The generation of novel phenotypes in a young radiation of soda lake cichlid fishes

India Grace Tyrala Lawson

Master of Science by Research

University of York

Department of Biology

January 2024

Abstract

Adaptive radiations are fantastic systems to study the origin of biodiversity as they generate considerable phenotypic and genetic diversity when given ecological opportunity. *Alcolapia* are small subgenus of four cichlids species each with considerable eco-morphological diversity. Chapter one provides background on adaptive radiations, the history and ecology of *Alcolapia* and describes polymorphisms within the species flock. Chapter two investigates the phylogenomic relationship of site 14, a population of individuals in the east of Lake Natron that have unusual dental morphologies but mimic *A. alcalica* and *A. latilabris* head morphology. RAxML predicts site 14 to be the sister lineage to Lake Natron taxa (100 bs). Furthermore, a genetic PCA, trait measurements, fineRADstructure, RAxML and Dsuite analysis give evidence site 14 contains two morphs which correspond to field categorisations. Furthermore, a GWAS on the two morphs identifies a candidate gene, *OSR2,* which affects mouse craniofacial and dental development, so may also cause the morphological differences within the head region of site 14 morphs. Further phenotyping of site 14 teeth is needed as previous research indicates they are atypical and reassessment of the size of the *Alcolapia* flock is required. Chapter three investigates the genetic basis of upturned and terminal morphologies within *A. alcalica*. A GWAS on mouth orientation, identifies two genes, *ANKMY2a* and *SOSTDC1a* which are related to craniofacial development in cichlids. All extremely upturned individuals were female, indicating a sexual basis for this trait and could be related to mouth brooding ability. Further investigation of the ecology of upturned *A. alcalica* is needed as upturned *A. alcalica* have been shown to consume fish scales, hinting at cannibalistic behaviour. Chapter four present conclusions and outlines the direction of further research including wider sampling, testing sexual selection for mouth orientation, KO experiments of candidate genes and investigating other polymorphisms.

Author's declaration

I declare that the following thesis is my own work and where I have had assistance from other people, I have correctly identified and cited their contribution. This work has not been submitted prior to this to any other qualification or programme of study.

India Lawson

Acknowledgements

Firstly, I would like to give a special thank you to my supervisor Kanchon Dasmahapatra, for guiding me through challenging bioinformatics pipelines, brainstorming ideas and getting just as excited about my research as me. I have learnt so much this past year. Thank you to Julia Day, for sending samples and being so hospitable when I visited UCL for sample collection and photography. I am also very grateful to Yacine Ben Chehida for providing his spare time including coaching me on bioinformatics techniques and troubleshooting my code. I also owe thanks to Michael Lawson for his valuable advice and analysis suggestions. Thank you to Antonia Ford, Julia Day, George Turner, Asilatu Shechonge and Lewis White who collected all the *Alcolapia* samples during a previous expedition in 2017, enabling this project. Both Michael Lawson and Antonia Ford also performed DNA extractions on samples utilised in this project. Thank you also to Edward Page who patiently taught me multiple laboratory techniques that I mastered for this project.

Funding for whole genome sequencing was from the British Ecological Society and National Geographic.

In memory of my mum, Patricia, who always believed in me.

Table of Contents

Chapter 1: Introduction

Adaptive radiations and speciation

Adaptive radiations generate biodiversity and occur when multiple species diverge rapidly from a common ancestor, each with distinct trait-environment correlations (Jønsson et al., 2012). Changes in abiotic or biotic factors due to extinctions, colonisation or following the evolution of key innovations can provide organisms with ecological release and the opportunity to exploit vacant niches (Stroud and Loses, 2016). Following this, different populations may be exposed to various selection pressures and over generations they may gradually become more genetically distinct to one another as certain alleles are favoured or purged by selection, gained by mutation, or lost/fixed by genetic drift. Well documented examples include Dawin's finches, *Anolis* lizards (Loses, 2011) and the Lake Victorian Cichlids, the latter of which is the most speciose clade known to date, with up to 500 species emerging in 15,000-250,000 years (Verheyen et al., 2003; Stiassny and Meyer, 1999; Johnson et al., 2000). However not all clades radiate, and some do so at much faster rates than others (Tao et al., 2021).

Several factors are linked with the ability to radiate, including dispersal ability (Tao et al., 2021) and capacity to hybridize in nature (Seehausen, 2004). Recent evidence finds that the propensity and speed of a clade to radiate is related to the recycling of ancient polymorphisms that arose in the ancestral species of extant taxa, leading to genetic and phenotypic diversity for selection to act upon (Jae Young Choi et al., 2021; Marques et al., 2019). Moreover, hybridisation between divergent lineages drove the explosive speciation of Lake Victoria cichlids as gene flow reintroduced lost alleles, introduced new variants and generated novel allelic combinations for selection to act upon (Meier et al., 2023). Hybridisation can also have beneficial impacts in by breaking down gene incompatibilities by recombination (Aguillon et al., 2022), promoting speciation.

Adaptive radiations provide us with useful models for understanding the complexity of speciation, especially in the face of gene flow as hybridisation can have different evolutionary outcomes. As most species are exposed to a combination of allopatric and sympatric conditions (Butlin et al., 2008), secondary contact often occurs between diverging species. Hybridisation is particularly common in incipient species as reproductive barriers are often weak (Lawson et al., 2023). This can cause two species to collapse into one forming hybrid swarms (Taylor et al., 2005), initiate hybrid speciation (Barrera-Guzmán et al., 2017) or it can reinforce species boundaries as hybrids suffer a fitness cost relative to conspecifics and are selected against (Butlin et al., 2008). Although gene flow can reduce genetic differences between two populations, importantly, if a hybrid is reproductively viable, adaptive genes can also cross the species boundary via horizontal gene transfer and become selected for (Burgarella et al., 2019). Patterns of adaptive introgression are prevalent across multiple radiations from beak morphology in Darwin's finches (Grant and Grant, 2015) to mimicry in *Heliconius* butterflies (Thawornwattana et al., 2023), highlighting the importance of interspecific breeding in maintaining biodiversity. Therefore, hybridisation plays an important role in transferring adaptive genes (Burgarella et al., 2019), initiating speciation and adaptive radiations (Seehausen, 2004).

