alignment.

1) HAIRPIN_TO_HAIRPIN+MATURE_TO_HAIRPIN+MATURE_TO_MATURE

These are homologs of known teleost miRNAs, possibly with single base polymorphisms:

a) Mature sequence identical to the best match (ignoring ± 2bp length difference at 3′ end)
b) Mature sequence different from the best match but seed sequence (bases 2-8) identical
c) Mature sequence different from the best match with a difference in the seed
d) 5′ isomiRs, leading to seed shifting

2) HAIRPIN_TO_HAIRPIN+MATURE_TO_HAIRPIN+MATURE_TO_MATURE

Alternative processing of a known hairpin - arm switching

3) HAIRPIN_TO_HAIRPIN+MATURE_TO_HAIRPIN

There is a homologous hairpin but the mature miRNA has changed substantially. This could be a sign of divergent evolution of the mature miRNA sequence.

4) HAIRPIN_TO_HAIRPIN+MATURE_TO_MATURE

There is no significant alignment for the hairpin but the mature sequence is highly similar to known teleost miRNA. This could be a sign of convergent evolution of mature miRNA sequence.

5) HAIRPIN_TO_HAIRPIN+MATURE_TO_MATURE

In this category are novel miRNAs likely to have arisen de-novo since the divergence between cichlid and medaka ancestors.

2.4.2 Target prediction by `RNAhybrid`

miRNA target prediction for the Cichlid genome consortium publication was done using `RNAhybrid` [115]. For each miRNA, the calculation of p-values was calibrated using the tool `RNAcalibrate` before using the `RNAhybrid` tool for the calculation of minimum free energy hybridisation between the miRNA and all annotated 3′UTRs in all five species. For both `RNAcalibrate` and `RNAhybrid` I forced perfect matches between bases 2 and 7 of the miRNA (the seed) using the `-f 2,7` option.

2.4.3 Target prediction with `PACMIT`

The false discovery rate for computational target prediction is often quoted to be approximately 50% [122]. It is difficult to to choose the best among the different methods because of a lack of studies comparing the available algorithms using a set of experimentally validated miRNA targets. Marín and Vanícek conducted one of only a few such studies [117], comparing several prediction methods on a set of experimentally validated miRNA targets in human (Figure 2.11). It is on the basis of this study that I chose the 'Prediction of Accessible MicroRNA Targets (`PACMIT`)' algorithm for miRNA target prediction analyses outside the Cichlid Genome Consortium.

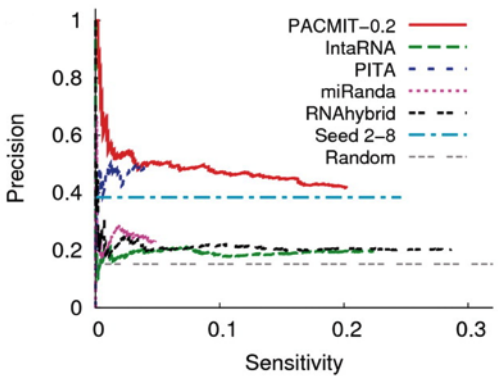


Fig. 2.11 **Precision versus sensitivity of a selection of miRNA target prediction algorithms**. Figure from [117].

PACMIT uses predictions of secondary structure of the 3′UTR sequence to determine whether any locus with sequence complementarity to the miRNA seed would be accessible for binding. I used the ViennaRNA Package [111] to predict the secondary structures, excluding UTRs shorter than 22bp and also UTRs with undetermined bases (N characters). With the accessibility information available, I executed PACMIT to always require perfect pairing between bases 2-8 of the miRNA and the 3′UTR and accessibility of the bases at miRNA positions 2-5, as suggested by the authors in personal communication and in [123] as follows:

```
paccmit.pl -utrs UTRs.fa -mirs miRNAs.fa -prop 3 -nuclend 5 -nmer 4 -pcutoff 0.2
-output out.txt
```

Chapter 3

New genome sequence datasets

3.1 Introduction

3.1.1 Overview of whole-genome data

The genome assemblies and annotations from the Cichlid Genome Consortium provide a key resource facilitating in-depth studies of the individual radiations in the Great Lakes of East Africa. Taking advantage of this opportunity, we obtained almost 3,000 Gbp of raw whole genome sequence data from 281 individuals covering 88 species from in and around Lake Malawi (Table 3.1). This rich dataset has facilitated all the cichlid work described in the rest of this thesis.

Table 3.1 **Whole genome sequencing at Sanger Institute - an overview.**

Sample type	Number of individuals	Primary goals	Coverage per individual	Overall amount of sequence
Mother - father - offspring trio	9 (three trios)	De-novo assembly Haplotype phasing	~40X	~360Gbp
Lake Malawi HC	83	Assessing between species diversity	~15X	~1,245Gbp
Lake Malawi LC	34	Assessing within species diversity	~5X	~170Gbp
Astatotilapia calliptera populations	19	Diversity in rivers and lakes around Lake Malawi	~15X	~285Gbp
Crater lake *Astatotilapia*	130	Detailed view of divergence between incipient species	~15X - n:12 ~5X - n:118	~180Gbp ~590Gbp
Other	6	Various; see text	~15X - n:5 ~60X - n:1	~75Gbp ~60Gbp
Total:	281			~2,965Gbp

The 281 cichlid samples were obtained from our collaborators on this project George Turner (University of Bangor) and Martin Genner (University of Bristol), who have also assigned individual specimens to species. Sequencing progressed in three phases:

a first batch of samples was sequenced in Spring 2013, second batch in Autumn 2014, and the third in Summer 2015. Tables 3.2 and 3.3 describe all the individual specimens sequenced and their assignment to sequencing batches. Samples from the Summer 2015 batch are not included in any analysis because they were sequenced only a few weeks before the completion of this thesis.

Lake Malawi harbours five species of tilapiine cichlids (tribe Tilapiini): four *Oreochromis* species and *Tilapia rendalli*. These fish, although present in Lake Malawi, do not form a part of the main Lake Malawi adaptive radiation and represent a separate evolutionary lineage. Therefore, they were not included in our sampling.

We have focussed on the rapidly radiating tribe Haplochromini. Our Lake Malawi samples (Table 3.2) cover all its major haplochromine lineages, including specimens from the following groups:

1. Nine species from the 'mbuna' group of mainly shallow-water rock-dwelling cichlids. Our specimens represent much of the diversity of this group, covering 7 out of 10 genera defined by Ribbink *et al.* [124]. Mbuna is a common name given to these fish by the Tonga people of Malawi [124]. Since the Ribbink *et al.* detailed classification of 196 mbuna species [124] found in the Malawian waters, dozens new species have been described [125], members of the *Pseudotropheus tropheops* species-complex have been assigned to a new (sub)genus *Tropheops* [126], and members of the *Pseudotropheus zebra* species-complex have been assigned to a new genus *Metriaclima* [127].
2. Ten species of pelagic (open-water) cichlids. Cichlids are primarily bottom-dwellers, but members of three genera (*Diplotaxodon*, *Rhampochromis*, and *Pallidochromis*) have undergone extensive changes in morphology and behaviour to become pelagic piscivores (ecotype h in Figure 2.1), feeding mainly on crustaceans and lake sardines [125]. We have sequenced three out of seven scientifically described species of *Diplotaxodon* [87, p. 198], and three undescribed species: *D.* 'macrops black dorsal' [128], *D.* 'ngolube' [87, p. 239], and *D.* 'white back similis' (M. Genner, pers. comm.). There are eight described species of *Rhampochromis* [129], or which we have sampled three. Finally, *Pallidochromis* is a single species genus - *P. tokolosh* is morphologically intermediate between the other two genera and is a slightly more benthic form [87, pp. 198-199].
3. Twelve species of deep water benthic haplochromines, generally caught at depths of 50m, a 'twilight' zone with very little visible light. These include members of the genera *Alticorpus* and *Aulonocara*, characterised by greatly enlarged sensory openings of their heads and lateral lines, several of the species currently assigned

to the genus *Lethrinops*, and *Otopharynx speciosus*. There are also 47 (described and undescribed) deep water species of *Placidochromis* [87, pp. 104-197]. and three or four deep water species of *Stigmatochromis* [125, pp. 405-408]. However, we have not obtained any deep-water specimens from these two genera.

4. Forty-five species of cichlids found predominantly in shallow waters close to the shore (like 'mbuna'), but on sandy or muddy lake floor and the transition zones between sandy and rocky habitats. This is a very diverse group of cichlids with hundreds of described species [125], including for example large (over 35cm) predators such as *Buccochromis nototaenia*, the small plankton-feeding shoaling *Copadichromis*, and mollucivores such as *Chilotilapia rhoadesii* and *Mylochromis anaphyrmus*. We refer to this group as 'sand dwellers'.
5. Two haplochromine cichlids found in Lake Malawi are able to cross the lake-river barrier: *Astatotilapia calliptera* and *Serranochromis robustus*. A versatile, relatively small cichlid (~10-15cm) common in the rivers throughout Lake Malawi catchment, *A. calliptera* in Lake Malawi frequents shallow sheltered bays with muddy sediment and aquatic plants, often feeding on snails [125, p. 281]. It has been suggested that it may be related to the ancestral lake-river generalist species that seeded most or perhaps all of the Lake Malawi haplochromine radiation [86, 125]. For this and other reasons to be discussed later, we sampled *A. calliptera* genetic variation extensively. *S. robustus* is a large predator often seen in very shallow water near river estuaries [125, p. 277] and is a common species in rivers to the south-west of Lake Malawi, including the Zambezi system [130]. Eccles and Trewavas [129, pp. 24-26] suggest that some Lake Malawi genera, especially among the larger sand-dwellers, may have ancestors allied to *S. robustus* or other members of an ancestral group of riverine species of the Zambezi system.

Table 3.2 **Lake Malawi sequencing.** Colours indicate sequencing batches: blue - Spring 2013, brown - Autumn 2014, green - Summer 2015. Symbols indicate common species groups: ∗ - 'mbuna', ● - open-water (pelagic) cichlids, ▼ - deep water sand-dwellers, ▲ - shallow water sand-dwellers, □ - lake-river generalists

Panel A: Population genomics samples

Species	Samples 15X	Samples 5X	Species	Samples 15X	Samples 5X
Alticorpus geoffreyi▼	1	0	*Hemitilapia oxyrhynchus*▲	1	0
Alticorpus macrocleithrum▼	1	0	*Iodotropheus sprengerae*⋆	1	0
Alticorpus peterdaviesi▼	1	0	*Labeotropheus trewavasae*⋆	1	0
Aulonocara 'blue chilumba'▼	1	0	*Lethrinops albus*▲	1	0
Aulonocara 'gold'▼	1	0	*Lethrinops auritus*▲	1	0
Aulonocara 'minutus'▼	1	0	*Lethrinops gossei*▼	1	0
Aulonocara 'yellow'▼	1	0	*Lethrinops lethrinus*▲	1	0
Aulonocara steveni▲	1	0	*Lethrinops* 'longimanus redhead'▼	1	0
Aulonocara stuartgranti 'maisoni'▲	1	0	*Lethrinops* 'oliveri'▼	1	0
Buccochromis nototaenia▲	1	0	*Metriaclima zebra*⋆	1	0
Buccochromis rhoadesii▲	1	0	*Mylochromis anaphyrmus*▲	1	4
Champsochromis caeruelus▲	2	0	*Mylochromis ericotaenia*▲	1	0
Chilotilapia rhoadesii▲	1	3	*Mylochromis melanotaenia*▲	1	0
Copadichromis borleyi▲	1	0	*Nimbochromis linni*▲	1	0
Copadichromis likomae▲	1	0	*Nimbochromis livingstoni*▲	1	0
Copadichromis mloto▲	1	0	*Nimbochromis polystigma*▲	1	0
Copadichromis quadrimaculatus▲	1	0	*Otopharynx* 'brooksi nkhata'▼	1	0
Copadichromis trimaculatus▲	1	0	*Otopharynx lithobates*▲	1	0
Copadichromis virginalis▲	1	4+2[1]	*Otopharynx speciosus*▼	2	0
Ctenopharynx intermedius▲	1	2	*Pallidochromis tokolosh*●	1	0
Ctenopharynx nitidus▲	1	0	*Petrotilapia genalutea*⋆	1	0
Ctenopharynx nitidus▲	1	0	*Placidochromis electra*▲	1	0
Cyathochromis obliquidens⋆	1	0	*Placidochromis johnstoni*▲	1	0
Cynotilapia afra⋆	1	0	*Placidochromis longimanus*▲	1	4
Cynotilapia axelrodi⋆	1	0	*Placidochromis milomo*▲	1	0
Dimidiochromis compressiceps▲	1	0	*Placidochromis subocularis*▲	0	8
Dimidiochromis dimidiatus▲	1	0	*Protomelas ornatus*▲	2	0
Dimidiochromis kiwinge▲	1	0	*Rhamphochromis esox*●	1	0
Dimidiochromis strigatus▲	1	0	*Rhamphochromis longiceps*●	1	0
Dimidiochromis strigatus▲	1	0	*Rhamphochromis woodi*●	1	0
Diplotaxodon 'ngulube'●	1	0	*Serranochromis robustus*□	1	0
Diplotaxodon 'white back similis'●	1	0	*Stigmatochromis 'guttatus'*▲	1	0
Diplotaxodon greenwoodi●	1	0	*Stigmatochromis modestus*▲	1	0
Diplotaxodon limnothrissa●	1	0	*Taeniochromis holotaenia*▲	1	0
Diplotaxodon macrops●	1	0	*Taeniolethrinops furcicauda*▲	1	0
Diplotaxodon 'macrops black dorsal'●	1	0	*Taeniolethrinops macrorhynchus*▲	1	0
Fossorochromis rostratus▲	1	3	*Taeniolethrinops praeorbitalis*▲	1	0
Genyochromis mento⋆	1	0	*Tremitochranus placodon*▲	1	4
Hemitaeniochromis spilopterus▲	1	0	*Tropheops tropheops*⋆	1	0
Hemitaeniochromis spilopterus▲	1	0	*Tyrannochromis nigriventer*▲	1	0
Hemitilapia oxyrhynchus▲	1	0			

Panel B: Deep coverage samples for genome assembly

Species	Sampling location	Relationship	Coverage paired-end	Coverage mate-pair
Astatotilapia calliptera□	Salima region	father	~40X	0
Astatotilapia calliptera□	Salima region	mother	~40X	0
Astatotilapia calliptera□	Salima region	offspring	~40X	~5X
Aulonocara stuartgranti▲	Usisya region	father	~40X	0
Aulonocara stuartgranti▲	Usisya region	mother	~40X	0
Aulonocara stuartgranti▲	Usisya region	offspring	~40X	~5X
Lethrinops lethrinus▲	Mazinzi reef	father	~40X	0
Lethrinops lethrinus▲	Mazinzi reef	mother	~40X	0
Lethrinops lethrinus▲	Mazinzi reef	offspring	~40X	~5X

[1]from Lake Malombe

In the summer of 2011, Martin Genner and George Turner explored the cichlid fish fauna of crater lakes in the Rungwe District of Tanzania, approximately 40km north of Lake Malawi (Figure 3.1). They discovered haplochromine cichlids derived from *Astatotilapia calliptera* in six of the lakes, and a pair of incipient species forming within one - Lake Massoko. We have obtained whole genome sequence data from 100 *Astatotilapia* individuals from Lake Massoko, 30 individuals from Lake Itamba, one from Lake Kingiri, four from the Itupi stream - the closest water body upstream of Lake Massoko, and one from the Mbaka river - a major river downstream from Lake Massoko. Furthermore, to explore the geographical context of the crater-lake radiation and given the potential importance of *A. calliptera* ancestors in the Lake Malawi radiation, we have added 13 more *A. calliptera* from the wider Lake Malawi catchment from locations shown in Figure 3.2 (also see Table 3.3 - Panel A). The *A. calliptera* radiation in the crater lakes of Tanzanian Rungwe District is explored in detail in chapter 6.

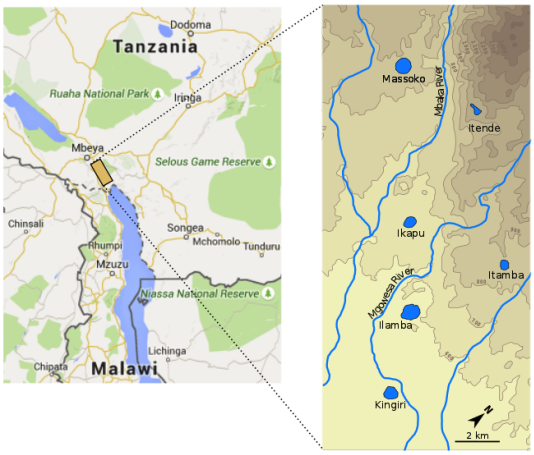


Fig. 3.1 **Map of the crater lake region in Southern Tanzania**

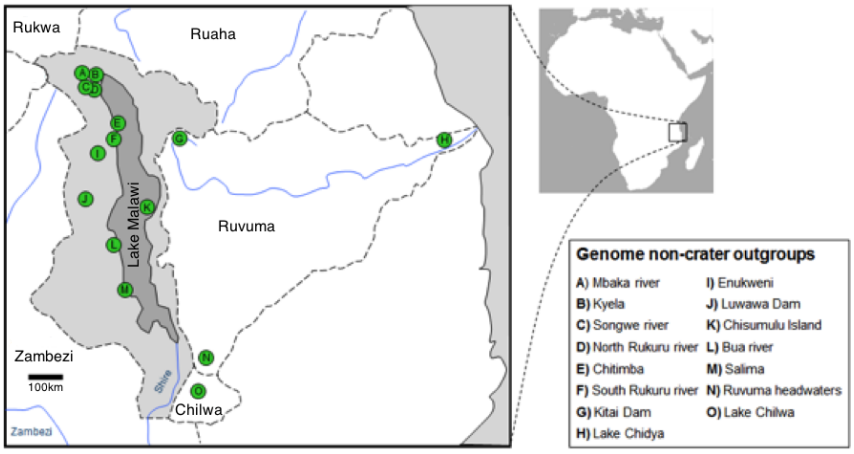


Fig. 3.2 **Collection sites of non-crater-lake *Astatotilapia calliptera* specimens.** Dotted lines represent catchment boundaries, with the Lake Malawi catchment shaded in grey. Figure from Martin Genner.

Table 3.3 **Cichlid samples from outside Lake Malawi.** Colours indicate sequencing batches: blue - Spring 2013, brown - Autumn 2014, green - Summer 2015.

Panel A: An overview of *Astatotilapia calliptera* and crater lake samples

Sampling location (ecomorph)	Samples 15X	Samples 5X	Latitude S	Longitude E
Lake Massoko (benthic)	6	32	9°20'00	33°45'18
Lake Massoko (littoral)	6	25	9°20'00	33°45'18
Lake Massoko (small unassigned)	0	31	9°20'00	33°45'18
Itupi stream	4	0	9°19'47	33°44'40
Lake Itamba	0	30	9°21'04	33°50'39
Lake Kingiri	1	0	9°25'00	33°51'00
Chitimba	1	0	10°34'37	34°10'14
North Rukuru	1	0	9°55'01	33°55'39
Songwe River	1	0	9°35'14	33°46'10
South Rukuru	1	0	10°45'42	34°07'33
Enukweni	1	0	11°11'14	33°52'52
Lake Chidya	1	0	10°35'49	40°09'19
Kitai Dam	1	0	10°42'22	35°11'46
Ruvuma river	1	0	14°22'22	35°32'54
Near Kyela	1	0	9°33'05	33°53'11
Luwawa Dam	1	0	12°06'57	33°43'23
Bua	1	0	13°18'30	33°32'51
Chisumulu island	1	0	~12°00'00	~34°37'00
Mbaka River	1	0	9°20'27	33°47'04
Lake Chilwa	1	0	15°22'15	35°35'30

Panel B: Other Sanger Institute sequenced samples from outside Lake Malawi

Species	Sampling location	Coverage	Notes
Astatotilapia 'rujewa'	Ruaha river	15X	*A.* 'rujewa' = *A.* 'ruaha', a taxon discovered in 2012 in the Ruaha river in Tanzania
Astatotilapia tweddlei	Lake Chiuta	15X	*A. tweddlei* is a common species to the East of Lake Malawi
Otopharynx tetrastigma	Lake Ilamba	15X	*O. tetrastigma* is a Lake Malawi endemic, but this sample is from a similar looking species found in Lake Ilamba
Rhamphochromis 'kingiri brevis'	Lake Kingiri	15X	Similar to *R. brevis* of Lake Malawi
Rhamphochromis 'kingiri dwarf'	Lake Kingiri	15X	Similar to a small *Rhamphochromis* species found in Lake Chilingali
Andinoacara coeruleopunctatus	Panama	60X	A cichlid from Central America, sequenced to provide a *de novo* genome assembly as an outgroup to African cichlids

Understanding the evolutionary origins of the Lake Malawi haplochromine radiation would facilitate discussion about its early stages and tests to distinguish selection on standing variation from adaptation driven by novel genetic variants arising within the lake. A study by Joyce *et al.* (2011) is inconclusive as to whether *A. calliptera* is an outgroup or a member of the Lake Malawi flock [86]. In search for a sister species, we have sequenced two candidate species: *Astatotilapia* 'rujewa' and *Astatotilapia tweddlei* (Table 3.3 - Panel B). In a more recent study, Genner *et al.* claim on the basis of mitochondrial DNA sequence that *Astatotilapia* 'rujewa' is "immediate sister taxon" to the Lake Malawi flock [131].

In the crater lake Kingiri, Genner and Turner also discovered a pair of species of *Rhampochromis* (Table 3.3 - Panel B). The two forms have dramatically different morphology but share mtDNA haplotypes (M. Genner, pers. comm). In conjunction with the sequencing of *Rhampochromis* from Lake Malawi, the nuclear DNA of the Kingiri samples will enable us to test if the two forms have invaded Kingiri independently or if they have diverged from a common ancestor inside Lake Kingiri, representing another case of sympatric speciation in addition to Lake Massoko *Astatotilapia*.

We have also obtained data from a haplochromine species of the crater lake Ilamba whose morphology is reminiscent of *Otopharynx tetrastigma* (Table 3.3 - Panel B). The data will enable us to investigate the origin of this species.

The Cichlid Genome Project did not provide a reference genome for non-African cichlids. Furthermore, at the time of writing reference genomes are not available for any non-cichlid members of the suborder Labrodei. The closest available genome is that of medaka, sharing common ancestor with cichlids well over a hundred million years ago [93]. We have obtained high coverage data for *Andinoacara coeruleopunctatus*, a Central American cichlid (Table 3.3 - Panel B), with the aim of generating a draft reference genome, which would enable us to address questions relating to the origin of the African radiation.

3.2 Alignment, variant calling, filtering, and genotype refinement

3.2.1 DNA extraction and sequencing

I extracted DNA from fin clips using PureLink® Genomic DNA extraction kit (Life Technologies). Genomic libraries for paired-end sequencing on the Illumina HiSeq 2000 machine were prepared by the Sanger Institute sequencing core teams according to the Illumina TruSeq HT protocol to obtain 100bp (Spring 2013 batch) and 125bp (Autumn 2014 and Summer 2015 batches) *paired-end* reads. In paired-end sequencing, both ends of a DNA fragment are read - e.g. the first 100bp and the last 100bp of a 300bp fragment. The mean fragment length (also called 'insert size') for paired-end sequencing has been 300-500bp. Three special *mate-pair* libraries with large insert sizes (~2,000bp) were generated by the Sanger Institute's Illumina Bespoke team to support *de novo* genome assemblies, as indicated in Table 3.2 - Panel B.

3.2.2 Alignment

Reads from samples of *A. calliptera*, *A. stuartgranti*, and *L. lethrinus* which were sequenced to ~40X coverage for genome assembly were down-sampled for studies relying on alignment to a reference genome to ~15X coverage and then processed identically to other samples.

All reads were aligned to *Metriaclima zebra* reference genome [93] using the `bwa mem v0.7.10` algorithm [132] using default options. Duplicate reads were marked on both per-lane and per sample basis using the `MarkDuplicates` tool from the `Picard` software package with default options (`http://broadinstitute.github.io/picard`) and local realignment around indels performed on both per lane and per sample basis using the `IndelRealigner` tool from the `GATK v3.3.0` software package [133].

3.2.3 Sample call-sets

Samples were divided into two partially overlapping sets:

1. **Crater lake set:** comprising all *A. calliptera* (crater lake and all other)
2. **Lake Malawi set:** comprising all Lake Malawi samples, crater lake *Rhampochromis* and *O. tetrastigma*, and all *Astatotilapia* except from crater Lakes Massoko and Itamba

Differences against the reference genome (variants) were determined (called) independently for these two sets.

3.2.4 Variant calling, filtering, genotype refinement, and haplotype phasing

Briefly, SNP and short indel variants against the *M. zebra* reference were called independently using `GATK v3.3.0` haplotype caller [134] and `samtools/bcftools v1.1` [135]. Variant filtering was then performed on each set of variants separately using hard filters based on overall depth, overall quality score, strand/mapping bias, and inbreeding coefficient (see below). Multiallelic sites were excluded. After filtering, I selected consensus sites (i.e. performed intersection of GATK and samtools sites). At a particular locus, if the GATK and samtools alleles differed, I kept the GATK allele. Finally, I used genotype likelihoods output by GATK at consensus sites to perform genotype refinement, imputation, and phasing in `BEAGLE v.4.0` [136]. The output of this process were filtered variants and phased genotypes against the *M. zebra* reference in the `VCF` format[2].

[2]Specification is available here: `https://github.com/samtools/hts-specs`

The particular commands and parameters used were:

samtools calling (multisample):

```
samtools mpileup -t DP,DPR,INFO/DPR -C50 -pm2 -F0.2 -ugf REFERENCE.fa SAMPLE1.bam
SAMPLE2.bam ...  | bcftools call -vmO z -f GQ -o samtools_VARIANTS.vcf.gz
```

GATK haplotype caller (per sample), later combined using GATK's `GenotypeGVCFs` tool:

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R REFERENCE.fa --emitRefConfidence
GVCF --variant_index_type LINEAR --variant_index_parameter 128000 -I SAMPLEn.bam
-o GATK_SAMPLEn.g.vcf
```

Hard filters applied to both datasets:

```
Minimal inbreeding coefficient:  -0.05

Minimum overall read depth:  600

Maximum overall read depth:  1700 (except for mtDNA: scaffolds 747,2036)
```

Hard filters applied to the GATK dataset:

```
Maximum phred-scaled p-value using Fisher's exact test to detect strand bias:
20 (except for mtDNA: scaffolds 747,2036)

Minimum accepted variant quality score:  300
```

Hard filters applied to the samtools dataset:

```
Minimum p-value for Mann-Whitney U test of Mapping Quality vs.  Strand Bias:
0.0001 (except for mtDNA - scaffolds 747,2036)

Minimum accepted variant quality score:  30
```

The consensus GATK and samtools call set was obtained using the `bcftools isec` tool:

```
bcftools isec -c indels -O z GATK_filtered_calls.vcf.gz samtools_filtered_calls.vcf.gz
-p GATK_samtools_intersect/
```

BEAGLE genotype refinement (per scaffold):

```
java -jar beagle.r1398.jar gl=GATK_samtools_consensus.vcf.gz phase-its=8 impute-its=8
out=beagle_GATK_sam_consensus
```

3.3 Coverage and cross-contamination estimates from data

As a part of preliminary quality control I used the `verifyBamID v1.0` [137] software to check for cross-contamination (whether the reads are contaminated as a mixture of two samples) and also to estimate genome coverage over filtered variant sites.

Cross-contamination can arise for example because multiple samples are processed at the same time and a small amount of tissue or extracted DNA is physically carried over from one sample to another, or when multiple samples are sequenced together on the same sequencing lane (this is known as 'multiplexing') and the sequencing machine is unable to decode correctly the 'tag' (short DNA sequence) that distinguishes the samples.

To check for cross-contamination, the software compares the original reads with the final variant calls and then: "Using a mathematical model that relates observed sequence reads to an hypothetical true genotype, verifyBamID tries to decide whether sequence reads match a particular individual or are more likely to be contaminated (including a small proportion of foreign DNA)".

Samples with `verifyBamID` estimated contamination >3% had considerable proportions of erroneously called variants in previous human whole-genome sequencing studies (R. Durbin, pers. comm.), so we excluded such samples from further analysis.

3.3.1 Lake Malawi call set

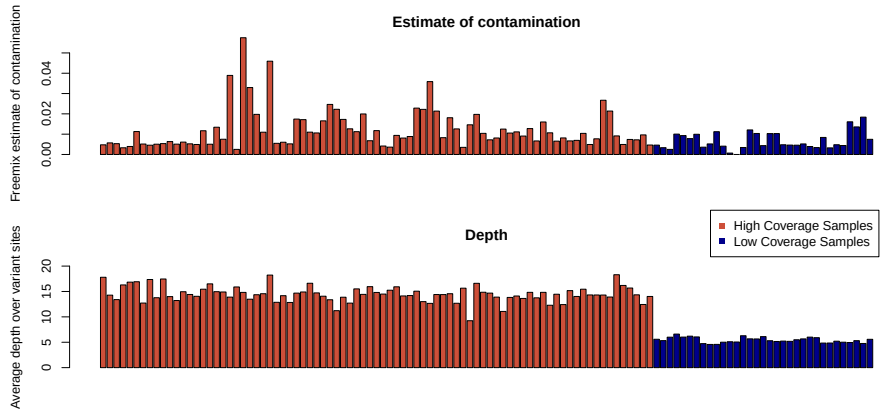


Fig. 3.3 **Cross-contamination and read-depth estimates for the Lake Malawi variant call-set.** Top: `verifyBamID` cross-contamination estimates. Bottom: read-depth over filtered variant sites.

The `verifyBamID` results for the Lake Malawi set of samples (as defined in Section 3.2.3) are shown in Figure 3.3. There are five high coverage (~15X) samples with estimated contamination scores >3%: *Aulonocara* 'blue chilumba' 5.7%, *Aulonocara stuartgranti* 'maisoni' 4.6%, *Alticorpus peterdaviesi* 3.9%, *Fossorochromis rostratus* 3.6%, *Aulonocara* 'gold' 3.2%. I eliminated these samples from all downstream analyses.

Direct estimates of read depth over filtered sites revealed that all samples were sequenced approximately to the intended coverage.

3.3.2 Crater lake call set

The `verifyBamID` results for the Crater lake set of samples (as defined in Section 3.2.3) are shown in Figure 3.4. Two samples have estimated contamination scores >3%: one from Lake Itamba 3.33% and one benthic individual from Lake Massoko 4.49%. I eliminated these two samples from all downstream analyses.

Direct estimates of read depth over filtered sites revealed that all samples were sequenced approximately to the intended coverage.

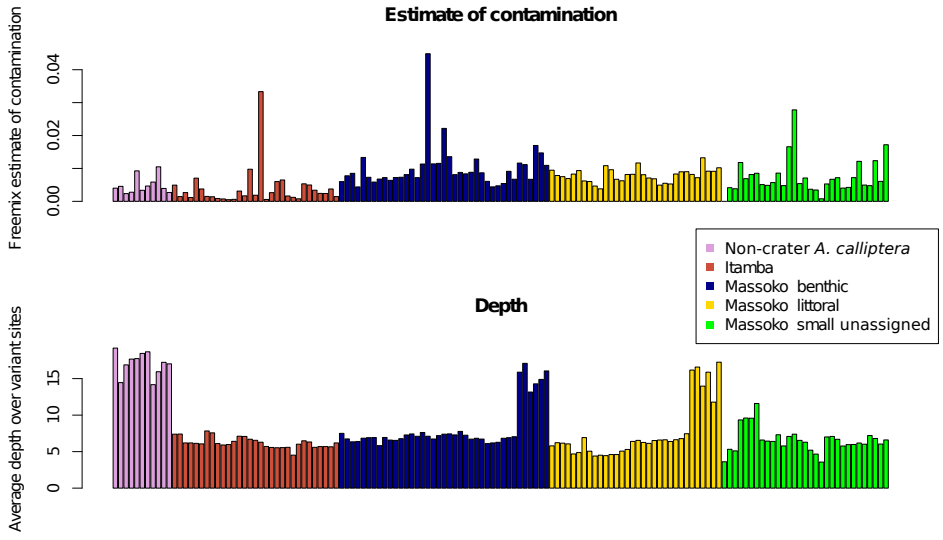


Fig. 3.4 **Cross-contamination and read-depth estimates for the Crater lake variant call-set.** Top: `verifyBamID` cross-contamination estimates. Bottom: read-depth over filtered variant sites.

3.4 Whole genome alignments

Overview

I generated a number of pairwise and multiple whole-genome alignments, following the 'UCSC paradigm' [138]. First I generated all possible pairwise alignments between the assemblies generated by the Cichlid Genome Consortium: the Lake Malawi *M. zebra* reference genome, the Lake Victoria *P. nyererei*, Lake Tanganyika *A. burtoni* and *N. brichardi*, and the Nile Tilapia *O. niloticus*. Then I added pairwise alignments between the genomes of these cichlids and the reference genomes of three other teleost species (medaka, stickleback, and zebrafish). Finally, I generated new contiguous genome-wide multiple alignments of these eight species in *M. zebra*, *P. nyererei*, *A. burtoni*, and *N. brichardi* genomic coordinates. A multiple alignment in *O. niloticus* coordinates is available from the Cichlid Genome Consortium [93].

Applications

Apart from being useful in their own right for studies of sequence evolution between the five cichlid species included, the alignments facilitate important analyses for my Lake Malawi population genomics study. Specifically, the alignments enable me to:

1. Distinguish ancestral vs. derived alleles at variant (segregating) sites in the Lake Malawi and Crater lake call sets
2. Assess long term evolutionary conservation of genomic regions of interest identified from the crater lake and Lake Malawi data
3. Use the Lake Victoria *P. nyererei* as a clear outgroup to root the phylogenetic tree of species within the Lake Malawi catchment
4. Find homologous sequences in zebrafish of regions of interest identified in the *M. zebra* genome, for example for follow-up functional studies

Producing (multiple) whole-genome alignments requires computational resources and expertise that are not available to a typical research group and was enabled by the strong computational facilities available at the Sanger Institute. Therefore, the multiple alignments in *P. nyererei*, *A. burtoni*, and *N. brichardi* genomic coordinates can be a valuable resource to research groups focussed on Lake Victoria and Lake Tanganyika cichlid radiations. Furthermore, the alignments will facilitate translation between genomic coordinates and thus enable comparisons between our findings based on alignment to the *M. zebra* genome, results produced by Lake Victoria researchers who use *P. nyererei* as the reference, and Lake Tanganyika results based on alignment to *A. burtoni* or *N. brichardi*.

Methods

Pairwise alignments of genome assemblies listed in Tables table:CichlidGenomes and table:alignGenomes were generated using `lastz v1.02` [139], with the following parameters:

For cichlid-cichlid alignments:

```
B=2 C=0 E=150 H=0 K=4500 L=3000 M=254 O=600 Q=human_chimp.v2.q T=2 Y=15000
```

For cichlid-other teleost alignments:

```
B=2 C=0 E=30 H=0 K=3000 L=3000 M=50 O=400 T=1 Y=9400
```

This was followed by using Jim Kent's `axtChain` tool with `-minScore=5000` for cichlid-cichlid and `-minScore=3000` for cichlid-other teleost alignments. Additional tools with default parameters were then used following the UCSC whole-genome alignment paradigm (`http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto`) in order to obtain a contiguous pairwise alignment.

Multiple alignment were generated from pairwise alignments using the `multiz v11.2` [140] program using default parameters and the following pre-determined phylogenetic tree: (((((((*M. zebra*, *P. nyererei*), *A. burtoni*), *N. brichardi*), *O. niloticus*), medaka), sticleback), zebrafish), in agreement with [93].

To obtain ancestral allele information for single nucleotide variants called against the *M. zebra* genome, indels were removed from the *M. zebra-P. nyererei* pairwise alignment (in *M. zebra* genomic coordinates), and ancestral allele information for variants filled into the `VCF` file using my custom C++ program `evo` (Available from `https://github.com/millanek/evo`) with the `aa-seq` and `aa-fill` options.

Table 3.4 **Versions of non-cichlid teleost assemblies used in whole-genome alignments.**

Species	UCSC version of assembly	URL used to download	Notes
medaka	oryLat2	ftp://hgdownload.soe.ucsc.edu/goldenPath/oryLat2/bigZips/oryLat2.fa.gz	NIG v1.0 assembly
stickleback	gasAcu1	ftp://hgdownload.soe.ucsc.edu/gbdb/gasAcu1/gasAcu1.2bit	Broad Institute v1.0
zebrafish	danRer7	http://hgdownload.cse.ucsc.edu/gbdb/danRer7/danRer7.2bit	Sanger Zv9 assembly

3.5 Cichlid genome browser

Introduction

Interactive visual exploration is a key component of research with genomics datasets, complementing computational approaches [141]. Genome browsers enable users to examine a portion of the genome and various annotation tracks (e.g. assembly gaps, genes, repetitive sequence annotation, multiple sequence alignments) at arbitrary scale, from individual DNA bases up to a whole genome view. Several genome browsers with Web interfaces have been developed, originally hosting data and annotations related to the Human Genome Project [142]. Two of the most popular websites, Ensembl [17] (`http://www.ensembl.org`) and the UCSC Genome Browser [143] (`https://genome.ucsc.edu`) have since grown to host reference genomes and annotations for 77 and 69 vertebrate species respectively. However, of the reference genomes generated by the Cichlid Genome Consortium, only the *O. niloticus* assembly was deemed to be of sufficiently high quality for inclusion in the Ensembl and UCSC browsers. BouillaBase (`http://bouillabase.org`) at University of Maryland hosts genomes and annotations, and data from the Cichlid Genome Consortium, but its capabilities are limited compared to Ensembl or UCSC browsers, and at the time of writing it is very slow - to the point that I found it virtually unusable in support of my research.

An alternative to Web based global services has emerged in the form of stand-alone genome browsers enabling exploration of genomics datasets on standard desktop computers [141]. Rather than remotely presenting a pre-defined set of genomes and data, desktop browsers display datasets on users' computers and thus are much more flexible. The Integrative Genomics Viewer (IGV) [144] is perhaps the most popular of these tools and I have used it a number of times during the PhD. However, even these tools have significant drawbacks, including limitations on the amount of data that can be loaded (visualised genomes and data sets generally must be to be loaded in RAM memory), the need to re-load all datasets every time the program is restarted, and the inability to share data with collaborators.

To overcome the above difficulties and enable high quality visualisation for all cichlid genomes and data generated during this PhD, I set up the Cambridge Cichlid Browser (CCB). The CCB site runs on the UCSC Genome Browser engine, offers the majority of its functions, and is currently hosted on a server computer at the Gurdon Institute in Cambridge: `http://cichlid.gurdon.cam.ac.uk`.

Datasets and functions

Cambridge Cichlid Browser (CCB) hosts reference genomes and annotations generated by the Cichlid Genome Consortium and multiple datasets generated during my PhD. The browser is driven by an underlying MySQL database and the total volume of data available at the moment is ~100GB. CCB offers the majority of function of the UCSC browser, reviewed in [145]. In addition to exploring the five genomes and their annotations with zoom and scroll functions, it is possible to search by specifying genomic coordinates, search for genes by name, and search for homologous regions to a DNA or protein sequence with BLAT [146]. Other useful functions include 'Table Browser' for access to the underlying database, 'In-silico PCR' for fast design of PCR primers, 'LiftOver' for quick translation of genomic coordinates between reference genomes (based on whole-genome alignments), and PDF output of browser graphics. Figure 3.5 displays a screenshot from the browser's *M. zebra* genome gateway.

Fig. 3.5 **Cambridge Cichlid Browser *M. zebra* genome gateway.**

Figure 3.6 shows an example of CCB graphics, showing ~900kb section of the 'scaffold 26' fragment of the *M. zebra* genome, with annotation tracks including assembly gaps, genes, and multiple alignment with cichlids and other teleosts.

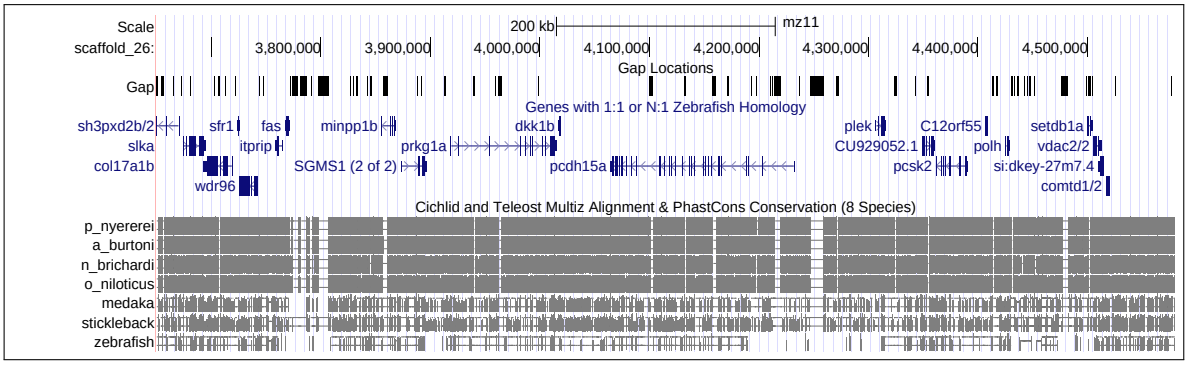


Fig. 3.6 **Cambridge Cichlid Browser - example browser graphic.**

Worldwide usage

The browser has been online since 1st March 2014 and has already made an impact beyond my immediate collaborators. Over 3,000 visits to the site have been recorded by the Google Analytics code I have linked to the website (`www.google.com/analytics/`). Limiting the statistics to session where users truly interacted with the browser (i.e. visited at least two pages, as opposed to just viewing the front page), there have been 657 browser sessions with a total of 19,607 page views initiated by 113 unique users (unique IP addresses). The average session duration has been 20 minutes and 37 seconds with an average 29.84 pages viewed per session.

The majority of the sessions have been initiated from the United States, UK, Germany, and Switzerland, but a number of other countries are also represented. Figures 3.7 and 3.8 show user locations and provide an insight into the user base. The locations correspond to major cichlid laboratories with a focus on genomics: for example, Craig Albertson's laboratory is currently located in Amherst, Massachusetts; Jeffrey Streelman's laboratory in Atlanta, Georgia; Russel Fernald's laboratory in Stanford; George Turner's laboratory in Bangor, Wales; and Axel Meyer's laboratory in Konstanz, Germany. In conclusion, these statistics make it clear that the Cambridge Cichlid Browser has already proven to be a valuable resource for the cichlid genomics research community.

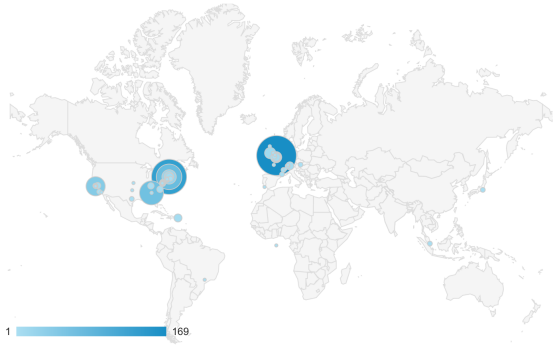


Fig. 3.7 **A map showing the location of CCB users.**

City	Sessions	Sessions	Contribution to total: Sessions
No spam	657 % of Total: 17.44% (3,767)	657 % of Total: 17.44% (3,767)	
1. Amherst	224	34.09%	
2. Cambridge	169	25.72%	
3. Atlanta	62	9.44%	
4. Stanford	42	6.39%	
5. London	29	4.41%	
6. Bangor	12	1.83%	
7. Greenfield	10	1.52%	
8. Konstanz	8	1.22%	
9. Manati	8	1.22%	
10. Ware	7	1.07%	

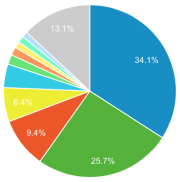


Fig. 3.8 **Ten cities with the highest contribution to CCB sessions.**

Chapter 4

De novo genome assembly

4.1 Introduction

All existing DNA sequencing technologies are are limited by short read lengths, orders of magnitude shorter than whole chromosomes (see e.g. [147]). All the DNA sequence data generated during this PhD is in the form of 100 or 125bp long reads (section 3.2.1). The goal of *de novo* genome assembly is to reconstruct the underlying genome sequence from the reads by finding and following overlaps between sequences. It is often helpful to visualise and conceptualise this process by drawing an 'assembly graph', where each vertex corresponds to a sequence and edges between vertices correspond to overlaps between reads, as in Figure 4.1. If error-free reads could be obtained from a single underlying genome sequence with randomly generated bases, the assembly task would be a trivial problem given 100bp read length. A 50bp overlap between two reads would then effectively guarantee that the two DNA sequences indeed originated from overlapping loci in the underlying genome sequence (with a negligible error rate of 7.8×10^{-31}). In the real world, however, genome assembly is a difficult problem. This is due to a combination of three factors: 1) DNA sequencing errors; 2) heterozygosity; 3) the highly repetitive nature of vertebrate genome sequences (see section 1.1.2).

When sequencing DNA from a diploid organism (this includes humans and the majority of animals), most DNA sequence is found in two copies, one contributed by the mother and the other by the father. Heterozygous sites, the differences between the maternally and paternally contributed sequences (chromosomes), are usually represented within the assembly graph topology as 'bubbles' (Figure 4.1).

Genome assembly typically involves three major stages: error correction, assembly, and *scaffolding*. First, the correction process attempts to eliminate the majority of sequencing errors. Second, sequence overlaps are found and sequences merged into

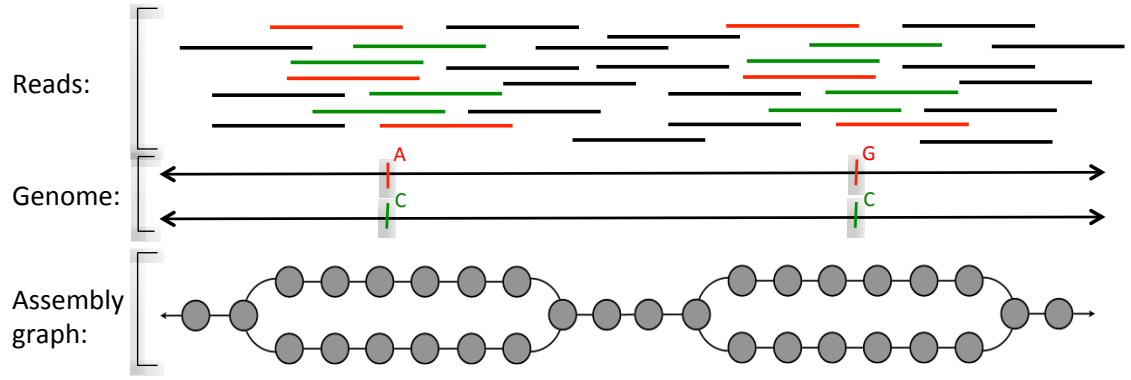


Fig. 4.1 **An example of genome assembly graph with two bubbles due to heterozygous sites**.

contiguous segments (*contigs*). During assembly, the assembly algorithms attempt to to resolve repeats and remove additional sequencing errors and heterozygous variants by examining the structure of the assembly graph. The contig ends if the underlying genome sequence cannot be resolved due to ambiguity (caused by repeats alone, or by any combination of repeats, errors, and heterozygous variants), or if there is not any read/sequence with sufficient overlap to add to the end of the contig. Third, genome assemblers use the long range information provided by paired-end or mate-pair reads whose inserts (see section 3.2.1) span unresolvable sequences. Thus, assemblers generate *scaffolds* - multiple contigs linked together and separated by gaps. Unresolvable sequence between the contigs is denoted by runs of the symbol `'N'` and insert sizes are used to determine approximate lengths of the unresolvable sequences.

The quality of a genome assembly is often evaluated based the distribution of lengths of assembled sequences (contigs and scaffolds). A commonly used summary statistic for assessment of contiguity of a genome assembly is N50. N50 is the length of the longest sequence s where half the length of the genome is assembled in sequences greater than or equal in length to s [148] (Figure 4.2).

When the true length of the genome is known (e.g. has been estimated by flow cytometry), it is also possible to assess the assembly by comparing the total length of the assembled fragments with the expected length of the genome. Finally, the options for assessing the accuracy for a *de novo* assembly are limited by the fact that in most cases the true answer is not known. Internal consistency of the assembly can be evaluated for example by assessing the consistency of alignment of paired-end reads to the completed assembly [148].

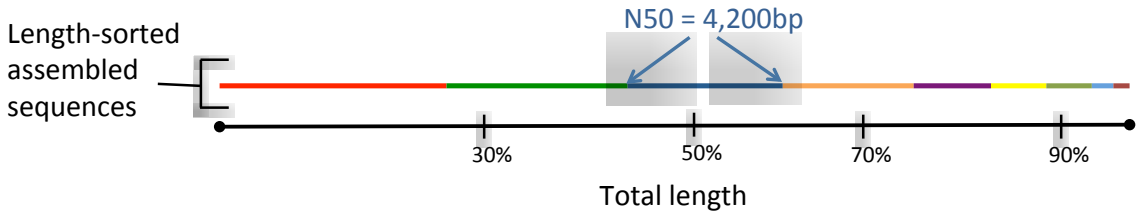


Fig. 4.2 **An illustration of the N50 measure of assembly contiguity**. Analogous measures can be defined to assess the length of assembled sequences covering a given proportion of the genome, e.g. N30 at 30%, or N70 at 70%.

I have completed all assemblies presented in this thesis using the `sga v0.10.13` genome assembler [149], and using my own extension `'trio-sga'` which takes advantage of mother-father-offspring trio sequence data to reduce problems associated with heterozygosity. The majority of sga's algorithms are based on efficient queries over an *FM-index*, a compressed index of a set of reads [150, 151]. Specifically, the FM-index data structure facilitates finding all occurrences of a sequence of length k (a k-mer; k must be shorter than the read length) in a set of reads in time that is independent of how many reads are searched.

Formally, let $\Re$ be a set of $|\Re|$ DNA reads, $\Re = \{R_0, R_1, ..., R_{|\Re|-1}\}$. Reads are indexed by $i = \{0, 1, 2, ..., |\Re| - 1\}$, and the length of each read is denoted by $|R_i|$. Then let Q be a query sequence of length $|Q|$, where $\exists i$ such that $|Q| \leq |R_i|$. Finding all occurrences of Q within $\Re$ by exhaustive search requires

$$\sum_{i=0}^{|\Re|-1} (|R_i| - |Q| + 1)$$

string comparisons. For example, searching for a 30-mer (k-mer of size 30bp) in the *Andinoacara coeruleopunctatus* dataset would require approximately 38billion comparisons. The FM-index based 'Backward Search' algorithm [150] counts all occurrences of a 30-mer within $\Re$ in 30 steps, regardless of the size of $\Re$. The number of CPU cycles in each step is comparable to the number of CPU cycles required for a single string comparison.

4.2 `trio-sga` - Trio-aware genome assembly

4.2.1 Overview

The majority of available genome assembly algorithms have been developed for organisms that are inbred, homozygous, or have low levels of heterozygosity and many genome projects reduced or even completely eliminated heterozygosity, by using inbred laboratory strains (e.g. the *C. elegans* genome project [152] and *D. melanogaster* genome project [153]), or by obtaining a homozygous form of the organism by other laboratory techniques (e.g. the potato genome [154]). However, in some cases such approaches are not feasible, for example if the organism in question is resistant to inbreeding due to strong inbreeding depression [155], or if the generation time is too long or the organism unsuitable for inbreeding in the laboratory. High heterozygosity can make genome assembly very challenging [156]. Algorithms specifically designed for organisms with moderate-to-high levels of heterozygosity have been developed [157, 158], but the methods still the lag in performance compared with assemblies of homozygous strains - simply due to the fundamental difficulty of assembling highly heterozygous genomes.

Communities of scientists have recently come together with the aims to *de novo* assemble the genomes of 10,000 vertebrate species [159], 5,000 arthropod species [160], and 7,000 (mainly marine) invertebrates [161]. Many species, especially in the latter two groups, will have high levels of heterozygosity.

Therefore, I have developed `trio-sga` - a set of three algorithms designed to facilitate better quality genome assembly for organisms with moderate-to-high levels of heterozygosity. `trio-sga` algorithms extend the `sga` genome assembler [149]. Two of the algorithms use haplotype phase information in mother-father-offspring trios to eliminate the majority of heterozygous sites even before the assembly itself (i.e. search for sequence overlaps) commences. The third algorithm is designed to reduce sequencing costs by enabling the use of parents' reads in the assembly of the genome of the offspring. In this section, I briefly describe `trio-sga`. In Section 4.3, I illustrate its performance by assembling highly heterozygous *Heliconius* butterfly genomes. Then, in Section 4.4, I demonstrate that the algorithms can improve assembly contiguity even at the lower levels of heterozygosity found in cichlids.

`trio-sga` software is available at `https://github.com/millanek/trio-sga`. It is written in C++, and can run multithreaded on UNIX-like systems. The core code implementing the logic of Algorithms 1 and 2 is in the file `TrioCorrectProcess.cpp` and the code for Algorithm 3 is in `FilterParentProcess.cpp`.

4.2.2 Algorithms

The input into trio-sga are three separate sets of DNA reads: reads from the mother, reads from the father, and reads from their offspring. trio-sga assembles the reads from the offspring, taking advantage of information present in the parents' reads with the aim to generate two separate assemblies of the offspring's genome: maternal (i.e. the haplotype inherited from the mother) and paternal (i.e. the haplotype inherited from the father).

The basic building block of all three trio-sga algorithms is a query over an FM-index about the number of occurrences of a given k-mer and of its reverse complement in a read set. For each trio we build three FM-indices (separately for the reads from the mother, father, and the offspring). K-mer occurrences in reads from the mother's DNA are denoted $\mathrm{C}_M(\mathrm{k})$, occurrences in reads from the father are denoted $\mathrm{C}_F(\mathrm{k})$, and from the offspring $\mathrm{C}_O(\mathrm{k})$. The three trio-sga algorithms are described below.

1. Pre-filtering the set of reads sequenced from the offspring in order to reduce heterozygosity. Two, usually overlapping, sets of reads are generated with the goal of eventually assembling the paternally and maternally contributed chromosomes separately. A conceptual overview of this algorithm, assuming error-free reads is in Algorithm 1 and Figure 4.3A.

Algorithm 1: Filtering the set of reads sequenced from the offspring in order to reduce heterozygosity (assuming error-free reads)

Data: FM-indices of mother and father reads; reads from the offspring
Result: Two partially overlapping sets of offspring reads for paternal and maternal haplotype assembly

```
foreach (read R from the offspring) do
    foreach (k-mer k in R) do
        if (C_F(k) > 0 and C_M(k) == 0) then
            assign R to paternal assembly read set; assigned = TRUE;
        end
        if (C_F(k) == 0 and C_M(k) > 0) then
            assign R to maternal assembly read set; assigned = TRUE;
        end
    end
end
if assigned = FALSE then
    assign R to both read sets
end
```

2. Improving error correction. Error correction used by `sga` and most other genome assemblers (e.g. ref [162]) is based on k-mer frequencies. It relies on the fact that the number of occurrences in the read set of error-containing k-mers is in general lower than the number of occurrences of k-mers that do not contain errors, i.e. the frequency distributions of correct and error-containing k-mers differ (Figure 4.3B and Figure 4.3C). In practice, an occurrence threshold is set to distinguish between correct and error-containing k-mers. However, in most data sets (depending on the error rate) there is a 'grey zone' where the two distributions overlap (Figure 4.3C), with low k-mer occurrences of correct sequence due to low coverage and/or non-random sampling and high k-mer occurrences of error-containing k-mers for example due to repeated (or systematic) errors. Using data from the parents helps to distinguish between error-containing and correct sequences within the grey zone and to prevent under-correcting (accepting as correct reads that contain errors) and over-correcting ('fixing' reads that are in fact correct). For example, if a k-mer fails the threshold in the offspring, but is present above threshold in one of the parents, it is unlikely to be an error and

correction is not attempted. Algorithm 2 outlines how parents' data are used in the decision on whether or not to attempt correction on a k-mer in the offspring.

Algorithm 2: Trio-aware error correction: deciding whether to attempt correcting a k-mer

Data: FM-indices of mother, father, offspring reads; reads from the offspring
Result: Corrected offspring reads

```
1  // Initialise occurrence thresholds for k-mers in offspring, mother, and
     father FM-Indices
   Init: set thresholds t_O, t_M, t_F;
2  // Initialise indicator variables for ensuring haplotype phase consistency of
     corrected reads
   Init: set mc=FALSE; set fc=FALSE;
3  foreach (read R from the offspring) do
4      foreach (k-mer k in R) do
5          if (C_F(k) < t_F and C_M(k) < t_M) then increase t_O;
6          // Test the offspring threshold
7          if (C_O(k) > t_O) then next;
8          else
9              // This k-mer failed offspring threshold - test it in the parents
10             if ((mc == fc) or (mc==TRUE and fc==FALSE)) then
11                 if (C_M(k) > t_M) then set mc.temp=TRUE;
12             end
13             if ((mc== fc) or (mc==FALSE and fc==TRUE)) then
14                 if (C_F(k) > t_F) then set fc.temp=TRUE;
15             end
16             if (mc.temp==TRUE or fc.temp==TRUE) then
17                 // Passed k-mer count threshold in the parental reads
18                 set mc = mc.temp; set fc = fc.temp; next;
19             end
20         end
21         // Call the correction algorithm (not shown)
22         correction(R,k);
23     end
24 end
```

Line 5: if a k-mer found in the offspring does not occur (above threshold) in either parent, it is likely to be an error (or a de-novo mutation, but these are exceedingly rare compared with errors). Therefore, I increase the offspring k-mer occurrence threshold for this k-mer.

Lines 16-19: It is necessary to ensure haplotype phase consistency in error-correction; for example, if a k-mer is not corrected thanks to passing the threshold in the mother (but not the father), I assume that the read comes from the maternal haplotype. I keep track of this information (i.e. set **mc**=TRUE) and only take the mother's reads into account when assessing k-mers from the remainder of the read.

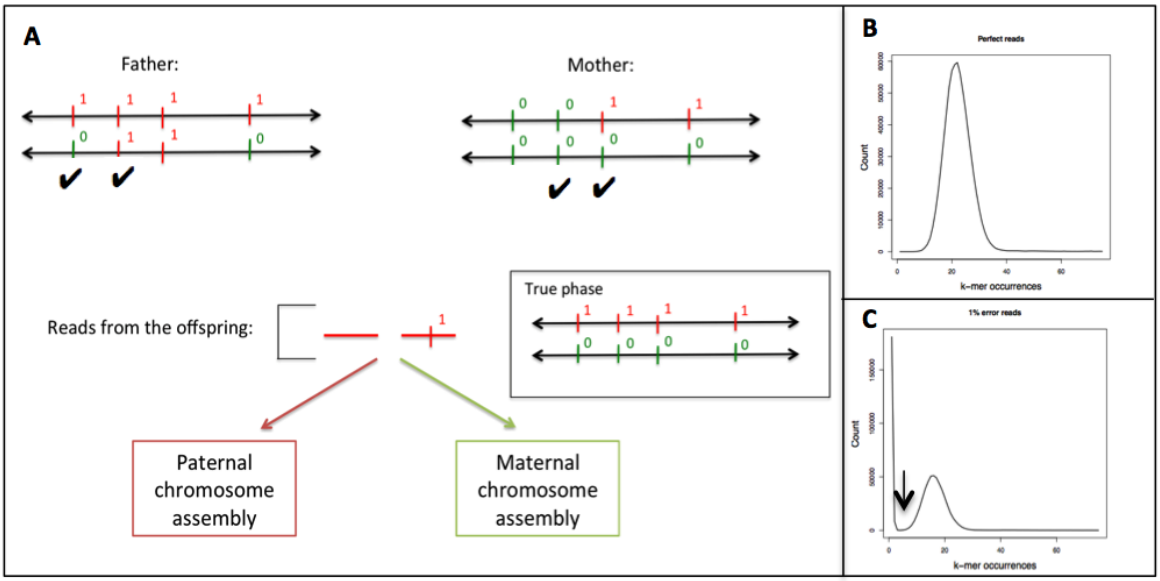


Fig. 4.3 **Trio aware read filtering and error correction.** **(A)** An example region of the genome with four segregating sites. The offspring inherited a haplotype with four derived alleles (denoted as 1) from the father and ancestral alleles from the mother. Read (or read pairs) from DNA containing the first or the second segregating site and the derived allele (as shown) can be phased and confidently assigned for paternal haplotype assembly. Similarly, reads from DNA containing the second or the third segregating site and the ancestral allele can be phased and confidently assigned for maternal haplotype assembly. **(B)** The distribution of 31-mer counts in simulated 100bp error-free reads with 30X genome coverage. **(C)** The distribution of 31-mer counts in simulated 100bp reads with uniform 1% error rate. There are now many more k-mers with low k-mer occurrences (<3); these are mainly errors, but there is a 'grey zone' (arrow), with kmers occurring 2-4 times being a mixture of correct and error-containing sequences. Figures **(B)** and **(C)** by Jared Simpson.

3. An algorithm to 'fill' regions of low coverage in the offspring by bringing in reads sequenced from the parents' DNA, thus using the parents' datasets to 'assemble through' these regions. Sequencing costs grow almost linearly with sequencing depth and in the trio assembly setting, three genomes need to be sequenced. This algorithm helps to keep the costs down. Reads from the father's DNA are checked for consistency with the offspring's paternal chromosome assembly and used to fill coverage gaps, and reads from the mother are used in the same way for the maternal chromosome assembly.

Algorithm 3: Check for consistency between error-corrected reads from the mother and error-corrected reads from the offspring's maternal haplotype or reads from the father and the paternal haplotype.

Data: FM-indices of error-corrected reads from one parent and the corresponding haplotype in the offspring
Result: Reads from the parent that are consistent with the offspring read set and can be used to bridge coverage gaps

```
1 foreach (read R from the parent) do
2 |   consistent=TRUE;
3 |   foreach (k-mer k in R) do
4 |   |   if (C_O(k) == 0) then
5 |   |   |   consistent=FALSE; break;
6 |   |   end
7 |   end
8 |   if (consistent) then mark R as consistent with the offspring;
9 end
```

4.3 *Heliconius* butterfly genome assemblies

I applied `trio-sga` to the deep coverage cichlid samples (Table 3.2 - Panel B), and will describe this work in section 4.4 below. However, the heterozygosity in these cichlid samples was not high enough for the trio approach to have a very large effect. To illustrate more dramatically the benefits of `trio-sga`, I will detour from cichlids to describe its application to *Heliconius* butterflies from South America, in collaboration with John Davey and Chris Jiggins from the Department of Zoology at Cambridge University.

With 43 species and a multitude of colour races that have radiated across the tropics of the South and Central America, *Heliconius* butterflies, like cichlids, provide outstanding opportunities to study a wide variety of evolutionary phenomena, including adaptation and speciation [163]. In common with other insects, such as *D. melanogaster*, *Heliconius* butterflies tend to have very large N_e and, therefore, very heterozygous

genomes [164, 165]. Sequencing libraries for three *Heliconius* mother-father-offspring trios were prepared and sequenced by the Sanger Institute sequencing core, obtaining 125bp paired-end reads with 300-500bp insert sizes.

The samples include:

1. A *Heliconius melpomene* trio
2. A *Heliconius cydno* trio
3. A cross between two *Heliconius* species: *H. cydno* mother, *H. melpomene* father, hybrid offspring

I obtained estimates of genome size and of heterozygosity using `sga preqc`, a recent extension of `sga` for estimating characteristics of a genome based on metrics derived from a random subset of reads [166]. The results (Figure 4.4) revealed that both *Heliconius* species species have genome sizes of ~280Mb, consistent with 292Mb flow cytometry estimate for *H. melpomene* [167]. The `sga preqc` genome size estimate for the hybrid offspring of the cross is ~560Mb, twice the true genome size. The erroneous estimate stems from the very high level of heterozygosity in the hybrid. On average, one in every 33bp is heterozygous in the hybrid. Heterozygosity in *H. cydno* is estimated to be $\sim\frac{1}{50}$bp and in *H. melpomene* $\sim\frac{1}{70}$bp. For comparison, heterozygosity in the cichlid *L. lethrinus* is estimated to be approximately an order of magnitude lower at $\sim\frac{1}{450}$bp.

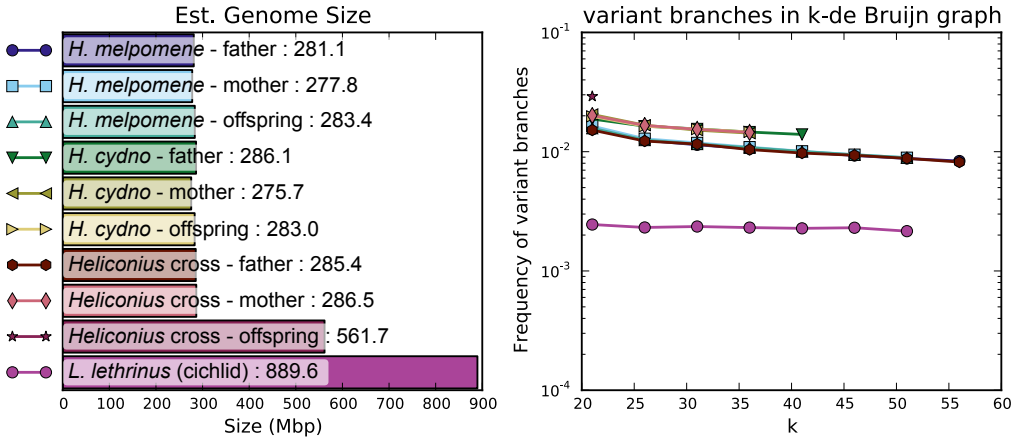


Fig. 4.4 **Estimates of genome sizes and heterozygosity for *Heliconius* genomes.** The cichlid species *L. lethrinus* has been included for comparison.

All read pairs with undetermined bases (`N` characters) were removed with the `sga preprocess` command. Then the FM-Indices were built and the `trio-sga` read filtering (phasing) and error-correction algorithms were called as follows:

```
sga correct-trio --paired --phase -x 3 -k 41 --mother-kmer-threshold=3
```

```
--father-kmer-threshold=3 offspring.fastq.gz mother.fastq.gz father.fastq.gz
```

The `trio-sga` read filtering (phasing) algorithm reduces heterozygosity in *Heliconius* data by approximately two orders of magnitude. Estimates by `sga preqc` show that the filtered datasets for *H. melpomene* and *H. cydno* have one heterozygous site approximately every 1700bp. Average heterozygosity in the hybrid was reduced even further to $\sim\frac{1}{3300}$bp (Figure 4.5).

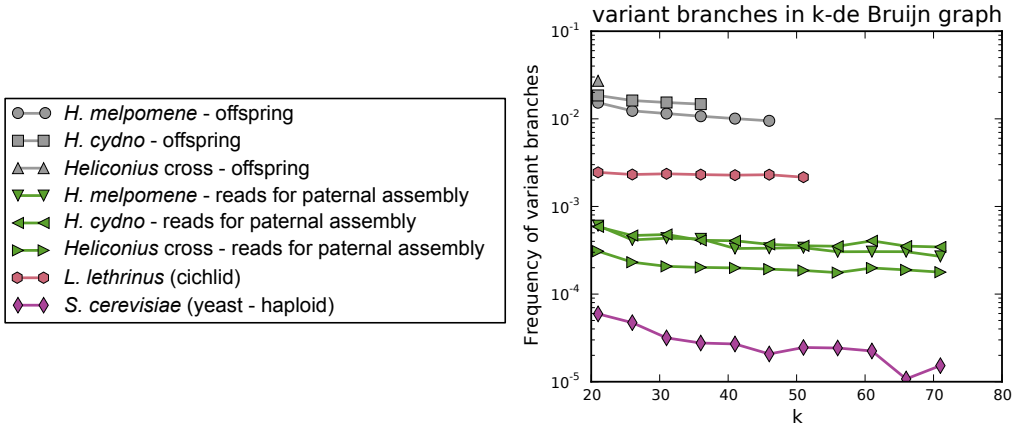


Fig. 4.5 **Read phasing reduces heterozygosity in *Heliconius* data.** Heterozygosities for the offspring samples (grey) and the cichlid *L. lethrinus* estimated from the same data as in Figure 4.4. The phased datasets (green; only paternal haplotype shown) have heterozygosity estimates up to two order of magnitude lower. Heterozygosity values for the haploid yeast species represent a misclassification rate ($<10^{-4}$) observed in `sga preqc` estimates [166].

Reads from the father consistent with the paternal haplotype were obtained by first error-correcting the father reads independently with `sga` and then using the following `trio-sga` command:

```
sga filter-parents --do-not-correct --paired -k 41 father_corrected.fastq.gz
offspring_paternal.fastq.gz
```

The reads were merged with the paternal haplotype reads, and then assembled into contigs using `sga`. This will be referred to as *trio assembly* in the rest of this section. I also assembled the offspring reads without using the parent's data. This will be referred to as *normal assembly*. In all cases, the `-r 10` parameter was used for the `sga assemble` subprogram, and assemblies were attempted with minimum overlap required between reads set to 70, 80, 90, 95, 100, 105, and 110bp. Then, I choose the assembly with the highest contig N50 statistic.

Table 4.1 ***Heliconius* contig assembly statistics.** Contigs of less than 500bp excluded.

Species	Assembly	Best minimum overlap (bp)	Total length (Mb)	Contiguity stats (bp) N30	N50	N70	N90
H. melpomene	normal	80	367	1,908	1,258	879	618
H. melpomene	trio	95	269	12,185	7,626	4,181	1,225
H. cydno	normal	80	369	1,456	1,030	773	586
H. cydno	trio	90	263	13,516	8,675	4,951	1,467
Heliconius cross	normal	70	432	1,678	1,175	848	611
Heliconius cross	trio	90	259	17,297	11,259	6,844	2,731

Assembly statistics in Table 4.1 demonstrate that normal assemblies of the highly heterozygous *Heliconius* genomes are very challenging for `sga`. Contig N50 statistics are barely above 1kb and the total length of the assemblies is greater than the genome size estimate, suggesting that in many cases two copies of a single genomic region have been retained. In contrast, trio assemblies have contig N50 between 7.5 and 11.2kb and assembly lengths correspond to genome sizes.

Given that the normal assemblies were clearly very poor, I generated scaffolds only for the trio assemblies. The paired-end reads used for the trio assembly were aligned to the contigs (excluding contigs of less than 200bp) using `bwa mem v0.7.10` [132], and alignments processed by `samtools v1.1` [135] to generate sorted `bam` files as follows:

```
bwa mem -p -M contigs.fa reads.fa.gz | samtools fixmate -O sam - - | samtools
view -b -h -F 256 - > alignment.bam

samtools sort -@ 4 -T temp -O bam -o alignment_sorted.bam alignment.bam
```

and scaffolds then generated with the requirement for evidence from at least five pairs of reads before joining two contigs:

```
sga-bam2de.pl --prefix n5 -n 5 -m 200 alignment_sorted.bam

sga-astat.py -m 200 alignment_sorted.bam > alignment_sorted.astat

sga scaffold -m 200 -a alignment_sorted.astat --pe n5.de -o n5_scaffolds contigs.fa

sga scaffold2fasta --write-unplaced -m 200 -o n5_scaffolds.fa --use-overlap
-a contigs-graph.asqg.gz n5_scaffolds

sga gapfill -o n5_scaffolds_gapfilled.fa --prefix=reads n5_scaffolds.fa
```

Scaffold assembly statistics are shown in Table 4.2. Again, all the results are for paternal haplotypes. The paternal haplotype in the cross comes from the *H. melpomene* father.

Table 4.2 ***Heliconius* scaffold assembly statistics.** Scaffolds of less than 500bp excluded.

Species	Assembly	Total length (Mb)	Scaffold number	Gaps	Contiguity stats (bp) N30	N50	N70	N90
H. melpomene	trio	273	44,740	14,944	31,535	19,406	10,276	2,408
H. cydno	trio	267	35,196	13,286	37,454	23,575	12,900	3,472
Heliconius cross	trio	260	22,284	10,331	45,640	29,456	17,738	6,802

It is interesting that in both contig and scaffold assemblies, the best (most contiguous) trio assembly is for the cross, followed by *H. cydno* and then *H. melpomene*. This pattern suggests that the more heterozygous the individual/species, the better trio assembly can be achieved.

The first set of *Heliconius* data was sequenced in February 2015, with average genome coverage ~40-50X per individual (~120-150X per trio). The coverage of the offspring reads drops when the `trio-sga` read phasing algorithm divides reads into two sets, with the magnitude of the drop depending on how many reads can be phased (the drop was 47% for the cross, 39% for *H. cydno*, and 34% for *H. mepomene*). Coverage is later recovered when reads from the father consistent with the paternal haplotype and reads from the mother consistent with the maternal haplotype are brought in. Nevertheless, I was concerned that the drop in coverage caused by `trio-sga` read phasing algorithm could lead to breaks in the assembly caused by insufficient coverage. Therefore, we sequenced more *Heliconius* DNA from the same samples in May 2015, doubling the coverage per individual to ~80-100X.

After doubling the coverage, I obtained trio assemblies in the same way as described above. Table 4.3 - Panel A lists statistics for the contig assemblies. The increase in coverage enabled me to increase minimum required overlap between reads from 90-95bp to 105-110bp. However, improvements in contig lengths have been small. The N50 of the paternal contigs increased from 8.7 to 9.3kb (~7%) for *H. cydno* and from to 11.3 to 12.7kb (~12.4%) for the cross. The *H. melpomene* paternal contigs could not be compared because of technical problems with this assembly. So far, two scaffold assemblies with high coverage data have been finished (Table 4.3 - Panel B). The N50 of the paternal scaffolds of the cross increased from 29.4 to 33.4kb, an increase of ~13.6%.

Table 4.3 ***Heliconius* assembly statistics - high coverage data.** Only trio assemblies were generated.

Panel A: Contig assemblies; contigs of less than 500bp excluded.

Species	Haplotype	Best minimum overlap (bp)	Total length (Mb)	Contiguity stats (bp) N30	N50	N70	N90
H. melpomene	maternal	110	263	12,909	7,842	4,170	1,161
H. cydno	maternal	105	253	14,667	9,501	5,586	1,813
H. cydno	paternal	110	268	14,577	9,344	5,306	1,535
Heliconius cross	maternal	105	263	19,278	12,457	7,454	2,744
Heliconius cross	paternal	105	263	19,460	12,674	7,614	2,905

Panel B: Scaffold assemblies; scaffolds of less than 500bp excluded.

Species	Haplotype	Total length (Mb)	Scaffold number	Gaps	Contiguity stats (bp) N30	N50	N70	N90
H. cydno	maternal	258	32,547	16,069	42,515	27,047	15,281	3,898
Heliconius cross	paternal	265	22,370	11,513	51,986	33,374	19,980	7,343

4.4 Cichlid trio genome assemblies

As shown previously in Figure 4.4 and discussed in more detail in chapter 5, the levels of heterozygosity in cichlids are approximately an order of magnitude lower than in *Heliconius* butterflies. Therefore, it was interesting to see if reducing heterozygosity using `trio-sga` can deliver improvements in cichlid genome assemblies. Deep coverage cichlid samples (Table 3.2) were assembled using both the normal and trio methods as described above for *Heliconius*. Because the read-length was 100bp, minimum overlap required between reads was set to 65, 70, 75, and 80bp.

Cichlid contig assembly statistics are shown in Table 4.4. Compared with normal `sga` assemblies, using trio data with `trio-sga` algorithms increases contig N50 by 35% to 45% in all three cichlid species.

Table 4.4 **Cichlid contig assembly statistics.** Contigs of less than 500bp excluded.

Species	Assembly	Haplotype	Best min. overlap	Length (Mb)	Contiguity stats (bp) N30	N50	N70	N90
A. calliptera	normal	---	70bp	654	4,532	2,980	1,867	930
A. calliptera	trio	maternal	80bp	681	6,361	4,118	2,540	1,159
A. calliptera	trio	paternal	80bp	681	6,381	4,153	2,564	1,167
A. stuartgranti	normal	---	70bp	656	4,842	3,125	1,935	949
A. stuartgranti	trio	maternal	75bp	677	6,579	4,222	2,571	1,164
A. stuartgranti	trio	paternal	75bp	679	6,644	4,253	2,588	1,167
L. lethrinus	normal	---	65bp	640	4,054	2,691	1,723	890
L. lethrinus	trio	maternal	75bp	673	5,873	3,852	2,407	1,126
L. lethrinus	trio	paternal	75bp	673	5,888	3,861	2,410	1,130

Scaffold assemblies with paired-end reads were obtained in the same way as described above for *Heliconius*. Paired-end scaffold assembly statistics are shown in Panel A of Table 4.5. The statistics reveal that the most of `trio-sga` N50 improvements at the contig level are carried forward to the paired-end scaffolds, with increases of 25% to 35%. At this stage, with N50 of ~9-12kb, the cichlid assemblies are much more fragmented than the corresponding *Heliconius* assemblies which have N50 of ~20-33kb. This may be partly due to lower genome coverage and shorter read lengths, but also probably due to more complex repeat structure of cichlid genomes.

For scaffolding with mate-pair reads, I aligned them to the paired-end scaffolds using `bwa mem v0.7.10` [132], processed the alignments by `samtools v1.1` [135] to generate sorted `bam` files as follows:

```
bwa mem -p -M paired-end-scaffolds.fa mate-reads.fa.gz | samtools fixmate -O
sam - - | samtools view -b -h -F 256 - > mate-alignment.bam

samtools sort -@ 4 -T temp -O bam -o mate-alignment_sorted.bam mate-alignment.bam
```

and then generated mate-pair scaffolds with the requirement for evidence from at least three mate-pairs before joining two paired-end scaffolds:

```
sga-bam2de.pl --prefix n3-mate -n 3 -m 200 mate-alignment_sorted.bam
```

```
sga-astat.py -m 200 mate-alignment_sorted.bam > mate-alignment_sorted.astat

sga scaffold -m 200 -a mate-alignment_sorted.astat --mate n3-mate.de
-o n3_mate-scaffolds n5_scaffolds_gapfilled.fa

sga scaffold2fasta --write-unplaced -m 200 -o n3_mate-scaffolds.fa
-f n5_scaffolds_gapfilled.fa n3_mate-scaffolds