Due to mutation, drift and selection, pre and postzygotic reproductive barriers can emerge over time and accumulate between separated populations, causing speciation. Speciation can evolve in allopatry, whereby populations become reproductively isolated due to non-overlapping geographical range (Butlin et al., 2008). Alternatively, new species can emerge in sympatry when populations exist in the same territory but diverge due to ecological niche partitioning, differences in pheromones, reproductive cycles, courtship behaviours and/or mechanical isolation (Butlin et al., 2008). During sympatric and allopatric speciation, prezygotic factors have been shown to play a larger role in isolating emerging lineages than postzygotic factors (Widmer, Lexer and Cozzolino, 2008) and are thought to emerge more quickly in animals (Coyne and Orr, 1989). However, post zygotic barriers are also extremely important for maintaining species boundaries and proliferate non-linearly over time (Orr, 1995).

Cichlids

Cichlidae are a species rich family of teleost fishes, with up to 1,600 known species which inhabit lakes and rivers in Africa, South and Central Americas, Sri Lanka, India and Madagascar (Stiassny and Meyer, 1999). Although contested in literature, increasing evidence suggests that the cichlid family emerged after Gondwana separated (Matschiner et al., 2020; Friedman et al., 2013). The East of Africa is exceptionally species rich for Cichlidae, with certain Tribes, such as the haplochromines being famous for explosive speciation (Elmer et al., 2009). Therefore, as this clade readily speciates, it provides a good model for understanding the origin of biodiversity.

Cichlids are famous for sexual selection and male nuptial colouration is a common mechanism for causing reproductive isolation in Lake Victoria cichlids (Seehausen et al., 1997). The presence of sexual dichromatism has been attributed to why certain clades are more speciose than others (Portik et al., 2019) as seen in cichlids (Wagner et al., 2012). In some species, it is so important for maintaining species boundaries, that a reduction in water clarity obscures visual clues that are required for assortative mating, leading to non-discriminate breeding between species (Seehausen et al., 1997).

Within cichlids from the African Great Lakes Tanganika, Victoria and Malawi, parallel evolution has emerged independently multiple times for several life history traits, including brood and egg size (Duponchelle et al., 2008). For pelagic species, fewer but larger eggs are selected for which may be a strategy to avoid predation on young fry (Duponchelle et al., 2008). Meanwhile, benthic species have evolved guarding like behaviours, enabling a greater number of smaller fry to be raised (Duponchelle et al., 2008). Parallel evolution has also occurred in cichlids from Lakes Apoyo and Xiloá in Nicaragua, Central America whereby pelagic species have convergently and independently evolved elongated bodies in comparison to benthic species which have deeper bodies (Elmer et al., 2014). Furthermore, melanic stripe patterning has convergently evolved independently within the Lake Victoria and Malawi cichlid radiations due to two different de novo mutations at the same locus (Urban et al., 2020).

Variation in diversification rates between cichlid lineages is correlated to lake depth and age but not to lake area (Wagner et al., 2012). Deeper lakes are associated with larger adaptive radiations in cichlids as different depths provide more available heterogenous niches which can facilitate ecological speciation (Wagner et al., 2012). Older lakes may also contribute to adaptive radiations in cichlids, as there is more time for mutations to arise and persist, which can then generate new species (Wagner et al., 2012).

Introducing *Alcolapia*, a small diverse species flock

The Lake Victoria radiation is often used to understand speciation in the face of gene flow as many of the sister species are at the early stages of the speciation continuum (Kocher, 2004). However, the large size of this radiation makes understanding evolutionary processes complex as there are multiple interactions between species. Meanwhile, this paper will study a smaller species flock with considerable polymorphism, ideal for understanding the emergence of biodiversity.

Situated in the Gregory Rift Valley of East Africa are three endorheic basins; Lake Natron found in northern Tanzania and nearby Lakes Magadi and Little Magadi which are in south Kenya. These lakes are inhabited by the *Oreochromis Alcolapia* species flock, also known as the ‘soda tilapias’ (Kevrekidis et al., 2020). *Alcolapia* has been reclassified as a subgenus of *Oreochromis* (Ford et al., 2015), the latter of which contains 37 species (Ford et al., 2019). There is evidence for widespread ancient gene flow across the Oreochromis phylogeny (Ciezarek et al., 2023). *Alcolapia* are a small, young radiation that exhibit considerable phenotypic diversity, most noticeably in their feeding apparatus including teeth, male nuptial colouration and craniofacial morphology (Ford et al., 2015; Seegers and Tichy, 1999; Seegers et al., 2001). This species flock consists of four species as young as 10,000 years (Williamson *et al*. 1993; Tichy and Seegers 1999; Roberts et al., 1993). *Alcolapia grahami* is the only species to inhabit the margins of the smaller Lake Magadi and its satellite Lake Little Magadi, while *Alcolapia latilabris*, *Alcolapia ndalalani* and *Alcolapia alcalica* are found at high densities inhabiting various lagoons around the perimeter of Lake Natron (Julia Day field notes, 2019; Figure 1, Ford 2015). Lake Natron species differ in their distribution with *A. alcalica* found universally around the lake, while *A. latilabris* and *A. ndalalani* are restricted to the south of Lake Natron. *A. ndalalani* has the smallest known distribution (Chapter 4, Lawson, 2023).