sga gapfill -o n3_mate-scaffolds_gapfilled.fa --prefix=reads n3_mate-scaffolds.fa
```

The effect of repeat structure in cichlid genomes of assembly contiguity becomes apparent when mate-pair reads with larger insert sizes are used for generating scaffolds. The contiguity of mate-pair scaffolds of the trio assemblies (Table 4.5 - Panel B), as measured by N50, is approximately 5-fold better than when scaffolding with paired-end reads alone. Cichlid mate-pair insert sizes are generally around 1-2kb; full distributions are shown in Figure 4.6. It is therefore clear that the long range information present in mate-pair reads was enabled the assemblies to span over a large number of repeat sequences in this size range. It is also interesting to note that *A. stuartgranti*, the species with the longest mate-pair inserts, has the best mate-pair scaffold N50.

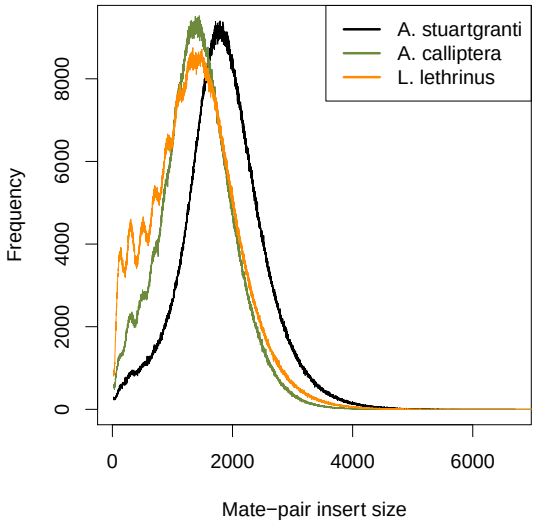


Fig. 4.6 **The distributions of cichlid mate-pair insert sizes**

Table 4.5 **Cichlid scaffold assembly statistics.** Scaffolds/contigs < 500bp excluded.

Panel A: Scaffolds assembled with paired-end reads (300-500bp insert size).

A. calliptera	normal	---	Not assembled				
A. calliptera	trio	maternal	690	19,720	12,155	6,811	2,432
A. calliptera	trio	paternal	689	20,075	12,310	6,876	2,428
A. stuartgranti	normal	---	671	13,679	8,485	4,855	1,835
A. stuartgranti	trio	maternal	686	17,579	10,838	6,136	2,251
A. stuartgranti	trio	paternal	687	17,781	10,879	6,115	2,233
L. lethrinus	normal	---	655	11,039	6,980	4,089	1,601
L. lethrinus	trio	maternal	679	15,034	9,364	5,355	2,006
L. lethrinus	trio	paternal	679	15,034	9,391	5,374	2,008

Panel B: Scaffolds after adding mate-pair reads (~1-2kb insert size).

A. calliptera	trio	maternal	720	97,843	54,823	26,262	5,366
A. calliptera	trio	paternal	719	98,689	55,879	26,524	5,366
A. stuartgranti	trio	maternal	732	104,746	59,945	29,520	5,985
A. stuartgranti	trio	paternal	733	102,317	59,588	29,285	5,801
L. lethrinus	trio	maternal	722	90,820	51,728	25,624	5,020
L. lethrinus	trio	paternal	722	90,316	51,956	25,647	5,038

4.5 *Andinoacara coeruleopunctatus* genome assembly

The genome of the Central American cichlid *Andinoacara coeruleopunctatus* was assembled using normal `sga` pipeline as described in Section 4.3. Assembly statistics are shown in Table 4.6. The total length of the assembly is similar to East African cichlids, while assembly contiguity is better, likely due to higher coverage (~60X vs. ~40X) and longer read lengths (125bp vs 100bp).

Table 4.6 ***A. coeruleopunctatus* assembly statistics.** Scaffolds/contigs < 500bp excluded.

Panel A: Contig assemblies.

Species	Best minimum overlap (bp)	Total length (Mb)	Contiguity stats (bp) N30	N50	N70	N90
A. coeruleopunctatus	90	687	7,780	5,123	3,219	1,487

Panel B: Paired-end scaffold assemblies.

Species	Total overlap (bp)	Contiguity stats (bp) N30	N50	N70	N90
A. coeruleopunctatus	699	26,236	16,890	10,174	4,132

Figure fig:PanamaScaffoldDist shows the full distribution of scaffold lengths in the *A. coeruleopunctatus* assembly.

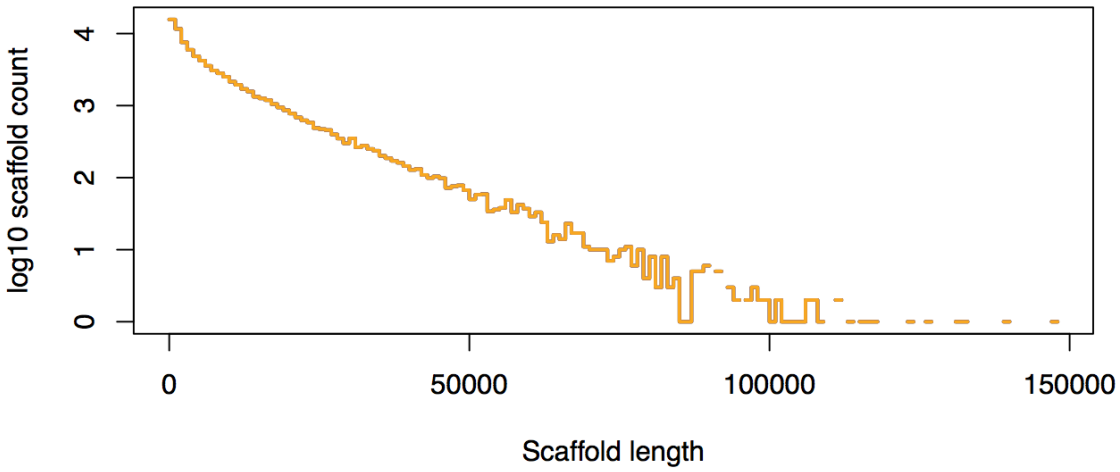


Fig. 4.7 **Distribution of scaffold lengths in *A. coeruleopunctatus* assembly.**

Chapter 5

Lake Malawi genetic diversity

5.1 Introduction

The Lake Malawi radiation has generated over 500 species in less than five million years, involving divergence in habitat use, size, shape, diet, feeding apparatus, sex determination, male breeding colour, and many other traits [83]. These phenomena present an opportunity to observe hundreds of varied, recent, and in some cases ongoing, speciation events at different stages along the speciation continuum, ranging from newly forming varieties to pairs of well defined species with very strong or complete reproductive isolation. Studying hundreds of 'snapshots' of speciation at different stages along the continuum, all in a similar genetic background, will give us unprecedented insight into the genomic patterns of divergence that underlie the build-up of reproductive isolation this system. We should be able to see whether the genomic patterns are consistent or vary between sub-groups of species, and also if there are genes or genomic regions that tend to be involved repeatedly and thus qualify as 'speciation genes' in cichlids.

Genomic divergence continues to accumulate after speciation is complete, both due to neutral processes and due to continued adaptive evolution in the newly formed species. In fact, reproductive isolation may not be sufficient for achieving a speciation event of lasting effect; it is a widely held view that in order for species to co-exist, they must achieve sufficient ecological differentiation to reduce direct competition [168, 169, 170, 171] (for an opposing view and evidence see [172]). Therefore, post-speciation adaptive evolution may be as important as pre-speciation divergence for generating and maintaining species diversity, and can be observed in Lake Malawi by comparing older fully-isolated species.

Overall, Lake Malawi cichlids promise to provide a rich picture of genome evolution on a timescale that bridges the gap between micro- and macro-evolutionary studies, with a particular focus on genetic basis of functional diversification. Thus, this study contributes to our understanding of how vertebrate genomes evolve and function and the long term impact of knowledge of genome function includes implications for animal and human health.

Previous studies demonstrate the utility of Lake Malawi cichlids in addressing major questions in evolutionary ecology and genomics, but only scratch the surface, focussing on a limited number of species and/or genes. The progress of research in the large cichlid radiations of lakes Malawi, Tanganyika, and Victoria has been hampered by difficulties in identifying species relationships, in reconstructing past geographical situations, and in controlling for possible introgression from non-sister taxa [86]. Therefore, a thorough characterisation the genomic diversity of a complete large adaptive radiation, and reconstructing of its evolutionary history, including the relative timing, frequency and sequence of evolution of major adaptive innovations are fundamental steps for making this complex and fascinating system tractable for in-depth investigation.

Here I describe initial analysis of the Lake Malawi samples collected in Spring 2013 and Autumn 2014 as listed in chapter 3. This will be extended by adding further data from Summer 2015 and additional analysis to form the basis of a future publication.

5.2 Genomic diversity

Variant calling for the Lake Malawi samples (as defined in section 3.2.3) against the *Metriaclima zebra* reference genome, from which divergence was 0.2-0.3% (0.9% for *A. rujewa*), resulted (after filtering) in 20,673,877 SNPs and 2,859,560 short insertions and deletions (1.2 - 1.8 million variants per individual, except the outgroup *A. rujewa* with 5.7 million variants; Figure 5.1).

I used the frequency of heterozygous sites as a simple summary statistic to estimate within-species nucleotide site diversity (π). Heterozygosity is indicative of long-term effective population size (N_e) over the past of order N_e generations [173]. As shown in Figure 5.2, π in all Lake Malawi species is within a relatively narrow range between ~1/2,000 and ~1/700 (except for the *A. calliptera* sample from Kitai Dam which has a much lower π estimate of ~1/12,000, presumably due to a strong population bottleneck/founder effect).

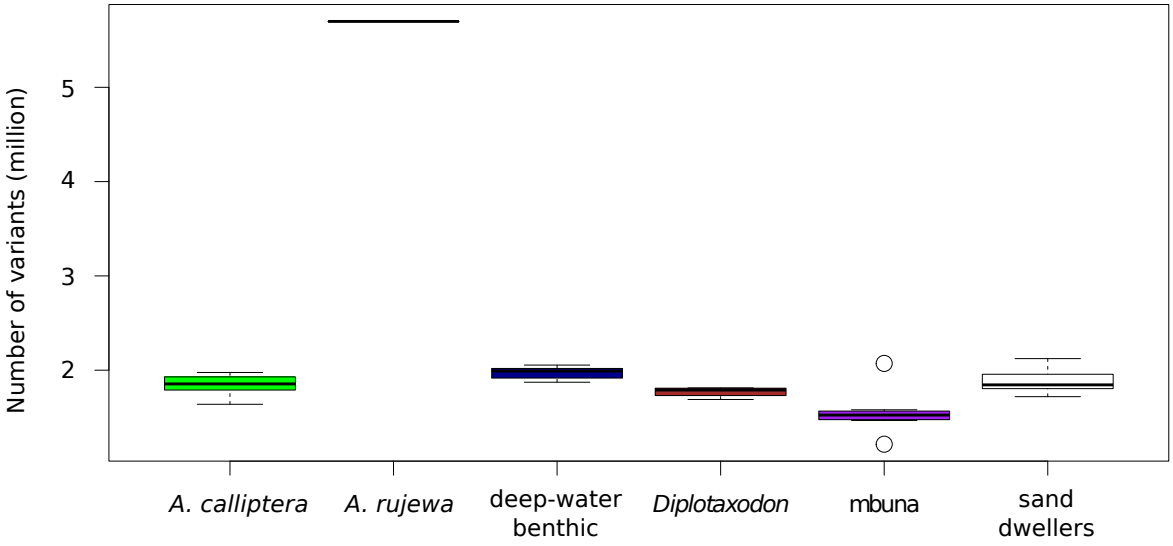


Fig. 5.1 **Variants called against *M. zebra* genome in Lake Malawi samples.**

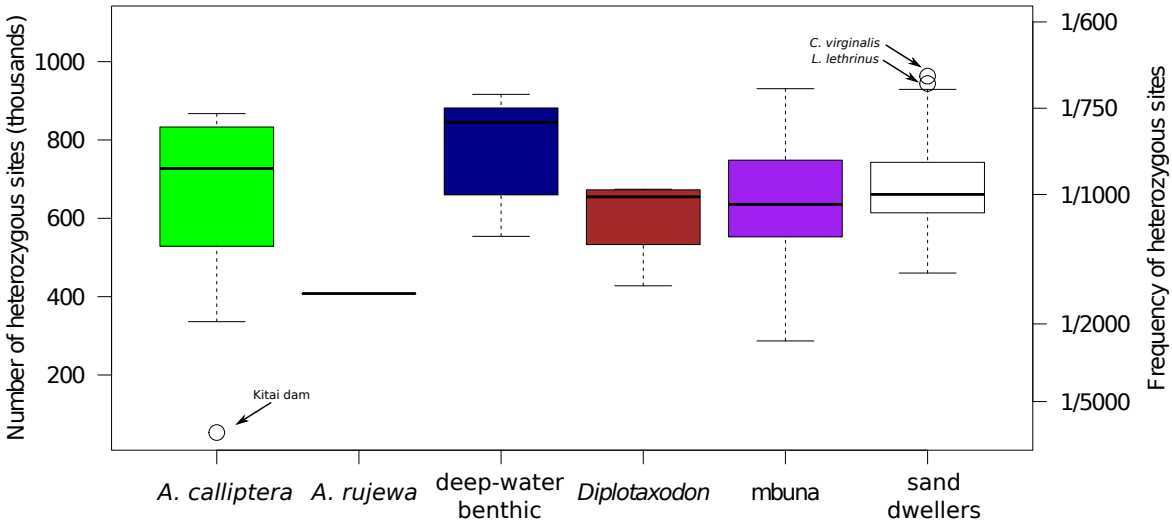


Fig. 5.2 **Frequency of heterozygous sites in Lake Malawi samples.**

The relationship between π and N_e in an ideal population (as defined in section 1.1.3) is:

$$\pi = 4N_e\mu \tag{5.1}$$

where μ is the per-generation mutation rate.

A direct estimate of μ in cichlids is not available. However, if we make the assumption that cichlid μ is similar to human ($\mu \approx 1.5 \times 10^{-8}$ [15]), we can obtain long term N_e estimates between ~7,200 and ~24,300 (~1,300 for the *A. calliptera* sample from Kitai Dam).

To complement measures of within-species genetic diversity (π), I also estimated between-species divergence by calculating the F_{ST} statistic for all pairs of species. The results, shown in Figure 5.3, revealed that:

1. *Diplotaxodon macrops* and *Diplotaxodon* 'macrops black dorsal', a putative new species studied by Genner *et al.* [128], are genetically virtually undistinguishable, at least in terms of genome-wide average F_{ST}. It is possible that their genetic differentiation is limited to a very small number highly diverged genomic loci, as seen for example between German carrion and Swedish hooded crows [77].
2. F_{ST} between the putative outgroup species *A. rujewa* and the rest of the Lake Malawi species varies from 0.629 to 0.857, suggesting that some shared variation remains between *A. rujewa* and Lake Malawi, despite the relatively high level of divergence.
3. Excluding the two special cases above, F_{ST} between species within the Lake Malawi sample set varies between 0.036 to 0.661. The lower value is between *Diplotaxodon limnothrissa* and *Diplotaxodon macrops*, and similar values are also estimated for other pairs of species (e.g. *Copadichromis virginalis* vs. *Copadichromis quadrimaculatus* and *Ctenopharynx intermedius* vs. *Ctenopharynx nitidus*). Such F_{ST} levels are virtually the same as between the two ecomorphs of Lake Massoko described in chapter 6, suggesting that for example the divergence between the two sympatrically occurring *Diplotaxodon* species could be studied with a similar approach as applied for Massoko. At 0.661, the highest F_{ST} estimate is similar to divergence between two wild and phenotypically virtually undistinguishable zebrafish strains from southern India and Bangladesh (F_{ST}=0.64) [174].

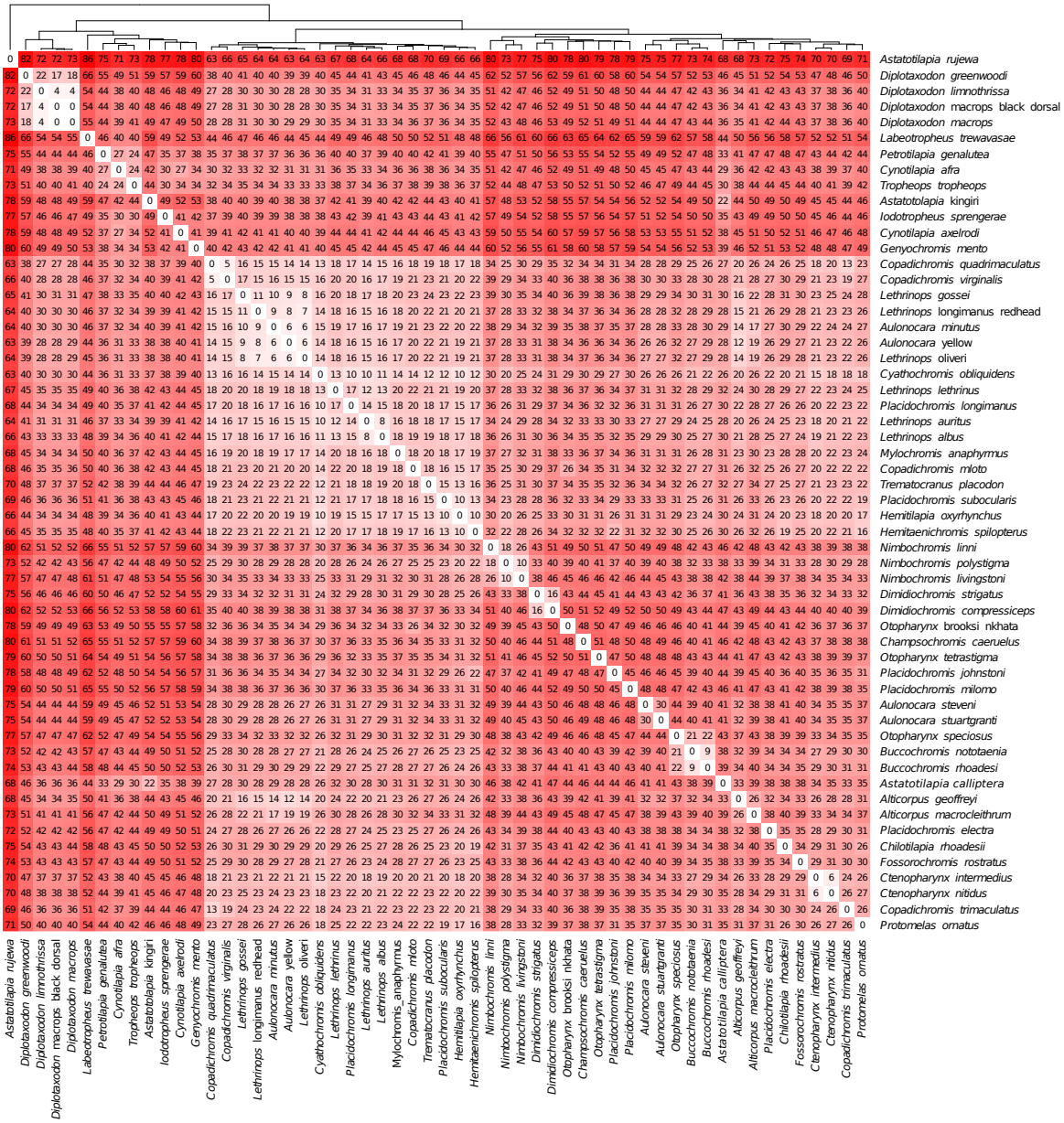


Fig. 5.3 **F_{ST} divergence between Lake Malawi species.** The tree (top) corresponds to simple hierarchical clustering of F_{ST} values by the `hclust` method in `R` [175]. F_{ST} values are given as percentages (%).

For the majority of Lake Malawi species, we have not sequenced enough individuals to study the genome-wide pattern of genetic differentiation and be able to identify loci that are under positive selection. Among exceptions to the above are *Placidochromis subocularis* and *Trematocranus placodon*. To test whether a genome scan for high F_{ST} outliers could be informative in this case, I calculated F_{ST} divergence between eight *P. subocularis* individuals and five *T. placodon* in non-overlapping sliding windows of 100 variants each.

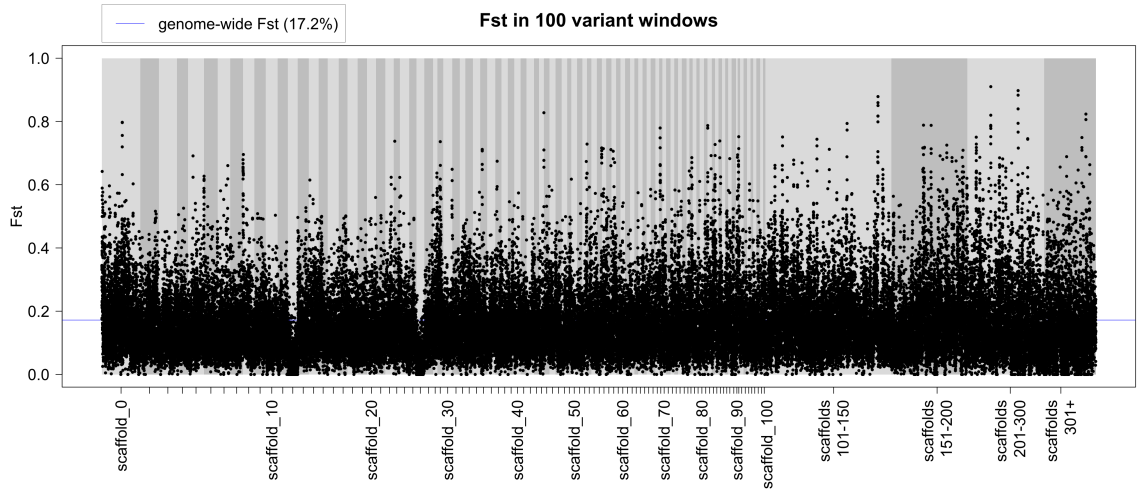


Fig. 5.4 **Genome-wide F_{ST} profile between *P. subocularis* and *T. placodon*.**

The results (Figure 5.4) suggest that neutral divergence between these two species, as measured by F_{ST}, is sufficiently low to allow for detecting selection. Also interesting is the pronounced dip in F_{ST} on scaffolds 12 and 26, which could suggest for example balancing selection, positive selection on the same ancestral variant in both species, or recent introgression between the two species. Nevertheless, it is worth keeping in mind that the numbers of individuals used in the analysis are likely too low to be able to reliably distinguish selection at specific loci from other causes of variation

Finally, to obtain an overview of the Lake Malawi dataset which includes both within- and between-species diversity I performed Principal Component Analysis (PCA). Figure 5.5 shows the position of each specimen along the three main principal components (three main axes of variation) which together explain 15.13% of the total genetic variation in the dataset. Within the space of this PCA projection, all specimens collected from the same species form closely knit clusters, as do some other groups comprised of several species: namely the mbuna, the genus *Diplotaxodon*, and all the deep water benthic specimens. Almost all shallow water sand-dwellers are located on a line along the second and third eigenvectors, except *Copadichromis virginalis* and

C. quadrimaculatus, both of which have evolved a more plankton feeding rather than bottom dwelling habit, and the shallow water members of the genus *Aulonocara*, who also are not typical sand-dwellers in that they prefer the intermediate habitat at the interface between sandy and rocky bottom areas.

The arrangement of species along a line in a PCA plot, as for the sand dwellers, is suggestive of patterns seen in modern human populations with varying degrees of admixture, such as Native Americans with European admixture. However, in this case we are considering separated species which may perhaps be subject to hybridisation and introgression at a low rate, rather than subpopulations of one species which have admixed due to population migrations in recent generations. I will return to evidence for hybridisation later in section 5.4.

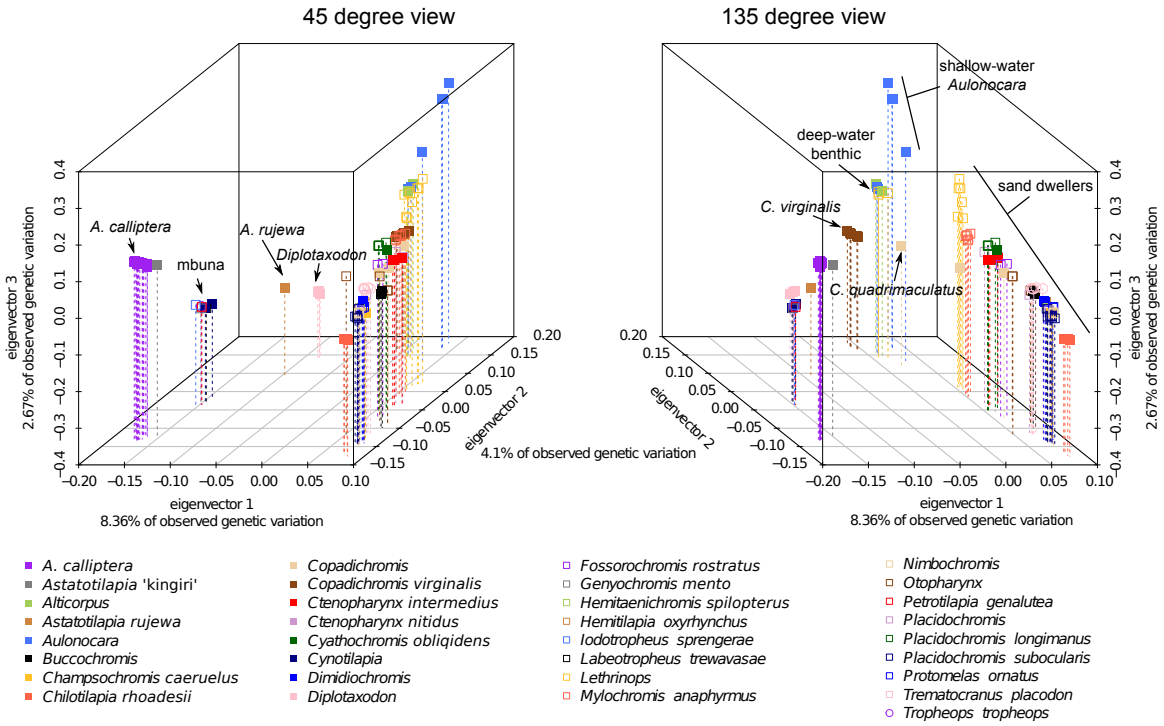


Fig. 5.5 **Principal Component Analysis of genetic variation in the Lake Malawi sample set.** Two views of the same 3D plot are shown, projecting the position of each sample along the first three principal components.

5.3 Phylogenetic relationships

The relative position of individuals in the PCA plot has given us some indication of the relationships between them. As a first step in trying to reconstruct the evolutionary

history or the Lake Malawi radiation, I used all the single nucleotide variants in the genome (except mtDNA) to build a 'consensus' phylogenetic tree with all the individuals in the sample set.

The tree (Figure 5.6), rooted using a whole genome alignment with the Lake Victoria *P. nyererei* (see section 3.4) reveals that:

1. All individuals assigned to the same species form monophyletic clades, strengthening the evidence for them being 'good' species and suggesting that, when averaging across all genomic loci, a phylogenetic approach to the study of Lake Malawi radiation can provide reliable information.
2. The whole genome signal confirms *Astatotilapia rujewa* as the closest available sister species to the Lake Malawi radiation, whereas the *Astatotilapia calliptera* clade is nested within the Lake Malawi radiation as a sister clade to the mbuna.
3. A number of genera, namely *Aulonocara*, *Copadichromis*, *Lethrinops*, *Otopharynx*, and *Placidochromis* are polyphyletic, revealing discrepancies between species relationships implied by traditional taxonomy and the whole genome DNA signal. For *Aulonocara* and *Lethrinops*, the split is between the shallow water (generally <40m) and deep water (generally >50m) species, with each group appearing to have different evolutionary histories. Samples from the genera *Otopharynx* and *Placidochromis* are polyphyletic within the sand dweller group, where taxonomists themselves (Eccles and Trewavas [129]) recognised they had difficulties in finding reliable phylogenetic signals in the morphology. The genus *Placidochromis* is defined by the absence of horizontal melanin pattern, presence of vertical bars, and absence of other defining characteristics [129]. The genus *Otopharynx* is defined on the basis of large melanin spots above the pectoral and/or anal fins, and again the absence of other derived characters [129]. The whole genome DNA phylogeny presented here suggests that these melanin patterns do not reliably indicate phyletic relationships and therefore are not good generic characters.
4. All *Copadichromis* samples, except *C. mloto*, are outside the sand dweller clade, suggesting that this genus of plankton feeders is indeed distinct from the other bottom dwelling fish.

Overall, the phylogeny presented here, while generally consistent with current nomenclature, suggests that for some genera a revision is required in the light of new molecular data. Interestingly, it suggests that ecological traits such as preferred water depth, previously unused in taxonomic revisions [129], may be useful phylogenetic characters.

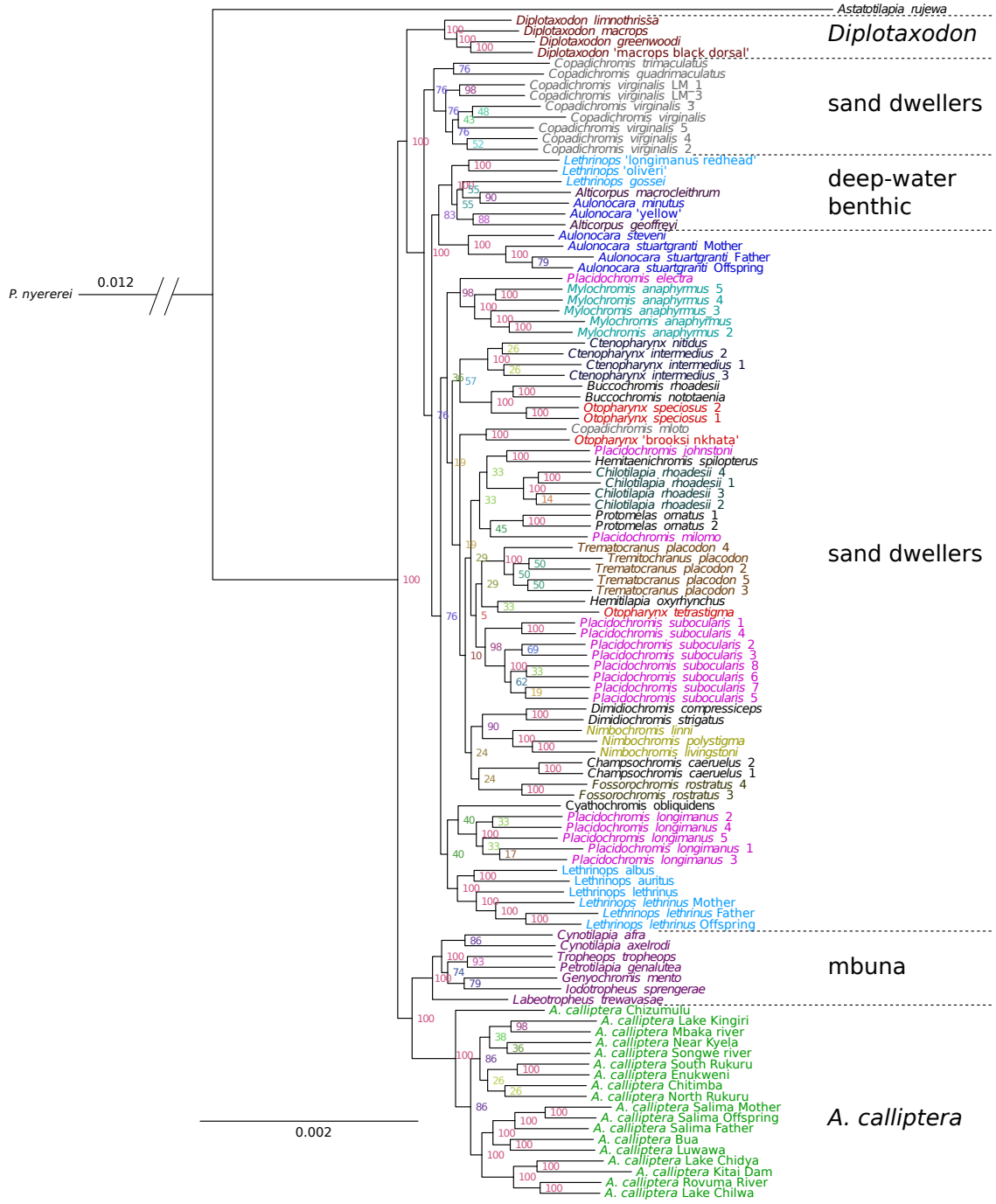


Fig. 5.6 **Lake Malawi whole genome maximum likelihood phylogeny.** Branch support values are based on 42 bootstrap replicates. Samples are coloured according to genera, except mbuna which are all displayed in a single colour.

Most previously published Lake Malawi phylogenies relied on mitochondrial DNA (e.g. ref [81, 86, 176]) or on a limited number (generally hundreds and up to two thousand) of nuclear variants generated with amplified fragment length polymorphisms (AFLP) [177] (e.g. ref [86, 176, 178, 179]).

Using the batch of Lake Malawi samples sequenced in Spring 2013 (see Table 3.2), I compared the mtDNA and whole genome consensus phylogenies. The results (Figure 5.7) show clearly that there are major differences between topologies of the two phylogenetic trees. The whole genome tree resolves species and known groups of species as monophyletic, although assumptions of the likelihood model are likely to be violated (this will be discussed in more detail later). In contrast, the major clades in the mtDNA tree do not correspond to any known or proposed groupings. Strong phylogenetic discordance between Lake Malawi mitochondrial and nuclear DNA phylogenies has been observed previously and interpreted as evidence for hybridisation and introgression between distinct taxa [86, 176].

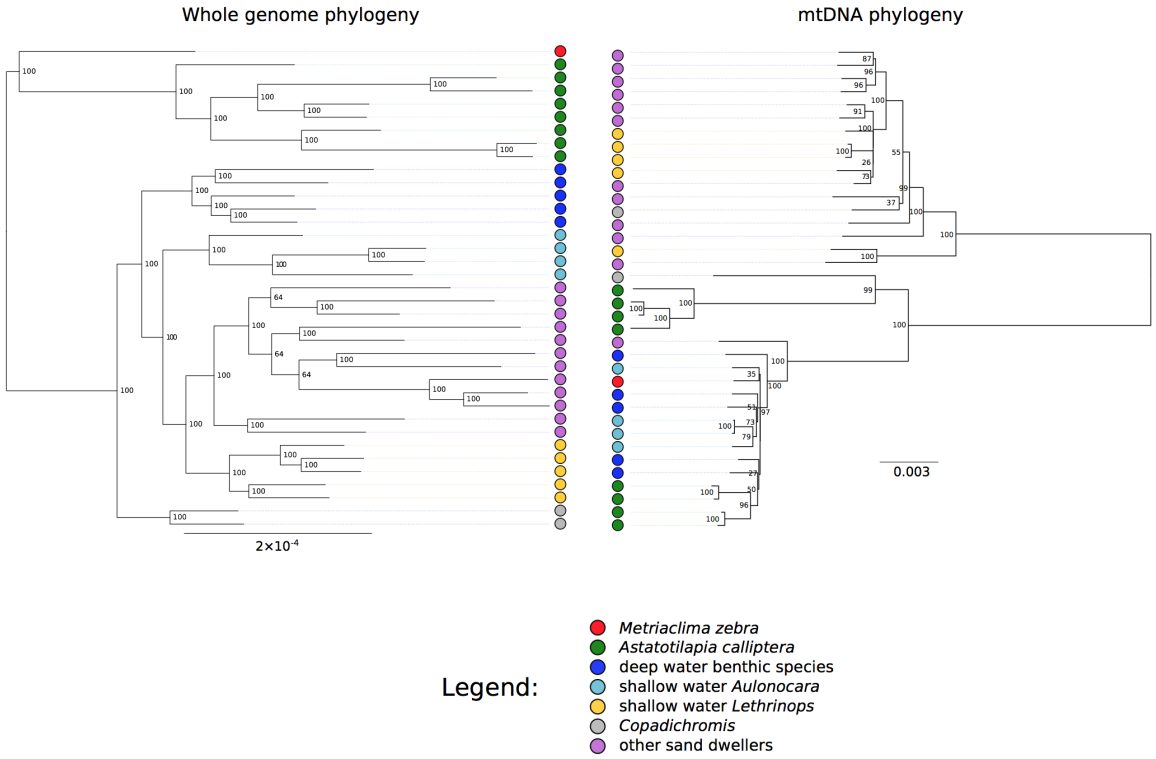


Fig. 5.7 **Contrast between whole genome and mtDNA phylogenies.** Branch support values are based on 281 bootstrap replicates for the whole genome phylogeny and 150 replicates for the mtDNA phylogeny. Both phylogenetic trees are rooted arbitrarily at their midpoints.

Next, I divided the genome into 1,460 regions, each comprising 5,000 segregating variants (mean region length 459kb, s.d. 141kb), and generated an independent phylogeny for each of these genomic regions. These *local phylogenies* are shown in Figure 5.8. Average bootstrap (over all branches) for 1,209 trees was above 40% and for 684 above 50% (Figure A.1), indicating that 5,000 variants were sufficient to provide an informative phylogenetic signal for the majority of the trees.

I found that, even if branch lengths are ignored, all trees differ from each other in their topologies (i.e. no two trees imply the same relationships between the species), implying pervasive effects of incomplete lineage sorting and/or introgression (see section 1.2). It is also notable that all 1,460 trees differ in their topology from the overall whole-genome consensus tree. However, given that both the mean and the standard deviation in the time to coalescence are $2N_e$ generations (90,000 years with N_e=15,000 and average generation time of three years), it is not unexpected that lineages reflecting within-species variation tend not to coalesce within the duration of the species.

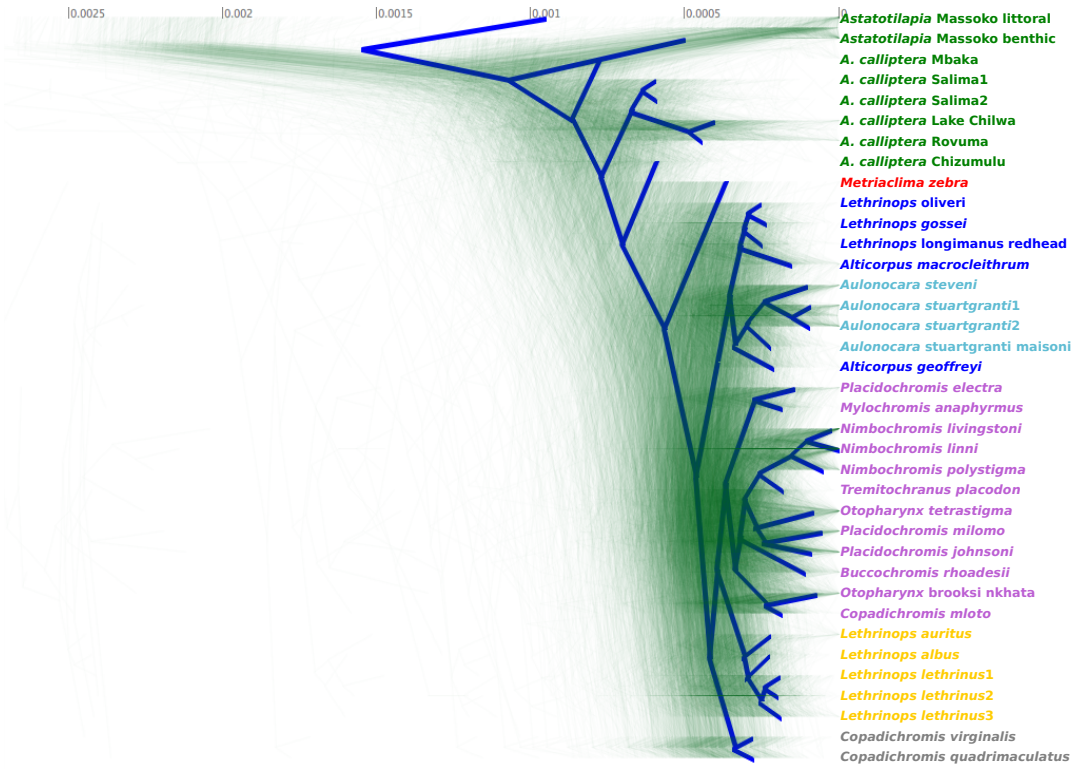


Fig. 5.8 **Variation in phylogenies between 1,460 regions along the genome.** Each tree is based on 5,000 consecutive variants. The tree were drawn using the `DensiTree v2.1` software [180] and the blue consensus tree has been obtained using the 'root canal' function of DensiTree.

It is worth noting here that the large amount of incomplete lineage sorting (ILS) described above means that the assumptions of the likelihood model used to build the whole genome phylogeny are likely to be substantially violated. The model assumes that the whole sequence has been generated by a single bifurcating tree; therefore, every instance of ILS is assumed to be the result of a repeated mutation at the same site, a highly improbable event. Therefore, future phylogenetic analysis in this system will need to also consider other approaches, such as the Neighbour-Joining method [181].

The Robinson-Foulds (RF) dissimilarity metric offers an objective procedure for comparing phylogenetic trees [182]. I used the `RF.dist` function implemented in the `phangorn` [183] package for `R` to calculate this metric for a random sample of 1,460 pairs of local phylogenies, and also to calculate distances between local phylogenies and the mtDNA tree. Figure 5.9A compares the two distributions of distances, showing clearly that the mtDNA tends to be more dissimilar (i.e. its topology tends to be more different from the local phylogenies based on nuclear DNA than they are from each other). I also calculated the RF metric between the mtDNA and whole genome phylogenies (distance 60). Only four out of the 1,460 local trees based on nuclear DNA have RF distance from the whole genome phylogeny $\geq$60.

From the point of view of population genetics, the main difference between mtDNA and nuclear DNA is that mtDNA is inherited only through the female lineage and is haploid. Assuming that there are equal numbers of reproducing males and females, this means that the effective population size for mtDNA is four times lower that for nuclear DNA (for a detailed discussion, including exceptions, see [184]). The lower N_e results in a different distribution of coalescence times (Figure 5.9) so mtDNA alleles tend to coalesce in different ancestral species than nuclear DNA alleles.

Another factor contributing to the mtDNA phylogeny being an outlier is the lack of recombination in the mitochondria; the mtDNA tree is just a single genealogy drawn from the distribution of many possible genealogies shaped by the population history and speciation events. On the other hand, a tree from even just a ~500kb segment of the genome is likely to be an average of several genealogies separated by ancestral recombination events.

Finally, mtDNA may be more likely to introgress between closely related species than nuclear DNA. A large number of studies have reported extensive mtDNA introgression with little or no evidence of introgression in the nuclear genome, e.g in fish [185], amphibians [186], birds [187], mammals [188], and fruit flies [189]. The factors that can explain these discrepancies are reviewed in ref [184]. While some of the cited studies rely on just a few microsatellite nuclear markers [186, 187], Good *et al.* [188] use

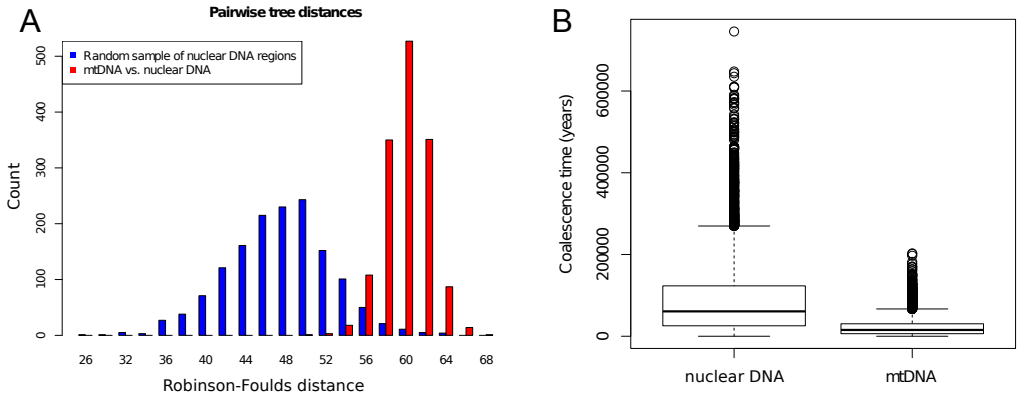


Fig. 5.9 **Difference between local phylogenies based on nuclear DNA and the mtDNA phylogeny. (A)** Robinson-Foulds distances. **(B)** Simulated coalescence times for nuclear DNA and mtDNA with exponential distributions, assuming an ideal population, N_e=15,000 and that the average generation time is three years. Coalescence times for nuclear DNA are then distributed as ~exp($\frac{1}{90{,}000}$) and mitochondrial ~exp($\frac{1}{22{,}500}$). I simulated 5,000 observations from each distribution.

sequence data from 10,500 genes to convincingly rule out "all but very minor levels of interspecific gene flow" despite complete mtDNA replacement between the two rodent species in their study. Therefore, in order to understand the importance and extent of introgression in the Lake Malawi radiation it is necessary to re-evaluate the evidence from a genome-wide perspective, rather than rely on mtDNA evidence.

5.4 Exploring evidence for interspecific introgression

5.4.1 ABBA-BABA tests

The ABBA-BABA statistic (also known as the *D statistic* or *Patterson's D*) tests for an excess of shared derived alleles between one of two populations (P_1 and P_2) and their outgroup (P_3) [190, 191]. I defined derived alleles with respect to the Lake Victoria species *P. nyererei*. Under the null hypothesis of no differential gene flow from P_3, the two populations P_1 and P_2 are expected to share derived alleles with P_3 equally often, while introgression from P_3 to P_1 or from P_3 to P_2 would result in excess sharing, as illustrated in Figure A.2.

I used the ABBA-BABA statistic to test if the phenotypic similarity between the shallow and deep water species of the genus *Lethrinops* could be due to introgression from shallow water *Lethrinops* into a deep water relative of the genus *Alticorpus*. In the

Lake Malawi whole genome phylogeny (Figure 5.6), shallow and deep water *Lethrinops* form two distinct groups, but their anatomy is sufficiently similar for them to be assigned to the same genus by Eccles and Trewavas [129]. I therefore tested for a signature of gene flow from the shallow water *L. lethrinus* into the deep water species. If that were the case, I expected to find an excess of the ABBA pattern, as illustrated by the test design shown in Figure 5.10.

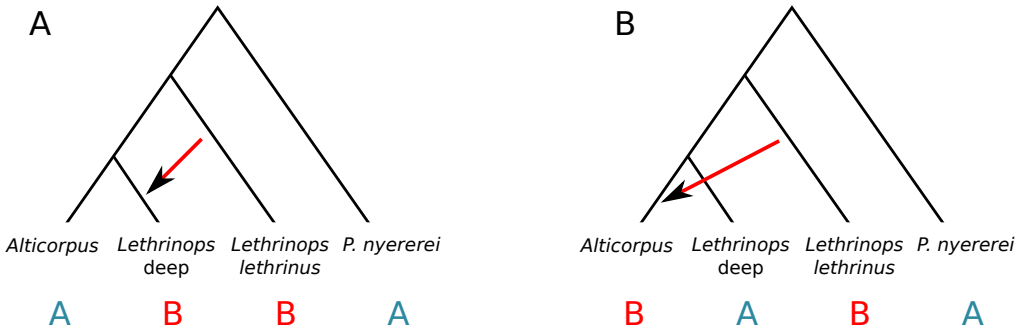


Fig. 5.10 **The ABBA-BABA test for introgression from shallow to deep water species of the genus *Lethrinops*.** The ancestral allele is denoted by the A character and the derived allele by the B character.

Overall, I did not find any excess of shared derived alleles between *L. lethrinus* and deep water *Lethrinops* when compared with the deep water genus *Alticorpus* (Patterson's D=0.37%; 1.28 SD from 0% or P=0.101). Therefore, significant levels of gene flow specifically into the deep water *Lethrinops* species can be ruled out. To explore whether smaller regions of the genome might have introgressed, perhaps due to selection following low levels of hybridisation, I also calculated the D statistic for non-overlapping genomic regions of 100 informative variants each (informative in the sense that they change the numerator of the D statistic) (Figure 5.11).

The regions for local D statistic calculation were on average 137kb long, but with large variation (from 4kb to 911kb; s.d. 63kb). I found several outlier regions with D statistic more than ±4.5 s.d. away from the mean. If introgression happened, these would be the best candidates for introgressed regions. However, it is important to note that there are outlier loci in both directions, i.e. suggesting introgression both into deep water *Lethrinops* and into *Alticorpus*. While this is not impossible, it is advisable to be cautious as the locally calculated D statistic has been shown by simulations to have large variance [192].

For all of the above ABBA-BABA analysis I had access only to the samples sequenced in Spring 2013, so three *L. lethrinus*, three deep water *Lethrinops* and two

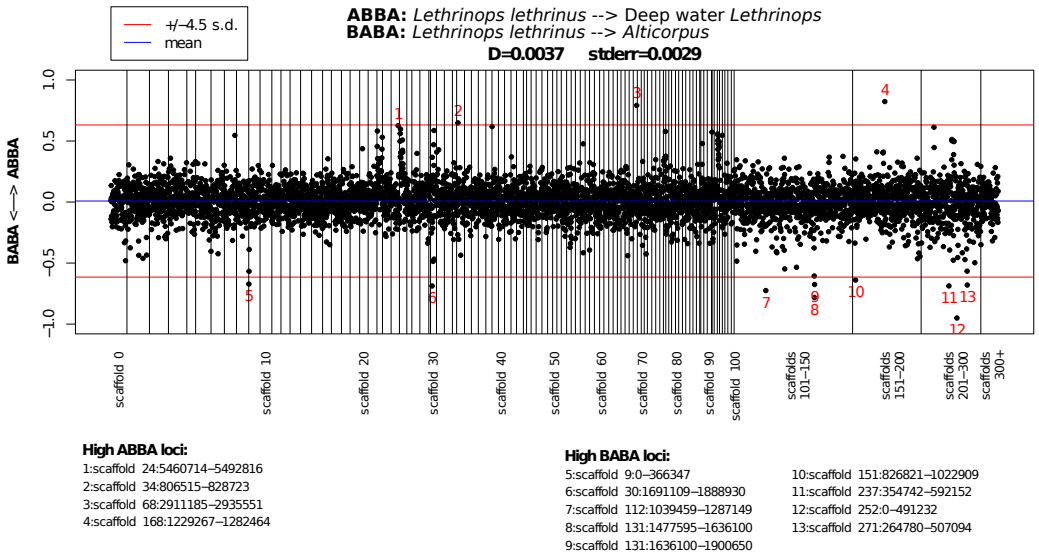


Fig. 5.11 **The results of the ABBA-BABA test for introgression from shallow water to deep water species of the genus *Lethrinops*.** Each point corresponds to the value of D statistic calculated over a window of 100 informative variants. Outlier regions are listed below the plot.

Alticorpus individuals. This was sufficient when calculating the D statistic over the entire genome. However, reducing neutral variance in the local introgression analysis will require adding more samples per group and using new statistics specifically designed to locate introgressed loci such as f_d [192] and f_{dM} (Methods - section 5.5).

5.4.2 Chromopainter and fineSTRUCTURE

I used the `Chromopainter` program [193] to 'paint' the chromosomes of each individual based on haplotype similarity with chunks of DNA from the other individuals. Each chromosome was conceptually segmented into regions, bounded by ancestral recombination sites, each of which is matched with another chromosome which is assigned as the 'donor' or the local 'closest relative'. This is done statistically, resulting in a measure of co-ancestry to each other chromosome in the data set. The results of this 'chromosome painting' for all Lake Malawi samples are summarised in the co-ancestry matrix shown in Figure 5.12. Individuals have been clustered by the `fineStructure` software [193] based on the amounts of co-ancestry they share. Overall, the co-ancestry matrix provides a view of Lake Malawi species relationships based on haplotype relationships, which is complementary to the phylogenetic analyses and the F_{ST} distances presented previously.

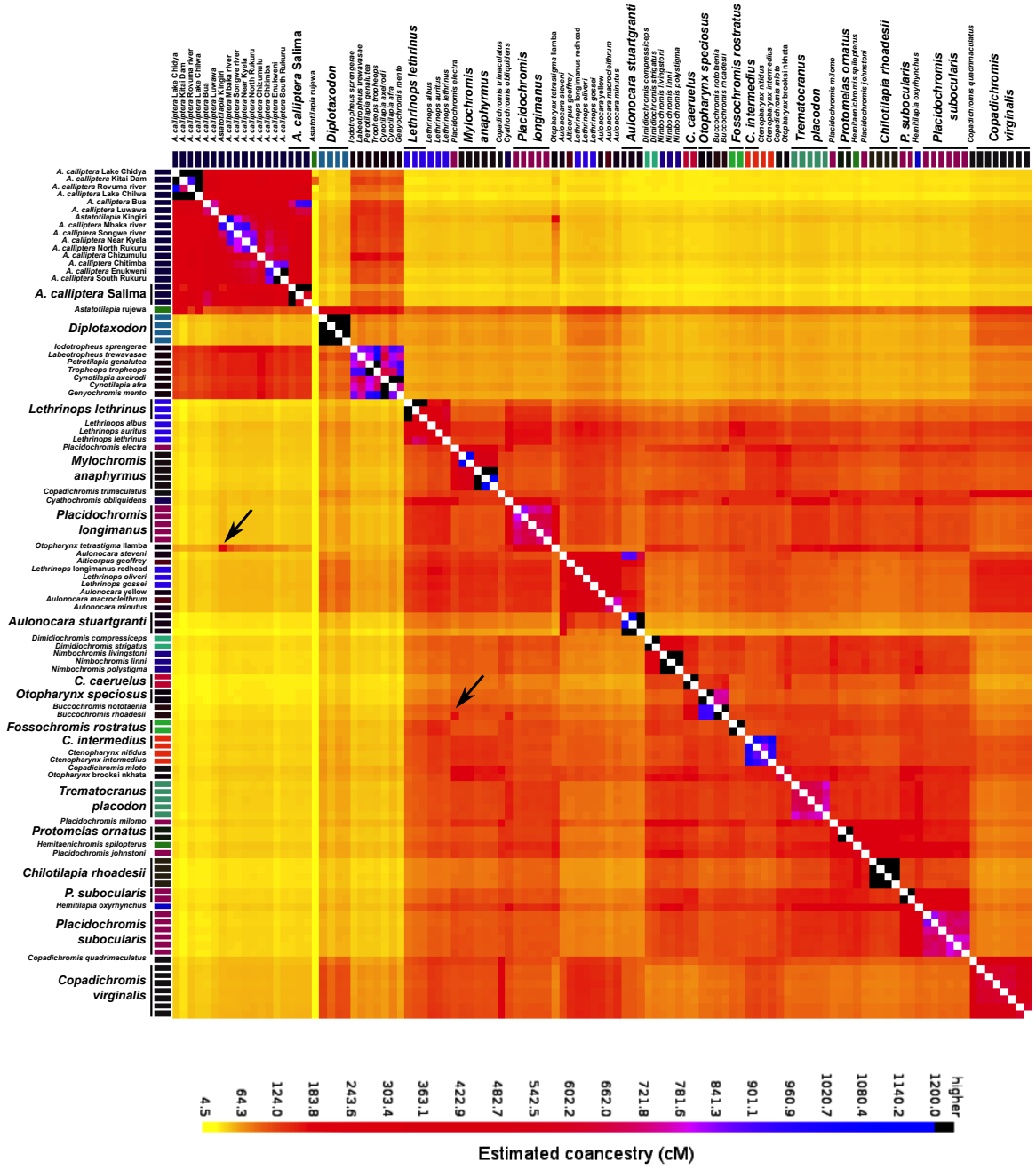


Fig. 5.12 **Co-ancestry between Lake Malawi samples measured by the Chromopainter software.** Each row corresponds to a 'recipient' and each column to a 'donor'. Thus the values indicate the total length (in cM) along which a 'donor' haplotype was inferred to be the closest relative of a 'recipient' haplotype. Clustering was done by the fineStructure software [193]. Examples of possible introgression are indicated by arrows.

Sharing of chunks of chromosomes between otherwise distantly related species would be an indication of recent admixture or introgression. The Chromopainter co-ancestry matrix provides several such hints. For example, there is unexpectedly high co-ancestry between the *Otopharynx tetrastigma* from Lake Ilamba and the *Astatotilapia* from Lake Kingiri (Figure 5.12), suggesting the *O. tetrastigma* and *Astatotilapia* may hybridise in Ilamba (we have not yet sequenced an Ilamba *Astatotilapia*, so we see the strongest signal in the closely related Kingiri sample). Other, although not as strong, hints of introgression can be seen within the shallow water sand dweller group. For example, *Buccochromis rhoadesii* has unexpectedly high co-ancestries with *Placidochromis electra* and with *Cyathochromis obliquidens*.

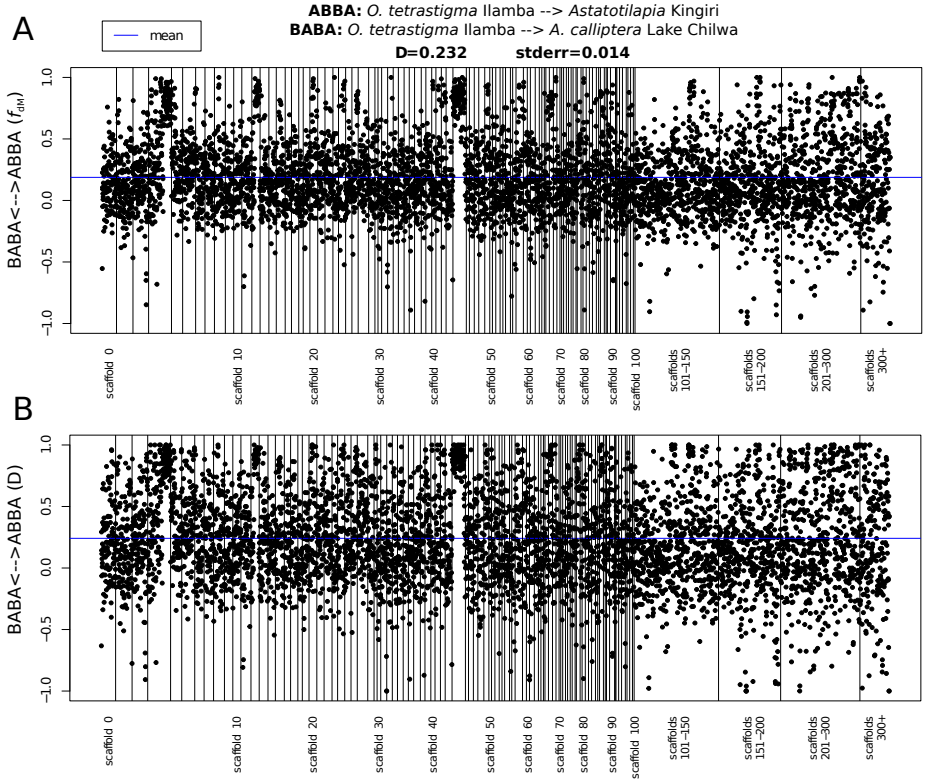


Fig. 5.13 ***O. tetrastigma* Ilamba introgression into crater lake *Astatotilapia*.** **(A)** Point correspond to the f_{dM} statistic calculated over windows of 50 informative variants. **(B)** The D statistic calculated over windows of 50 informative variants.

I followed up on the hypotheses of introgression generated by the Chromopainter software by calculating the ABBA-BABA statistics. First, I found a large excess of shared derived alleles between *O. tetrastigma* from Ilamba and *Astatotilapia* Kingiri, when compared with *A. calliptera* from Lake Chilwa (Patterson's D=23.2%; 16.6 SD from 0% or $P<3.4\times10^{-62}$). The proportion of admixture f [190] was estimated at 29.8±1.7%. Next, I calculated local D and f_{dM} statistics for non-overlapping genomic

regions of 50 informative variants each (Figure 5.13). The results confirm that the f_{dM} statistics has lower variance (s.d. = 0.31 against 0.37 for D) and suggest that scaffolds 3, 12, and 44 harbour regions with especially strong signatures of introgression.

Next, I tested for possible introgression from *Placidochromis electra* into *Buccochromis rhoadesii*. I found a small excess of shared derived alleles between these two species, when compared to sharing between *P. electra* and *B. notoaenia* (Patterson's D=1.31%; 2.52 SD from 0% or P=0.006). The proportion of admixture f was estimated at 1.3±0.9%. Therefore, the genome-wide statistics suggest that there has been a small amount of introgression from *P. electra* into *B. rhoadesii*. However, unlike in the previous case of *O. tetrastigma* hybridisation in Ilamba, local D and f_{dM} statistics do not provide clear pointers to specific introgressed regions (Figure 5.14).

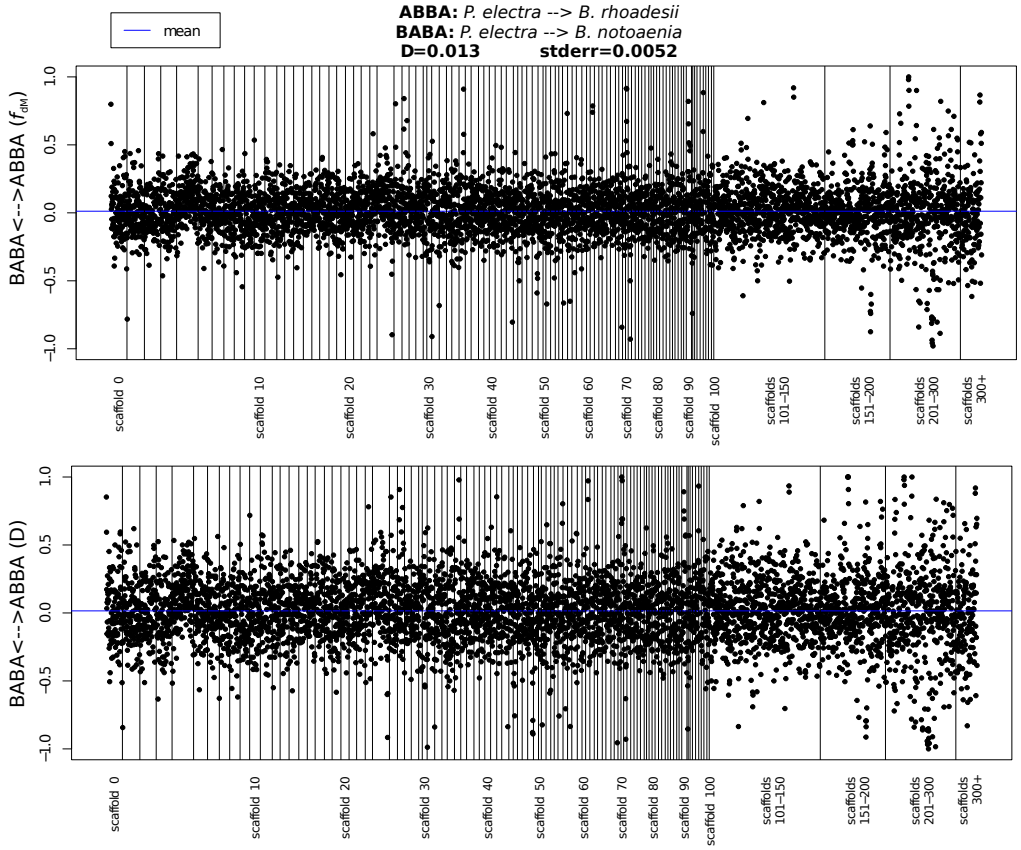


Fig. 5.14 ***Placidochromis electra* introgression into *Buccochromis rhoadesii*.** **(A)** Point correspond to the f_{dM} statistic calculated over windows of 50 informative variants. **(B)** The D statistic calculated over windows of 50 informative variants.

5.5 Methods

Basic statistics

Basic statistics (number of variants called, numbers of heterozygous sites) were obtained using the `bcftools v1.1 stats` tool, with the `-s -` option to include all samples. Results from the 'Per-sample counts' (PSC) lines were then plotted using the `R` software environment [175] .

Principal Component Analysis

SNP variants (no indels) with minor allele frequency >= 0.05 were selected using `vcftools v0.1.12b` options `--remove-indels --maf 0.05` and exported in `PLINK` format [194]. The variants were LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Principal Component Analysis on the resulting set of variants was performed using the `smartpca` program from the `eigensoft v5.0.1` software package [195] with default parameters.

The 3D plots were generated with the using the `scatterplot3d` [196] package for `R`.

Genome-wide F_{ST} calculations

All SNP variants were selected and LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Next I used the `smartpca` program from the `eigensoft v5.0.1` software package [195] with adding the parameter `outliersigmathresh: 11` to prevent the removal of outliers. The `smartpca` program calculates genome-wide F_{ST} for all pairs of populations specified by the sixth column in the `.pedind` file using the Hudson estimator, as defined by Bhatia, Patterson *et al.* [197] in equation 10, and using 'ratio of averages' to combine estimates of F_{ST} across multiple variants, as recommended in their manuscript [197].

Local F_{ST} calculations

For sliding-widow based analysis, I used my own implementation of the Bhatia, Patterson *et al.* [197] F_{ST} calculation in the C++ program `evo` (available from `https://github.com/millanek/evo` with the `fst --vcf` option).

Phylogenetic trees

For the whole genome maximum likelihood phylogeny in Figure 5.6, I generated consensus genome sequences using the `bcftools v1.2 consensus` tool. For each sample, I selected the sequence of one haplotype (as assigned by `beagle` haplotype phasing - see section 3.2.4) by using the `--haplotype=1` option in `bcftools`. All scaffolds except the mtDNA sequence (scaffolds 747, 2036) were concatenated into a single sequence and phylogenetic trees then inferred using `RAxML v7.7.8` [198] under the GTRGAMMA model (General Time Reversible model of nucleotide substitution with the Γ model of rate heterogeneity). The maximum likelihood tree was obtained as the best out of five alternative runs on distinct starting maximum parsimony trees (using the `-N 5` option). Forty two bootstrap replicates were obtained using `RAxML's` rapid bootstrapping algorithm [199]. It was my intention to run more bootstrap replicates, enough to satisfy `RAxML -N autoFC` frequency-based bootstrap stopping criterion, but this has proven computationally infeasible on a dataset of this size (obtaining the 42 replicates required thousands of hours of CPU time). Still 42 replicates provide a reasonable indication of bootstrap support for the maximum likelihood tree. Bipartition bootstrap support was drawn on the maximum likelihood tree using `RAxML -f b` option.

For all other phylogenies using samples from Spring 2013 (Figures 5.7, 5.8), I generated consensus sequences using unphased genotype data with my own C++ program `evo` (available from `https://github.com/millanek/evo`) using the `getWGSeq --whole-genome` options for the whole genome data, `getWGSeq --mtDNA` for mitochondrial sequences, and `getWGSeq --split 5000` for local phylogenies. Heterozygous variants were represented by IUPAC codes. The phylogenies were then built in the same way as described above, with one difference: the number of bootstrap replicates was always determined by the `RAxML -N autoFC` frequency-based bootstrap stopping criterion.

Chromopainter and fineSTRUCTURE

Singleton SNPs were excluded using `bcftools-1.1 -c 2:minor` option, before exporting the remaining variants in PLINK format [194]. The `chromopainter v0.0.4` software [193] was then run for 150 largest genomic scaffolds. Briefly, I created a uniform recombination map using the `makeuniformrecfile.pl` script, then estimated the effective population size (N_e) for a subsample of 20 individuals using the chromopainter inbuilt expectation-maximization procedure [193], averaged

over the 20 N_e values using the provided `neaverage.pl` script. The chromopainter program was then run for each scaffold independently, with the `-a 0 0` option to run all individuals against all others. Results for individual scaffolds were combined using the `chromocombine` tool before running `fineSTRUCTURE v0.0.5` with 1,000,000 burn in iterations, and 200,000 sample iterations, recording a sample every 1,000 iterations (options `-x 1000000 -y 200000 -z 1000`). Finally, the sample relationship tree was built with fineSTRUCTURE using the `-m T` option and 20,000 iterations.

Patterson's D (ABBA-BABA) and related statistics

I calculated the D statistics using equation S15.2 of Green et al. [190], allowing me to use allele-frequency information from all benthic and littoral individuals. I also estimated *f*, the admixture fraction following Green *et al.* equation S18.5, and calculated the standard error for both estimates by a weighted block jackknife, using blocks of 5,000 informative variants (i.e. variants with ABBA or BABA patterns).

I also calculated a version of the f_d statistic designed by Martin *et al.* specifically to detect introgressed loci [192, equation 6]. One of the limitations of the f_d statistic is that it is not symmetric; it is distributed on the interval (-Infinity,1]. Therefore, I define a closely related statistic which I call f_{dM}. Compared with the f_d statistic, f_{dM} has the advantages that it is bounded on a the interval [-1,1], and under the null hypothesis of no introgression is symmetrically distributed around zero.

Following the notation from Martin et al. [192], I consider three populations and an outgroup with the relationship (((P$_1$, P$_2$), P$_3$), O). Let:

$$S(P_1; P_2; P_3; O) = \sum_i ((1 - p_{i1})p_{i2}p_{i3}(1 - p_{i4})) - \sum_i (p_{i1}(1 - p_{i2})p_{i3}(1 - p_{i4})) \quad (5.2)$$

where p_{ij} is the frequency of the derived allele at site i in population j.

f_{dM} is then defined as follows:

$$f_{dM} = \begin{cases} \dfrac{S(P1; P2; P3; O)}{S(P1, PD, PD, O)} = f_d, & \text{if } p_{i2} \geq p_{i1} \\ \dfrac{S(P1; P2; P3; O)}{-S(PD, P2, PD, O)}, & \text{otherwise} \end{cases}$$

where P_D is the population (either P$_1$ or P$_3$) that has the higher frequency of the derived allele. For a detailed discussion of the f_d statistic see Martin et al. [192].

The D and f statistics were calculated genome-wide and D and f_{dM} also in non-overlapping windows of 50 or 100 informative variants each, as indicated in the main text. To add ancestral allele information (i.e. the outgroup variants) to the Lake Malawi set VCF file, I used whole genome alignment between *M. zebra* and *P. nyererei*, as described in section 3.4.

Chapter 6

Incipient speciation in Lake Massoko, Tanzania

6.1 Introduction

6.1.1 Publication note

The work described in this chapter, except section 6.3, has been published as M. Malinsky et al., 'Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake'. *Science*. 350, 1493-1498 (2015) [200].

6.1.2 Background

In the summer of 2011, my collaborators Martin Genner and George Turner conducted a survey of fish fauna in six crater lakes in the Rungwe District of Tanzania (Tables 6.1, B.1)[1]. In all six lakes, they found endemic haplochromine cichlids of the genus *Astatotilapia*, closely related to *Astatotilapia calliptera* (Figure 6.1A), a species widely-distributed in the rivers, streams and shallow lake margins of the region, including in Lake Malawi itself. Thus, the Rungwe District *Astatotilapia* are close relatives of the of Lake Malawi endemic haplochromine cichlids.

In Lake Massoko, the benthic zone in deep waters (~20-25m) is very dimly lit and populated by cichlids with phenotypes clearly different to those typical of shallow waters (~<5m) close to the shore (littoral). Deep-water males are dark blue-black, while most males collected from the shallow waters are yellow-green, similar to riverine

[1]Except for a brief mention of the presence of *Tilapia squamipinnins* and *Astatotilapia calliptera* in Lake Massoko by Ricardo [201], there were to our knowledge no published records about this fish fauna.

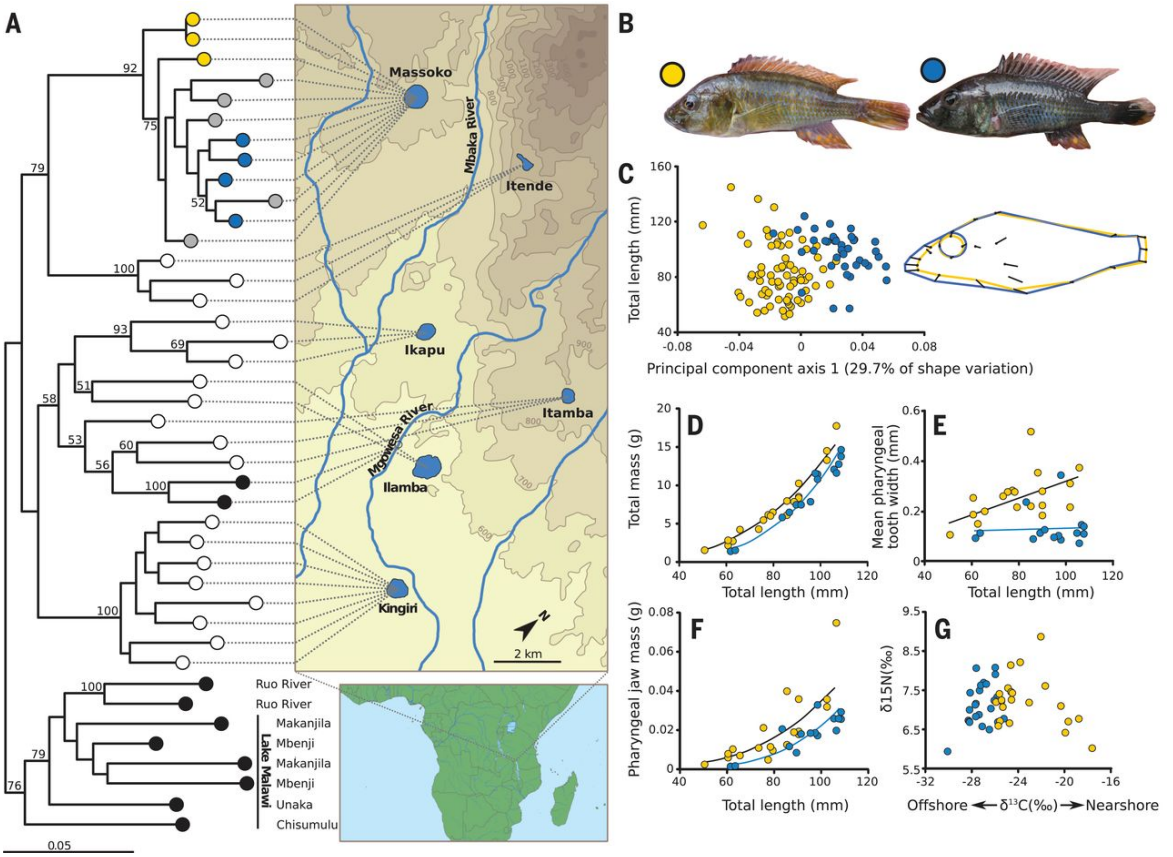


Fig. 6.1 **Cichlid radiation in the crater lakes of southern Tanzania.** **(A)** A phylogeny of the crater lake *Astatotilapia* based on reference-aligned RAD data (9,206SNPs). It demonstrates reciprocal monophyly between the populations in each lake except for Itamba, and close relationship to *A. calliptera* from rivers and from Lake Malawi. Within Lake Massoko, yellow symbols indicate the littoral morph, blue symbols indicate the benthic, and grey symbols denote small, phenotypically ambiguous, and thus unassigned individuals. Additional *Astatotilapia* individuals from other crater lakes are denoted by open circles. *A. calliptera* from rivers and Lake Malawi are denoted by black circles. Bootstrap values are displayed for nodes with >50% support. **(B)** Breeding males of the yellow littoral and blue benthic morphs of Lake Massoko. The symbols next to the photographs correspond to symbols used in (C-G). **(C-F)** Morphological divergence between the two morphs of Lake Massoko. Relative to the littoral, the benthic morph has relatively longer head and jaw **(C)**, lower body mass **(D)**, narrower 'papilliform' pharyngeal teeth **(E)**, and lighter lower pharyngeal jaws **(F)**. The benthic fish have stable isotope ratios that tend to be more depleted in C^{13} than the littoral, indicative of a more offshore-planktonic diet **(G)**.

forms (Figure 6.1B; Table 6.2). We also collected small (<65mm standard length) males that were not readily field-assigned to either ecomorph (Methods - section 6.8.1). The benthic and littoral morphs are reminiscent of the species pair of *Pundamilia* cichlids from Lake Victoria [89], but within a potentially simpler historical and geographical context. Lake Massoko is steep-sided, has a strong thermocline at ~15m, and an anoxic boundary at ~25m [202]. The estimated time of lake formation is ~50,000 years ago [203].

Table 6.1 **Location and geographical characteristics of crater lakes with haplochromine cichlid fauna in Rungwe District, Tanzania.** Data from [202], except Ikapu, estimated from Google Earth and own survey of depth.

Lake	Latitude	Longitude	Altitude (m)	Surface area (km^2)	Maximum depth (m)	Volume ($\times 10^6 m^3$)
Kingiri	9°25' S	33°51' E	515	0.27	34	5.37
Ilamba	9°24' S	33°50' E	548	0.42	23	7.01
Ikapu	9°22' S	33°48' E	653	0.28	3	0.85
Itamba	9°21' S	33°51' E	821	0.12	18	0.69
Itende	9°19' S	33°47' E	1020	0.14	2	0.28
Massoko	9°20' S	33°45' E	845	0.38	37	8.91

Table 6.2 **Depth distribution of ecomorphs in Lake Massoko.** Based on collections using a variety of methods (Methods - section 6.8.1) in July-August and December 2014, and August 2015. There is a significant association between bottom depth and morph frequencies ($\chi^2_{4df} = 207.1$, P<0.001).

	0-5m	**5-10m**	**10-15m**	**15-20m**	**20-25m**	**Total**
Benthic	0	6	11	25	75	117
Littoral	98	54	15	21	0	188
Total	98	60	26	46	75	305
% Benthic	0	10	42.3	54.3	100	
% Littoral	100	90	57.7	45.7	0	

Ecomorph separation

To examine relationships between crater lake and riverine *A. calliptera* of southern Tanzania, Genner and Turner obtained restriction site associated DNA (RAD) [204] data from 30 fish from the Rungwe District, and 11 outgroup *Astatotilapia* from the broader Lake Malawi catchment (Figure 6.2, Table B.2). A maximum likelihood phylogeny constructed on the basis of these data demonstrates monophyly of all

specimens from Lake Massoko (Figure 6.1A; Methods - section 6.8.2). Thus, the RAD phylogeny provides evidence that Massoko morphs might have evolved in primary sympatry, as previously proposed for crater lake cichlid radiations of Cameroon [205] and Nicaragua [206].

Morphological analyses of these two colour morphs revealed significant differences in head and body shape, body mass, the shape of pharyngeal teeth, and pharyngeal jaw mass (Figure 6.1C-F; Table 6.3; ANCOVA tests, all P<0.001). Genner and Turner also found significant differences in stable isotope ratios (Figure 6.1G; Table S5; ANCOVA test, P<0.001), indicative of dietary differences. Together these results demonstrate ecomorph separation and adaptation to different ecological environments

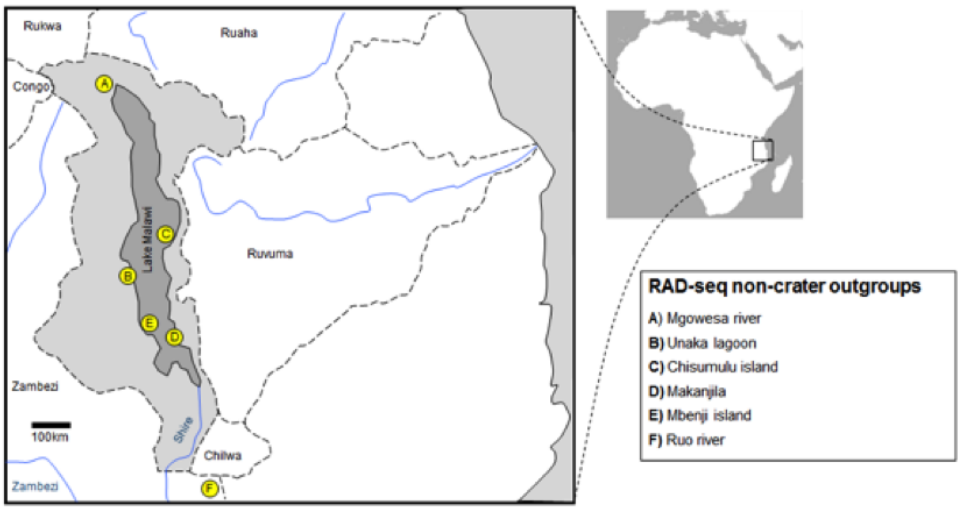


Fig. 6.2 **Collection sites of RAD sequenced non-crater-lake *Astatotilapia calliptera* for phylogenetic analysis.** Figure by M. Genner.

Table 6.3 **Results of morphological and stable isotope analysis.** Analysis of Covariance (ANCOVA) tests of morphological and stable isotope differences among benthic and littoral morphs. In each case total length (TL) was employed as a covariate.

	N benthic	N littoral	F TL	P TL	F ecomorph	P ecomorph
External morphology (PC1)	41	73	5.749	0.018	166.884	< 0.001
Body mass*	15	19	677.780	< 0.001	34.170	< 0.001
Pharyngeal jaw mass*	15	19	110.432	< 0.001	18.337	< 0.001
Pharyngeal jaw tooth width	15	19	4.037	0.053	25.121	< 0.001
Stable isotopes (δ^{13}C)	24	22	3.296	0.076	46.834	< 0.001
Stable isotopes (δ^{15}N)	24	22	0.516	0.476	0.636	0.430

*log10 transformed

The genomic causes and effects of divergent ecological selection during speciation are still poorly understood, in part because of the scarcity of well-characterised examples.

As I discussed in chapter 5, investigation of early stages of speciation in the large cichlid radiations of lakes Malawi, Tanganyika, and Victoria has been hampered by the complexity of those radiations, leading to difficulties in identifying sister species relationships, in reconstructing past geographical situations, and in controlling for possible introgression from non-sister taxa.

Therefore, while also working on improving our knowledge of species relationships in the Lake Malawi radiation (chapter 5), I focussed during my PhD on using whole genome DNA sequence data for a detailed study of ecomorph divergence in the much simpler Lake Massoko system. I investigated the geographical basis of the divergence, tested the 'genomic islands' model of ecological speciation, and explored functional correlations between highly divergent genomic regions and key traits likely to be involved in speciation, including mate choice and visual pigment spectral sensitivities.

6.2 Whole-genome evidence of Massoko divergence

To study the genome-wide pattern of Massoko ecomorph divergence and to further clarify its geographical context, we obtained whole-genome sequence data at ~15X coverage for 6 individuals each of the yellow littoral and blue benthic ecomorphs and 16 additional *A. calliptera* from the wider Lake Malawi catchment (Fig. S1), supplemented by lower coverage (~6X) data from 87 specimens from Lake Massoko (25 littoral, 32 benthic, and 30 small unassigned) and 30 individuals from Lake Itamba. This whole genome data has been described in chapter 3 (Figure 3.2; Table 3.3). The divergence from the *Metriaclima zebra* reference assembly was 0.2-0.3%, and variants were called at 4,755,448 sites (1.2-1.6 million sites per individual).

A maximum likelihood phylogeny built from whole genome sequence data confirmed reciprocal monophyly of *Astatotilapia* within Lakes Massoko and Itamba, and revealed the sister group of Massoko fish to be an *A. calliptera* population from the nearby Mbaka river (Figure 6.3A). All specimens of the benthic ecomorph formed a monophyletic clade derived from the littoral ecomorph (Figure 6.3A). Principal component analysis (PCA) showed strong population structure (Tracy-Widom statistics: $P<1\times10^{-12}$), with benthic and littoral individuals separated by the first eigenvector and forming separate clusters (Figure 6.3B). In contrast, within Lake Itamba, PCA did not reveal significant population structure (Tracy-Widom statistics: $P=0.11$). Individuals from Massoko that were not field-assigned to either of the ecomorphs did not form a monophyletic clade in the phylogeny (Figure 6.3A) or a distinct cluster in PCA (Figure 6.3B).

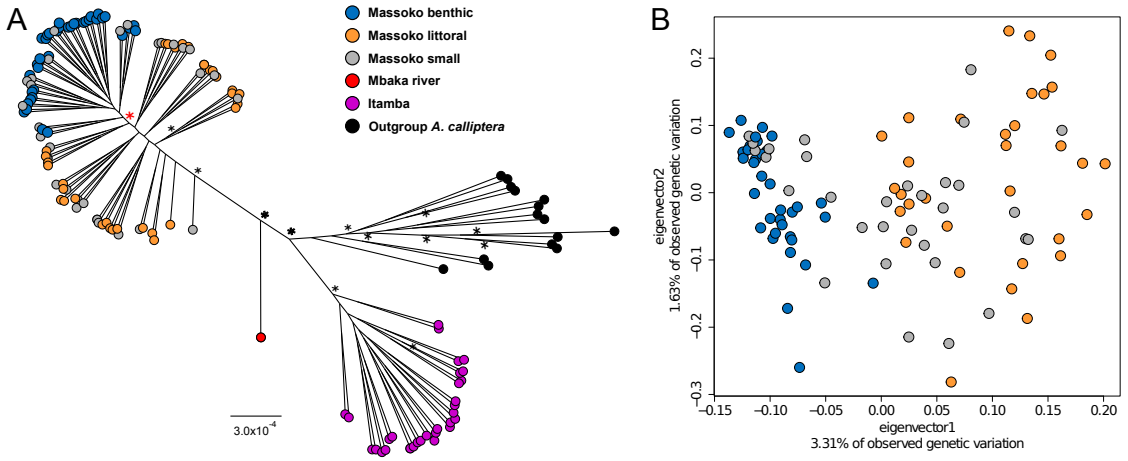


Fig. 6.3 **Lake Massoko divergence with whole genome sequence data. (A)** A maximum likelihood whole-genome phylogenetic tree. Black stars indicate nodes with 100% bootstrap support. The red star highlights the branch that separates all the Massoko benthic samples from the rest of the phylogeny (50% bootstrap support). **(B)** Principal Component Analysis of genetic variation within Lake Massoko.

I estimated individual ancestries for all Massoko specimens with the ADMIXTURE software [207] (Methods - section 6.8.3). Eleven of the 31 samples field-assigned as littoral, and 10 of the 30 unassigned individuals were identified as admixed (with the benthic gene pool), with admixture fraction >25% (Figure 6.4). On the other hand, no individuals identified as benthic were estimated to be admixed to the same extent. Therefore, recent gene flow may be biased from deep to shallow waters. We suggest that the remaining 20 unassigned samples represent sub-adult individuals of both benthic and littoral ecomorphs.

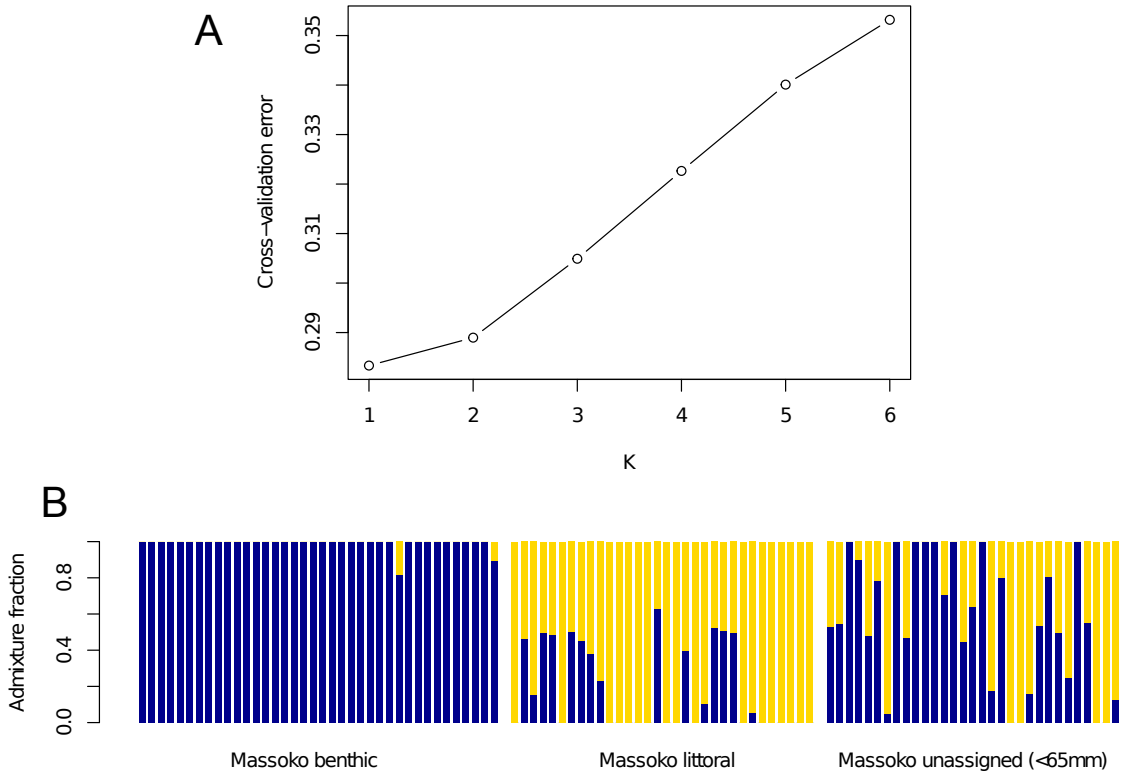


Fig. 6.4 **ADMIXTURE estimates of individual ancestries. (A)** ADMIXTURE cross-validation approach to choosing the `K` parameter - the postulated number of ancestral populations. The error estimates are based on 10-fold cross-validation. The lowest error is observed with `K=1`, suggesting that population differentiation between the ecomorphs is subtle [193] (Methods - section 6.8.3). **(B)** With two postulated ancestral populations (`K=2`), benthic individuals form a virtually homogenous group. Eleven of the samples field-assigned as littoral appear to be >25% admixed. The unassigned samples are a mixture of benthic, littoral, and admixed individuals.

Analysis of fine-scale genetic relationships with fineSTRUCTURE [193] supports the monophyly of the benthic ecomorph within the littoral, but also suggests that compared with the benthic population, the littoral population has greater co-ancestry with other *A. calliptera*; in particular with the Mbaka river sample (Figure 6.5).

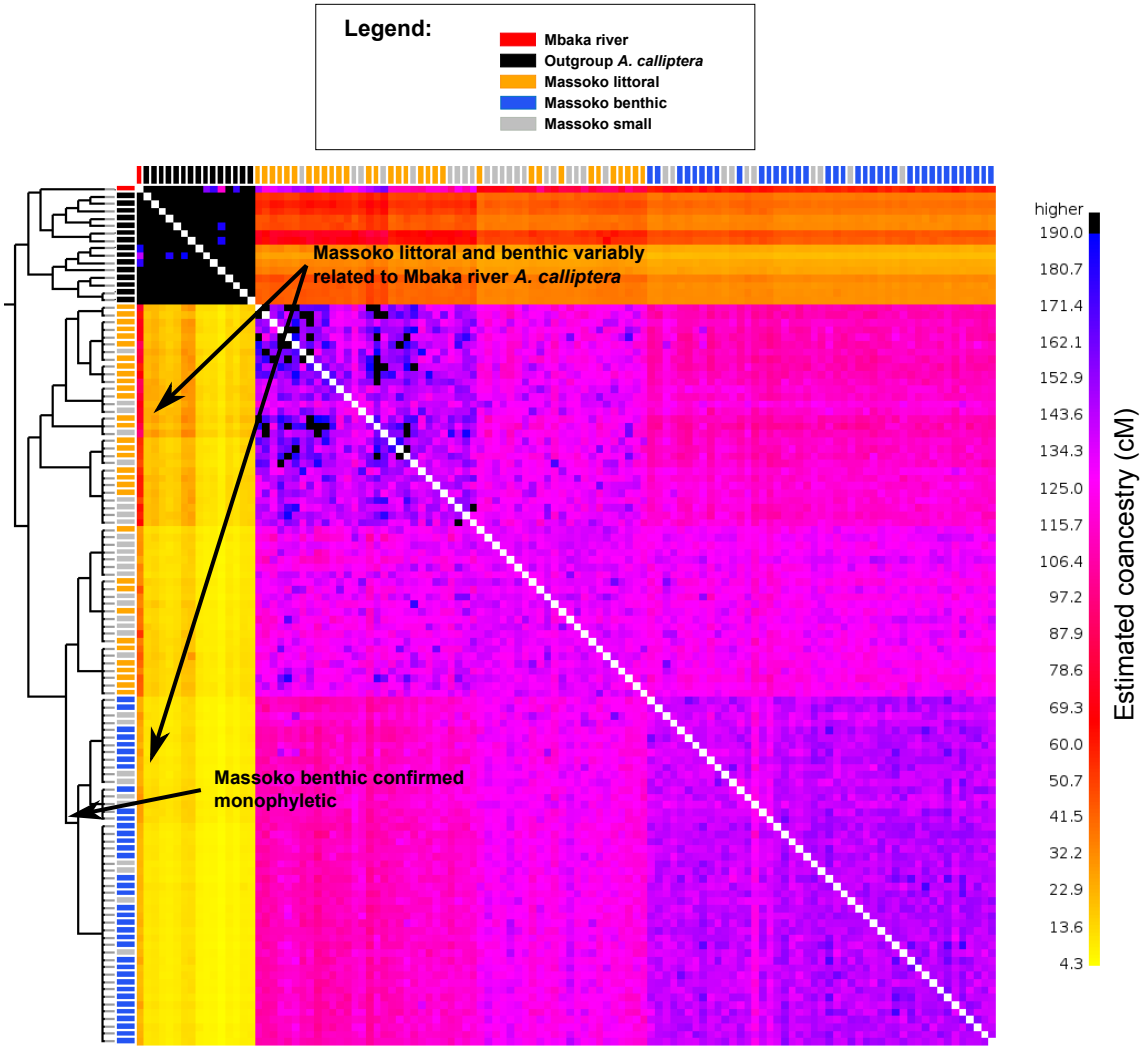


Fig. 6.5 **Massoko fineSTRUCTURE results.** Co-ancestry matrix with the tree showing inferred relationships between samples. Each tip and label correspond to an individual, with labels coloured according to the population/ecomorph as indicated in the legend. The results show tight clustering and monophyly of the benthic ecomorph, greater population structure within the littoral ecomorph, and a difference between the ecomorphs with respect to co-ancestry with Mbaka river *A. calliptera*, as indicated.

Therefore, I tested for evidence of secondary gene flow, as seen in cichlid populations from Cameroonian crater lakes [208]. Under the null hypothesis of no differential gene

flow into Massoko, *A. calliptera* from Mbaka river should share derived alleles equally often with the littoral and with the benthic populations [190, 191]. Instead, we found an excess of shared derived alleles between *A. calliptera* from the Mbaka river and the littoral population, when compared with the benthic population (Patterson's D=1.1%; 4.86 SD from 0% or P<5.8×10^{-7}) (Methods - section 6.8.3). The proportion of admixture f with Mbaka was estimated at 0.9±0.2%. However, this value is low, at a proportion that is approximately half of the Neanderthal introgression into non-African humans [190] and cross-coalescence rate analysis with MSMC [209] (Methods - section 6.8.3) indicates an average separation time of both Massoko ecomorphs from other *A. calliptera* samples (including Mbaka river) approximately ten times earlier than the split between the two ecomorphs (Figure 6.6). Thus, it is unlikely that a secondary invasion from the neighbouring river systems contributed to the divergence of the ecomorphs.

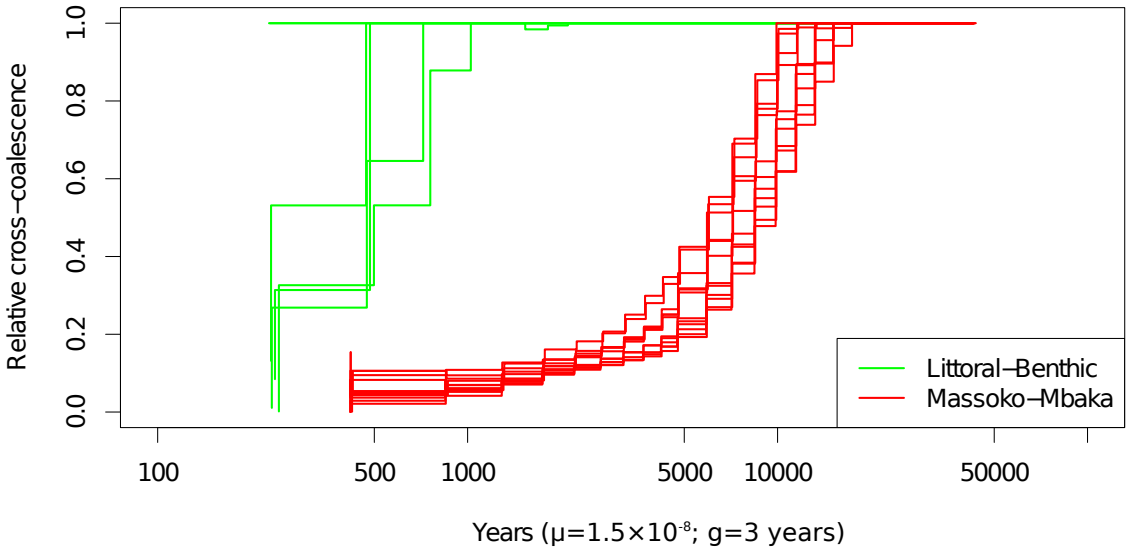


Fig. 6.6 **MSMC cross-coalescence between littoral and benthic ecomorphs (green) and between all Massoko high-coverage individuals and the sample from Mbaka river.** All Massoko individuals split from Mbaka approximately ten times earlier than any separation between benthic and littoral ecomorphs is observed. The time axis values assume: 1) average generation time g=three years and 2) per generation mutation rate μ=1.5×10^{-8}, making the assumption that the μ in cichlids is similar to μ estimated in human studies [15]. Direct estimate of μ in cichlids is not available.

6.3 Population size estimates

The number and frequency of heterozygous sites is a simple summary statistic estimating the level of genetic polymorhism in a population, and thus is indicative of long-term effective population size (N_e) over the past of order N_e generations [173]. Figure 6.7 shows heterozygosity in Itamba, Massoko, and in additional *A. calliptera* populations. Itamba individuals are considerably more heterozygous than fish from Massoko. This is interesting, because Itamba is smaller; it covers only 32% of the surface area, and has only 8% of the water volume of Massoko. Within Massoko, genome-wide average heterozygosity is lower in the benthic individuals compared with the littoral, consistent with benthic being the derived morph. There are also statistically significant differences between heterozygosity between the low-coverage and high-coverage samples, both in Massoko benthic (Welch two sample, two sided t-test: $p = 0.02$) and in Massoko littoral fish ($p = 4.6 \times 10^{-5}$). This result suggests that my variant calling pipeline under-calls heterozygous sites in low-coverage samples on average by approximately 3% in the benthic samples and 5% in the littoral samples. As already discussed in chapter 5, heterozygosity in the additional *A. calliptera* varies by more than an order of magnitude, reflecting their complex and disparate population histories.

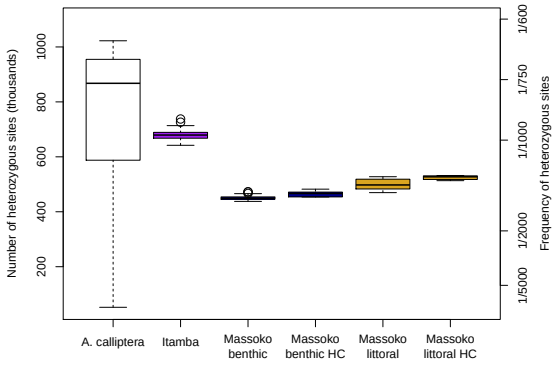


Fig. 6.7 **Heterozygosity in Itamba, Massoko, and additional *A. calliptera* populations** Lake Massoko individuals sequenced to high coverage (see Section 3.1) are shown separately, denoted by 'HC'.