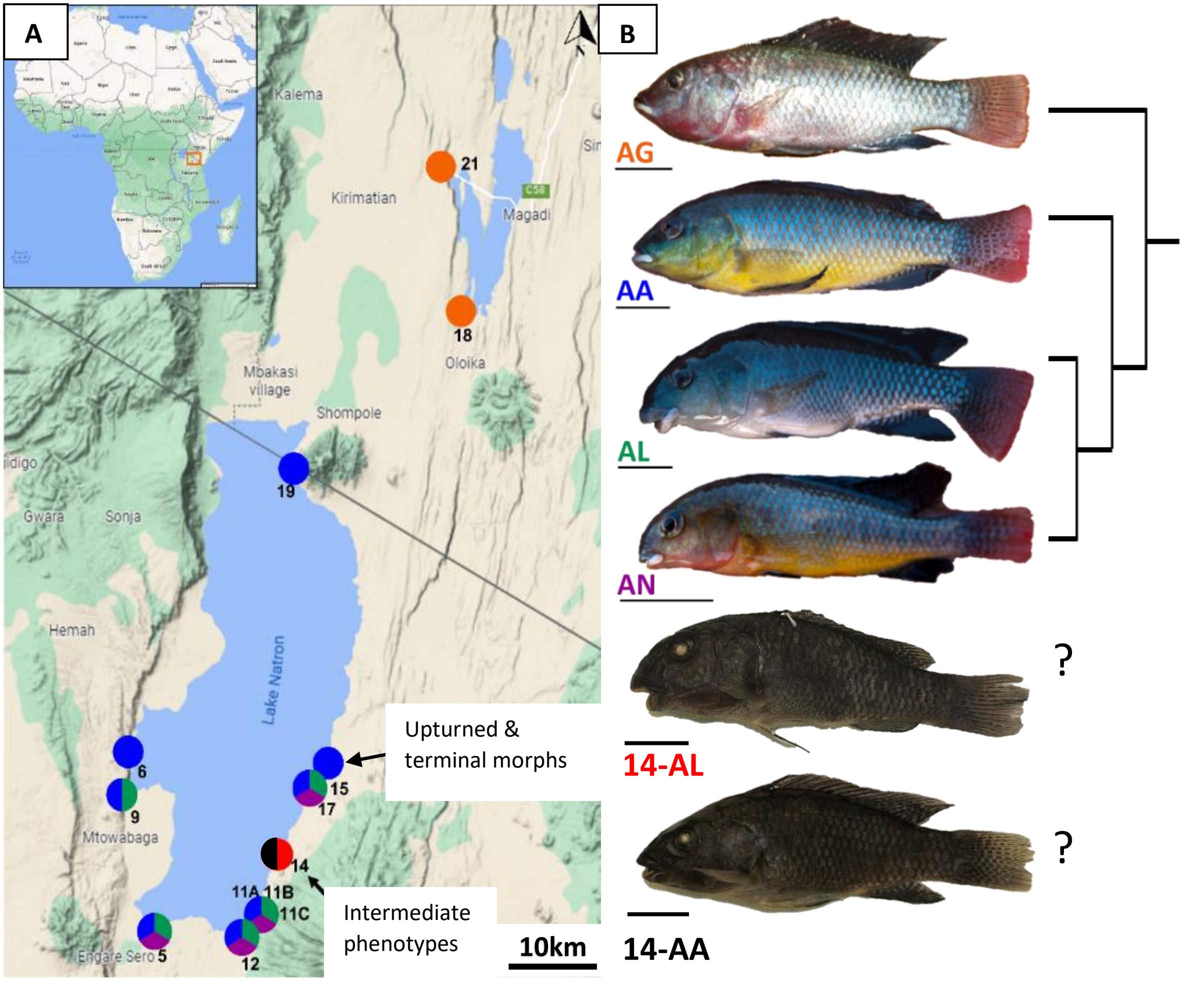


Figure 1.1: A) Geographic distributions of *Alcolapia* species. Lakes Natron and Magadi are physically separated so there is no migration between *A. grahami* and Lake Natron species. Numbers correspond to sites sampled by Ford et al., 2016 which were also used in this study. Pie charts show presence of various species/populations at each site and are not ratios**. B)** Current phylogenetic relationship of *Alcolapia* species with *A. grahami* the outgroup to Lake Natron species. The phylogenomic relationship of 14-AA and 14-AL is unknown. Black scale bars = 10mm. 14-AL and 14-AA are preserved specimens so lack pigment. Top four fish photographs taken by Ford (2015).

Alcolapia possess several unique adaptations which may be seen as "key innovations" as they enable this subgenus to inhabit some of the most hostile soda lakes in the world. The ability to tolerate highly alkaline conditions evolved once in African cichlids (Ford et al., 2019) at least 10-13Mya in the Tungen Hills, Central Kenya. This unique trait is shared between

Alcolapia species and *Oreochromis amphimelas*, the latter being the closest extant sister taxon to *Alcolapia* (Kevrekidis et al., 2020; Ford et al., 2015) which inhabits Lakes Eyas, Manyara, Singida, Kitangiri and the Yaeda basin (Bayona, 2006). During hypoxic conditions, *A. grahami* have been observed to air breathe (Narahara et al., 1996). Furthermore, to be able to feed in alkaline waters, *A. grahami* have evolved a trifurcating oesophagus and intestine that minimises alkaline water from neutralising stomach acid which is essential during digestion (Bergman et al., 2003). In addition, to survive the caustic conditions within Lakes Magadi, *A. grahami* only excretes urea (Wood et al., 1989). Similar mechanisms have evolved *in A. alcalica* (White et al., 2022). Furthermore, *A. grahami* survive vast fluctuations in temperature, dissolved O_2, salinity, $C0_2$ concentration and has the greatest ability to tolerate alkalinity fluctuations of any fish species known (Reite et al., 1974). During further experiments, *O. niloticus* died within 1 hour of being exposed to Lake Magadi conditions (Wright et al., 1990). Given that *Alcolapia* are the only fish to survive and persist in the hostile Lakes Natron and Magadi (White, 2021), this highlights the unique evolutionary mechanisms *Alcolapia* have evolved, which may have contributed to its success at radiating.

Lake geology and climate

Around 700 Ka a paleolake called Orolonga existed which subsequently split around 8-10ka forming Lakes Magadi and Natron (Tichy and Seegers, 1999; Seegers et al., 2001). This was due to major climatic shift towards more arid conditions following substantial changes in global hydrological cycles which is known as the Younger Dryas Event (Williamson et al., 1993; (Garcin et al. 2007). As East Africa received less rainfall (Williamson et al., 1993), water levels diminished in Lakes Natron and Magadi to around 1-2m deep and the conditions within the lakes intensified, evidenced by increased salinity in sediment levels 7,000 years ago (Roberts et al., 1993; Butzer et al., 1972, cited by Seegers et al., 1999). This likely increased the selective pressures on the *Alcolapia* species inhabiting these lakes. Today Lake Natron is reported to have a pH of 9-11.5, high ammonia emissions, moderate salinity, and hypoxic conditions during the night and an excess of oxygen during the day (Clarisse et al., 2019; Bergman et al., 2003; Wood et al., 1989; (Lameck, et al., 2023). Lake Magadi has similar conditions to Lake Natron (Table1, Getenet et al., 2023) with dissolved $C0_2$ of 180 mmol l^{-1}, and recorded osmolality of 525 mOsm/kg (Wood et al., 1989) and daily fluctuations in temperatures of 25-40°C (Narahara et al., 1996).