The amount of linkage disequilibrium (LD) observed in a population can be used to estimate N_e over more recent history than sequence heterozygosity. LD between variants further apart from each other reflects more recent N_e than LD between variants that are closer together [210]. Figure 6.8 shows the decay of LD with distance in Itamba, and Massoko benthic and littoral ecomorphs. Interestingly, Massoko benthic fish have both the highest short distance LD, and the lowest long distance LD (beyond 42kb), suggesting a recent increase in N_e, compared to the there two populations. The Itamba samples, on the other hand, have lower short range LD (consistent with the high heterozygosity in Itamba), but have the highest level of LD for SNPs separated by more than 20kb. This pattern persists beyond the 100kb distance. The high level of long range LD in Itamba is evidence for low recent N_e.

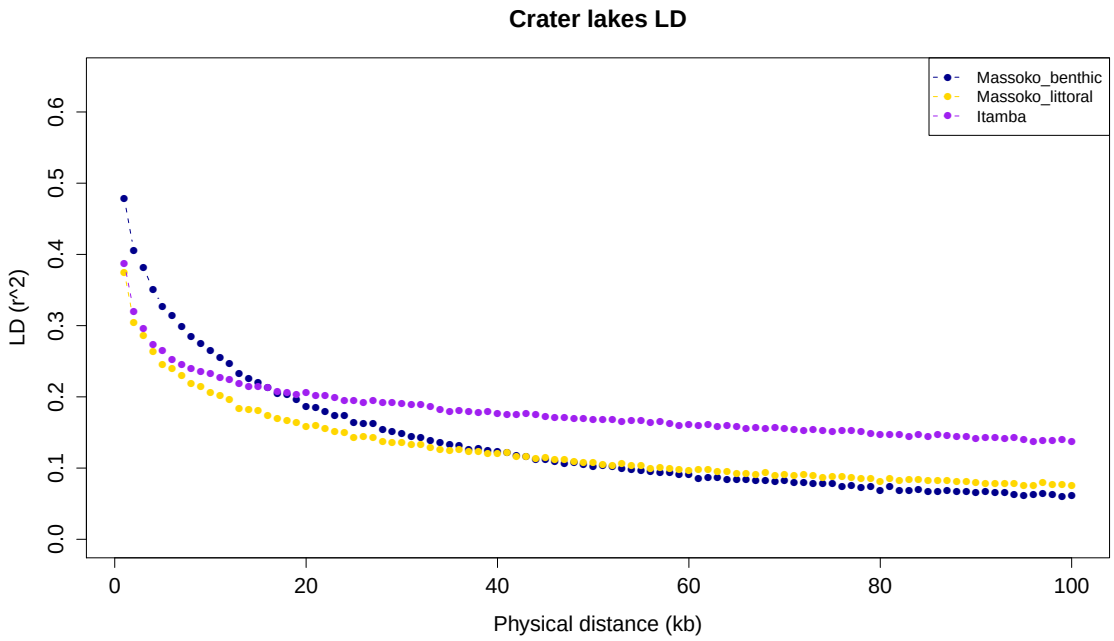


Fig. 6.8 **Decay of linkage disequilibrium in Itamba and Massoko ecomorphs** The figure shows average r^2 in 1kb windows.

MSMC estimates population size history

A more detailed picture of population size history may provide vital clues regarding the evolutionary history of the studied *Astatotilapia* populations. Therefore, I applied the Multiple sequentially Markovian coalescent (MSMC) model, developed by Schiffels and Durbin [209], to infer the history of N_e from the distribution of times since the most recent common ancestor between two alleles in an individual (2 haplotypes) or the first coalescence (common ancestor) for a pair of alleles among a set of four haplotypes. Using four haplotypes enables the algorithm to infer more recent N_e, because the first coalescence among the set of four haplotypes is typically more recent than when only two haplotypes are used.

Figure 6.9 shows MSMC estimates of N_e history for Massoko ecomorphs, Itamba, and *A. calliptera* from Mbaka river. For both Massoko ecomorphs, we see a pronounced drop in N_e (by approximately an order of magnitude), starting at 10,000 years ago, corresponding to the time of split from the Mbaka river population (Figure 6.6), and reaching a minimum at ~3,000. The drop could be related to a nearby volcanic eruption that threw out a layer of pumice that likely floated on the surface and affected the lake ecology, as suggested by M. Genner (pers. comm.) based on evidence from core samples. Consistent with the LD evidence, recent N_e increase in Massoko is more

pronounced in the benthic ecomorph (starting at ~1,000 years ago, corresponding to the split time between the two ecomorphs: Figure 6.6). Also consistent with the LD evidence, the recent (~1,000 years ago) levels of N_e in Itamba are very low.

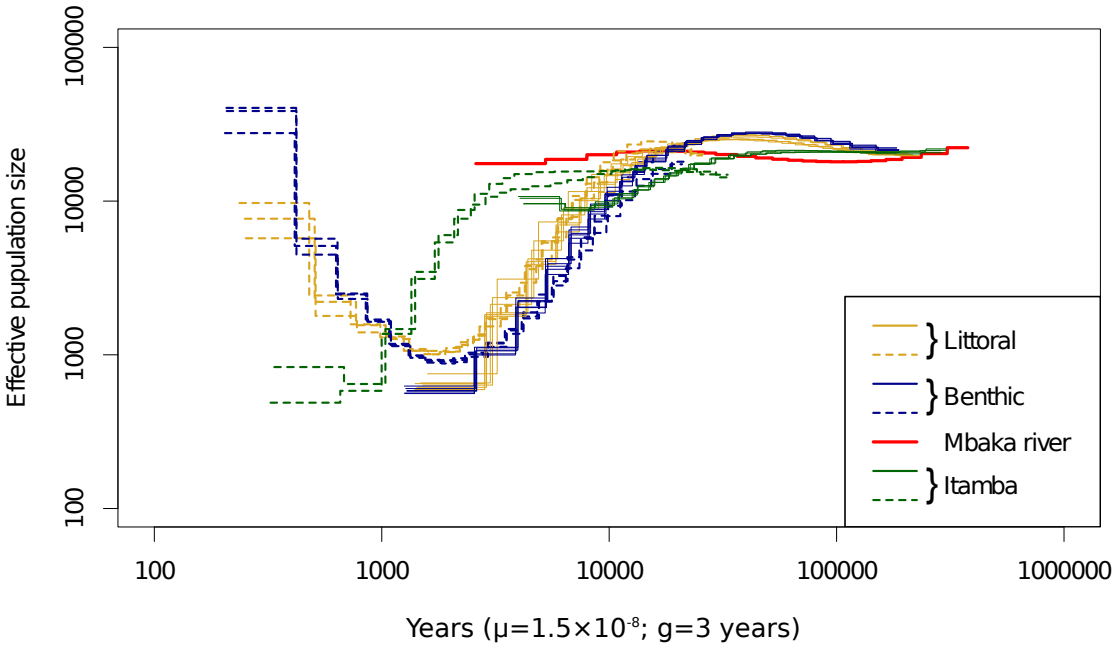


Fig. 6.9 **Inferred population size histories for Massoko ecomorphs, Itamba, and *A. calliptera* from Mbaka river.** Two haplotype inferences (solid lines) are based on six high coverage (HC) individuals each from the benthic and littoral ecomorphs, four Itamba individuals and the *A. calliptera* specimen collected from Mbaka river. Each solid line line corresponds to data from one individual. For the four-haplotype inference mode (inferring more recent N_e; dashed lines), I combined individuals into pairs (i.e. three pairs each for benthic and littoral ecomorphs; two pairs for Itamba). Thus each dashed line corresponds to data from two individuals. The four-haplotype inference mode could not be used for Mbaka river, as only one individual from Mbaka has been sequenced. Therefore, I do not have estimates of more recent population size history for Mbaka river. The time axis values assume: 1) average generation time g=three years and 2) per generation mutation rate $\mu = 1.5 \times 10^{-8}$, making the assumption that the μ in cichlids is similar to μ estimated in human studies [15]. Direct estimate of μ in cichlids is not available.

6.4 Islands of speciation

Interestingly, there are no fixed differences between Massoko benthic and littoral ecomorphs. Genome-wide divergence F_{ST} is 0.038, and almost half (47.6%) of the variable sites have zero F_{ST} (Table 6.4). Above the low background, a genome-wide F_{ST} profile shows clearly demarcated 'islands' of high differentiation (Figures 6.10A, 6.3C). For single sites, the maximum F_{ST} is 13.6 standard deviations (s.d.) above the mean, and 7,543 sites have F_{ST} over 6 s.d. above the mean. By contrast, comparisons of the combined Massoko population and Itamba population revealed a pattern of consistently high F_{ST} across the genome (Figure 6.10B). There are no statistical outliers (Figure 6.10C): not a single site has F_{ST} more than 3 s.d. away from the mean. Similar results were obtained when varying the window size to comprise 15, 50, 100, or 500 variants (Table 6.4; Figures B.1, B.2).

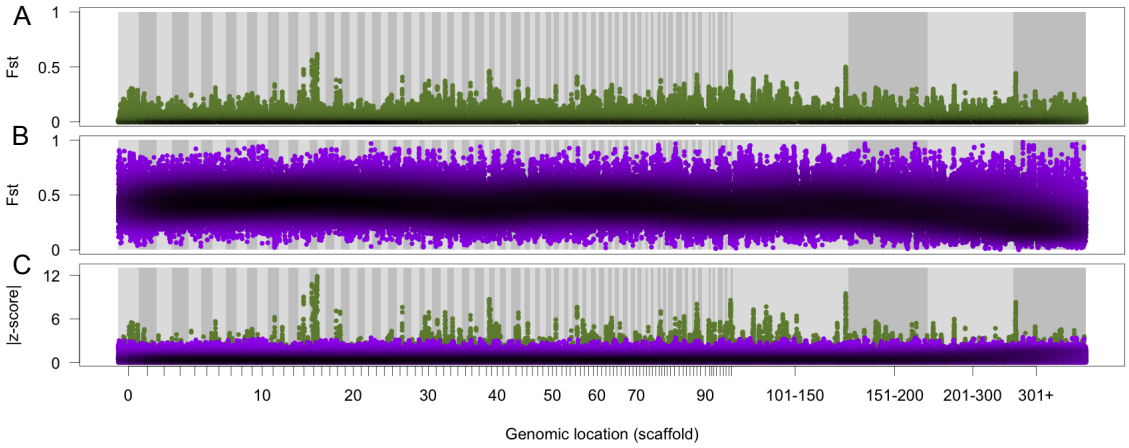


Fig. 6.10 **Genome-wide pattern of F_{ST} divergence in windows of 15 variants each.** Darker colour indicates greater density of datapoints. **(A)** Divergence between benthic and littoral ecomorphs within Massoko. **(B)** Divergence between combined Massoko and Itamba populations. **(C)** Absolute standard scores of Massoko-Itamba divergence (purple) overlaid on divergence between benthic and littoral ecomorphs (green).

Table 6.4 **A summary of sliding-window based F_{ST} calculations.**

Window size (variants)	Average length (bp)	F_{ST} range	Median F_{ST}	Proportion with zero F_{ST}	95th percentile	99th percentile
1	NA	0.00 - 0.72	0.003	0.476	0.126	0.247
15	5,369	0.00 - 0.66	0.016	0.258	0.134	0.24
50	17,839	0.00 - 0.62	0.018	0.208	0.129	0.231
100	35,455	0.00 - 0.60	0.019	0.171	0.126	0.225
500	174,390	0.00 - 0.46	0.024	0.064	0.115	0.197

A role of selection in generating genomic 'islands'

To evaluate the extent to which the observed peaks of divergence could be explained by neutral processes, I used the coalescent simulator `ms` [211] to generate neutral (i.e. without selection) samples under two models of species formation: 'Isolation after migration' (IAM) and 'Isolation with migration' (IWM), as defined by Sousa and Hey [212] (Figure 6.11). Under both models, the divergence begins in the presence of gene-flow. Under the IWM model, gene-flow continues until the present, whereas under the IAM model it ceases at time T_1 (Figure 6.11).

Under both models I fixed the migration parameter `M=5` to simulate moderate bidirectional migration. The parameter `M` is is defined as $4N_0m$, where m is the fraction of each subpopulation made up of new migrants each generation. It was not intuitively clear to me how strong migration is implied by different values of `M`. Therefore, I calculated migration probabilities for a range values of values; these are presented in Table B.3. Next, under the IAM model I fixed the migration cessation time T_1 to equal half of the initial split time T_2. Finally, I fitted the split time parameter to obtain overall $F_{ST} = 0.038$ (over all the simulated sites), matching the overall F_{ST} divergence observed between the benthic and littoral ecomorphs (see Figure B.3).

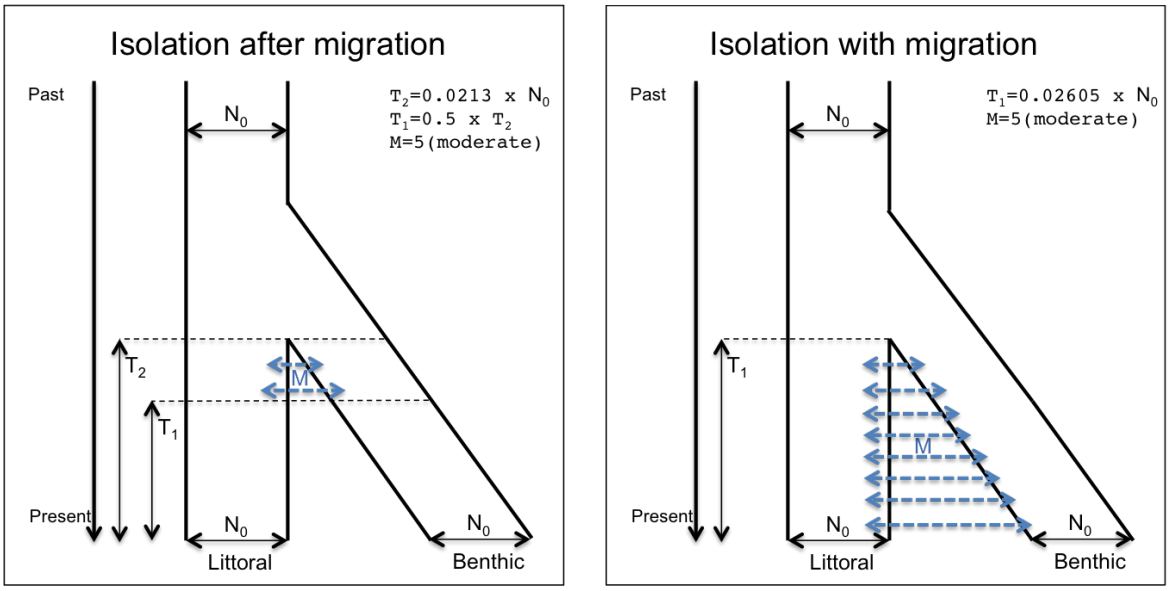


Fig. 6.11 **Models of species formation used for neutral coalescent simulations**

Comparing the distributions of F_{ST} values calculated from observed data (observed F_{ST}) and from the simulated data (simulated F_{ST}) using quantile-quantile plots (Figure 6.12) revealed that approximately the top 1% of observed F_{ST} values are higher than the corresponding simulated values, strengthening the evidence for the role of

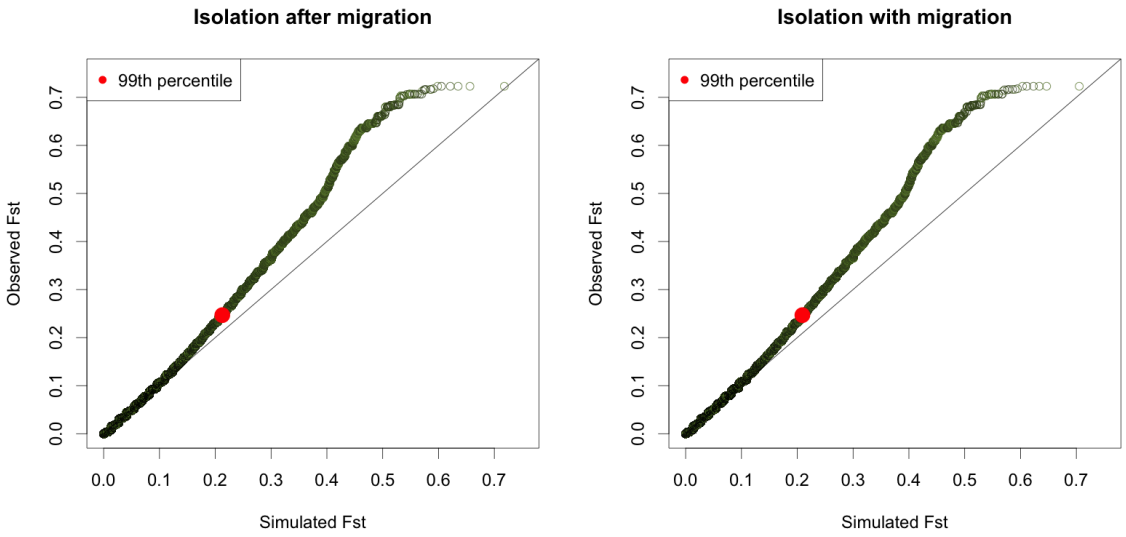


Fig. 6.12 **Comparing the distributions of observed and simulated F_{ST} values** Quantile-quantile plots comparing Massoko benthic-littoral F_{ST} observed in whole genome sequencing with F_{ST} from neutral coalescent simulations under two models of species formation. Darker colour indicates greater density of datapoints, and the position of the 99th percentile of both distributions is indicated in red. Very similar patterns are observed for simulations under both models.

divergent selection at these sites or at sites linked to them. The same conclusion can be drawn from simulations under both models.

To assess how the comparisons of observed and simulated F_{ST} values are affected by changes in simulated population sizes and the migration rates, I explored additional demographic scenarios for the IWM model. I simulated the IWM model with a period of reduced population size (bottleneck) around the split time between the two forms (Figure 6.13). Briefly, under this model, population sizes are reduced to $0.1\times N_0$ between the time points $T_1 = 0.5\times T_2$ and $T_3 = 1.5\times T_2$, where T_2 is the split time and N_0 is the current population size in both populations. For times further back in time (before $1.5\times T_2$), the population size is $0.8\times N_0$. These parameters were selected to approximately mimic population size changes inferred by MSMC (section 6.3). I also attempted to radically increase the migration parameter to `M=10`, `M=20` and `M=40`, significantly increasing the migration probabilities (Table B.3). As in all simulations, the split times (T_2) were fitted to match the overall observed F_{ST}.

The quantile-quantile plots (Figure 6.13) comparing F_{ST} from the additional simulations with F_{ST} values calculated from observed data show very little qualitative difference between the different demographic scenarios, confirming previous studies

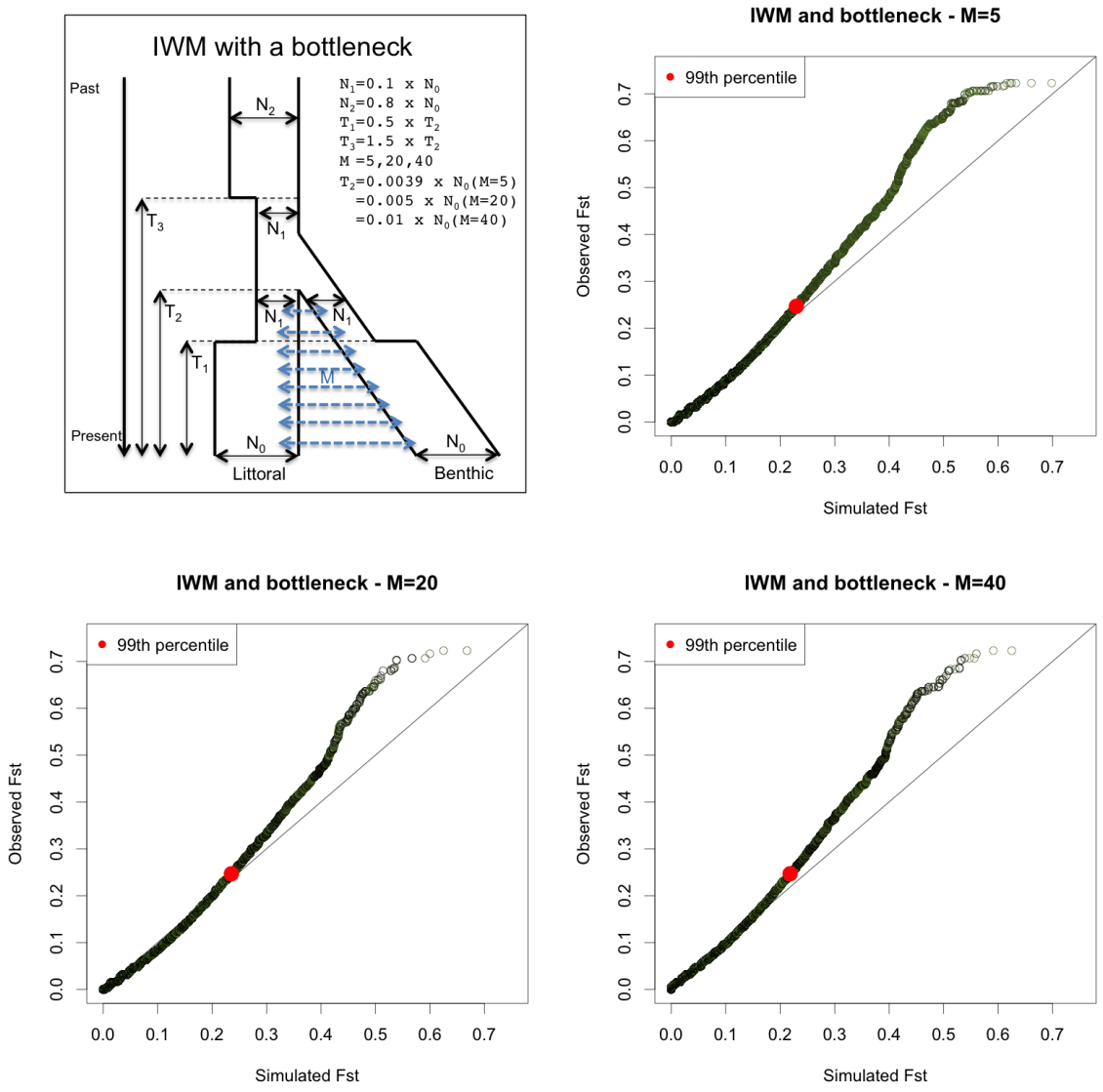


Fig. 6.13 **Comparing the distributions of observed and simulated F_{ST} values for additional demographic models** The isolation with migration (IWM) model of species formation with a strong population bottleneck (top left) was used for neutral coalescent simulations, with a range of values for the migration parameter M. Quantile-quantile plots compare the distributions of simulated FST values under this model with observed benthic-littoral divergence (top-right and bottom). Darker colour indicates greater density of datapoints, and the position of the 99th percentile of both distributions is indicated in red. Similar patterns are observed for all simulations indicating approximately the top 1% of observed values are higher than corresponding simulated values.

and theoretical predictions asserting the distribution of F_{ST} values tends to be robust to demography [213]. Overall, these results suggests the following two conclusions: a) given an overall genome-wide level of F_{ST}, demography has a limited effect on the distribution of F_{ST} simulated under neutrality; b) the pattern of F_{ST} observed in the Massoko benthic-littoral divergence is very unlikely to have been generated by neutral processes alone. Specifically, the results are consistent with divergent selection acting on sites with approximately the top 1% of observed F_{ST} values (approximately $F_{ST} \geq 0.25$).

Loci underlying isolating traits in speciation with gene-flow

I identified genomic regions with observed benthic-littoral $F_{ST} \geq 0.25$ (i.e. with F_{ST} above maximum levels seen in neutral simulations) (Methods - section 6.8.3). For this I used windows of 15 variants each - providing a balance between fine genomic resolution and reducing stochastic variation by averaging over variants. Small gaps arising from brief dips of F_{ST} below the threshold were eliminated by merging regions within 10kb of one another. I found 344 such regions, with total length of 8.1Mb (~1% of the genome). Next, to focus on the more significant outliers, I narrowed the list down to a set of 98 highly diverged regions (HDRs) for further characterisation (Table B.4) by adding the requirement that at least one 10kb window must have reached $F_{ST} \geq 0.3$. The HDRs vary in length from 4.4kb to 285kb (median 36.1kb), with total length of 5.5Mb.

To investigate if the high F_{ST} divergence in HDRs is caused by reduced diversity or by reduced gene-flow, I calculated Nei's d_{XY}, an absolute measure of sequence divergence [66]. In contrast to studies re-examined by Cruickshank and Hahn [66], I found that, overall, d_{XY} in Massoko is significantly higher in the confirmed HDRs compared with the rest of the genome ($P<2.2\times10^{-16}$, two-tailed Mann-Whitney test; Figure 6.14A). Individually, 55 HDRs have d_{XY} above the 90th percentile of the genome-wide distribution. Post-split selective sweeps or other types of linked selection in the benthic or littoral populations would not be expected to generate such 'islands' of high differentiation in d_{XY} [66]. Therefore, these 55 regions (listed in Table 6.5) are the best candidates for loci underlying isolating barriers and being true 'islands of speciation'.

A key prediction of speciation with gene-flow models is that loci causing speciation should be located in relatively few linked clusters within the genome [71, 72]. As described by Cruikshank and Hahn: "With more than even a few islands that do not introgress because of selection, the number of recombinant individuals containing

Table 6.5 **Candidate 'islands of speciation'** The maximum (**max**) d_{XY}, π_{diff}, and F_{ST} values for each HDR, together with the quantile in the corresponding distributions ($\mathbf{F}_x$). Linkage group (**LG**) assignment for each scaffold (**sc**) is based on a linkage map published in [214] - the **?** sign signifies that the scaffold could not be placed to a linkage group.

LG	sc	start	end	max d_{XY} $\times 10^{-3}$	$\mathbf{F}_x$ (max d_{XY})	max π_{diff} $\times 10^{-4}$	$\mathbf{F}_x$ (max π_{diff})	max F_{ST}	$\mathbf{F}_x$ (max F_{ST})
	15	1934238	1967068	1.53	95.44%	5.33	97.62%	0.52	99.97%
	15	2912637	2961336	1.35	92.98%	1.83	75.74%	0.43	99.93%
	15	4641580	4880808	1.68	96.54%	2.92	88.81%	0.56	99.99%
LG5	15	4907565	5049805	2.09	98.09%	5.23	97.47%	0.45	99.95%
	15	6705210	6818468	2.07	98.06%	11.2	99.84%	0.59	99.99%
	15	7208463	7304325	1.60	96.01%	5.12	97.31%	0.62	100.00%
	15	7507678	7592890	1.75	96.98%	4.86	96.82%	0.63	100.00%
	18	6702155	6728393	1.48	94.86%	9.39	99.66%	0.38	99.84%
	0	11529594	11540402	1.92	97.66%	2.53	85.30%	0.31	99.65%
	0	11994849	12015103	1.55	95.62%	2.92	88.87%	0.31	99.61%
LG7	32	4518067	4589712	1.80	97.21%	3.17	90.64%	0.39	99.88%
	32	4886114	4908718	1.66	96.41%	6.35	98.68%	0.32	99.67%
	99	355072	639642	1.41	94.03%	7.50	99.25%	0.49	99.97%
	14	3582492	3609841	1.36	93.22%	6.43	98.74%	0.32	99.70%
LG12	14	3661853	3697260	1.45	94.47%	6.01	98.41%	0.51	99.97%
	43	3655049	3696771	1.62	96.21%	1.41	67.26%	0.34	99.77%
	57	46109	77869	1.36	93.19%	6.82	98.99%	0.33	99.72%
LG20	108	814090	941030	2.04	97.96%	4.08	94.84%	0.39	99.87%
	164	0	113596	1.27	91.55%	5.26	97.51%	0.30	99.58%
	164	196412	276496	1.60	96.02%	7.91	99.40%	0.31	99.62%
	30	183937	257768	1.62	96.19%	8.06	99.44%	0.35	99.79%
	30	797497	844062	1.31	92.39%	3.92	94.25%	0.32	99.71%
	31	4041909	4078026	1.47	94.77%	2.39	83.83%	0.32	99.68%
	82	2236206	2273645	1.75	96.99%	7.26	99.17%	0.36	99.80%
LG23	88	819852	845401	1.27	91.42%	4.70	96.51%	0.34	99.76%
	88	1194601	1316288	1.50	95.08%	11.3	99.89%	0.41	99.90%
	88	1372483	1527476	1.86	97.48%	11.3	99.88%	0.46	99.95%
	88	1732907	1868455	1.38	93.55%	10.1	99.76%	0.35	99.79%
	95	1001404	1044619	1.35	93.01%	6.99	99.08%	0.33	99.71%
	51	1450783	1493272	1.53	95.44%	3.93	94.31%	0.31	99.67%
LG8	113	1062779	1122847	3.11	99.27%	5.08	97.25%	0.43	99.93%
	190	805453	832175	1.21	90.13%	1.08	58.40%	0.33	99.73%
LG3	126	420889	439080	3.78	99.53%	26.6	99.99%	0.32	99.69%
	186	693304	711017	1.38	93.60%	3.08	90.02%	0.30	99.60%
LG19	120	918534	962612	1.99	97.86%	4.35	95.64%	0.36	99.81%
	162	1227615	1263777	1.47	94.78%	1.49	69.10%	0.31	99.63%
?	39	465841	688825	2.37	98.59%	5.94	98.35%	0.47	99.96%
	39	2323506	2340670	1.22	90.40%	1.86	76.19%	0.31	99.66%
?	148	1458116	1644136	2.48	98.75%	10.3	99.74%	0.50	99.97%
	148	1669247	1754172	2.28	98.45%	6.05	98.45%	0.33	99.74%
LG18	6	2399603	2417150	1.46	94.65%	9.49	99.68%	0.30	99.61%
LG2	11	5426321	5452278	1.42	94.17%	2.97	89.21%	0.33	99.72%
LG13	26	5297874	5318387	1.59	95.97%	2.40	83.92%	0.30	99.59%
LG4	55	3423595	3500130	1.61	96.12%	4.76	96.63%	0.42	99.91%
LG11	64	55966	175700	1.42	94.08%	8.53	99.53%	0.34	99.75%
LG15	78	6039	59940	1.49	94.99%	2.28	82.59%	0.38	99.85%
LG14	84	2399084	2517997	1.40	93.81%	11.0	99.79%	0.38	99.85%
LG6	97	2188270	2212097	1.31	92.28%	2.25	82.16%	0.32	99.68%
LG9	229	470627	578767	1.87	97.53%	7.42	99.23%	0.39	99.86%
?	45	2785077	2828731	1.74	96.94%	3.24	91.09%	0.31	99.61%
?	91	129230	153938	1.62	96.20%	0.87	51.23%	0.34	99.75%
?	112	1966090	2014162	1.29	91.87%	7.81	99.38%	0.32	99.71%
?	114	1902474	2005892	2.10	98.13%	11.2	99.79%	0.32	99.70%
?	206	177202	290266	1.46	94.63%	8.66	99.56%	0.38	99.84%
?	304	0	70114	1.14	100.00%	18.9	99.98%	0.41	99.91%

the correct combination of parental alleles...becomes vanishingly small." [66, p. 3151]. Instead of a large number of scattered islands, the theory predicts a smaller number of clusters that grow in size due to the 'divergence hitchhiking' process. We tested this prediction using a recently generated linkage map [214] and found that at least 27 out of the 55 putative speciation islands are co-localised on five linkage groups (LGs), with 26 of them clustered within their respective LGs (Figure 6.14B; Table 6.5). These potential speciation clusters extended for approximately 25cM on LG5, 40cM on LG7, 30cM on LG12, and 5cM on LG20 and 45cM on LG 23. In total, these regions account for under 7% of the genome, suggesting that divergence hitchhiking may play a role in shaping the observed pattern of genomic differentiation.

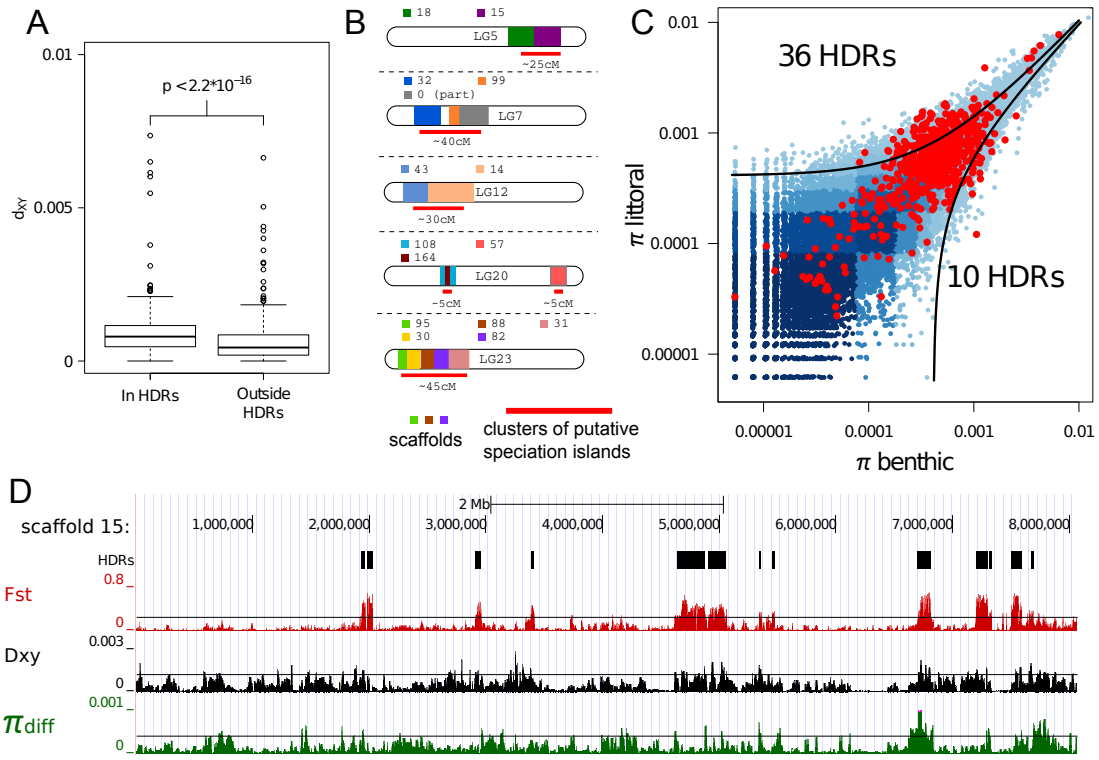


Fig. 6.14 **Islands of speciation between benthic and littoral ecomorphs. (A)** d_{XY} is significantly higher within HDRs ($p < 2.2 \times 10^{-16}$ one-tailed Mann-Whitney test), compared with the rest of the genome. **(B)** Clustering of putative speciation islands on five linkage groups. **(C)** Nucleotide diversity (π) within HDRs (red points) and outside HDRs (blue with shading corresponding to density). Each point corresponds to a 10kb window (therefore, there may be multiple points per HDR). 95% of observations lie between the two curves ($y=x \pm 4.1 \times 10^{-4}$). Putative sweeps in the benthic ecomorph are in the top left corner and putative sweeps in the littoral in the bottom right corner. **(D)** Patterns of F_{ST}, d_{XY}, and π_{diff} in a speciation cluster on scaffold 15.

Although genomic islands within these clusters are often separated only by a few hundred kb, F_{ST} divergence between HDRs generally drops to background levels (see Figure 6.14D), with one exception on scaffold 88 where a broader 'continent' of divergence has formed (Figure 6.15).

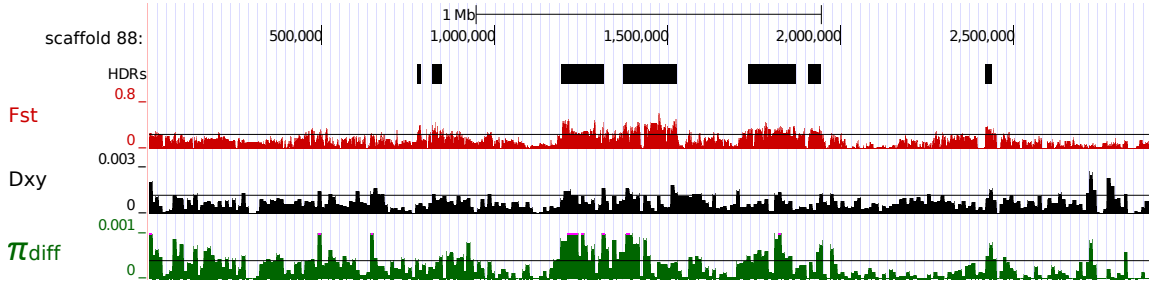


Fig. 6.15 **Patterns of F_{ST}, d_{XY}, and π_{diff} in a speciation cluster on scaffold 88.** In this region of the genome, F_{ST} appears to be elevated above or close to the 99th percentile (black line, representing maximum observed neutral divergence in simulations) over a large distance (2Mb), suggesting that divergence hitchhiking might have started forming a broader 'continent' of divergence.

Signals of adaptation

A reduced level of genetic polymorphism in one subpopulation may be indicative of a recent selective sweep. Overall, the magnitude of difference in nucleotide diversity (π) between benthic and littoral ecomorphs (π_{diff}) is significantly higher in the HDRs than in the rest of the genome ($P<2.2\times10^{-16}$, two-tailed Mann-Whitney test; Figure 6.16A) (Methods - section 6.8.3). Individually 46 HDRs have π_{diff} above the 95th percentile of the genome-wide distribution and are likely to have been under recent positive selection in one of the two ecomorphs. There is a significant overlap between HDRs with high d_{XY} (putative 'speciation islands') and HDRs with high π_{diff} (putative recent selective sweeps) - 35 of 55 high d_{XY} islands also have high π_{diff} (Figure 6.16B; $P=3\times10^{-5}$, hypergeometric test). On the other hand, the 11 putative sweeps that did not lead to elevated d_{XY} are indicative of adaptation not directly involved in reproductive isolation. Reduced nucleotide diversity in high π_{diff} regions, indicative of selective sweeps, was significantly more prevalent in the benthic ecomorph (36 of 46; $P<1.6\times10^{-4}$, two tailed Binomial test; Figures 6.14C, top left; 6.16C), consistent with the benthic ecomorph being derived and undergoing more extensive adaptation. Nevertheless, there are also a small number of strong outliers suggesting selective sweeps in the littoral ecomorph (Figure 6.14C, bottom right).

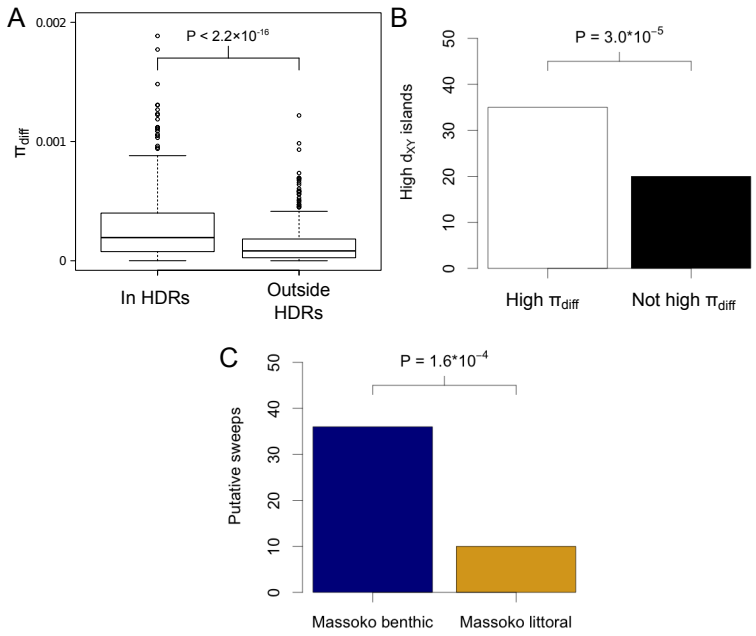


Fig. 6.16 **Characterisation of HDRs in terms of the magnitude of difference in nucleotide diversity between benthic and littoral ecomorphs (π_{diff}). (A)** π_{diff} is significantly higher within HDRs ($P < 2.2\times10^{-16}$, two-tailed Mann-Whitney test), compared with the rest of the genome. **(B)** The overlap between HDRs with high d_{XY} (putative 'speciation islands') and HDRs with high π_{diff} (putative recent selective sweeps) is significant. Thirty-five out of 55 high d_{XY} islands also have high π_{diff} ($P=3\times10^{-5}$, hypergeometric test). **(C)** Thirty-six out of the 46 putative selective sweeps are in the Massoko benthic morph, providing significant evidence that positive selection has been more prevalent in the benthic form ($P<1.6\times10^{-4}$, two tailed Binomial test).

Further support for sympatric divergence

We next tested whether the HDRs correlated with the signal of gene flow from the Mbaka river described above. Compared with the rest of the genome, the HDRs do not have elevated values of Patterson's D (P=0.22, two-tailed Mann-Whitney test; Figure 6.17C), nor elevated f statistics, which were recently proposed as an means by which one could identify introgressed loci [192] (Methods - section 6.8.3) (P=0.08, two-tailed Mann-Whitney test; Figure 6.17D). These results suggest that introgression from Mbaka river did not play a major role in generating the HDRs between the benthic and littoral ecomorphs within Lake Massoko (Figure 6.17), and strengthen the evidence that the ecomorph divergence has been happening within the lake.

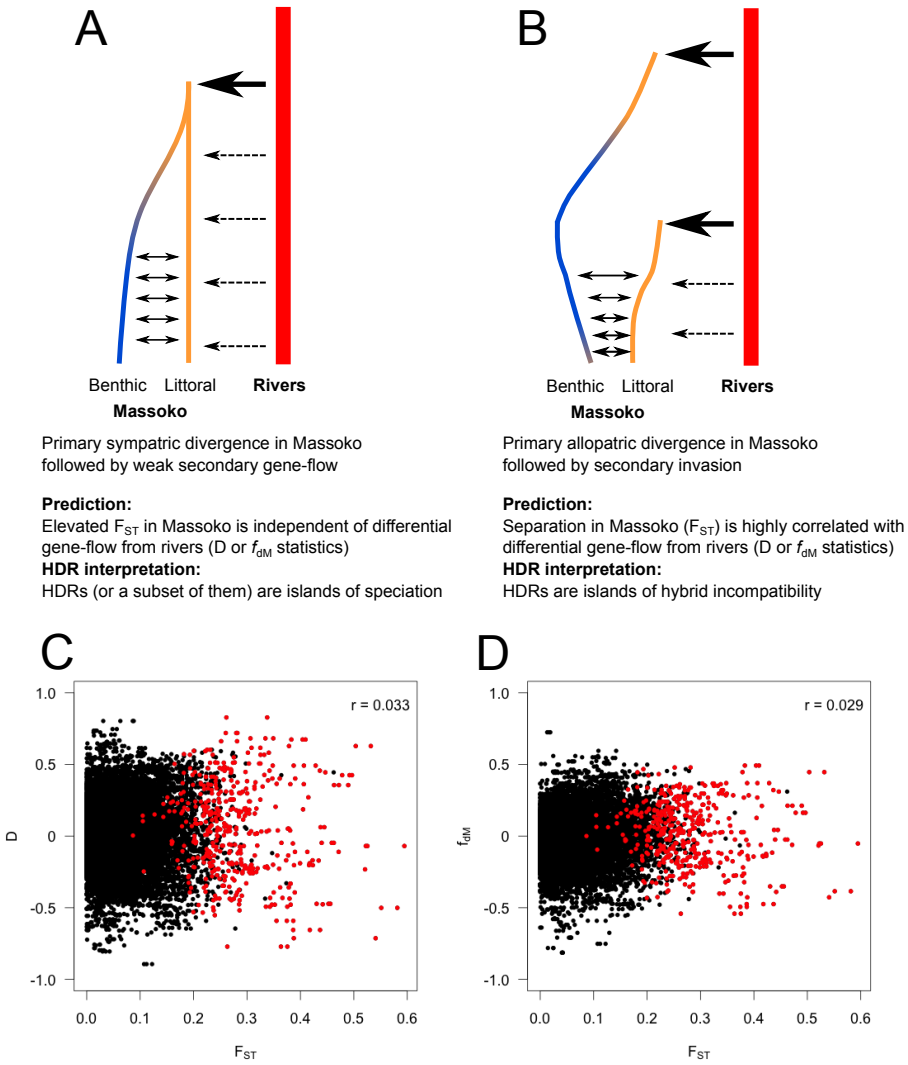


Fig. 6.17 **Evidence against allopatric (double-invasion) divergence between benthic and littoral ecomorphs. (A-B)** Two geographic models of divergence with predictions and interpretation of HDRs. **(A)** Primary sympatric divergence in Massoko. **(B)** Primary allopatric divergence. **(C-D)** F_{ST} between benthic and littoral ecomorphs is independent of levels of differential gene-flow measured by **(C)** Patterson's D, and **(D)** the f_{dM} statistic for locating introgressed regions. F_{ST} was averaged over windows with 15 variants, and D and f_{dM} were averaged over windows of 50 informative variants (i.e. variants with ABBA or BABA patterns). Values for windows within HDRs are shown in red, values for windows outside HDRs are shown in black. Pearson correlation coefficients r are displayed in the top right corner of each figure. The lack of correlation between either D or f_{dM} with F_{ST} is consistent with the predictions of the model of primary sympatric divergence.).

6.5 Divergent SNPs associated with mate choice

Many recently diverged taxa, particularly those not geographically isolated, show stronger pre-mating isolation than post-mating isolation [35, 215]. This also tends to be true with haplochromine cichlids [216, 217]. Therefore, a laboratory experiment was set up with George Turner and his team in Bangor to test for reproductive isolation resulting from direct mate choice (Methods - section 6.8.4). Fifty Massoko females were given a choice from sixteen males representing the variety of male phenotypes. In parallel, I designed a SNP assay with 117 polymorphic sites representing 44 (HDRs) identified from the first 12 high-coverage Massoko samples sequenced in February 2013 (section 3.1). The regions are listed in Table 6.6.

Table 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.** F_{ST} values are based only on the genotyped 18 Massoko benthic and 16 littoral males (the whole-genome sequenced individuals and other individuals used in the mate-choice trial are not included).

Variant coordinates	F_{ST}	Variant coordinates	F_{ST}	Variant coordinates	F_{ST}
scaffold 1:4365113	0.466	scaffold 40:1588650	0.292	scaffold 88:1786645	0.71
scaffold 1:4365291	0.466	scaffold 40:1588728	0.292	scaffold 88:1786888	0.677
scaffold 6:7294300	0.405	scaffold 40:1621279	0.273	scaffold 88:1825810	0.71
scaffold 7:3305104	0.291	scaffold 40:1882810	0.206	scaffold 88:1825839	0.742
scaffold 7:3305216	0.291	scaffold 49:3025847	0.503	scaffold 88:1886989	0.665
scaffold 7:3313226	0.323	scaffold 55:2299953	0.025	scaffold 88:1923134	0.71
scaffold 7:3318131	0.262	scaffold 55:3423696	0.419	scaffold 88:1940222	0.581
scaffold 7:3318579	0.262	scaffold 55:3424572	0.456	scaffold 88:1940476	0.677
scaffold 12:3793589	0.416	scaffold 55:3435034	0.382	scaffold 88:1942123	0.581
scaffold 12:3793994	0.416	scaffold 55:3435507	0.382	scaffold 88:2222087	0.243
scaffold 14:3663857	0.424	scaffold 55:3451516	0.303	scaffold 88:2222122	0.243
scaffold 14:3669941	0.424	scaffold 55:3453112	0.303	scaffold 88:2429606	0.665
scaffold 14:4168852	0.232	scaffold 55:3480538	0.345	scaffold 88:2470678	0.345
scaffold 15:2959443	0.396	scaffold 55:3483480	0.378	scaffold 88:2473383	0.345
scaffold 15:2962256	0.135	scaffold 67:1282346	0.171	scaffold 88:2487048	0.135
scaffold 15:5455316	0.22	scaffold 67:1284312	0.171	scaffold 91:11230	0.101
scaffold 15:5458471	0.174	scaffold 67:1288981	0.219	scaffold 91:54791	0.159
scaffold 15:7238850	0.659	scaffold 82:2709101	0.076	scaffold 91:55547	0.159
scaffold 15:7251797	0.701	scaffold 82:2724117	0.076	scaffold 91:117365	0.092
scaffold 15:7252309	0.657	scaffold 82:2731927	0.098	scaffold 91:528794	0
scaffold 15:7254754	0.659	scaffold 87:112	0.159	scaffold 91:530091	0.008
scaffold 15:7269475	0.659	scaffold 87:5003	0.159	scaffold 97:188537	0.145
scaffold 15:7269758	0.701	scaffold 87:50289	0.098	scaffold 97:193146	0.118
scaffold 18:4359768	0.082	scaffold 87:51344	0.072	scaffold 97:193356	0.19
scaffold 18:4362575	0.038	scaffold 88:1185176	0.198	scaffold 97:2055581	0.295
scaffold 19:377749	0.049	scaffold 88:1198991	0.496	scaffold 126:32880	0.279
scaffold 26:1611369	0.038	scaffold 88:1199168	0.382	scaffold 126:38797	0.249
scaffold 29:3346390	0.011	scaffold 88:1213550	0.496	scaffold 126:1275145	0.264
scaffold 30:4055115	0.458	scaffold 88:1213711	0.462	scaffold 126:1284768	0.377
scaffold 30:6447512	0.194	scaffold 88:1312010	0.616	scaffold 146:1672851	0