Lake Magadi has a much smaller area than Lake Natron, 108.8 km^2 (Coe, 1966) compared to 1059 km^2 (Dawson, 2008). The depth of the water in Lake Natron is 3-4m (Dawson, 2008) and both lakes are separated by a taller land barrier (Williamson et al., 1993), preventing *Alcolapia* from migrating between them. Lake Natron is fed by water from volcanic springs and streams, the latter of which are found at the edges of the lake, while Lake Magadi is solely dependent on volcanic springs and rainfall for water influx (Coe, 1966). Four rivers continuously supply Lake Natron with water, The Engare Nyiro from the north, the Peninj and Moinik rivers to the west and The Engare Sero in the southwest (Dawson, 2008). Springs are also found on the west and east of Lake Natron (Dawson, 2008), and together these water sources at the perimeters of the Lakes Natron and Magadi form pockets *Alcolapia* can inhabit. The rivers vary in the sediment in which they transport, River Engare Sero carries volcanic material while the other rivers carry (Dawson, 2008). Due to past eruptions, the chemical compositions of the sediments of Lake Natron vary (Joint Research Centre, 2013; Williamson et al., 1993), with magnesite found in the east of the lake and high natrocarbonite concentrations found towards the south (Dawson et al., 2008). Together, this may create variation in habitats and water clarity across the zones *Alcolapia* inhabit.

Given the proximity of Lake Magadi, ~20km from Lake Natron (Google Maps, 2024), it is likely both lakes experience similar patterns in precipitation and lake area changes throughout the year. Within the catchment of Lake Natron, there is seasonal variation in precipitation, with the least amount of rainfall in July, followed by a lag when lake levels are lowest in September and October (Clarisse et al., 2019). Large expanses within both lakes are frequently covered by a thick layer of trona, a mineral deposit consisting of sodium carbonatite, sodium bicarbonate and halite (Getenet et al., 2023). Certain lagoons remain during dry periods such as in the south, north and west of Lake Natron and these provide a refuge for *Alcolapia* populations (Clarisse et al., 2019). During dry periods, the lagoons become disconnected, causing habitat fragmentation (Seegers and Tichy, 1999). Migration between the lagoons likely exists following periods of flash flooding (Coe, 1966; Zaccara et al., 2014). However, after periods of extensive rainfall, higher mortality rates have also been reported which may be related to deoxygenation caused by algal blooms

(Coe, 1963, cited by Tichy and Seegers, 1999). Therefore, migration between lakes may be costly, preventing different populations from being completely panmictic.

Ecology of *Alcolapia*

Within Lakes Magadi and Natron, species richness is low as there are few species that can survive in such hostile conditions (White, 2021). In fact, *Alcolapia* are the only fish to populate these soda lakes and their diet consists predominately of cyanobacteria which is found at high abundancies (Lawson et al., 2023). By modelling the average traits of descendant species, the ancestor of *Alcolapia* is predicted to be a generalist herbivore (Ford et al., 2016). *A. alcalica* living in sympatry ingest considerably more cellulose (43%) than other *Alcolapia* species (4-5%) (Figure 2D, Ford et a., 2016) as they consume more Poales, an order of vascular plants (Figure 3.3, Lawson, 2023). *A. alcalica* have also been found to prey on more juvenile fish than any other *Alcolapia*, indicating it has a more cannibalistic nature (Figure 2D, Ford et a., 2016). Although there is similarity between the diets of the benthic species *A. ndalalani and A. latilabris* living in sympatry, the mechanism by which each species obtains their food varies, avoiding competitive exclusion (Lawson et al., 2023; Ford et al., 2016). *A. latilabris* uses its enlarged lips, exposed teeth and ability to from suction to graze on the algae covering rock surfaces while *A. ndalalani* prefers to nibble, picking at algae patches on rocks (Ford et al., 2016; Lawson, 2023). These foraging strategies are reflected in grit consumption as *A. latilabris* engulfs more and larger grit particles in its stomach than *A. ndalalani* (Figure 2D, Ford et al., 2016). The tropic web is relatively simple in Lake Natron consisting of four levels (Ford et al., 2016) but several avian predators have been reported to eat *A. grahami* including the African Great white and pink backed pelicans (Johannsson et al., 2014). The presence of predators negatively decreases speciation rate in cichlids (McGee et al., 2020).

Evolutionary history of *Alcolapia*

Alcolapia is a monophyletic clade, with allopatric *A. grahami* predicted to be the sister species to Lake Natron *Alcolapia* which are evolving in sympatry (Ford et al., 2015). The divergence of Lake Natron species is predicted to predate the formation of lake Natron

(~10,000 years) (Chapter 4, Lawson, 2023). *A. grahami* are most closely related to northern *A. alcalica* populations as they have the lowest F_{ST} (Ford et al., 2015; Ford et al., 2019) and share more alleles when compared to other populations and species (Chapter four, Lawson, 2023). F_{ST} remains low across most of the genome for *Alcolapia* species comparisons, indicating shallow divergence between taxa (Figure 4, Ford et al., 2015). Extensive gene flow within Lake Natron species (Zaccara et al., 2014) makes it harder to discern between incomplete lineage sorting and introgression (Lawson et al., 2023; Ford et al., 2015). However, it is suggested that *A. ndalalani* and *A. latilabris* are sister species based on maximum likelihood analysis (Ford et al., 2015; Chapter Four, Lawson, 2023). There is also distinctive population structure between northern and southern *A. alcalica* (Ford et al., 2015).

According to Nevado et al., (2019), species with small population sizes are more likely to radiate as alleles are fixed or lost at a faster rate due to genetic drift and selection. Frequent habitat contraction and expansion from water level change within Lakes Natron and Magadi likely place intense selection pressures and population bottlenecks on *Alcolapia* species. From demographic modelling, *A. alcalica* is predicted to have the largest known population size (Ne) of 79,000, followed by *A. grahami* 44,857 and *A. latilabris* being 24,985 (Chapter 4, Lawson, 2023). The population size estimation of *A. ndalalani* is unknown but given that is has the smallest distribution (Chapter 4, Lawson, 2023), it may have a smaller effective population size than *A. latilabris*. Furthermore, as interspecific breeding becomes more common when mating opportunities are limited (Seifert et al., 2010; Rieseberg, 1997), habitat fragmentation within Lake Natron during the dry season may encourage hybridisation and subsequent introgression. Gene flow between species is prevalent within *Alcolapia* (Ford et al., 2015) and hybridisation is a known driver of adaptive radiations in Lake Victoria cichlids (Meier et al., 2023). Therefore, the small, frequently fragmented populations of *Alcolapia* may have contributed to the relatively high speciation rate within this species flock.

Polymorphisms and aims

It is important to investigate the genetic basis of morphological polymorphisms to increase our understanding of how ecological guilds are filled and whether these adaptations are convergent, giving us greater knowledge of the mechanisms of diversification. Within *Alcolapia* several morphological polymorphisms persist with unknown function. In the East of Lake Natron exists a site (Figure 1.1A, site 15) where *A. alcalica* possess both extremely upturned and terminal mouths in the absence of population structure at this location (Ford et al., 2015). More subtle upturned phenotypes also exist in *A. grahami* from Little Magadi as opposed to terminal morphs inhabiting Lake Magadi (Figure S6 from Kavembe et al., 2016). No candidate genes have been identified with a connection to upturned phenotypes in *Alcolapia* (Kavembe et al., 2016). Likewise, there is no population structure between blue and yellow male *A. alcalica* that coexist at a location in the south of Lake Natron (Figure 2.4 A, site 11; Ford et al., 2015), which makes understanding the genetic basis of this phenotype easier with fewer samples This is as the vast majority of the genome will be highly similar between the two groups so that any genomic difference will likely be due to the polymorphisms which is especially useful when performing GWAS. If there is high population structure, then you need many samples to reduce the noise of your data to be able to identify the genomic regions which are associated with the trait of interest. Furthermore, blue-yellow colour morphs have also been found within *A. latilabris* populations (Julia Day field trip photographs, 2015). As assortative mating has been found between *Alcolapia* living in sympatry (Lawson et al., 2023), it would be interesting to know whether male colouration may affect this. Melanism has also been found within Lake Natron populations (Seegers et al., 1999 and 2001; Julia Day field notes, 2015).

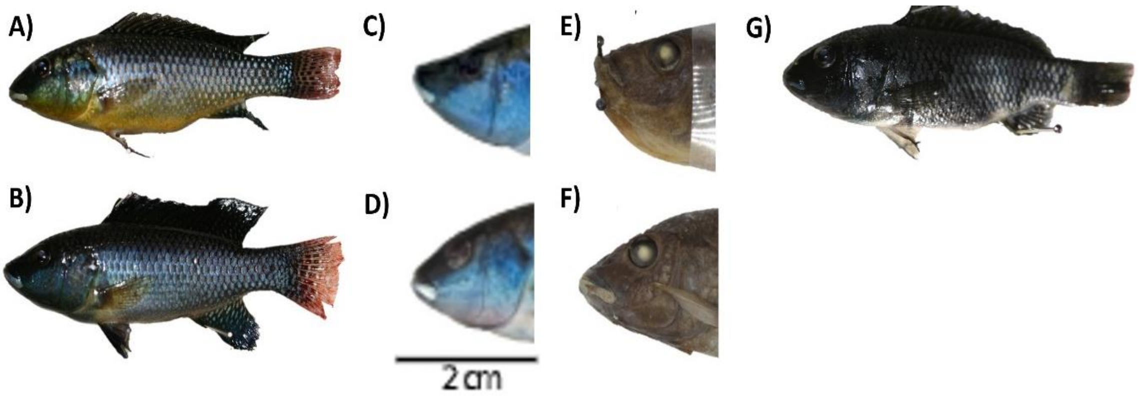


Figure 1.2: Polymorphisms within *Alcolapia*: **A)** and **B)** show male nuptial **blue/yellow colour morphs** in *A*. *alcalica* from t site 15 (see Figure 1.1A for site locations). This polymorphism has also been identified at sites 5, 11, and 12 for *A*. *latilabris* (Day and Ford field notes, 2017. **C)** and **D)** show **upturned** (Little Magadi) and **terminal** (Lake Magadi) *A*. *grahami* (edited from Kavembe et al., 2016) **E)** and **F)** show **upturned** (sampled from a pool) and **terminal** (sampled from a nearby stream within the main lagoon) *A*. *alcalica* from site 15, East of Lake Natron. **G)** Suspected hybrid from site 17, showing possible **melanism**. Apart from C and D, photos are without scales**. A, B and G photographed by Ford, 2017.**

Blue-yellow colour polymorphism has not yet been quantified to see if it is continuous. There is also an additional site in the east of Lake Natron (Figure 1.1A, site 14) which contains *Alcolapia* that are phenotypically similar to *A. alcalica* and *A. latilabris*, particularly in regard to their similar head shape morphology but with notable differences in lip thickness, colouration and mouth orientation. This may imply that they are of hybrid origin or are a population under divergence. Currently, the genetic identity of these individuals is unknown.

This thesis aims to investigate the origin of novel morphological traits within a young emerging adaptive radiation that has been shown to have considerable gene flow between species (Ford et al., 2015). Chapter two will explore the phylogenetic and evolutionary relationship of a population of *Alcolapia* inhabiting the eastern shores of Lake Natron (Figure 1.1, site 14) which exhibit unusual morphotypes most similar to *A. alcalica* or *A. latilabris*. Meanwhile Chapter three aims to investigate the genetic basis of upturned and terminal mouth orientation within *A. alcalica,* establish whether this is a discrete or continuous trait and identify regions within the genome which are related to this polymorphism. Together this study will increase our understanding of the evolutionary history of the *Alcolapia* flock, the origin of new or shared phenotypes in the presence of gene flow and the generation of novel biodiversity. Furthermore, this study will also provide greater understanding of the genetic basis of cranial morphology within vertebrates, particularly in cichlids which exhibit extraordinary diversity in this trait (Powder and Albertson, 2016).

Chapter 2: Discovering and categorising a new species of *Alcolapia* within Lake Natron.