scaffold 30:6448182	0.307	scaffold 88:1312055	0.616	scaffold 155:1053156	0.232
scaffold 30:6452026	0.281	scaffold 88:1441936	0.71	scaffold 217:517341	0.419
scaffold 31:426382	0	scaffold 88:1484291	0.71	scaffold 241:14395	0.03
scaffold 31:1867496	0.104	scaffold 88:1488398	0.71	scaffold 241:16987	0.162
scaffold 34:2235172	0.112	scaffold 88:1539583	0.452	scaffold 241:32682	0.165
scaffold 34:2250784	0.152	scaffold 88:1647616	0.387	scaffold 259:243418	0.355
scaffold 34:2802787	0.026	scaffold 88:1654621	0.387	scaffold 259:249905	0.387
scaffold 35:1693366	0.071	scaffold 88:1675293	0.387	scaffold 316:212810	0.456
scaffold 38:642672	0.388	scaffold 88:1781770	0.71	scaffold 316:213151	0.456

I genotyped a reference sample of 18 benthic and 16 littoral males, demonstrating that the SNP assay can reliably separate the ecomorphs along the first principal component (PC1) in PCA (Figure 6.18A, top). 78 out of the 117 sites replicated with high F_{ST} (≥ 0.2) in these additional reference males (Table 6.6). I then genotyped all females and males participating in the mate-choice experiments (Figure 6.18A, bottom) and calculated an average of the PC1 distances between each female and the males she mated with during the experiment, as assayed by microsatellite paternity analysis by Alexandra Tyers at University of Bangor. Richard Challis at University of Bangor then found that compared with expectation under random mating (Methods - section 6.8.4), females had a moderate, but significant ($P=4.3\times10^{-5}$, paired t-test), preference for mating with males genetically similar to themselves (i.e. close to them along PC1) (Figure 6.18B), demonstrating association between HDR variants and mate choice. Assortative mating by genotype was strong among females with positive (littoral) PC1 scores ($P=5.9\times10^{-9}$, paired t-test), while no assortative mating was detected among females with negative (benthic) PC1 scores (Figure 6.18B).

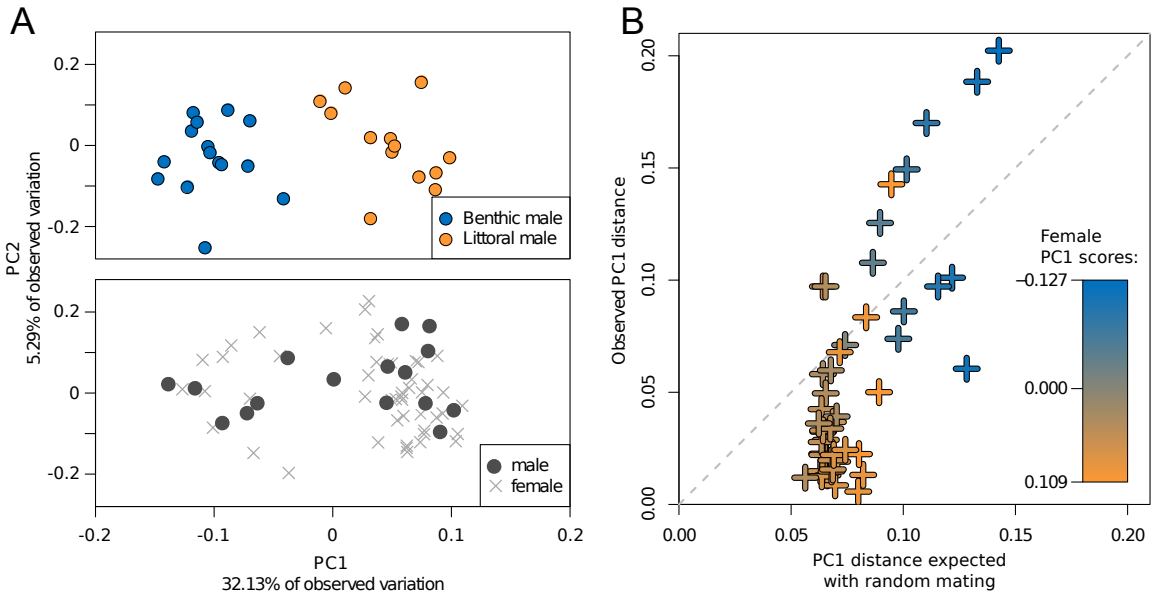


Fig. 6.18 **Mate-choice trials. (A)** PCA based on 117 genotyped SNPs. **Top:** The first axis of variation (PC1) in PCA reliably separates benthic and littoral males in a reference sample. **Bottom:** PC1 positions of females (N=50) and males (N=16) participating in mate-choice trials. **(B)** Results: Each point compares the average of absolute PC1 distances between a female and: males she mated with (observed PC1 distance) and all males she could have mated with (expected PC1 distance). Points are coloured according to the PC1 score of the female. Females below and to the right of the dashed diagonal line on average mate with males more like themselves in terms of PC1 score than would be true if they mated at random.

Stronger mating discrimination by ancestral populations compared to derived ones has been previously found in *Drosophila* and sticklebacks, possibly because low population density following a founder event favours less choosy individuals [218]. However, it is also possible that the benthic ecomorph only mates assortatively in the deep water environment; given that the experiments used wide-spectrum lighting characteristic of shallow water. Overall, the moderate assortative mating suggests a role for sexual selection in ecomorph divergence, but does not indicate that it is a primary force causing population-wide divergence.

6.6 Functions of adaptation

To explore the function of candidate adaptive genes, I performed Gene Ontology (GO) enrichment analysis [219]. Zebrafish (*Danio rerio*) has the most extensive functional gene annotation of any fish species. Therefore, I used the Broad Institute's assignment of orthologs between the *M. zebra* genome and zebrafish [93].

The numbers of genes available for GO analysis, genome-wide and in the enrichment gene-sets, are detailed in Table 6.7. Genome-wide, 13,230 (61.3%) of *M. zebra* genes have an assigned zebrafish ortholog, mapping to 11,810 unique zebrafish genes. Of these, ~7,000 (~33% of the total gene count) have useable GO annotation, the exact number depending on the GO category being assessed.

Table 6.7 **Numbers of genes available for Gene Ontology enrichment analysis** GO categories are denoted as follows: **MF** - molecular function; **CC** - cellular component; **BP** - biological process

Region	Total genes	Unique zebrafish orthologs	Orthologs with GO annotation MF	CC	BP
Whole genome	21,567	11,810	7,086	6,441	6,961
Speciation islands ±50kb	215	146	73	72	75
All HDRs ±10kb	207	132	65	58	64
All HDRs ±50kb	398	288	135	123	133

GO enrichment analysis was performed on three enrichment gene-sets: a) genes in putative 'islands of speciation' ±50kb (enrichment terms in Table B.5); b) genes in all HDRs ±10kb (Table B.6); c) genes in all HDRs ±50kb (Table B.7). There is often an overlap between gene-sets annotated with different GO terms, in part because the terms are related to each other in a hierarchical structure [219]. Therefore, I used the Enrichment Map [220] app for Cytoscape (`http://www.cytoscape.org`) to organise all the significantly enriched terms into networks where terms are connected if they have a high overlap, i.e. if they share many genes. The resulting network, combining results of all three analyses, revealed clear clusters of enriched terms related to: a)

morphogenesis (e.g. cartilage and pharyngeal system development, fin morphogenesis), consistent with morphological differentiation; b) sensory systems (e.g. photoreceptor cell differentiation), consistent with previous studies showing the role of cichlid vision in adaptation and speciation [89, 221]; and c) (steroid) hormone signalling (Figure 6.19).

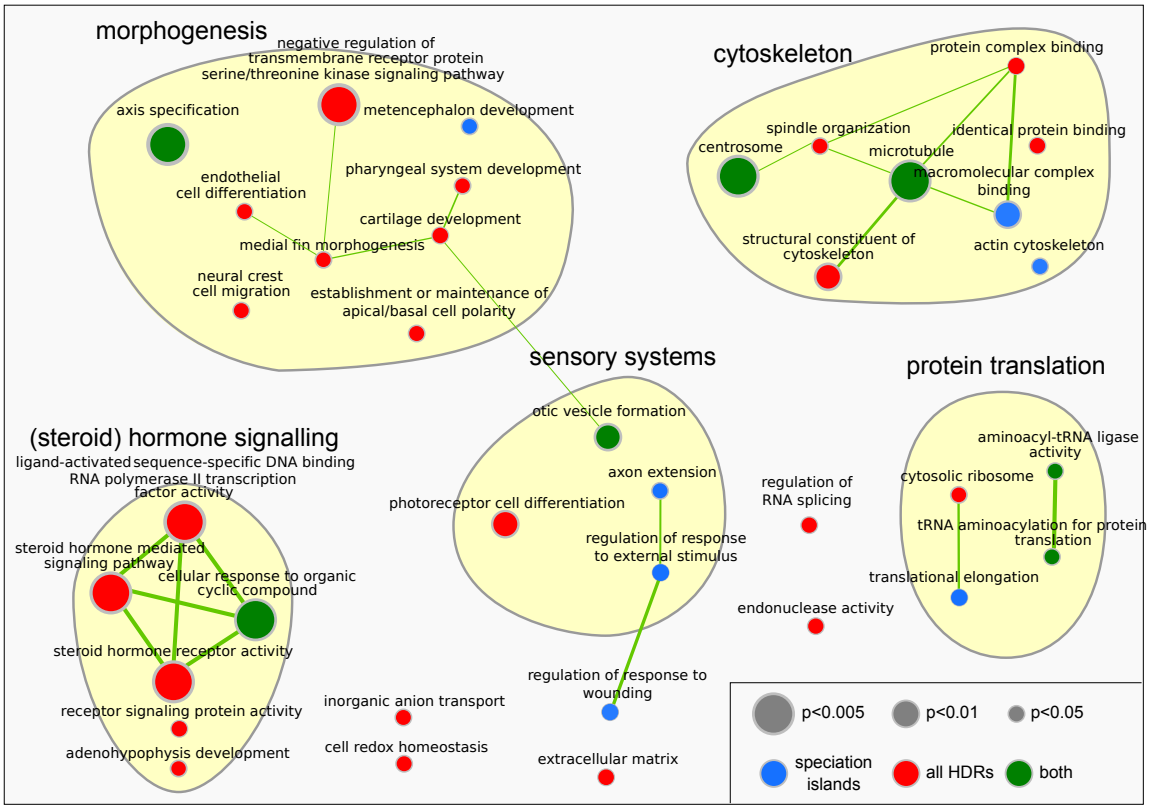


Fig. 6.19 **Enrichment Map for significantly enriched GO terms.** The level of overlap between GO enriched terms is indicated by the thickness of the edge between them. The size of the node indicates the best p-value for the term, and the colour of the node indicates the gene group for which the term was found significant (i.e. has P<0.05 in candidate 'speciation islands' ±50kb - blue; in all HDRs ±10kb or ±50kb - red; or in both groups - green). Broad functional groupings (morphogenesis, sensory systems...) were derived using automatic clustering using clusterMaker [222] and WordCloud [223], followed by manual editing.

The GO enrichment for terms related to sensory systems suggests a role of light environment heterogeneity in the divergence of the two forms. I examined in more detail the functions of candidate genes involved in photoreceptor function (Table 6.8), and two highly diverged alleles of the rhodopsin (*rho*) gene in Lake Massoko (alleles H4 and H5, separated by four amino acid changes; $F_{ST} = 0.39$; Figure B.4). Blue-shifted rhodopsin absorption spectra are known to play a role in deep-water adaptation [221].

Therefore, I initiated a collaboration with Yohey Terai (SOKENDAI, Japan) who expressed rhodopsins from H4 and H5 alleles, reconstructed them with 11-*cis*-retinal, and measured their absorption spectra (Methods - section 6.8.5). The results demonstrate that the H5 allele, associated with the deep-water benthic ecomorph, has a blue-shifted absorption spectrum (Figure 6.20A). The retina-specific retinol dehydrogenase *rdh5* (Table 6.8) produces 11-*cis*-retinal, the visual pigment binding partner of rhodopsin [224], and thus likely has a direct role in dark adaptation. Finally, a mouse ortholog of *rp1l1b* affects photosensitivity and morphogenesis of the outer segment (OS) of rod photoreceptor cells, locating to the axoneme of the OS and of the connecting cilia [225] (Figure 6.20B). Together, these results suggest divergent selection on *rho*, *rdh5*, and *rp1l1b* may facilitate the adaptation of scotopic (twilight) vision to the darker conditions experienced by the benthic ecomorph.

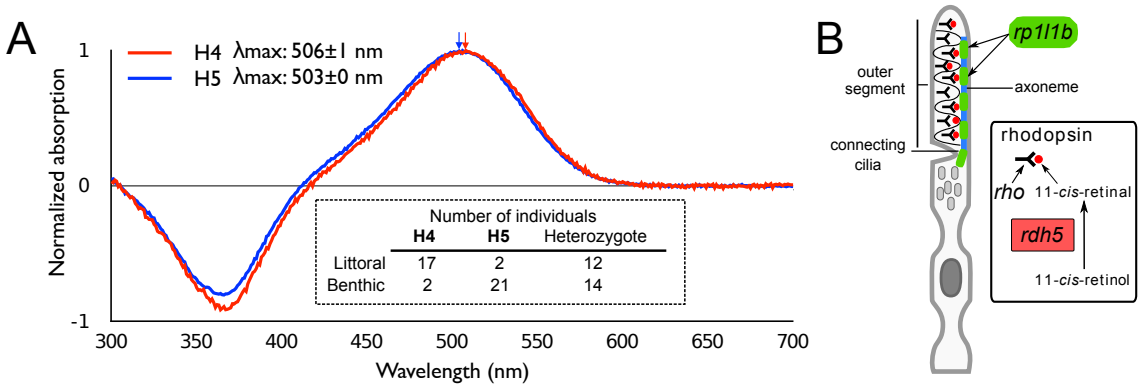


Fig. 6.20 **Rhodopsin and rod cells. (A)** The absorption spectrum of the H5 allele of rho, more prevalent in the benthic ecomorph, is shifted towards blue wavelengths. **(B)** An illustration of the joint roles of *rho*, *rdh5*, and *rp1l1b* in photoreceptor rod cells. *rdh5* produces the chromophore 11-*cis*-retinal that binds *rho*, while *rp1l1b*, located at the axoneme of the outer segment and connecting cilia, also contributes to photosensitivity.

The role of the visual system in driving speciation along light gradients mediated by water depth has also been demonstrated in the *Pundamilia pundamilia* and *Pundamilia nyererei* cichlids of Lake Victoria [89]. In this system, the long-wavelength-sensitive opsin gene (LWS) variation is associated with water depth and male coloration, such that "more red-shifted alleles occur at greater depth" and "populations with predominantly red-shifted opsin alleles have predominantly red males". However, the LWS gene does not exhibit any protein-coding variation in Lake Massoko and it's location on scaffold 202 is not in proximity to Massoko HDRs.

Table 6.8 **Genes contributing to GO enriched terms in sensory perception**

sc	start	end	Gene symbol	Gene description	Entrez ID
15	5522881	5527067	rdh5	retinol dehydrogenase 5 (11-cis/9-cis)	556528
15	6719099	6756944	gnas	GNAS complex locus	557353
15	7474462	7493463	dctn2	dynactin 2 (p50)	394141
30	789904	799043	enpp4	ectonucleotide pyrophosphatase/phosphodiesterase 4	550586
30	853997	869417	rp1l1b	retinitis pigmentosa 1-like 1	101885561
57	14523	18913	mmp9	matrix metalloproteinase 9	406397
64	212132	217320	oprd1b	opioid receptor delta 1b	336529
66	676755	696647	bmper	BMP binding endothelial regulator	338246
84	2552689	2568793	chd	chordin	30161
99	277429	315316	cep290	centrosomal protein 290	560588
112	1985779	2057991	nsmfb	NMDA receptor synaptonuclear signaling and neuronal migration factor b	569891
164	58200	148583	plxnb1a	plexin b1a	561012

Several other HDR associated genes involved in vision have been studied in detail. The *dctn2* gene encodes the p50 subunit of the dynactin complex involved in microtubule-dependent intracellular transport. The strongest effects of *dctn2* morpholino knockdown in zebrafish were observed in observed in photoreceptor cells and in retinal radial glia. In addition, *dctn2* is involved in maintenance of the neuronal projection between the retina and the tectum of the midbrain, and in the survival of mechanosensory hair cells the in inner ear, and the lateral line [226]. A morpholino knockout of *cep290* revealed "defects in retinal, cerebellar, and otic cavity development" in zebrafish and immunogold labelling in mouse photoreceptor cells showed the greatest concentration in their connecting cilium, "supporting a possible ciliary role in the eye" [227].

Chordin and *bmper* both play key roles in many developmental processes by regulating bone morphogenesis proteins, including key roles in the formation of the otic vesicle (an embryonic structure that develops into the inner ear) [228, 229]. While the role of vision in general and opsin genes in particular has been studied extensively, these results suggest that fish divergence involving depth/light gradients may also involve hearing adaptations. The potential for hearing adaptation in fish divergence has so far been largely ignored, but it is likely that low light conditions at greater depths increase the importance of the sense of hearing, for example in predator/prey detection [230]. On the morphological side, hearing adaptations could be reflected for example in the size or shape of the otolith, or of the sensory hair cells [231].

The 14 genes driving GO enrichment in the morphogenesis functional grouping are listed in Table 6.9, and the 13 genes driving enrichment in steroid hormone signalling in Table 6.10. A detailed examination of the functions of these genes, possibly followed by further experiments, could shed light on the roles they may be playing in Massoko ecomorph divergence.

Table 6.9 **Genes contributing to GO enriched terms related to morphogenesis.**

sc	start	end	Gene symbol	Gene description	Entrez ID
15	7349839	7393790	gli1	GLI-Kruppel family member 1	352930
15	7474462	7493463	dctn2	dynactin 2 (p50)	394141
18	2784308	2814518	skib	nuclear oncoprotein skib	30113
18	3389782	3436376	sema3fa	semaphorin 3fa	544658
30	246439	259707	lypd6	LY6/PLAUR domain containing 6	447932
40	1538464	1598298	faf1	Fas associated factor 1	406243
57	1672259	1711225	rarga	retinoic acid receptor gamma a	30606
66	676755	696647	bmper	BMP binding endothelial regulator	338246
84	2552689	2568793	chd	chordin	30161
88	2438486	2477875	acvr2aa	activin A receptor type IIAa	553359
93	2203936	2236454	fbln1	fibulin 1	30240
93	2237233	2253337	fbln1	fibulin 1	30240
99	277429	315316	cep290	centrosomal protein 290	560588
114	1233887	1259975	fn1a	fibronectin 1a	100005469

Table 6.10 **Genes contributing to GO enriched terms in hormone signalling.**

sc	start	end	Gene symbol	Gene description	Entrez ID
0	14000583	14107901	rxraa	retinoid X receptor alpha a	555578
15	7349839	7393790	gli1	GLI-Kruppel family member 1	352930
15	7631396	7638449	tarbp2	TAR (HIV) RNA binding protein 2	336141
39	2242698	2298569	esr2b	estrogen receptor 2b	317733
40	1609941	1613581	dmrta2	doublesex and mab-3 related transcription factor like family A2	474350
57	1658903	1665415	nr1d4a	nuclear receptor subfamily 1 group D member 4a	563150
57	1672259	1711225	rarga	retinoic acid receptor gamma a	30606
64	77011	78338	nr0b2a	nuclear receptor subfamily 0 group B member 2a	403010
74	637448	647727	vtg3	vitellogenin 3 phosvitinless	30518
88	778189	875341	ahr2	aryl hydrocarbon receptor 2	30517
88	2438486	2477875	acvr2aa	activin A receptor type IIAa	553359
108	2974	18121	zgc:171775	zgc:171775	562552
186	675492	721013	rgs12b	regulator of G-protein signaling 12b	378970

6.7 Comparisons to other systems

Overall, the results presented in this chapter suggest a pair of incipient species undergoing divergence with gene flow within the crater lake Massoko. Their overall level of divergence ($F_{ST} = 0.038$) is low compared with background F_{ST} observed in other recent studies of speciation with gene flow in *Anopheles* mosquitoes (S and M form; $F_{ST} = 0.21$) [66], *Ficedula* flycatchers ($F_{ST} = 0.36$) [70], and *Heliconius* butterflies ($F_{ST} = 0.18$) [69], highlighting that we are looking at an early stage of divergence. The MSMC analysis suggests that median effective divergence occurred within the last 500-1,000 years (~200-350 generations), following separation of lake fish from the Mbaka river population around 10,000 years ago (Fig. S4). However, divergence may have started considerably earlier than these times, masked by subsequent gene flow.

Among populations at similar levels of divergence to Lake Massoko ecomorphs are *Timema* stick insects ($F_{ST} = 0.015$ for adjacent and $F_{ST} = 0.03$ for geographically isolated population pairs), where thousands of regions of moderately elevated divergence were found all across the genome [232], and German carrion and Swedish hooded crows ($F_{ST} = 0.017$), that have strongly diverged with fixed differences, but at fewer than five loci [77]. In Massoko, we observe an intermediate pattern between these two extremes, with a few dozen moderately elevated islands, clustering within the genome indicating close linkage, and no fixed differences. A genome-wide pattern with multiple loci of moderate divergence suggests a genomic architecture similar to the ecological divergence of a sympatric threespine stickleback pair in Paxton Lake, Canada [233], and the sympatric divergence of dune-specialist sunflowers, Helianthus [234].

6.8 Detailed methods

6.8.1 Field sampling and eco-morphological analysis

Field sampling for genetic, morphological and stable isotope samples

Astatotilapia samples from Lake Massoko were collected on 17th July 2011, and from 19th to 25th November 2011 by Martin Genner, George Turner and their teams. Fish were collected using fixed gill nets and SCUBA. On being brought to the surface, fish were given an overdose of anaesthetic (MS-222). From each fish we collected a genetic sample (fin clip) that was stored in ethanol, and cut a fillet of the flank for stable isotope analyses that was sun-dried and stored with desiccant. Samples of potential food sources were also collected and dried, including epilithic algae, sponges and bivalves. Whole fish were preserved in formalin (~4%). Genetic samples (fin clips) of outgroup *Astatotilapia calliptera* were collected opportunistically between 2009 and 2014.

Field sampling for assessment of ecomorph frequency with depth

Astatotilapia samples from Lake Massoko were collected from 28th July to 7th August 2014, 10th to 15th December 2014, and 5th - 24th August 2015. Fish were collected using fixed gill nets, angling and SCUBA. I was part of the field team during the first two trips, in charge of the gill net sampling. Depth was assessed using a plumbline, surface depth meter, and dive gauges. Fish were photographed on collection. All adult males >65mm Standard Length were assigned to an ecomorph on the basis of the colour of body and fins, and gross morphology. Depth was recorded as bottom depth, which means that fish caught in the water column may sometimes be included. This, along with drift of passive fishing gears may perhaps have led to an over-representation of shallow water fish in deeper water records

6.8.2 RAD-seq data processing and analysis

Note: The work on RAD-seq data described in this section was done by Richard Challis, University of Bangor. The methods are included here (in this thesis) for completeness.

DNA extraction and sequencing

DNA was extracted from ethanol-preserved fin tissue from 56 wild caught fish using a standard CTAB-Chloroform extraction method including an RNAase treatment step. This was sent to Floragenex (`http://www.floragenex.com/`) for library preparation using the Sbf1 enzyme and sequencing on an Illumina HiSeq2000 platform, providing 100bp single end reads. The samples were sequenced in two rounds. In the first round sequencing was 28 samples per lane, but 43 individuals obtained less than 1M reads each. In a second round 41 of these 43 individuals were reprepped, and sequenced at 20 and 21 samples per lane. Raw data have been deposited at the NCBI Sequence Read Archive under BioProject PRJNA286304 (Accessions SAMN03768857 to SAMN03768912).

Variant calling and filtering

Samples with fewer than 300000 reads (approximately 20X coverage per tag) were removed. Raw reads for the remaining 42 samples were demultiplexed and adaptor trimmed leaving 89 base reads for use in reference guided RAD tag analysis. Reads were aligned to the Mbaka River *Astatotilapia calliptera* consensus sequence (see section 6.8.3) using `bwa-mem v0.7.12` [132]. An average of 96.2% (±0.3%) of reads mapped to the reference and these mapped reads were filtered to remove reads with terminal alignments and reads that were not uniquely mappable leaving an average of 90.3% (±1.1%) of the original reads in the filtered read set. SNPs were called using the stacks [235] `ref_map.pl` pipeline with a minimum stack depth (`-m`) of 5. The full dataset was filtered to remove SNPs that had been called in less than 75% of samples and the resulting matrix contained 7,906 SNPs and was 82.3% complete.

Phylogenetic trees and constraint tests

Phylogenetic model testing using `ModelGenerator v0.85` [236] supported the use of the GTR + Γ model of sequence evolution with an estimated transition/transversion ratio of 2.65. A maximum likelihood (ML) phylogeny was produced using `RAxML v8.0.22` [198] using the GTRGAMMA model. Support for the ML tree topology was inferred using 100 rapid bootstrap samples [199]. The phylogeny was rooted on *A. tweddlei* and has been deposited in treebase (accession: TB2:S18241).

6.8.3 Whole genome data processing and analysis

Gene annotation

For analyses concerning gene content and function, I used the `V1` gene annotations generated at the Broad Institute as a part of the cichlid genome project [93]. The includes assignment of orthologs between cichlid and medaka, tetraodon, stickleback, and zebrafish genomes.

Whole genome phylogenetic tree

Consensus genome sequences were generated using the `bcftools v1.2 consensus` tool. For each sample, I selected the sequence of one haplotype (as assigned by `beagle` haplotype phasing - see section 3.2.4) by using the `--haplotype=1` option in `bcftools`. All scaffolds except the mtDNA sequence (scaffolds 747, 2036) were concatenated into a single sequence and phylogenetic trees then inferred using `RAxML v7.7.8` [198] under the GTRGAMMA model (General Time Reversible model of nucleotide substitution with the Γ model of rate heterogeneity). The maximum likelihood tree was obtained as the best out of five alternative runs on distinct starting maximum parsimony trees (using the `-N 5` option). Sixty six bootstrap replicates were obtained using `RAxML's` rapid bootstrapping algorithm [199]. It was my intention to run more bootstrap replicates, enough to satisfy `RAxML -N autoFC` frequency-based bootstrap stopping criterion, but this has proven computationally infeasible on a dataset of this size (obtaining the 66 replicates required ~7,647 hours of CPU time). Still 66 replicates provide a reasonable indication of bootstrap support for the maximum likelihood tree. Bipartition bootstrap support was drawn on the maximum likelihood tree using `RAxML -f b` option.

Principal Component Analysis

SNP variants (no indels) with minor allele frequency >= 0.05 were selected using `vcftools v0.1.12b` options `--remove-indels --maf 0.05` and exported in `PLINK` format [194]. The variants were LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Principal Component Analysis on the resulting set of variants was performed using the `smartpca` program from the `eigensoft v5.0.1` software package [195] with default parameters.

Linkage Disequilibrium

First, I obtained a random subsample of approximately 10% of all SNP variants (no indels) from the full joint Massoko, Itamba and *Astatotilapia calliptera* variant set. Then, linkage disequilibrium (LD) was calculated for each variant within 1Mbp window using `vcftools v0.1.12b` options `--hap-r2 --ld-window 1000000`. The plots use the R^2 measure of LD, and show averages within 1000bp windows calculated in the `R` software environment for statistical computing [175].

ADMIXTURE ancestry estimation

All SNP variants were exported in PLINK format [194] and LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. The `ADMIXTURE v1.23` program was then run with default parameters. The postulated number of ancestral populations `K` was set to 1, 2, 3, 4, 5, and 6. From a statistical standpoint, the authors of the software suggest choosing the value of `K` with the lowest cross-validation error. We performed 10-fold cross-validation (`--cv=10`) and found the lowest cross-validation error is with `K=1` (Figure 6.4). ADMIXTURE cross-validation implying `K=1` is a common phenomenon when population differentiation is subtle, but meaningful results can still be obtained with higher values of `K` - see for example the application of ADMIXTURE to HGDP human European data in [193] and Figure S12 therein.

Accessible genome

To obtain an estimate of the length of the genome accessible for accurate variant calling using Illumina short reads I used Heng Li's `SNPable` tool (available from `http://lh3lh3.users.sourceforge.net/snpable.shtml`). The SNPable tool divides the reference genome into overlapping k-mers (sequences of length k - I used k=50), and then the extracted k-mers are aligned back to the genome (I used `bwa aln -R 1000000 -O 3 -E 3`). Then I only kept regions where the majority of overlapping 50-mers were mapped back uniquely and without 1-difference. Excluding gaps (runs of `N` in the reference), this SNPable 'mask' excludes approximately 7.4% of the reference, resulting in an 'accessible genome' of 660,796,086bp.

Heterozygosity

The number of heterozygous sites per individual was calculated using the custom C++ program `evo` (available from `https://github.com/millanek/evo`), with the `stats`

`--hets-per-individual` option, and then divided by the length of the 'accessible genome' (see above) to obtain the frequency of heterozygous sites.

MSMC cross-coalescence analysis

Because results of MSMC rely, in part, on detecting the density of heterozygous sites, we restricted this analysis to high coverage (~15X) samples. Genomic regions on which short reads cannot be uniquely mapped were masked out by a) excluding genomic regions where mapped depth was higher than 35X (more than twice the average genome coverage); b) using Heng Li's SNPable tool (see 'accessible genome' above).

Running MSMC without the `-fixedRecombination` parameter for 100 iterations indicated that the rho/mu parameter is approximately 2. This value was used for all following MSMC runs (i.e rho/mu parameter set to 2: `msmc --rhoOverMu=2 --fixedRecombination -P 0,0,1,1`). Each run of the cross-coalescence analysis used four haplotypes, two from each ecomorph for the benthic-littoral split, and two from Mbaka and two from Massoko for the Massoko-Mbaka split.

Since MSMC relies on long-range haplotype phasing, we re-phased the data using the `shapeit v2.r790` haplotype phasing method [237] including the use of phase-informative reads [238]. Because of the need for long-range phase information, we restricted the analysis to 50 largest genomic scaffolds, comprising ~390Mb of sequence.

MSMC N_e history estimation

The analysis was performed as the above cross-coalescence runs but without the `-P` parameter (i.e. `msmc --rhoOverMu=2 --fixedRecombination`), except that low coverage (~6X) samples had to be used for Lake Itamba. For these samples, I excluded (masked out) genomic regions where mapped depth was higher than 20X.

Coalescent simulations

I used the coalescent simulator `ms` [211] to simulate the divergence of two subpopulations, sampling 74 chromosomes from the first population corresponding to 37 *Massoko benthic* samples and 64 chromosomes from the second population corresponding to 32 *Massoko littoral* samples (`-I 2 74 64` parameter). The simulations were performed under a range of models and demographic scenarios, as described in the main text. Migration rate for the Isolation with migration (IWM) model was included directly in the `-I` parameter (e.g. `-I 2 74 64 5` and for the Isolation after migration migration model was adjusted using the

For each model/scenario: a) the between-population split time (`-ej` parameter) was adjusted to match the overall observed benthic littoral F_{ST} of 3.89%; b) I simulated 500,000 independent samples, each sample with one segregating site (effectively simulating 500 thousand unlinked loci). Therefore, the basic command line for the IWM model looked as follows:

```
ms 138 500000 -s 1 -I 2 74 64 M -ej splitT 1 2
```

where `M` is the migration parameter (see Table B.3), and `splitT` stands for the split time.

Calculating F_{ST} and defining HDRs

F_{ST} was calculated both for simulations and for the cichlid data using my own code implemented in the C++ program `evo` (available from **`https://github.com/millanek/evo`**), with the `fst --ms` option for simulations and `fst --vcf` option for cichlid data. The F_{ST} calculation implements the Hudson estimator, as defined by Bhatia, Patterson *et al.* [197, equation 10], using 'ratio of averages' to combine estimates of F_{ST} across multiple variants, as recommended in their manuscript.

For defining HDRs, I used windows of 15 variants each, which I found to provide good balance between fine genomic resolution and reducing stochastic variation by averaging over variants. Nevertheless, I found some cases where F_{ST} between neighbouring regions dipped briefly below the threshold, which I believe to be in most cases due to remaining stochastic variation. The length (the extent) of HDRs was defined by merging windows with $F_{ST} >= 0.25$ that were next to each other or within 10,000bp of one another using `bedtools v2.16.2` [239]: `mergeBed -d 10000 -i windows_fst_above0.2.bed`. F_{ST} was also calculated in 10kb windows and each HDR must contain at least one window with $F_{ST} >= 0.3$, as described in the main text.

Characterisation of HDRs in terms of d_{XY}, and π_{diff}

Both d_{XY} and nucleotide diversity (π) were calculated for 10kb windows. The d_{XY} statistic was calculated as defined by Wakeley [240, equation 3]. Both calculations are implemented in my C++ program `evo` (available from **`https://github.com/millanek/evo`**), and were obtained by using the `fst --vcf` option.

Average nucleotide diversity in each window was calculated separately for the benthic (π_B) and littoral (π_L) ecomorphs and π_{diff} was then calculated as the absolute value of the difference between π_B and π_L; i.e. $\pi_{diff} = |\pi_B - \pi_L|$. The 'direction' of the 'sweep' is in the morph with lower π; i.e. if $\pi_B < \pi_L$ then the potential 'sweep' was inferred to be in the benthic morph.

Gene Ontology enrichment analysis

Gene Ontology (GO) enrichment for genes found within HDRs was calculated in `R` [175] using the `topGO` package [241] from the Bioconductor project [242]. The GO hierarchical structure was obtained from the `GO.db` annotation [243] and linking zebrafish gene identifiers to GO terms was accomplished using the `org.Dr.eg.db` annotation package [244].

Chromopainter and fineSTRUCTURE

Singleton SNPs were excluded using `bcftools-1.1 -c 2:minor` option, before exporting the remaining variants in PLINK format [194]. The `chromopainter v0.0.4` software [193] was then run for 150 largest genomic scaffolds. Briefly, I created a uniform recombination map using the `makeuniformrecfile.pl` script, then estimated the effective population size (N_e) for a subsample of 20 individuals using the chromopainter inbuilt expectation-maximization procedure [193], averaged over the 20 N_e values using the provided `neaverage.pl` script. Estimated N_e values ranged from 1,046 to 6,015 (mean 3914, sd. 990). The chromopainter program was then run for each scaffold independently, with the `-a 0 0` option to run all individuals against all others. Results for individual scaffolds were combined using the `chromocombine` tool before running `fineSTRUCTURE v0.0.5` with 1,000,000 burn in iterations, and 200,000 sample iterations, recording a sample every 1,000 iterations (options `-x 1000000 -y 200000 -z 1000`). Finally, the sample relationship tree was built with fineSTRUCTURE using the `-m T` option and 20,000 iterations.

Patterson's D (ABBA-BABA) and related statistics

To test for possible gene-flow between surrounding rivers and Massoko, I calculated the ABBA-BABA statistic (16, 17). The ABBA-BABA test (also known as *D statistic* or *Patterson's D*) tests for introgression an excess of shared derived alleles between one of two populations and an outgroup. Formally, I calculated D(benthic, littoral, Mbaka river, *P. nyererei*) using equation S15.2 of Green et al. [190], allowing me to use

allele-frequency information from all benthic and littoral individuals. I also estimated *f*, the admixture fraction following Green *et al.* equation S18.5, and calculated the standard error for both estimates by a weighted block jackknife, using blocks of 5,000 informative variants (i.e. variants with ABBA or BABA patterns).

The D and f statistics were calculated genome-wide and D and f_{dM} also in non-overlapping windows of 50 informative variants each. To add ancestral allele information (i.e. the outgroup variants) to the crater lakes VCF file, I used whole genome alignment between *M. zebra* and *P. nyererei*, as described in section 3.4. The f_{dM} statistic has been calculated as defined previously in section 5.5.

6.8.4 Mate choice experiments

Note: The mate choice trials were done by Alexandra Tyers, University of Bangor. The methods are included here (in this thesis) for completeness. As also noted in the main text, I designed the SNP assay that formed the basis for similarity scoring during the analysis of these trials, and contributed to the analysis.

Experimental setup, and aquarium work

A single 4m long tank with a gravel/sand substrate was divided into eight sections by 'partial partition' grids [245]. Each section contained a terracotta plant pot (to function as territorial focal points for the males) and a selection of plastic plants. Water was filtered and heated (to ~26 C) externally and the tank lit from above by white and UV enhanced fluorescent tube lamps. Fish were fed daily with algae flake and 2-3 times weekly with frozen bloodworm.

Two female mate choice trials were carried out using two different sets of eight males and a total of 50 females. All fish were wild caught and shipped to the UK in December 2011. Trial 1 ran from the beginning of November 2012 to the end of January 2013 (3 months) and trial 2 started at the beginning of February 2013, ending in June 2013 (4.5 months). Each set/trial comprised 3-4 large littoral, 1-2 large benthic and 3-4 'small' males. Forty-five of the 50 females produced broods in both trials. All of the larger littoral and benthic males within each set were of a comparable size and unable to fit through the partial partition grids. The large males were placed in every-other section, leaving the territories in-between available to the small males which, being of a similar size to some of the larger females, were also able to move freely between sections. Before introduction of the females to the experimental tank, males were left until the smaller ones had settled into the 'empty' territories between the bigger males.

As with other haplochromine cichlid fish, *Astatotilapia* are maternal mouth-brooders, egg are picked up by the female during spawning and protected in the buccal cavity during development before release as free-swimming young approximately three weeks later. Females were removed from the experimental tank after spawning and isolated in small tanks on a recirculating system during the brooding phase. After the first trial, offspring were gently removed from the females mouths after 10 days and euthanised by anaesthetic (clove oil) overdose. Females were kept in their individual tanks to allow for rest and recovery before the second trial. Once all females had spawned in the first trial, the males were changed. Allele diversity at the chosen microsatellite loci (Ppun5, 7 & 21) [246] was sufficiently high to allow for the identification of all individual females by their microsatellite profile, it was therefore possible to return all females to the experimental tank at the same time for the second trial and re-identify individuals later during the second round of paternity testing. After spawning in the second trial, females were again isolated, but left to brood to term. Five offspring from each brood were euthanised for paternity testing.

Paternity testing

475 offspring from 95 broods (five per brood/trial), produced over the two replicates were genotyped for paternity analysis (250 from 50 broods in trial 1; 225 from 45 broods in trial 2). Tissue was taken from ethanol preserved fry samples and DNA obtained by salt extraction. DNA samples from offspring, mothers, and all potential fathers were used for assigning paternity by allele sizing after PCR multiplex (Qiagen multiplex kit) of three microsatellite markers (Ppun5, 7 & 21) [246]. Genotyping of the amplified samples was carried out on an Applied Biosystems (ABI) 3130xl genetic analyzer using LIZ 500(-250) (ABI) size standard. The genotype of each individual (males, females, offspring) were determined by manual scoring of alleles in `Peak Scanner v2`. 447 (94%) of the genotyped offspring were successfully assigned to an individual male. Due to allele sharing among males used in trial 2, 23 offspring could not be assigned unambiguously. Seven offspring could not be assigned due to problems with amplification or disagreement between microsatellite loci (possible cross-contamination). Overall, 8-10 offspring from each female that produced more than one brood, were unambiguously assigned to father.

Forty-six of the 50 females spawned with more than one male during the course of the experiment and some females were found to have spawned with up to four males in total (44% of individual broods were sired by more than one male).

Data analysis

I designed a Sequenom MassARRAY SNP genotyping assay [247] for 117 SNPs, over four Sequenom plates. The assay was performed by the Wellcome Trust Sanger Institute core genotyping team. Analysis of the SNP data was performed in `R` using the `Bioconductor` [248] package `SNPRelate` [249] to account for linkage disequilibrium (LD) between the SNPs. SNPs were filtered using a recursive sliding window approach (`snpgdsLDpruning`) with an LD threshold of 0.2. Principal components analysis of the filtered dataset was used to obtain a score for each of the individuals used in the mate choice experiments.

The expected distance between female and male PC1 scores (under the null hypothesis of no assortative mating) was calculated as follows:
Expected value = mean absolute distance between female PC1 score and PC1 scores all the possible combinations of males she might have mated with.
The **observed value** is the mean absolute distance between female PC1 score and PC1 score of all males mated with.

The above calculations are based on the total number of males a female actually mated with and the number of trials she took part in. The values are therefore different for each female because some did not take part in both trials (or did not spawn in both trials), and there was variation in the total number of males mated with over the course of the experiment (between 2-4).

For each female, the number of potential mates is a product of the possible combinations in trial 1 (T1) and trial 2 (T2). The number of combinations (choosing r males out of n males) in each trial are n!/r!(n-r)!. For example, the number of combinations is 7 for a female that mated with a single male in T1 and 588 for a female that mated with 2 males in T1 and 2 males in T2.

6.8.5 Measuring rhodopsin absorption spectra

Note: The work described here was done by Yohey Terai, SOKENDAI, Japan, starting with DNA samples sent by myself. The methods are included here (in this thesis) for completeness.

Reconstruction and measurement of absorption spectra of visual pigments

Production, reconstruction, purification, and measurement of the visual pigments were performed as described Ueyama *et al.* [250] with minor modifications. Briefly, the sequences of *rho* (also known as RH1) H4 and H5 alleles were amplified by PCR

using genomic DNA of Lake Massoko cichlids as a template with a pair of specific PCR primers [221] designed to produce a fusion protein with a FLAG-tag (Sigma-Aldrich) at its C terminus. The amplified DNA fragments were digested with restriction enzymes and cloned into the expression vector pFLAG-CMV-5a (Sigma-Aldrich). The visual pigments were reconstituted with A1-derived retinal. Absorption spectra of the pigment solutions in the presence of hydroxyl-amine (<100mM) before and after photobleaching were recorded using a spectrophotometer (UV-2400, Shimadzu, Japan). The measurements were taken 5 times before and after photobleaching. We determined the mean peak spectral values (maximum absorption spectra: λmax) and standard errors from multiple preparations and measurements for each pigment. All procedures after reconstitution of the pigments were performed under dim red light (>680 nm) conditions.

Chapter 7

Conclusions

7.1 Conclusions

In this thesis, I describe the some of the first steps in the application of whole genome sequence data to study the exceptional diversity of East African cichlids. I generated an annotation of microRNA loci for five reference genomes, predicted which protein coding genes may be regulated by them, and explored sequence evolution both in the genes themselves and their targets. Then I took advantage of one of the reference genomes (the *M. zebra* genome for Lake Malawi) and aligned whole genome sequences from 239 individuals to it to obtain two detailed catalogues of variation: over 20 million genetic variants for Lake Malawi and almost 5 million variants to study incipient speciation of *Astatiotilapia* cichlids in the isolated crater lake Massoko. Another 29 Lake Malawi and 4 crater lake region *Astatotilapia* individuals have been sequenced but their genomes have not yet been analysed. To further facilitate research on this fascinating system, I constructed whole genome alignments between the reference genome assemblies, assigned ancestral alleles to genetic variants in Lake Malawi, and built a genome browser to visualise genome wide datasets for East African cichlids.

During an initial analysis of the Lake Malawi dataset I found that heterozygosity based N_e estimates range from ~7,200 to ~24,300, and that F_{ST} between Lake Malawi species varies between 0.04 and 0.66. A phylogenetic approach to the study of species relationships provides a strong signal when averaging across all genomic loci, but there is a lot of discordance between local phylogenies, and especially between phylogenies built using nuclear and mitochondrial DNA data. The discordance is due to high prevalence of incomplete lineage sorting, but interspecific introgression may also play a role. In three cases of possible introgression I formally tested the hypothesis using the ABBA-BABA statistic. In the first case, where introgression was suggested by

Chapter 7

Conclusions

7.1 Conclusions

In this thesis, I describe the some of the first steps in the application of whole genome sequence data to study the exceptional diversity of East African cichlids. I generated an annotation of microRNA loci for five reference genomes, predicted which protein coding genes may be regulated by them, and explored sequence evolution both in the genes themselves and their targets. Then I took advantage of one of the reference genomes (the *M. zebra* genome for Lake Malawi) and aligned whole genome sequences from 239 individuals to it to obtain two detailed catalogues of variation: over 20 million genetic variants for Lake Malawi and almost 5 million variants to study incipient speciation of *Astatiotilapia* cichlids in the isolated crater lake Massoko. Another 29 Lake Malawi and 4 crater lake region *Astatotilapia* individuals have been sequenced but their genomes have not yet been analysed. To further facilitate research on this fascinating system, I constructed whole genome alignments between the reference genome assemblies, assigned ancestral alleles to genetic variants in Lake Malawi, and built a genome browser to visualise genome wide datasets for East African cichlids.

During an initial analysis of the Lake Malawi dataset I found that heterozygosity based N_e estimates range from ~7,200 to ~24,300, and that F_{ST} between Lake Malawi species varies between 0.04 and 0.66. A phylogenetic approach to the study of species relationships provides a strong signal when averaging across all genomic loci, but there is a lot of discordance between local phylogenies, and especially between phylogenies built using nuclear and mitochondrial DNA data. The discordance is due to high prevalence of incomplete lineage sorting, but interspecific introgression may also play a role. In three cases of possible introgression I formally tested the hypothesis using the ABBA-BABA statistic. In the first case, where introgression was suggested by

phenotypic similarity, I did not find any evidence. Therefore, the phenotypic similarity in denture between shallow and deep water *Lethrinops* may be a case of parallel evolution. In the other two cases, introgression was suggested by unexpectedly high co-ancestry (sharing of haplotypes). In both cases, the ABBA-BABA statistic confirmed introgression, from *P. electra* into *B. rhoadesii* (weak; f=1.3%) in Lake Malawi and from *O. tetrastigma* into *Astatotilapia* (strong; f=29.8%) in the crater lake Ilamba. While the co-ancestry relationships in the Lake Malawi dataset suggest several other possible cases of introgression between relatively closely related species, it is perhaps not as common as previously thought, and seems to be very rare between more distinct lineages - perhaps possible only under special conditions, such as in the isolation and turbid waters of Lake Ilamba.

I described a detailed analysis of of early-stage adaptive divergence of two cichlid fish ecomorphs in Lake Massoko, a small (700m diameter) isolated crater lake in Tanzania. The ecomorphs differ in depth preference, male breeding colour, body shape, diet and trophic morphology. With whole genome sequences of 145 fish, I showed multiple lines of evidence that the divergence between the ecomorphs has happened in sympatry within Lake Massoko. I identified 98 clearly demarcated genomic 'islands' of high differentiation on the basis of F_{ST} and characterised these highly diverged regions (HDRs) in terms of local absolute sequence divergence, nucleotide diversity, and genomic location, showing that the results are consistent with the 'islands of speciation' model of speciation with gene-flow. I also designed a SNP assay to support aquarium mate choice experiments that demonstrate association of genotypes across these islands to divergent mate preferences. The HDRs contain candidate adaptive genes enriched for functions in sensory perception (including rhodopsin and other twilight vision associated genes), hormone signalling and morphogenesis. The possible adaptive role of rhodopsin is further supported by results of a collaboration confirming a shift in light absorption spectra between the two alleles of rhodopsin in Lake Massoko.

7.2 Future work

The Lake Massoko study is one of the most extensive genomic and ecological studies of divergence in the initial phases of speciation to date, especially in a sympatric setting. Next, we would like to gain a better understanding of the significance and functions of the discovered HDRs. Funding has been obtained and work is under way to collect ~200 Massoko males across several depth transects around the lake, measure anatomical and ecological traits and sequence the whole genomes for each individual fish. We will then

test for association between genetic variants and measured traits. The significance of associations will be tested in an independent sample. This work will be complemented by additional mate-choice trials under varying lighting conditions to approximate the different light environments in the lake. The ecomorphs of Lake Massoko show clear differences in traits normally associated with adaptive radiation in cichlid fishes, including body shape, pharyngeal jaw morphology, diet, microhabitat preference, retinal pigment sensitivity, male colour and mate preference [82, 83, 93, 206, 221]. Therefore, this study suggests processes and specific genomic regions that may be involved in speciation events within the great cichlid radiations of Lakes Malawi, Victoria, and Tanganyika.

After more than a century of work on the taxonomy and evolution of Lake Malawi cichlids, involving thousands of measurements of individual morphological characters, the relationships between individual species still remain unclear. Today, with whole genome sequence data, we have an unprecedented opportunity to reconstruct the evolutionary history of the radiation, as demonstrated by the initial results presented in this thesis. Understanding the causes and consequences of speciation and subsequent adaptation requires investigation of taxon pairs at different levels of divergence. Knowledge of species relationships, including the relative timing, frequency and sequence of evolution of major adaptive innovations will open up this complex and fascinating system to a new generation of research.

The first immediate steps will be adding data from the 32 samples that have been sequenced in Summer 2015 to the existing analysis. I expect this data to provide additional specific insights, in particular in relation to 1) The position of the genus *Rhampochromis* within the radiation; 2) The relative relationships of *Astatotilapia rujewa*, *A. tweddlei*, and *Serranochromis robustus* to the Lake Malawi flock and their contributions to the Malawi haplochromine radiation; 3) The origin of the 'large predator', and 'molluscivore' phenotypes in Lake Malawi; 4) The potential source of gene flow to Lake Massoko, with four individual *A. calliptera* sequenced from the Itupi stream, the closest existing water body upstream of Lake Massoko; 5) The geographical context of the origin of two very different forms of *Rhampochromis* in Lake Kingiri.

Our data now contains multiple samples from each of the major evolutionary lineages and covers 34 genera, as described in chapter 3. There are currently 53 recognised endemic genera in Lake Malawi, several of which are polyphyletic as shown by the initial analysis in this thesis. Therefore, I estimate that ~50 additional species will need to be sampled for a satisfactory overview of Lake Malawi evolutionary history down to the generic level.

As alluded to on several occasions previously in this thesis, I am especially interested in studying the genetic basis of parallel evolution. Within Lake Malawi, there are also several adaptations present in species that are not believed to be very closely related, for example the 'thick lip' phenotype 7.1, or the *Lethrinops*-like denture (demonstrated in this thesis).

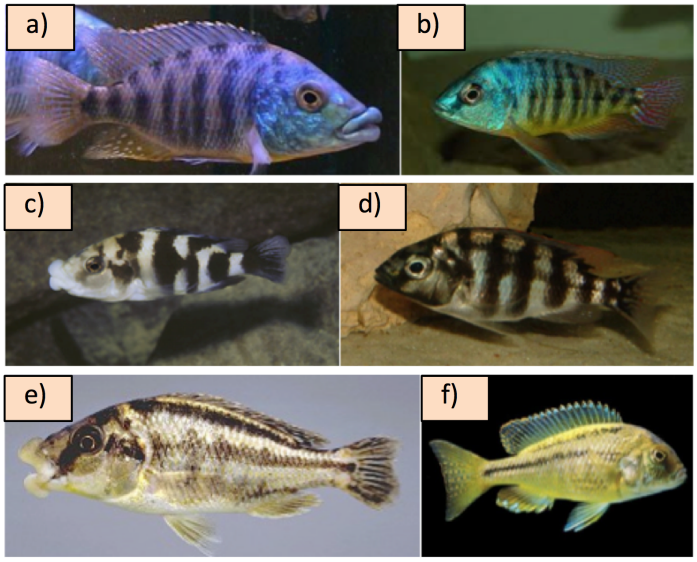


Fig. 7.1 **Parallel evolution of the 'thick lip' phenotype in Lake Malawi.** Contrast 1: **(a)** *Protomelas ornatus* **(b)** *Protomelas fenstratus*; Contrast 2: **(c)** *Placidochromis milomo* **(d)** *Placidochromis* 'johnstoni solo'; Contrast 3: **(e)** *Chilotilapia euchilus* **(f)** *Chilotilapia rhoadesii*; Species **a**, **c**, and **e** are not believed to be particularly closely related but all possess the 'thick lip' phenotype. On the other hand the pairs **a-b**, **c-d**, and **e-f** are closely related but **b**,**d**, and **f** lack the 'thick lip' adaptation. Images from George Turner.

Do such parallel phenotypes indicate genetic parallelism? Does the adaptation involve changes in the same genes, genomic regions, or even in the same genetic variants [251]? If the same genetic variants are involved, is it because mutations that occurred independently in different species, because the allele was already present in a shared ancestral population, or because it was transferred from one species to another by introgression [252]?

In section 3.4, I describe whole genome alignments that will facilitate direct comparisons of genomic regions identified in Lake Malawi studies to genomic regions underlying speciation or adaptive divergence in lakes Tanganyika and Victoria. Thus, it will be possible to draw more general conclusions about speciation and adaptive divergence in East African cichlids, and to address the genomic basis of parallel phenotypic evolution

across lakes, which is often even more striking than within Malawi (Figures 1.12, 7.2). Walter Salzburger and his team at the University of Basel in Switzerland are currently sequencing the whole genomes of two individuals each from all of the recognised species of Lake Tanganyika with all data due to become available by summer 2016 (W. Salzburger, pers. comm.). The combination of this dataset with the Malawi dataset presented in this thesis could provide an excellent starting point for the first whole genome across-lake study in East African cichlids.

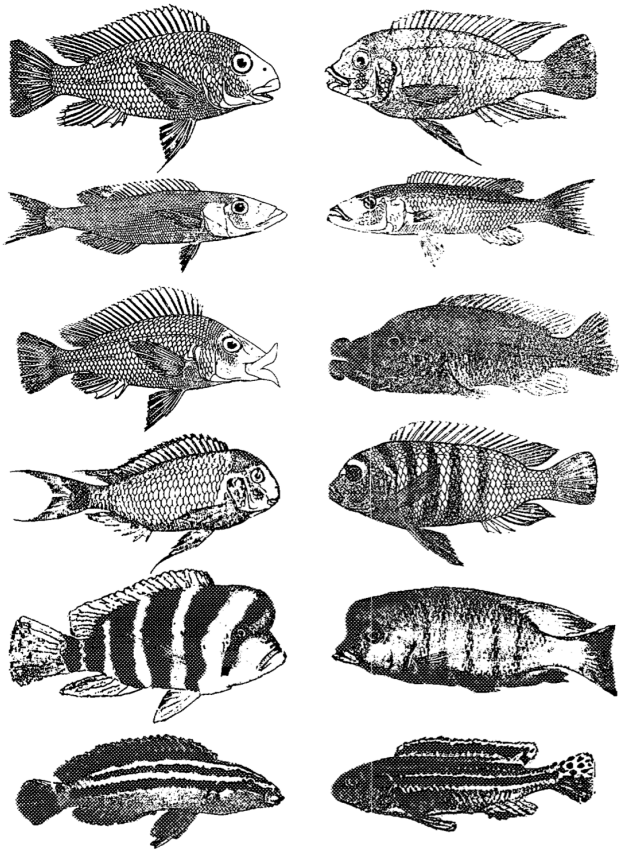


Fig. 7.2 **Parallel evolution between lakes Tanganyika and Malawi.** Six pairs of morphologically similar species from Lake Tanganyika (left) and Lake Malawi (right). The specific shared features are (from top to bottom): rasping jaw morphology, elongated body typical of pursuit predators, 'thick lips', mbuna habit, enlarged nuchal hump, and strong horizontal striping. Image from Kocher *et al.* [90].

References

[1] Darwin, C. *On the Origin of Species* (Oxford University Press, 1859). ISBN 978-0-19-921922-3.

[2] Huxley, J., Pigliucci, M., Müller, G. B. *Evolution: The Modern Synthesis*. The Definitive Edition (MIT Press, 2010).

[3] Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., *et al. Introduction to Genetic Analysis* (Macmillan Higher Education, 2010).

[4] Miko, I. Gregor Mendel and the principles of inheritance. *Nature Education Knowledge*, 1(1), 2008.

[5] Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature genetics*, 47(7):pp. 702–709, July 2015.

[6] Kimball, J. W. Germline vs. Soma. `http://biology-pages.info`. Accessed 20th August 2015.

[7] O'Connor, C., Miko, I. Developing the chromosome theory. *Nature Education*, 1(1), 2008.

[8] Avery, O. T., Macleod, C. M., McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of experimental medicine*, 79(2):pp. 137–158, February 1944.

[9] Hershey, A. D., Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of general physiology*, 36(1):pp. 39–56, May 1952.

[10] Sabour, D., Schöler, H. R. Reprogramming and the mammalian germline: the Weismann barrier revisited. *Current opinion in cell biology*, 24(6):pp. 716–723, December 2012.

[11] Buckley, B. A., Burkhart, K. B., Gu, S. G., *et al.* A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. *Nature*, 489(7416):pp. 447–451, September 2012.

[12] Heard, E., Martienssen, R. A. Transgenerational epigenetic inheritance: myths and mechanisms. *Cell*, 157(1):pp. 95–109, March 2014.

[13] Devanapally, S., Ravikumar, S., Jose, A. M. Double-stranded RNA made in C. elegans neurons can enter the germline and cause transgenerational gene silencing. *Proceedings of the National Academy of Sciences of the United States of America*, 112(7):pp. 2133–2138, February 2015.

[14] Scally, A., Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nature reviews. Genetics*, 2012.

[15] Ségurel, L., Wyman, M. J., Przeworski, M. Determinants of mutation rate variation in the human germline. *Annual review of genomics and human genetics*, 15:pp. 47–70, 2014.

[16] Gregory, R. T. Animal Genome Size Database. , 2015.

[17] Flicek, P., Amode, M. R., Barrell, D., *et al.* Ensembl 2014. *Nucleic acids research*, 42(D1):pp. D749–D755, December 2013.

[18] Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nature reviews. Genetics*, 6(9):pp. 699–708, September 2005.

[19] Lander, E. S. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):pp. 187–197, February 2011.

[20] Mikkelsen, T. S., Wakefield, M. J., Aken, B., *et al.* Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, 447(7141):pp. 167–177, May 2007.

[21] Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nature reviews. Genetics*, 12(12):pp. 846–860, December 2011.

[22] Sayed, D., Abdellatif, M. MicroRNAs in Development and Disease. *Physiological Reviews*, 91(3):pp. 827–887, January 2011.

[23] Weick, E.-M., Miska, E. A. piRNAs: from biogenesis to function. *Development (Cambridge, England)*, 141(18):pp. 3458–3471, September 2014.

[24] ENCODE Project Consortium, Bernstein, B. E., Birney, E., *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):pp. 57–74, September 2012.

[25] Graur, D., Zheng, Y., Price, N., *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome biology and evolution*, 5(3):pp. 578–590, 2013.

[26] Siepel, A., Bejerano, G., Pedersen, J. S., *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):pp. 1034–1050, August 2005.

[27] Zuckerkandl, E., Pauling, L. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*, 1965.

[28] The Nobel Prize in Physiology or Medicine 1968. `http://www.nobelprize.org/nobel_prizes/medicine/laureates/1968/index.html`, 2014. Accessed 18th August 2015.

[29] Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, 267(5608):pp. 275–276, May 1977.

[30] Hedrick, P. W. *Genetics of Populations* (Jones & Bartlett Publishers, 2011).

[31] Wakeley, J. *Coalescent Theory*. An Introduction (Roberts Publishers, 2009).

[32] Maynard Smith, J., Haigh, J. The hitch-hiking effect of a favourable gene. *Genetical research*, 23(1):pp. 23–35, February 1974.

[33] Charlesworth, B., Morgan, M. T., Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):pp. 1289–1303, July 1993.

[34] Stephan, W. Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1544):pp. 1245–1253, April 2010.

[35] Coyne, J. A., Orr, H. A. *Speciation* (SINAUER ASSOCIATES, 2004). ISBN 978-0-87893-089-0.

[36] Mayr, E. What is a species, and what is not? *Philosophy of Science*, 1996.

[37] Safran, R. J., Nosil, P. Speciation: The Origin of New Species. *Nature Education Knowledge*, 3(10), 2012.

[38] Nosil, P. *Ecological Speciation* (OUP Oxford, 2012).

[39] Darwin, C. *On the Origin of Species* (OUP Oxford, 2008).

[40] Fitzpatrick, B. M., Fordyce, J. A., Gavrilets, S. What, if anything, is sympatric speciation? *Journal of evolutionary biology*, 21(6):pp. 1452–1459, November 2008.

[41] Mayr, E. *Animal Species and Evolution* (Belknap Press, 1963).

[42] Adams, J. DNA sequencing technologies. *Nature Education*, 2008.

[43] Metzker, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11(1):pp. 31–46, January 2010.

[44] Sanger, F., Air, G. M., Barrell, B. G., *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265(5596):pp. 687–695, February 1977.

[45] Sanger, F., Nicklen, S., Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):pp. 5463–5467, December 1977.

[46] ENA. Statistics. `http://www.ebi.ac.uk/ena/about/statistics`, 2015. European Nucleotide Archive, Accessed 15th August 2015.

[47] Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). `http://www.genome.gov/sequencingcosts`. Accessed 15th August 2015.

[48] Baker, M. De novo genome assembly: what every biologist should know. *Nature methods*, 9(4):pp. 333–337, January 2012.

[49] Nielsen, R., Paul, J. S., Albrechtsen, A., *et al.* Genotype and SNP calling from next-generation sequencing data. *Nature reviews. Genetics*, 12(6):pp. 443–451, June 2011.

[50] Ingram, V. M. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature*, 180(4581):pp. 326–328, August 1957.

[51] McDonald, J. H., Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):pp. 652–654, June 1991.

[52] Atrian, S., Sánchez-Pulido, L., Gonzàlez-Duarte, R., *et al.* Shaping of Drosophila alcohol dehydrogenase through evolution: relationship with enzyme functionality. *Journal of molecular evolution*, 47(2):pp. 211–221, August 1998.

[53] Smith, N. G. C., Eyre-Walker, A. Adaptive protein evolution in Drosophila. *Nature*, 415(6875):pp. 1022–1024, February 2002.

[54] Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):pp. 520–562, December 2002.

[55] Del Punta, K., Leinders-Zufall, T., Rodriguez, I., *et al.* Deficient pheromone responses in mice lacking a cluster of vomeronasal receptor genes. *Nature*, 419(6902):pp. 70–74, September 2002.

[56] Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):pp. 69–87, September 2005.

[57] Pollard, K. S., Salama, S. R., Lambert, N., *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, 443(7108):pp. 167–172, September 2006.

[58] Vitti, J. J., Grossman, S. R., Sabeti, P. C. Detecting Natural Selection in Genomic Data. *Annual review of genetics*, 47:pp. 97–120, August 2013.

[59] Fay, J. C., Wu, C. I. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):pp. 1405–1413, July 2000.

[60] Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):pp. 585–595, November 1989.

[61] Balding, D. J., Bishop, M., Cannings, C. *Handbook of statistical genetics* (John Wiley and Sons, 2007). ISBN 978-0-470-05830-5.

[62] Seehausen, O., Butlin, R. K., Keller, I., *et al.* Genomics and the origin of species. *Nature reviews. Genetics*, 15(3):pp. 176–192, February 2014.

[63] Saether, S. A., Saetre, G.-P., Borge, T., *et al.* Sex chromosome-linked species recognition and evolution of reproductive isolation in flycatchers. *Science (New York, N.Y.)*, 318(5847):pp. 95–97, October 2007.

[64] Kitano, J., Ross, J. A., Mori, S., *et al.* A role for a neo-sex chromosome in stickleback speciation. *Nature*, 461(7267):pp. 1079–1083, October 2009.

[65] Felsenstein, J. Skepticism Towards Santa Rosalia, or Why are There so Few Kinds of Animals? *Evolution; international journal of organic evolution*, 35(1):p. 124, January 1981.

[66] Cruickshank, T. E., Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23(13):pp. 3133–3157, June 2014.

[67] Turner, T. L., Hahn, M. W., Nuzhdin, S. V. Genomic islands of speciation in Anopheles gambiae. *PLoS biology*, 3(9):p. e285, September 2005.

[68] Savolainen, V., Anstett, M.-C., Lexer, C., *et al.* Sympatric speciation in palms on an oceanic island. *Nature*, 441(7090):pp. 210–213, May 2006.

[69] Nadeau, N. J., Whibley, A., Jones, R. T., *et al.* Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1587):pp. 343–353, February 2012.

[70] Ellegren, H., Smeds, L., Burri, R., *et al.* The genomic landscape of species divergence in Ficedula flycatchers. *Nature*, 491(7426):pp. 756–760, November 2012.

[71] Feder, J. L., Egan, S. P., Nosil, P. The genomics of speciation-with-gene-flow. *Trends in Genetics*, 28(7):pp. 342–350, July 2012.

[72] Via, S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367(1587):pp. 451–460, February 2012.

[73] Noor, M., Bennett, S. M. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, 2009.

[74] Turner, T. L., Hahn, M. W. Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, 19(5):pp. 848–850, March 2010.

[75] Charlesworth, B. Measures of divergence between populations and the effect of forces that reduce variability. *Molecular biology and evolution*, 15(5):pp. 538–543, April 1998.

[76] Pennisi, E. Disputed islands. *Science (New York, N.Y.)*, 345(6197):pp. 611–613, August 2014.

[77] Poelstra, J. W., Vijay, N., Bossu, C. M., *et al.* The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science (New York, N.Y.)*, 344(6190):pp. 1410–1414, June 2014.

[78] Barlow, G. W. *The Cichlid Fishes.* Nature's Grand Experiment in Evolution (Basic Books, 2002).

[79] Zardoya, R., Vollmer, D. M., Craddock, C., *et al.* Evolutionary conservation of microsatellite flanking regions and their use in resolving the phylogeny of cichlid fishes (Pisces: Perciformes). *Proceedings. Biological sciences / The Royal Society*, 263(1376):pp. 1589–1598, November 1996.

[80] Streelman, J. T., Karl, S. A. Reconstructing labroid evolution with single-copy nuclear DNA. *Proceedings. Biological sciences / The Royal Society*, 264(1384):pp. 1011–1020, July 1997.

[81] Meyer, A. Phylogenetic relationships and evolutionary processes in East African cichlid fishes. *Trends in ecology & evolution*, 8(8):pp. 279–284, August 1993.

[82] Wagner, C. E., Harmon, L. J., Seehausen, O. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, 487(7407):pp. 366–369, July 2012.

[83] Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nature reviews. Genetics*, 5(4):pp. 288–298, April 2004.

[84] Salzburger, W., Meyer, A., Baric, S., *et al.* Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. *Systematic biology*, 51(1):pp. 113–135, February 2002.

[85] Salzburger, W., Mack, T., Verheyen, E., *et al.* Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes. *BMC evolutionary biology*, 5:p. 17, 2005.

[86] Joyce, D. A., Lunt, D. H., Genner, M. J., *et al.* Repeated colonization and hybridization in Lake Malawi cichlids. *Current biology : CB*, 21(3):pp. R108–9, February 2011.

[87] Snoeks, J., Konings, A. *The cichlid diversity of Lake Malawi/Nyasa/Niassa* (Cichlid Press, 2004).

[88] Elmer, K. R., Reggio, C., Wirth, T., *et al.* Pleistocene desiccation in East Africa bottlenecked but did not extirpate the adaptive radiation of Lake Victoria haplochromine cichlid fishes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(32):pp. 13404–13409, August 2009.

[89] Seehausen, O., Terai, Y., Magalhaes, I. S., *et al.* Speciation through sensory drive in cichlid fish. *Nature*, 455(7213):pp. 620–626, October 2008.

[90] Kocher, T. D., Conroy, J. A., McKaye, K. R., *et al.* Similar morphologies of cichlid fish in Lakes Tanganyika and Malawi are due to convergence. *Molecular phylogenetics and evolution*, 2(2):pp. 158–165, June 1993.

[91] Muschick, M., Indermaur, A., Salzburger, W. Convergent evolution within an adaptive radiation of cichlid fishes. *Current biology : CB*, 22(24):pp. 2362–2368, December 2012.

[92] Laland, K., Uller, T., Feldman, M., *et al.* Does evolutionary theory need a rethink? *Nature*, 514(7521):pp. 161–164, October 2014.

[93] Brawand, D., Wagner, C. E., Li, Y. I., *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518):pp. 375–381, September 2014.

[94] Carthew, R. W., Sontheimer, E. J. Origins and Mechanisms of miRNAs and siRNAs. *Cell*, 136(4):pp. 642–655, February 2009.