Abstract

Adaptive radiations help us understand the processes by which biodiversity is generated. *Oreochromis Alcolapia* provides a simplified model for studying the emergence of traits and speciation in action as it is a young genus (~10,000 years) that is relatively simple, consisting of four known species. Within this species flock there is considerable phenotypic diversity in male coloration, denture morphology and head morphology. This study investigates a morphologically unusual population of *Alcolapia* from a previously sampled site (site 14) that are endemic to Lake Natron. Dental abnormalities have been reported at this site. ddRAD sequence data was obtained from sites scattered around the perimeters of Lakes Natron and Magadi. When collected, specimens at site 14 were categorised as either *A. alcalica* or *A. latilabris* due to their close morphological resemblance. However, RAxML, genetic PCAs and fineRadstructure analysis strongly contradict this theory and provide evidence for a new species of *Alcolapia*, consisting of two morphs. Maximum likelihood trees predict that this taxon from site 14 is the sister taxa to Lake Natron species with maximum support. In addition, population structure within site 14 corresponds almost exactly to morphological phenotyping, providing evidence for two morphs. Given that the site 14 population exhibits relatively high nucleotide diversity, it suggests a wider distribution of this species. Finally, a GWAS on site 14 morphs identified two SNPs ~25,000 bp from a candidate gene, *OSR2*, a transcription factor involved in dental and craniofacial development in mice. Therefore, this gene may be causing phenotypic differences within the head region between site 14 morphs. Further phenotyping of the teeth of each site 14 morph is needed and KO experiments of *OSR2* are required to confirm its role.

Introduction

Adaptive radiations are driven by ecological opportunity, which can arise when niches are vacant or become available following the extinction of competitor species and/or predators (Stroud and Losos, 2016). The emergence of biotic resources, such as plant or animal taxa, can also promote biodiversity by providing new ecological niches. During the Early Cretaceous, the arrival of angiosperms, provided new opportunities for multiple groups of organisms (especially the order Coleoptera) to exploit, radiate and co-evolve (Benton et al., 2021; Wilson et al., 2012). In turn this series of events spurred the growth of terrestrial species richness (Benton et al., 2021). Likewise, the spread of grassland habitats in Northern America is believed to have facilitated the rapid radiation and diversification in dental morphology of *Equidae* species (horses), most of which are now extinct (MacFadden, 2005). Relaxed evolutionary constraint from expanding habitats, colonisation of new territories, emergence of novel biota or absence of competitors therefore promotes phenotypic and genetic diversity for selection to act upon which in turn can generate adaptive radiations.

The evolution of new phenotypes can enable a species to exploit an existing niche in a novel way, which can lead to ecological release if there is an absence of species that perform a similar function in the ecosystem. In turn, these key innovations can initiate adaptive radiations due to the monopoly of resources (Stroud and Losos, 2016). In cichlids, a large and highly speciose family of fishes, the evolution of a second set of jaws is associated with their high diversity rates (Burress et al., 2020). The pharyngeal jaws decouple food processing and transport from the frontal set of jaws, freeing these to perform alternative tasks including food acquisition (Liem, 1973). Likewise, the evolution of elongated digits and a winged membrane has facilitated flight in *Chiroptera* (bats), allowing them to occupy a previously vacant niche, which in turn has enabled a large adaptive radiation of this order (Amador et al., 2019; Sadier et al., 2021).

The colonisation of vacant territories such as islands, newly formed mountain ranges or landlocked lakes can also promote adaptive radiations. These types of systems are invaded less often and so may lack competitors or predators. Hence, resources are readily available which facilitates ecological release (Stroud and Losos, 2016). For example, the Galápagos (and Cocos) islands were founded by an ancestral species of finch that originated from the mainland. This adaptively radiated to form 18 species, widely known as Darwin's finches (Rubin et al., 2021). Across this species flock there is high diversity in beak morphology. This allows the birds to avoid competitive exclusion as they can exploit different food resources from cracking large nuts, to piercing and drinking the blood of large sea birds (Abzhanov, 2010). Meanwhile, within cichlids, colonisation of Lake Victoria which had habitat heterogeneity has promoted explosive speciation (Ngoepe et al., 2023), with up to 16 different trophic levels reported by members of this radiation (Feller and Seehausen, 2022). Both the Galápagos islands and Lake Victoria are relatively young, estimated to be 5 Mya and Lake Victoria cichlids were void of competitor species which provided them ecological opportunity to speciate (Ngoepe et al., 2023).

It is still not clear why certain lineages are more speciose than others, but this could be linked to the ability of a species to hybridise. Interspecies hybridisation is a known driver of diversification as it increases genetic diversity within lineages through horizontal gene flow from other lineages (Meier et al., 2023; Patton et al., 2020). In cichlids, interspecific breeding has driven numerous adaptive radiations, particularly in Lake Victoria, where 500 species have emerged in the last 16,000 years (Meier et al., 2023). Ancient hybridisation between two distant lineages of cichlids generated extraordinary genetic and phenotypic variation for selection to act upon (Meier et al., 2017). In addition, the repeated coalescence and division of different cichlid lineages, further rearranged allele frequencies within different populations, contributing to their high species richness (Meier et al., 2023).

Furthermore, hybridisation can promote novel adaptations as recombination can breakdown allelic combinations which are acting antagonistically with one another, enabling different phenotypes to emerge (Seehausen, 2013). In some cases, a hybrid may even have a more extreme phenotype than either parental species, which is also known as transgressive hybridisation (Rieseberg et al., 2003) and this phenomenon is common in hybridising lineages (Rieseberg et al., 1999). Extreme negative or positive phenotypes occur due to the additive allelic effects at a particular quantitative trait locus (Rieseberg et al., 2003). If the hybrid phenotype is adaptive and the hybrids reproductively isolated from the parental species, hybrid speciation may occur.

Cichlids are famous for their high speciation rates which has resulted in numerous independent adaptive radiations, especially within the African Great Lakes. Since most cichlids readily hybridise, trait evolution and speciation events can be studied as recombination breaks down linkage. However, many of these adaptive radiations are large and intricate, which makes this process complex. Meanwhile, *Oreochromis Alcolapia*, a subgenus estimated to be as young as ~10,000 years old, is a smaller radiation consisting of four described species, each with considerable phenotypic and genetic diversity, making it a good model to understand the generation of biodiversity (Williamson *et al*. 1993; Tichy and Seegers 1999). *Alcolapia* are also contained in endorheic lakes, and so are isolated from gene flow from other cichlid genera (Ford et al., 2015). Therefore, high phenotypic diversity within *Alcolapia*, its simple size and insular nature make it an excellent model for studying the origin of biodiversity.

Despite its young age, *Alcolapia* have evolved distinctive eco-morphological traits. Each species occupies different environmental niches and have distinguished head (Ford et al., 2016) and teeth morphology (Seegers et al., 2001). *A. grahami* has an upturned mouth which may be association with surface predation, *A. alcalica* is terminal in appearance, which may enable it to be more of a generalist feeder, *A. ndalalani* has a subterminal mouth, with small jaws adapted to pick at food in the benthic region, while *A. latilabris* has a distinct subterminal mouth with large protruding teeth enabling it to scrape algae from rocks (Seegers et al., 2001; Lawson, 2023; Figure 1.1B). Additionally, a population of *A. alcalica* have both upturned and terminal mouth morphologies (Figure 1.2E, F). There is also high diversity in male colouration, with *A. latilabris* and *A. alcalica* possessing blue-yellow colour polymorphism which may be important in sexual selection, while *A. grahami* has distinct red colouration. Furthermore, melanism has been reported in some *A. alcalica* populations (Seegers et al., 1999; Seegers et al., 2001; Julia Day field notes, 2015).

Alcolapia are endemic to two soda Lakes in East Africa. *Alcolapia grahami* inhabit Lake Magadi and Little Magadi in south Kenya, while *Alcolapia latilabris*, *Alcolapia alcalica* and *Alcolapia ndalalani* live sympatrically in Lake Natron, northern Tanzania (Figure 1.1A). *A*. *alcalica* has the widest distribution, with distinct northern and southern populations, while *A*. *latilabris* and *A*. *ndalalani* are confined to the south of Lake Natron (Figure 1.1A). *Alcolapia* are known to inhabit the lake perimeters, near springs or streams which support permanent lagoons within the lakes where conditions are more favourable (Figure 2.2). Lakes Magadi and Natron frequently crystalise, forming a thick crust of trona which isolates several distinct lagoons (Figure 2.2) and the areas of the Lakes change throughout the year, coinciding with seasonal change (Clarisse et al., 2019; Google Earth Pro, Historical Maps, 1995, 2016 and 2017). When water levels are high following rainfall, lagoons become connected, which may facilitate movement between previously isolated populations within Lake Natron, Little Magadi and Lake Magadi (Google Earth Pro, Historical Maps, 1995, 2016 and 2017; Zaccara et al., 2014). There is no migration between each lake as these are separation by permanent land barriers. *A*. *grahami* are therefore currently evolving from Lake Natron species in allopatry while *Alcolapia* within Lake Natron exist in sympatry and may be subject to periodic isolation due to trona deposits. Currently it is believed that *A. ndalalani* is sister species to *A. latilabris*, while *A. grahami* diverged prior to the Lake Natron species (Ford et al., 2015).

Reproductive isolation is incomplete between *Alcolapia* species as they hybridise but there is evidence for weak assortative mating (Lawson et al., 2023; Castric et al., 2008). Extensive introgression has been documented within *Alcolapia* (Ford et al., 2015), which may in turn contribute to its high species richness by providing genetic diversity for selection.

Lakes Natron and Magadi were originally joined as one paleolake, called Olorongo (Figure 2.2) which separated ~10-11,000 years ago following major climatic change during the Younger Dryas (Roberts et al., 1993). During this period, water levels fell from 60m to 1-2m, separating the lakes (Roberts et al., 1993). Around 7Ka, the conditions within both lakes became more extreme as the salts and other compounds concentrated (Roberts et al., 1993), putting greater selective pressures on *Alcolapia* evolution. Currently, *Alcolapia* which inhabit Lakes Natron and Magadi are subject to similar extreme environmental conditions (albeit variation in mineral concentrations) with recorded temperatures of 25-43°C (Narahara et al., 1996; Wood et al., 2016), moderate salinity (Lameck et al., 2023), pH of 9-11.5, vast diurnal oxygen fluctuations and high ammonia deposits, making the lakes hostile to most organisms (Clarisse et al., 2019; Getenet et al., 2023; Getenet et al., 2023).

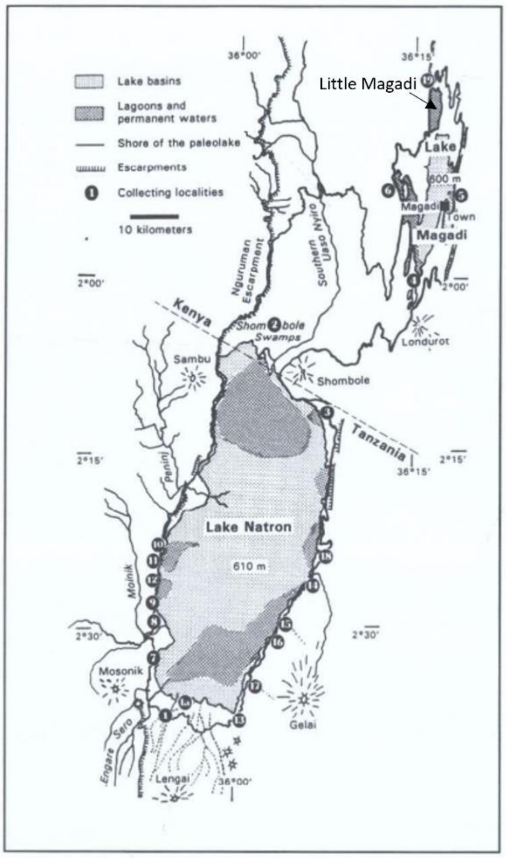


Figure 2.2: Annotated map from Seegers et al., (2001) highlighting the location of the permanent lagoons within Lakes Natron, Magadi and Little Magadi which support *Alcolapia* populations. Bold line indicates perimeter of the paleolake Orolonga.

In 2001, Seegers et al., reported that they had observed a new morphotype of *Alcolapia* inhabiting a bay supplied by streams at the foothills of the Gelai Mountain in the east of Lake Natron. Individuals from this population appeared to be intermediate of *A. alcalica* and *A. latilabris* but they did not believe them to be of hybrid origin as *A. latilabris* was not found in situ. They refer to these individuals as *A. sp. aff. latilabris*, found at a site they call TZ 99/27, with the coordinates 2°31'53.0"S 36°02'34.0"E (Seegers et al., 2001). The authors describe the teeth of *A. sp. aff. latilabris* as being less curved than *A. latilabris* and report that it has distinctive dentition. They also deduced that this newly described morphotype has a similar functional role to the *A. latilabris,* when *A. alcalica* and *A. ndalalani* are also present. This infers that this morph is also a bottom feeder, like *A. latilabris*.