[95] Krol, J., Loedige, I., Filipowicz, W. The widespread regulation of microRNA biogenesis, function and decay. *Nature reviews. Genetics*, 11(9):pp. 597–610, September 2010.

[96] Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):pp. 215–233, January 2009.

[97] Gunter, C., Dhand, R. The chimpanzee genome. *Nature*, 437:p. 47, 2005.

[98] Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):pp. 25–36, July 2008.

[99] Lee, R. C., Feinbaum, R. L., Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):pp. 843–854, December 1993.

[100] Reinhart, B. J., Slack, F. J., Basson, M., *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):pp. 901–906, February 2000.

[101] Lund, A. H. miR-10 in development and cancer. *Cell Death & Differentiation*, 17(2):pp. 209–214, May 2009.

[102] Bissels, U., Bosio, A., Wagner, W. MicroRNAs are shaping the hematopoietic landscape. *Haematologica*, 97(2):pp. 160–167, February 2012.

[103] Stark, A., Brennecke, J., Bushati, N., *et al.* Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3? UTR evolution. *Cell*, 2005.

[104] Li, Y., Wang, F., Lee, J.-A., *et al.* MicroRNA-9a ensures the precise specification of sensory organ precursors in Drosophila. *Genes & development*, 20(20):pp. 2793–2805, October 2006.

[105] Li, X., Cassidy, J. J., Reinke, C. A., *et al.* A microRNA imparts robustness against environmental fluctuation during development. *Audio and Electroacoustics Newsletter, IEEE*, 137(2):pp. 273–282, April 2009.

[106] Christodoulou, F., Raible, F., Tomer, R., *et al.* Ancient animal microRNAs and the evolution of tissue identity. *Nature*, 463(7284):pp. 1084–1088, February 2010.

[107] Fujimura, K., Okada, N. Development of the embryo, larva and early juvenile of Nile tilapia Oreochromis niloticus (Pisces: Cichlidae). Developmental staging system. *Development, growth & differentiation*, 49(4):pp. 301–324, May 2007.

[108] Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 2011.

[109] Friedländer, M. R., Mackowiak, S. D., Li, N., *et al.* miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research*, 40(1):pp. 37–52, January 2012.

[110] Langmead, B., Trapnell, C., Pop, M., *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):p. R25, 2009.

[111] Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., *et al.* ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, 6:p. 26, 2011.

[112] Griffiths-Jones, S., Grocock, R. J., van Dongen, S., *et al.* miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(Database issue):pp. D140–4, January 2006.

[113] Ro, S., Park, C., Young, D., *et al.* Tissue-dependent paired expression of miRNAs. *Nucleic acids research*, 35(17):pp. 5944–5953, 2007.

[114] Lu, J., Shen, Y., Wu, Q., *et al.* The birth and death of microRNA genes in Drosophila. *Nature genetics*, 40(3):pp. 351–355, February 2008.

[115] Rehmsmeier, M., Steffen, P., Hochsmann, M., *et al.* Fast and effective prediction of microRNA/target duplexes. *RNA (New York, N.Y.)*, 10(10):pp. 1507–1517, October 2004.

[116] Kapsimali, M., Kloosterman, W. P., de Bruijn, E., *et al.* MicroRNAs show a wide diversity of expression profiles in the developing and mature central nervous system. *Genome biology*, 8(8):p. R173, 2007.

[117] Marín, R. M., Vanícek, J. Efficient use of accessibility in microRNA target prediction. *Nucleic acids research*, 39(1):pp. 19–29, January 2011.

[118] Ko, C.-Y., Tsai, M.-Y., Tseng, W.-F., *et al.* Integration of CNS survival and differentiation by HIF2α. *Cell Death & Differentiation*, 18(11):pp. 1757–1770, November 2011.

[119] Long, J. M., Lahiri, D. K. Advances in microRNA experimental approaches to study physiological regulation of gene products implicated in CNS disorders. *Experimental neurology*, 235(2):pp. 402–418, June 2012.

[120] Loh, Y. H. E., Yi, S. V., Streelman, J. T. Evolution of MicroRNAs and the Diversification of Species. *Genome Biology and Evolution*, 3(0):pp. 55–65, January 2011.

[121] Kocher, T. D., Streelman, T., Seehausen, O., *et al.* Genetic Basis of Vertebrate Diversity: the Cichlid Fish Model. Technical report, March 2006. `http://www.genome.gov/pages/research/sequencing/seqproposals/cichlidgenomeseq.pdf`.

[122] Saunders, M. A., Liang, H., Li, W. H. Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences of the United States of America*, 2007.

[123] Marín, R. M., Voellmy, F., von Erlach, T., *et al.* Analysis of the accessibility of CLIP bound sites reveals that nucleation of the miRNA: mRNA pairing occurs preferentially at the 3?-end of the seed match. *RNA (New York, N.Y.)*, 2012.

[124] Ribbink, A. J., Marsh, B. A., Marsh, A. C., *et al.* A preliminary survey of the cichlid fishes of rocky habitats in Lake Malawi. *S. Afr. J. Zool.*, 18:pp. 149–310, August 1983.

[125] Konings, A. *Malaŵi Cichlids in Their Natural Habitat* (Cichlid Press, 2007).

[126] Trewavas, E. Nouvel examen des genres et sous-genres du complexe Pseudotropheus-Melanochromis du lac Malawi (Pisces, Perciformes, Cichlidae). *Revue française d'Aquariologie*, 10(4):pp. 97–106, 1983.

[127] Stauffer, J. R., Jr, Bowers, N. J., Kellogg, K. A. A revision of the blue-black Pseudotropheus zebra (Teleostei: Cichlidae) complex from Lake Malaŵi, Africa, with a description of a new genus and ten new species. In *Proceedings of the Academy of Natural Sciences of Philadelphia* (1997).

[128] Genner, M. J., Nichols, P., Carvalho, G. R., *et al.* Reproductive isolation among deep-water cichlid fishes of Lake Malawi differing in monochromatic male breeding dress. *Molecular Ecology*, 16(3):pp. 651–662, February 2007.

[129] Eccles, D. H., Trewavas, E. *Malawian Cichlid Fishes* (1989).

[130] Joyce, D. A., Lunt, D. H., Bills, R., *et al.* An extant cichlid fish radiation emerged in an extinct Pleistocene lake. *Nature*, 435(7038):pp. 90–95, May 2005.

[131] Genner, M. J., Ngatunga, B. P., Mzighani, S., *et al.* Geographical ancestry of Lake Malawi's cichlid fish diversity: Figure 1. *Biology Letters*, 11(6):p. 20150232, June 2015.

[132] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org*, March 2013.

[133] DePristo, M. A. M., Banks, E. E., Poplin, R. R., *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):pp. 491–498, April 2011.

[134] McKenna, A., Hanna, M., Banks, E., *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):pp. 1297–1303, August 2010.

[135] Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, 27(21):pp. 2987–2993, November 2011.

[136] Browning, S. R., Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, 81(5):pp. 1084–1097, November 2007.

[137] Jun, G., Flickinger, M., Hetrick, K. N., *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American journal of human genetics*, 91(5):pp. 839–848, November 2012.

[138] Miller, W., Rosenbloom, K., Hardison, R. C., *et al.* 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome research*, 17(12):pp. 1797–1808, December 2007.

[139] Harris, R. S. *Improved pairwise alignment of genomic DNA*. Ph.D. thesis, The Pennsylvania State University, 2007.

[140] Blanchette, M., Kent, W. J., Riemer, C., *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):pp. 708–715, April 2004.

[141] Thorvaldsdóttir, H., Robinson, J. T., Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):pp. 178–192, March 2013.

[142] Birney, E., Bateman, A., Clamp, M. E., *et al.* Mining the draft human genome. *Nature*, 409(6822):pp. 827–828, February 2001.

[143] Rosenbloom, K. R., Armstrong, J., Barber, G. P., *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic acids research*, 43(Database issue):pp. D670–81, January 2015.

[144] Thorvaldsdóttir, H., Robinson, J. T., Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):pp. 178–192, March 2013.

[145] Kent, W. J., Sugnet, C. W., Furey, T. S., *et al.* The human genome browser at UCSC. *Genome research*, 12(6):pp. 996–1006, June 2002.

[146] Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome research*, 12(4):pp. 656–664, April 2002.

[147] Pendleton, M., Sebra, R., Pang, A. W. C., *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12(8):pp. 780–786, August 2015.

[148] Earl, D., Bradnam, K., St John, J., *et al.* Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome research*, 21(12):pp. 2224–2241, December 2011.

[149] Simpson, J. T., Durbin, R. Efficient de novo assembly of large genomes using compressed data structures. *Genome research*, 22(3):pp. 549–556, March 2012.

[150] Simpson, J. T., Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics (Oxford, England)*, 26(12):pp. i367–73, June 2010.

[151] Ferragina, P., Manzini, G. Opportunistic data structures with applications. *Foundations of Computer Science*, 2000.

[152] C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, 282(5396):pp. 2012–2018, December 1998.

[153] Adams, M. D., Celniker, S. E., Holt, R. A., *et al.* The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)*, 287(5461):pp. 2185–2195, March 2000.

[154] Potato Genome Sequencing Consortium, Xu, X., Pan, S., *et al.* Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):pp. 189–195, July 2011.

[155] Charlesworth, D., Willis, J. H. The genetics of inbreeding depression. *Nature reviews. Genetics*, 10(11):pp. 783–796, November 2009.

[156] Zhang, G., Fang, X., Guo, X., *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 490(7418):pp. 49–54, October 2012.

[157] Vinson, J. P. J., Jaffe, D. B. D., O'Neill, K. K., *et al.* Assembly of polymorphic genomes: algorithms and application to Ciona savignyi. *Genes & development*, 15(8):pp. 1127–1135, July 2005.

[158] Donmez, N., Brudno, M. Hapsembler: An Assembler for Highly Polymorphic Genomes. In *Research in Computational Molecular Biology*, pp. 38–52 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011).

[159] Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of heredity*, 100(6):pp. 659–674, November 2009.

[160] i5K Consortium. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *The Journal of heredity*, 104(5):pp. 595–600, September 2013.

[161] GIGA Community of Scientists, Bracken-Grissom, H., Collins, A. G., *et al.* The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. *The Journal of heredity*, 105(1):pp. 1–18, January 2014.

[162] Li, R. R., Zhu, H. H., Ruan, J. J., *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genes & development*, 20(2):pp. 265–272, February 2010.

[163] Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487(7405):pp. 94–98, July 2012.

[164] Charlesworth, B. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature reviews. Genetics*, 10(3):pp. 195–205, March 2009.

[165] Keightley, P. D., Pinharanda, A., Ness, R. W., *et al.* Estimation of the spontaneous mutation rate in Heliconius melpomene. *Molecular biology and evolution*, 32(1):pp. 239–243, January 2015.

[166] Simpson, J. T. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics (Oxford, England)*, 30(9):pp. 1228–1235, May 2014.

[167] Jiggins, C. D., Mavarez, J., Beltrán, M., *et al.* A genetic linkage map of the mimetic butterfly Heliconius melpomene. *Genetics*, 171(2):pp. 557–570, October 2005.

[168] Hardin, G. The competitive exclusion principle. *Science (New York, N.Y.)*, 131(3409):pp. 1292–1297, April 1960.

[169] Zaret, T. M., Rand, A. S. Competition in tropical stream fishes: support for the competitive exclusion principle. *Ecology*, 1971.

[170] Turner, G. F., Burrows, M. T. A Model of Sympatric Speciation by Sexual Selection. *Proceedings of the Royal Society B: Biological Sciences*, 260(1359):pp. 287–292, June 1995.

[171] Weissing, F. J., Edelaar, P., van Doorn, G. S. Adaptive speciation theory: a conceptual review. *Behavioral ecology and sociobiology*, 65(3):pp. 461–480, March 2011.

[172] M'Gonigle, L. K., Mazzucco, R., Otto, S. P., *et al.* Sexual selection enables long-term coexistence despite ecological equivalence. *Nature*, 484(7395):pp. 506–509, April 2012.

[173] Wang, J. Estimation of effective population sizes from data on genetic markers. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1459):pp. 1395–1409, July 2005.

[174] Whiteley, A. R., Bhat, A., Martins, E. P., *et al.* Population genomics of wild and laboratory zebrafish (Danio rerio). *Molecular Ecology*, 20(20):pp. 4259–4276, October 2011.

[175] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. `http://www.R-project.org/`.

[176] Genner, M. J., Turner, G. F. Ancient hybridization and phenotypic novelty within Lake Malawi's cichlid fish radiation. *Molecular biology and evolution*, 29(1):pp. 195–206, January 2012.

[177] Vos, P., Hogers, R., Bleeker, M., *et al.* AFLP: a new technique for DNA fingerprinting. *Nucleic acids research*, 23(21):pp. 4407–4414, November 1995.

[178] Albertson, R. C., Markert, J. A., Danley, P. D., *et al.* Phylogeny of a rapidly evolving clade: the cichlid fishes of Lake Malawi, East Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 96(9):pp. 5107–5110, April 1999.

[179] Allender, C. J., Seehausen, O., Knight, M. E., *et al.* Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):pp. 14074–14079, November 2003.

[180] Bouckaert, R., Heled, J. DensiTree 2: Seeing Trees Through the Forest. *bioRxiv*, 2014.

[181] Saitou, N., Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):pp. 406–425, July 1987.

[182] Robinson, D. F., Foulds, L. R. Comparison of phylogenetic trees. *Mathematical Biosciences*, 1981.

[183] Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics (Oxford, England)*, 27(4):pp. 592–593, February 2011.

[184] Ballard, J. W. O., Whitlock, M. C. The incomplete natural history of mitochondria. *Molecular Ecology*, 13(4):pp. 729–744, April 2004.

[185] Wilson, C. C., Bernatchez, L. The ghost of hybrids past: fixation of arctic charr (Salvelinus alpinus) mitochondrial DNA in an introgressed population of lake trout (S. namaycush). *Molecular Ecology*, 1998.

[186] Zieliński, P., Nadachowska-Brzyska, K., Wielstra, B., *et al.* No evidence for nuclear introgression despite complete mtDNA replacement in the Carpathian newt (Lissotriton montandoni). *Molecular Ecology*, 22(7):pp. 1884–1903, April 2013.

[187] Pons, J.-M., Sonsthagen, S., Dove, C., *et al.* Extensive mitochondrial introgression in North American Great Black-backed Gulls (Larus marinus) from the American Herring Gull (Larus smithsonianus) with little nuclear DNA impact. *Heredity*, 112(3):pp. 226–239, March 2014.

[188] Good, J. M., Vanderpool, D., Keeble, S., *et al.* Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. *Evolution; international journal of organic evolution*, 69(8):pp. 1961–1972, August 2015.

[189] Bachtrog, D., Thornton, K., Clark, A., *et al.* Extensive introgression of mitochondrial DNA relative to nuclear genes in the Drosophila yakuba species group. *Evolution; international journal of organic evolution*, 60(2):pp. 292–302, February 2006.

[190] Green, R. E., Krause, J., Briggs, A. W., *et al.* A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, 328(5979):pp. 710–722, May 2010.

[191] Durand, E. Y., Patterson, N., Reich, D., *et al.* Testing for ancient admixture between closely related populations. *Molecular biology and evolution*, 28(8):pp. 2239–2252, August 2011.

[192] Martin, S. H., Davey, J. W., Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular biology and evolution*, 32(1):pp. 244–257, January 2015.

[193] Lawson, D. J., Hellenthal, G., Myers, S., *et al.* Inference of population structure using dense haplotype data. *PLoS genetics*, 2012.

[194] Purcell, S., Neale, B., Todd-Brown, K., *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):pp. 559–575, September 2007.

[195] Patterson, N., Price, A. L., Reich, D. Population structure and eigenanalysis. *PLoS genetics*, 2(12):pp. e190–e190, December 2006.

[196] Ligges, U., Mächler, M. Scatterplot3d - an R Package for Visualizing Multivariate Data. *Journal of Statistical Software*, 8(11):pp. 1–20, 2003. `http://www.jstatsoft.org`.

[197] Bhatia, G., Patterson, N., Sankararaman, S., *et al.* Estimating and interpreting FST: The impact of rare variants. *Genome research*, 23(9):pp. 1514–1521, September 2013.

[198] Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, 22(21):pp. 2688–2690, November 2006.

[199] Stamatakis, A., Hoover, P., Rougemont, J. A rapid bootstrap algorithm for the RAxML Web servers. *Systematic biology*, 57(5):pp. 758–771, October 2008.

[200] Malinsky, M., Challis, R. J., Tyers, A. M., *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science (New York, N.Y.)*, 350(6267):pp. 1493–1498, December 2015.

[201] Ricardo, C. K. The fishes of Lake Rukwa. *Journal of the Linnean Society of London*, 1939.

[202] Delalande, M. Hydrologie et géochimie isotopique du lac Masoko et de lacs volcaniques de la province active du Rungwe (Sud-Ouest Tanzanie). *www.theses.fr*, May 2008.

[203] Barker, Williamson, Gasse, *et al.* Climatic and volcanic forcing revealed in a 50,000-year diatom record from Lake Massoko, Tanzania. *Quaternary Research*, 60(3):pp. 9–9, November 2003.

[204] Davey, J. W., Blaxter, M. L. RADSeq: next-generation population genetics. *Briefings in Functional Genomics*, 2010.

[205] Schliewen, U. K., Tautz, D., Pääbo, S. Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature*, 368(6472):pp. 629–632, April 1994.

[206] Barluenga, M., Stölting, K. N., Salzburger, W., *et al.* Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, 439(7077):pp. 719–723, February 2006.

[207] Alexander, D. H., Novembre, J., Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genes & development*, 19(9):pp. 1655–1664, August 2009.

[208] Martin, C. H., Cutler, J. S., Friel, J. P., *et al.* Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. *Evolution; international journal of organic evolution*, May 2015.

[209] Schiffels, S., Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):pp. 919–925, August 2014.

[210] Tenesa, A., Navarro, P., Hayes, B. J., *et al.* Recent human effective population size estimated from linkage disequilibrium. *Genome research*, 17(4):pp. 520–526, April 2007.

[211] Hudson, R. R. R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):pp. 337–338, February 2002.

[212] Sousa, V., Hey, J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature reviews. Genetics*, 14(6):pp. 404–414, June 2013.

[213] Beaumont, M. A. Adaptation and speciation: what can F(st) tell us? *Trends in ecology & evolution*, 20(8):pp. 435–440, August 2005.

[214] O Quin, C. T., Drilea, A. C., Conte, M. A., *et al.* Mapping of pigmentation QTL on an anchored genome assembly of the cichlid fish, Metriaclima zebra. *BMC genomics*, 14(1):p. 287, April 2013.

[215] Price, T. D., Bouvier, M. M. The evolution of F1 postzygotic incompatibilities in birds. *Evolution; international journal of organic evolution*, 56(10):pp. 2083–2089, October 2002.

[216] Stelkens, R. B., Young, K. A., Seehausen, O. The accumulation of reproductive incompatibilities in African cichlid fish. *Evolution; international journal of organic evolution*, 64(3):pp. 617–633, March 2010.

[217] Van Der Sluijs, I., Van Dooren, T. J. M., Seehausen, O., *et al.* A test of fitness consequences of hybridization in sibling species of Lake Victoria cichlid fish. *Journal of evolutionary biology*, 21(2):pp. 480–491, March 2008.

[218] Kaneshiro, K. Y. Sexual Isolation, Speciation and the Direction of Evolution. *Evolution; international journal of organic evolution*, 34(3):p. 437, May 1980.

[219] Ashburner, M., Ball, C. A., Blake, J. A., *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25(1):pp. 25–29, May 2000.

[220] Merico, D., Isserlin, R., Stueker, O., *et al.* Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PloS one*, 5(11):p. e13984, 2010.

[221] Sugawara, T., Terai, Y., Imai, H., *et al.* Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from Lakes Tanganyika and Malawi. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):pp. 5448–5453, April 2005.

[222] Morris, J. H., Apeltsin, L., Newman, A. M., *et al.* clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC bioinformatics*, 12:p. 436, 2011.

[223] Oesper, L., Merico, D., Isserlin, R., *et al.* WordCloud: a Cytoscape plugin to create a visual semantic summary of networks. *Source code for biology and medicine*, 6:p. 7, 2011.

[224] Duester, G. Families of retinoid dehydrogenases regulating vitamin A function: production of visual pigment and retinoic acid. *European journal of biochemistry / FEBS*, 267(14):pp. 4315–4324, July 2000.

[225] Yamashita, T., Liu, J., Gao, J., *et al.* Essential and synergistic roles of RP1 and RP1L1 in rod photoreceptor axoneme and retinitis pigmentosa. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 29(31):pp. 9748–9760, August 2009.

[226] Jing, X., Malicki, J. Zebrafish ale oko, an essential determinant of sensory neuron survival and the polarity of retinal radial glia, encodes the p50 subunit of dynactin. *Development (Cambridge, England)*, 136(17):pp. 2955–2964, September 2009.

[227] Sayer, J. A., Otto, E. A., O'Toole, J. F., *et al.* The centrosomal protein nephrocystin-6 is mutated in Joubert syndrome and activates transcription factor ATF4. *Nature genetics*, 38(6):pp. 674–681, June 2006.

[228] Esterberg, R., Fritz, A. dlx3b/4b are required for the formation of the preplacodal region and otic placode through local modulation of BMP activity. *Developmental biology*, 325(1):pp. 189–199, January 2009.

[229] Kwon, H.-J., Riley, B. B. Mesendodermal signals required for otic induction: Bmp-antagonists cooperate with Fgf and can facilitate formation of ectopic otic tissue. *Developmental dynamics : an official publication of the American Association of Anatomists*, 238(6):pp. 1582–1594, June 2009.

[230] Fay, R. R., Popper, A. N. Evolution of hearing in vertebrates: the inner ears and processing. *Hearing research*, 149(1-2):pp. 1–10, November 2000.

[231] Popper, A. N., Ramcharitar, J., Campana, S. E. Why otoliths? Insights from inner ear physiology and fisheries biology. *Marine and freshwater Research*, 56(5):p. 497, 2005.

[232] Soria-Carrasco, V., Gompert, Z., Comeault, A. A., *et al.* Stick insect genomes reveal natural selection's role in parallel speciation. *Science (New York, N.Y.)*, 344(6185):pp. 738–742, May 2014.

[233] Arnegard, M. E., McGee, M. D., Matthews, B., *et al.* Genetics of ecological divergence during speciation. *Nature*, 511(7509):pp. 307–311, July 2014.

[234] Andrew, R. L., Rieseberg, L. H. Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution; international journal of organic evolution*, 67(9):pp. 2468–2482, September 2013.

[235] Catchen, J. M., Amores, A., Hohenlohe, P., *et al.* Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda, Md.)*, 1(3):pp. 171–182, August 2011.

[236] Keane, T. M., Creevey, C. J., Pentony, M. M., *et al.* Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC evolutionary biology*, 6:p. 29, 2006.

[237] Delaneau, O., Marchini, J., Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):pp. 179–181, February 2012.

[238] Delaneau, O., Howie, B., Cox, A. J., *et al.* Haplotype estimation using sequencing reads. *American journal of human genetics*, 93(4):pp. 687–696, October 2013.

[239] Quinlan, A. R., Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):pp. 841–842, March 2010.

[240] Wakeley, J. Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical population biology*, 49(3):pp. 369–386, June 1996.

[241] Alexa, A., Rahnenfuhrer, J. *topGO: topGO: Enrichment analysis for Gene Ontology*, 2010. R package version 2.20.0.

[242] Huber, W., Carey, *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, 12(2):pp. 115–121, 2015. `http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html`.

[243] Carlson, M. *GO.db: A set of annotation maps describing the entire Gene Ontology*, 2015. R package version 3.1.2.

[244] Carlson, M. *org.Dr.eg.db: Genome wide annotation for Zebrafish*, 2015. R package version 3.1.2.

[245] Turner, G. F., Seehausen, O., Knight, M. E., *et al.* How many species of cichlid fishes are there in African lakes? *Molecular Ecology*, 10(3):pp. 793–806, March 2001.

[246] Taylor, M. I., Meardon, F., Turner, G., *et al.* Characterization of tetranucleotide microsatellite loci in a Lake Victorian, haplochromine cichlid fish: a Pundamilia pundamilia x Pundamilia nyererei hybrid. *Molecular Ecology Notes*, 2(4):pp. 443–445, December 2002.

[247] Gabriel, S., Ziaugra, L., Tabbaa, D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, Chapter 2:p. Unit 2.12, January 2009.

[248] Huber, W., Carey, V. J., Gentleman, R., *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12(2):pp. 115–121, February 2015.

[249] Zheng, X., Levine, D., Shen, J., *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 2012.

[250] Ueyama, H., Kuwayama, S., Imai, H., *et al.* Novel missense mutations in red/green opsin genes in congenital color-vision deficiencies. *Biochemical and biophysical research communications*, 294(2):pp. 205–209, June 2002.

[251] Conte, G. L., Arnegard, M. E., Peichel, C. L., *et al.* The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, 279(1749):pp. 5039–5047, December 2012.

[252] Stern, D. L. The genetic causes of convergent evolution. *Nature reviews. Genetics*, 2013.

[253] Hudson, R. R. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 1990.

Appendix A

Lake Malawi genetic diversity

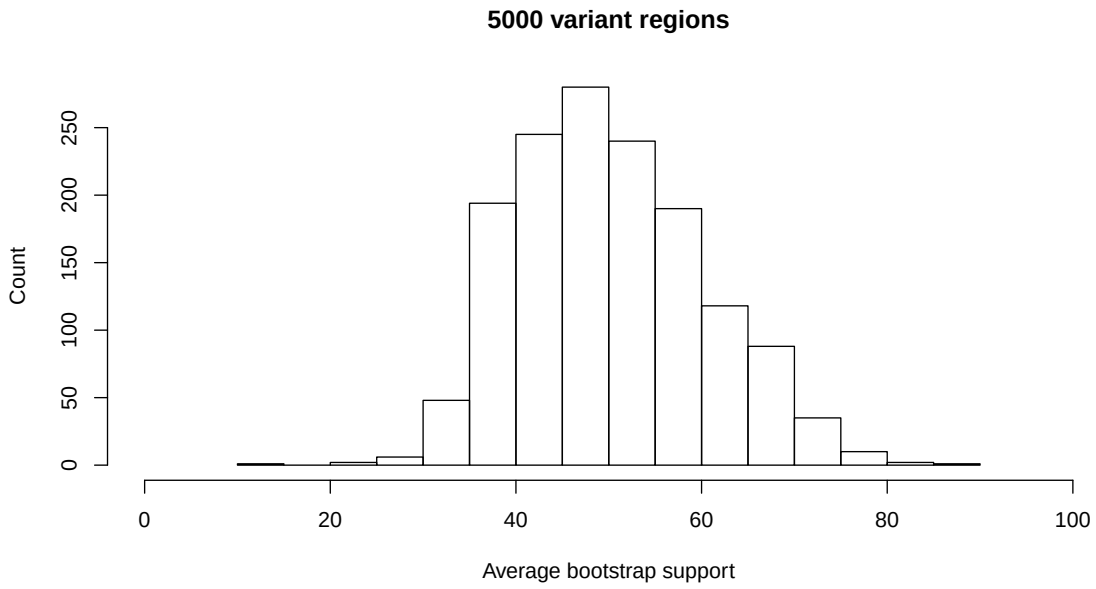


Fig. A.1 **Average bootstrap values for 1,460 phylogenies representing regions along the genome.**

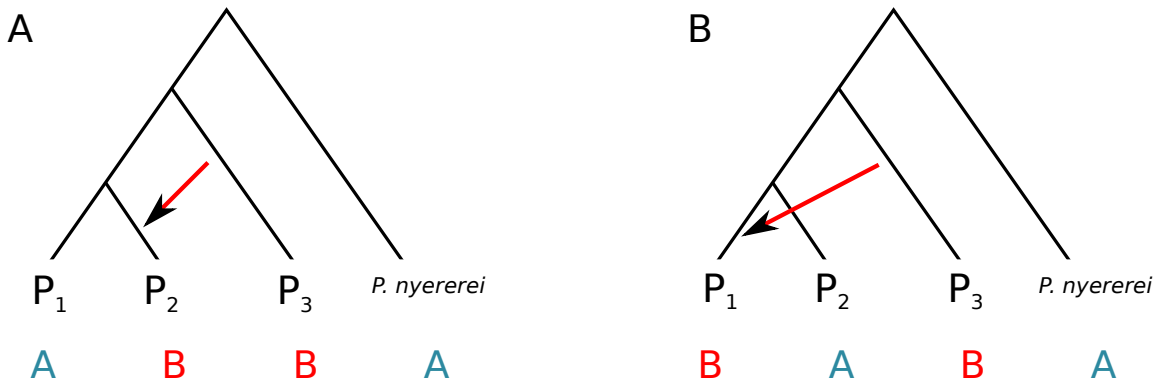


Fig. A.2 **The ABBA-BABA test for introgression.** The ancestral allele is denoted by the A character and the derived allele by the B character. **(A)** Differential gene flow P_3 to P_2 will result in an excess of shared derived alleles between them - the ABBA pattern. **(B)** Gene flow P_3 to P_1 will result in an excess of the BABA pattern.

Appendix B

Lake Massoko speciation

Table B.1 **Results of a survey of fish fauna in six crater lakes of Rungwe District, Tanzania.** Conducted in July and November 2011.

Lake	Species	Family	Tribe	Probable Status
Kingiri	*Astatotilapia* sp. 'kingiri black'	Cichlidae	Haplochromini	Endemic
	Rhamphochromis sp. 'kingiri dwarf'	Cichlidae	Haplochromini	Endemic
	Rhamphochromis sp. 'kingiri brevis'	Cichlidae	Haplochromini	Endemic
	Serranochromis robustus	Cichlidae	Haplochromini	Native
	Coptodon rendalli	Cichlidae	Tilapiini	Native
	Oreochromis shiranus	Cichlidae	Tilapiini	Native
	Oreochromis (Nyasalapia) squamipinnis	Cichlidae	Tilapiini	Native
	Clarias gariepinus	Clariidae		Native
	Micropanchax johnstoni	Poeciliidae		Native
	Barbus radiatus	Cyprinidae		Native
	Barbus trimaculatus	Cyprinidae		Native
Ilamba	*Astatotilapia* sp. 'ilamba black'	Cichlidae	Haplochromini	Endemic
	Otophraynx sp. 'Ilamba tetrastigma'	Cichlidae	Haplochromini	Endemic
	Oreochromis cf shiranus	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis cf squamipinnis	Cichlidae	Tilapiini	Native/Endemic?
	Clarias gariepinus	Clariidae		Native
	Mesobola cf. spinifer	Cyprinidae		Native
	Barbus paludinosos	Cyprinidae		Native
	Barbus trimaculatus	Cyprinidae		Native
	Barbus macrotaenia	Cyprinidae		Native
	Barbus radiatus	Cyprinidae		Native
Ikapu	*Astastotilapia* sp. 'ikapu dark'	Cichlidae	Haplochromini	Endemic
	Tilapia sparrmanii	Cichlidae	Tilapiini	Native/Introduced?
	Oreochromis 'golden chambo'	Cichlidae	Tilapiini	Endemic
	Clarias gariepinus	Clariidae		Native/Introduced?
Itamba	*Astatotilapia* sp. 'itamba dark'	Cichlidae	Haplochromini	Endemic
	Oreochromis cf. shiranus	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis (Nyasalapia) cf. karongae	Cichlidae	Tilapiini	Native/Endemic?
	Oreochromis niloticus	Cichlidae	Tilapiini	Introduced
Massoko	*Astatotilapia* sp. 'massoko benthic'	Cichlidae	Haplochromini	Endemic
	Astatotilapia sp. 'massoko littoral'	Cichlidae	Haplochromini	Endemic
	Coptodon rendalli	Cichlidae	Tilapiini	Introduced?
	Oreochromis (Nyasalapia) squamipinnis	Cichlidae	Tilapiini	Native/Endemic?
	Clarias gariepinus	Clariidae		Introduced?
Itende	*Astatotilapia* sp. 'itende'	Cichlidae	Haplochromini	Endemic
	Oreochromis (Nyasalapia) squamipinnis	Cichlidae	Tilapiini	Native/Endemic?

Table B.2 **An overview of *Astatotilapia* samples collected for RAD sequencing.**

Sampling location (ecomorph)	N	Sampling Dates	Collector(s)	Latitude S	Longitude E
Lake Massoko (benthic)	5	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Massoko (littoral)	3	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Massoko (small, unassigned)	4	17/07/2011	MG, BN, GT, SM, AS	9°20'0	33°45'18
Lake Itende	3	27/11/2011	MG, GT, AS	9°19'19	33°47'15
Lake Ikapu	3	20/07/2011	MG, BN, GT, SM, AS	9°22'12	33°48'25
Lake Itamba	2	19/07/2011	MG, BN, GT, SM, AS	9°21'04	33°50'39
Lake Ilamba	2	17/07/2011	MG, BN, GT, SM, AS	9°23'33	33°50'09
Lake Kingiri	8	15+21/07/2011	MG, BN, GT, SM, AS	9°25'08	33°51'29
Ruo river	2	22/05/2009	MG, AS, JS	15°50'77	35°11'69
Unaka lagoon	1	24/07/2011	MG, AS	12°23'59	34°05'17
Mbenji island	2	–/01/2011	MG (from imported wild stock)	13°26'	34°29'
Makanjila	2	18/01/2011	MG, PP, JB	13°41'35	34°50'51
Chisumulu island	1	–/01/2011	MG (from imported wild stock)	12°00'	34°37'
Mgowesa river	2	16/07/2011	MG, BN, GT, SM, AS	9°23'43	33°49'38
Astatotilapia tweddlei (Lake Chilwa)	1	19/05/2009	MG, AS, JS	15°22'18	35°35'20

MG = Martin Genner, GT = George Turner, BN = Benjamin Ngatunga, SM = Semvua Mzighani, AS = Alan Smith, JS = Jennifer Swanstrom, PP = Paul Parsons, JB = Jon Bridle.

Table B.3 **The migration parameter `M` and the probability of migration** At any point, the coalescent simulation is tracking n_1 alleles from population 1 and n_2 alleles from population 2, denoted as (n_1,n_2). Going back in time, 'events' occur with probability $\mathbb{P}(event)$ in which either: a) a common ancestor of two lineages is found within a population; or b) a lineage migrates to the other population. This table gives probabilities that the event is a migration, for a range of parameters `M`, and numbers of tracked lineages (n_1,n_2). The calculation is based Hudson [253]:

$$\mathbb{P}(migration|event) = \frac{n_1 \frac{M}{2}}{\binom{n_1}{2}+\binom{n_2}{2}+(n_1+n_2)\frac{M}{2}} + \frac{n_2 \frac{M}{2}}{\binom{n_1}{2}+\binom{n_2}{2}+(n_1+n_2)\frac{M}{2}}$$

M	Lineages tracked (n_1,n_2)		
	(74,64)	(37,32)	(18,16)
5	6.82%	12.93%	23.74%
10	12.76%	22.89%	38.37%
20	22.63%	37.26%	55.46%
40	36.91%	54.29%	71.35%

Table B.4 **Location and lengths of highly diverged regions (HDRs)**

scaffold	start coordinate	end coordinate	length (bp)	scaffold	start coordinate	end coordinate	length (bp)
0	10512411	10559800	47389	51	1450783	1493272	42489
0	10570498	10598504	28006	55	3423595	3500130	76535
0	11529594	11540402	10808	57	46109	77869	31760
0	11994849	12015103	20254	57	1615373	1638983	23610
0	14003832	14040483	36651	64	55966	175700	119734
0	18256071	18263999	7928	74	591451	600724	9273
5	1920004	1936600	16596	77	2089974	2164351	74377
6	2399603	2417150	17547	78	6039	59940	53901
11	5426321	5452278	25957	82	2236206	2273645	37439
12	3879628	3890173	10545	83	1379873	1425116	45243
14	2810211	2821278	11067	84	2355047	2388352	33305
14	3582492	3609841	27349	84	2399084	2517997	118913
14	3661853	3697260	35407	88	819852	845401	25549
15	1934238	1967068	32830	88	1194601	1316288	121687
15	1981201	2033263	52062	88	1372483	1527476	154993
15	2912637	2961336	48699	88	1732907	1868455	135548
15	3390209	3412823	22614	88	1908746	1943289	34543
15	4641580	4880808	239228	88	2418799	2435992	17193
15	4907565	5049805	142240	91	129230	153938	24708
15	5452492	5474330	21838	92	296055	342364	46309
15	6705210	6818468	113258	93	1295671	1314656	18985
15	7208463	7304325	95862	95	1001404	1044619	43215
15	7317980	7335547	17567	97	298706	313982	15276
15	7507678	7592890	85212	97	2158978	2172667	13689
15	7682086	7700855	18769	97	2188270	2212097	23827
18	2797554	2832090	34536	99	330458	339693	9235
18	6702155	6728393	26238	99	355072	639642	284570
26	5297874	5318387	20513	108	572788	738675	165887
26	5517553	5550216	32663	108	814090	941030	126940
30	183937	257768	73831	108	963573	1023559	59986
30	797497	844062	46565	112	1966090	2014162	48072
30	3978481	4066243	87762	113	1062779	1122847	60068
31	4041909	4078026	36117	114	505730	526658	20928
32	4518067	4589712	71645	114	1902474	2005892	103418
32	4616851	4625659	8808	120	918534	962612	44078
32	4886114	4908718	22604	121	1894243	1983613	89370
33	4163850	4200587	36737	126	420889	439080	18191
36	1010120	1029957	19837	143	1376555	1380941	4386
39	465841	688825	222984	148	1458116	1644136	186020
39	1004596	1047034	42438	148	1669247	1754172	84925
39	1207716	1236548	28832	162	1227615	1263777	36162
39	2153726	2185052	31326	164	0	113596	113596
39	2245171	2269911	24740	164	196412	276496	80084
39	2294746	2311616	16870	186	693304	711017	17713
39	2323506	2340670	17164	190	805453	832175	26722
40	1569517	1608679	39162	206	177202	290266	113064
40	1870518	1891875	21357	229	252128	283116	30988
43	3655049	3696771	41722	229	470627	578767	108140
45	2785077	2828731	43654	304	0	70114	70114

Table B.5 **GO enrichment terms in candidate 'islands of speciation' ±50kb**

GO ID	Term	Total	Found	Expected	p-value
GO:0005813	centrosome	59	4	0.66	0.0041
GO:0009798	axis specification	31	3	0.33	0.0043
GO:0044877	macromolecular complex binding	233	7	2.4	0.0098
GO:0005874	microtubule	77	4	0.86	0.0105
GO:0046530	photoreceptor cell differentiation	44	3	0.47	0.0116
GO:1903034	regulation of response to wounding	17	2	0.18	0.014
GO:0030916	otic vesicle formation	19	2	0.2	0.0174
GO:0043484	regulation of RNA splicing	20	2	0.22	0.0192
GO:0022037	metencephalon development	23	2	0.25	0.025
GO:0006414	translational elongation	25	2	0.27	0.0293
GO:0004812	aminoacyl-tRNA ligase activity	27	2	0.28	0.0311
GO:0006418	tRNA aminoacylation for protein translation	26	2	0.28	0.0315
GO:0015629	actin cytoskeleton	63	3	0.7	0.033
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	2	0.32	0.0402
GO:0014070	response to organic cyclic compound	94	4	1.01	0.0472
GO:0032101	regulation of response to external stimulus	33	2	0.36	0.0488

Table B.6 **GO terms significantly enriched in all HDRs ±10kb**

GO ID	Term	Total	Found	Expected	p-value
GO:0005874	microtubule	77	4	0.69	0.0049
GO:0005200	structural constituent of cytoskeleton	12	2	0.11	0.0052
GO:0043401	steroid hormone mediated signaling pathway	46	3	0.42	0.0085
GO:0003707	steroid hormone receptor activity	47	3	0.43	0.009
GO:0005057	receptor signaling protein activity	58	3	0.53	0.0159
GO:0006418	tRNA aminoacylation for protein translation	26	2	0.24	0.0235
GO:0004812	aminoacyl-tRNA ligase activity	27	2	0.25	0.0251
GO:0001755	neural crest cell migration	28	2	0.26	0.027
GO:0045454	cell redox homeostasis	29	2	0.27	0.0288
GO:0051216	cartilage development	74	3	0.68	0.0303
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	2	0.28	0.0325
GO:0048675	axon extension	37	2	0.34	0.0451

Table B.7 **GO terms significantly enriched in all HDRs ±50kb**

GO ID	Term	Total	Found	Expected	p-value
GO:0071407	cellular response to organic cyclic compound	79	8	1.51	0.00012
GO:0003707	steroid hormone receptor activity	47	5	0.9	0.0019
GO:0004879	ligand-activated sequence-specific DNA binding RNA polymerase II transcription factor activity	31	4	0.59	0.0027
GO:0090101	negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	17	3	0.32	0.00381
GO:0005813	centrosome	59	5	1.13	0.0051
GO:0030916	otic vesicle formation	19	3	0.36	0.00528
GO:0046530	photoreceptor cell differentiation	44	4	0.84	0.00958
GO:0031012	extracellular matrix	70	5	1.34	0.0105
GO:0006418	tRNA aminoacylation for protein translation	26	3	0.5	0.01286
GO:0007051	spindle organization	26	3	0.5	0.01286
GO:0004812	aminoacyl-tRNA ligase activity	27	3	0.51	0.0142
GO:0005874	microtubule	77	5	1.47	0.0155
GO:0004519	endonuclease activity	52	4	0.99	0.0168
GO:0009798	axis specification	31	3	0.59	0.02075
GO:0005200	structural constituent of cytoskeleton	12	2	0.23	0.021
GO:0035141	medial fin morphogenesis	12	2	0.23	0.0211
GO:0021984	adenohypophysis development	13	2	0.25	0.02462
GO:0042802	identical protein binding	60	4	1.14	0.027
GO:0060037	pharyngeal system development	14	2	0.27	0.02837
GO:0022626	cytosolic ribosome	15	2	0.29	0.0323
GO:0045446	endothelial cell differentiation	16	2	0.31	0.0365
GO:1903034	regulation of response to wounding	17	2	0.32	0.04086
GO:0015698	inorganic anion transport	41	3	0.78	0.04291
GO:0032403	protein complex binding	135	6	2.57	0.0437
GO:0035088	establishment or maintenance of apical/basal cell polarity	18	2	0.34	0.0454

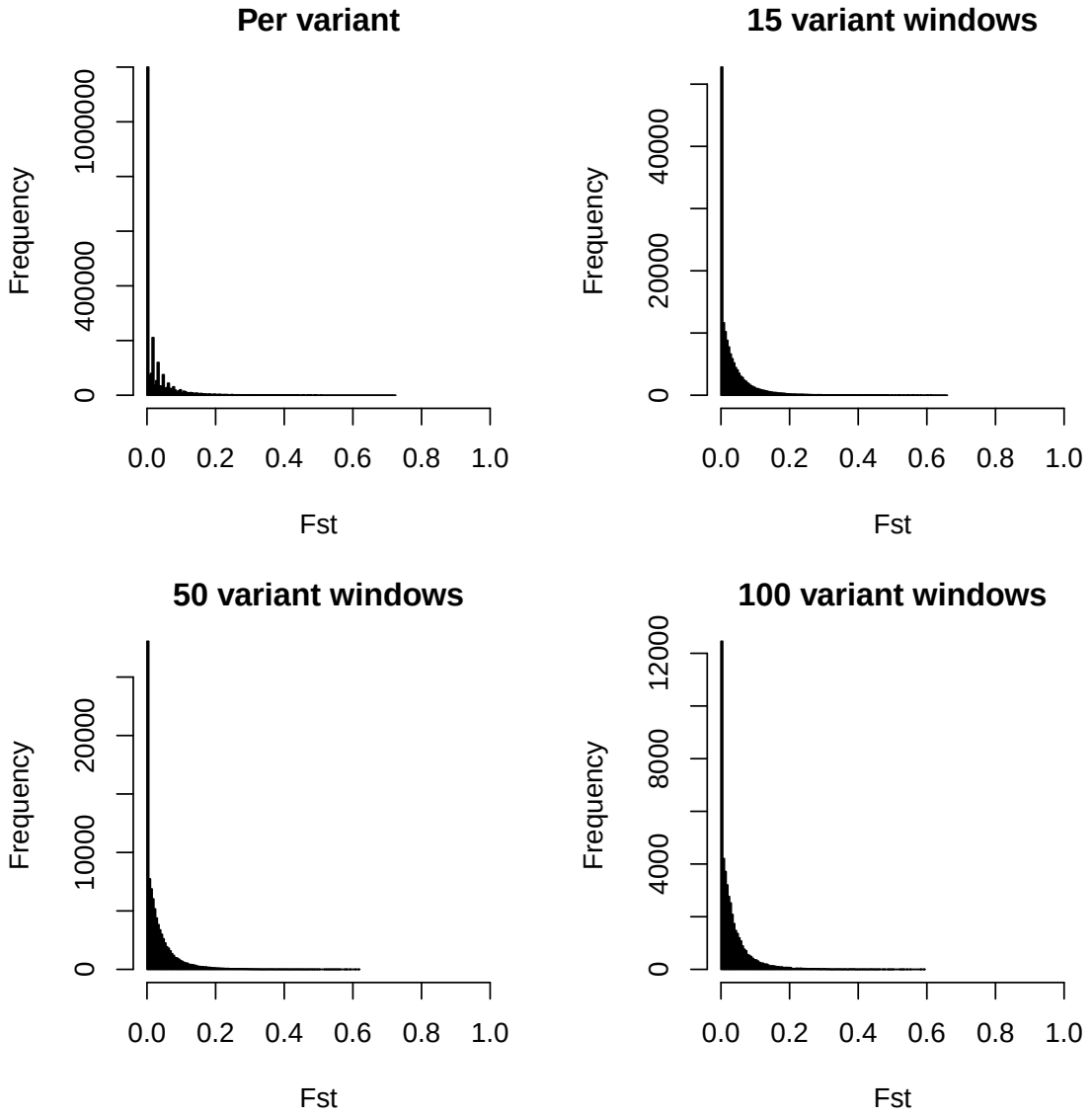


Fig. B.1 **Statistical distribution of within-Massoko F_{ST} divergence, per variant and in sliding windows of varying sizes** The distribution has a sharp L like shape, largely independent of the window size used, consistent with theoretical predictions about early stages of speciation with gene-flow [71].

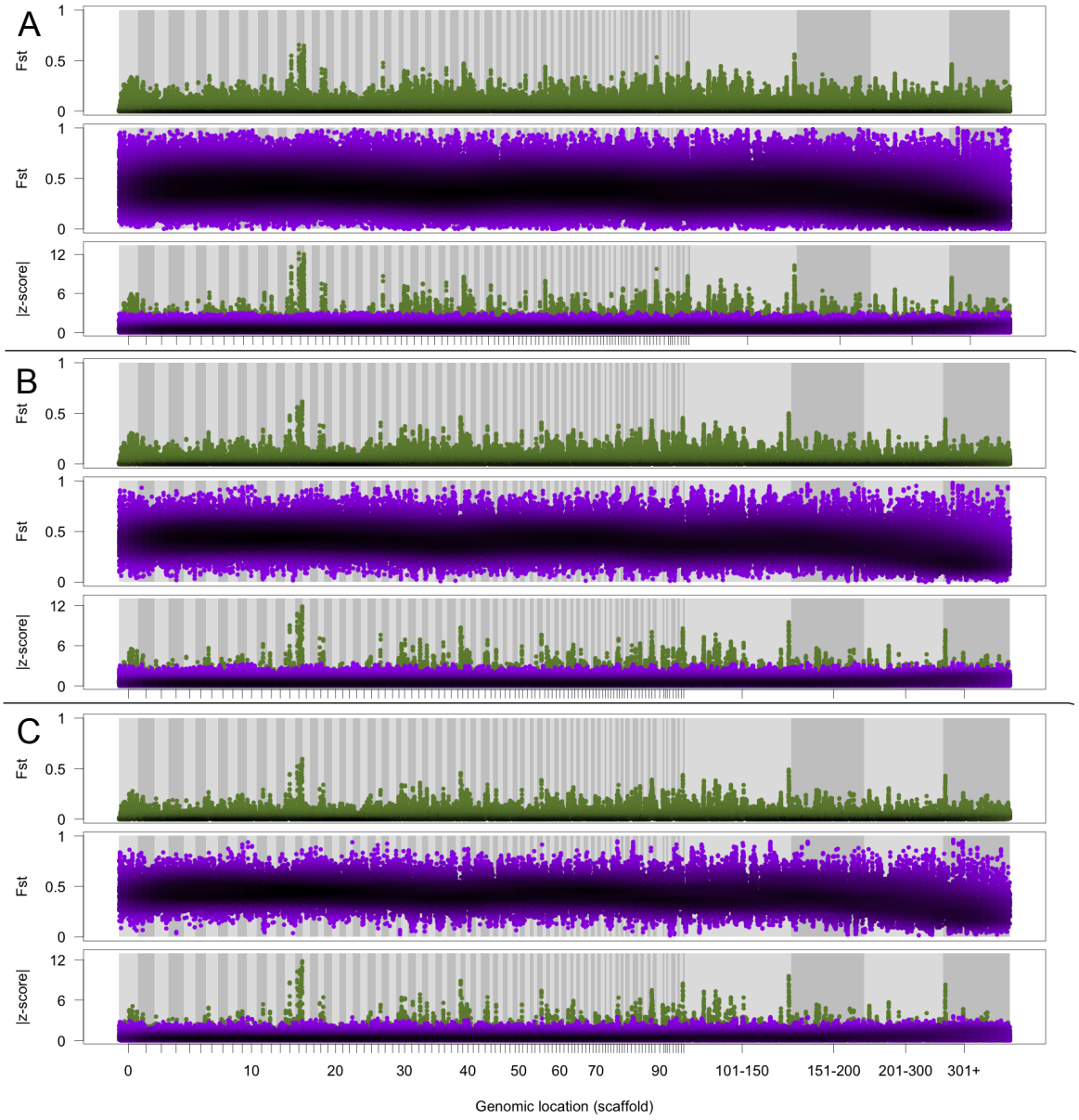


Fig. B.2 **Genome-wide pattern of F_{ST} divergence using sliding windows or varying sizes** The overall pattern of "genomic islands" raising above low background divergence is unaffected by varying the window size. Each figure shows genome-wide pattern of F_{ST} between *Massoko benthic* and *Massoko littoral* (green), and between combined Massoko and Itamba populations (purple), and absolute z-scores of Massoko-Itamba divergence (purple) and within-Massoko divergence (green). **(A)** Window size=15 variants; **(B)** Window size=50 variants; **(C)** Window size=100 variants.

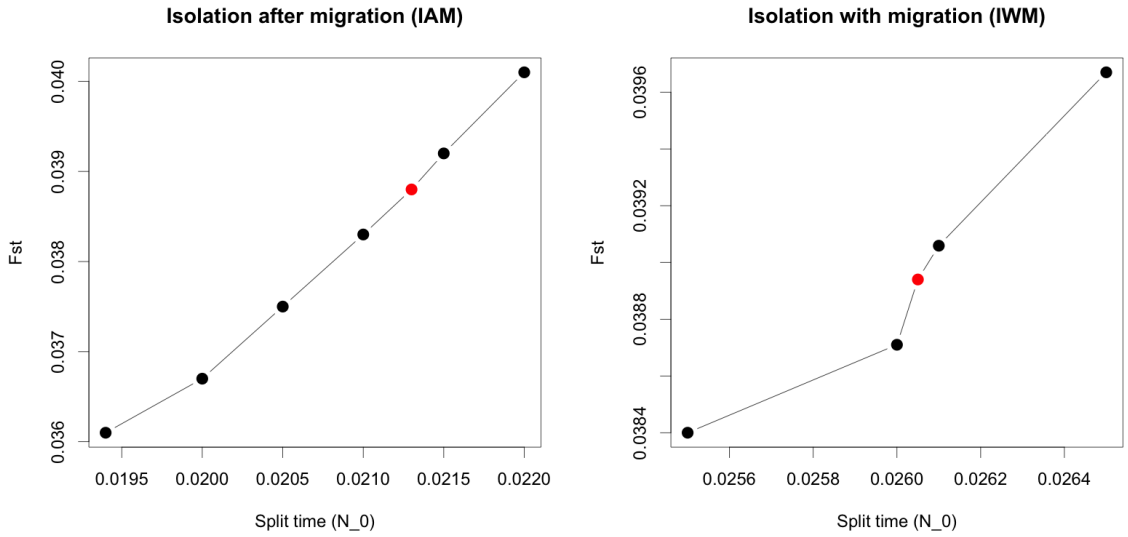


Fig. B.3 **Neutral simulations - fitting split time to match Massoko littoral-benthic F_{ST} divergence** The split time parameters were adjusted for simulations under both models of species formation to match the overall F_{ST} divergence observed between the *Massoko benthic* and *Massoko littoral* forms (0.0389). Several runs were tried (black points) until the optimal value was discovered for each model (red points).

AApos:	162	166	169	297	298	299
Ref:	V	S	T	G	A	A
H4:	L	A	A	S	S	S
H5:	.	.	A	.	.	S

Fig. B.4 **Amino-acid differences between the two haplotypes of the rhodopsin (*rho*) found in Lake Massoko *Astatotilapia*.** The differences are present at amino-acid positions 162, 166, 297, and 298. There are two additional amino-acid positions (169, 299) where both ecomorphs differ from the *M. zebra* reference.