In 2012, J. Day, A. Ford and G. Turner visited a site 600m away from TZ 99/27 and collected a series of specimens that Day and Ford phenotyped as either *A. alcalica* or *A. latilabris* due to their close phenotypic resemblance and referred to the site as site 14. Ford (2015) described 14-AA to have thicker lips than expected of *A. alcalica*, while 14-AL have less downturned mouth orientation than *A. latilabris*. Site 14 also have darker colouration than other *Alcolapia* species (Ford, 2015), although quantification of this phenotype has not yet been performed. As all three Lake Natron species have been documented at neighbouring locations to site 14 (Ford et al., 2015; Figure 1.1A), it is plausible the unusual phenotypes could be of hybrid origin (Ford, 2015).

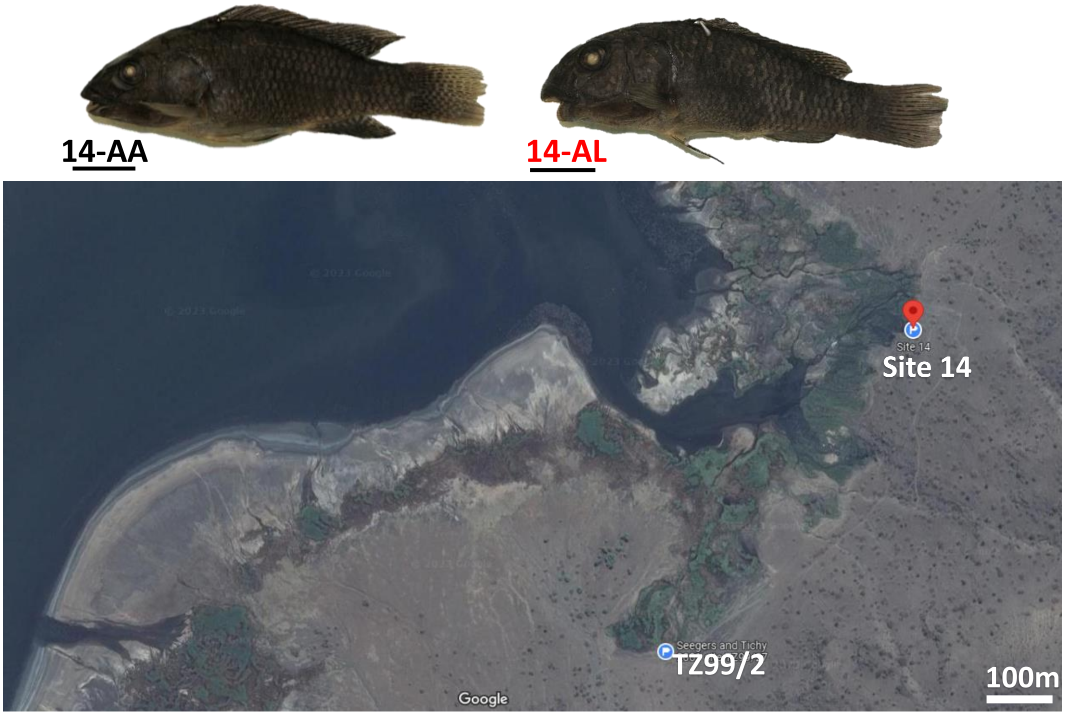


Figure 2.3: Site 14, situated in an inlet in the east of Lake Natron, is 600m from TZ99/27 where unusual *Alcolapia* morphs were found previously by Seegers et al., 2001 (Google Earth images, 2023).

Little is known about the genetics and ecology of fish inhabiting this location and so remains to be investigated to improve our understanding of the formation of adaptive radiations.

The aims of this chapter are to:

1) Investigate how population 14 is related to the *Alcolapia* flock.
2) Quantify morphological features within population 14.
3) Investigate the genetic basis of unusual mouth morphologies
4) Test for introgression between population 14 and other *Alcolapia* species.

Methods:

Samples were collected by Ford and Day, while Ford completed the DNA extractions.

In 2012*, Alcolapia* were collected from 14 sites from the perimeters of Lakes Natron, Magadi and Nakuru (see Figure 1.1A and Table S1)*. A. grahami* samples were introduced to Lake Nakuru during the 1960s (Vareschi, 1979). During the trip in 2012, specimens were collected from a location referred to as site 14 which is just over 600m northeast from the site TZ 99/27 described by Seegers et al., in 2001. Ford and Day observed considerable phenotypic variation in morphology at site 14 and classified the samples at this site as either *A. alcalica* or *A. latilabris* by lip thickness as the phenotypes strongly resembled either group. They noted that this site was more marshy than other locations.

DNA was extracted from 2-3mm of the pectoral fins of 136 samples with Qiagen DNeasy Blood & Tissue Kit. The resulting DNA was sent to CD Genomics for SbfI digested RAD-library preparation followed by 150bp Illumina pared-end sequencing. Two RAD datasets were generated from these fish collected: one containing nearby sites 11C and 14 which had higher coverage (x23, Ford and Day unpublished dataset) and another with lower coverage (x6.12, Lawson and Dasmahapatra, unpublished dataset) which contained sites: 5, 6, 9, 11A, 11B, 12, 15, 17, 18, 19, 21 and 24. The second larger batch was sequenced by Mike Lawson during his PhD*.* Morphometric analysis was conducted to identify if there were substantial differences or similarities in body shape between site 14 and established species. This could then be used to predict the evolutionary history of site 14 specimens. Previously samples from site 14 were morphologically intermediate of site 9 (see Figure 6.14 of Ford, 2015) *A. alcalica* and *A. latilabris* but *A. ndalalani* was not included in this comparison. Therefore, site 11C was chosen to compare morphometrics with site 14 as this was the closest nearby site which also contained all three Lake Natron species coinhabiting (Ford, 2015). Following photographing of 147 specimens form the same side, tpsUtil v1.58 converted the images into the correct format for tpsDig2 v2.17 (Rohlf, 2013) to enable landmarking to commence (Ferdous and Armbruster, n.d.). 16 whole-body landmarks were used (shown in Figure S1B) and converted into coordinates within TpsDig2 v2.17. Following this, MorphoJ v1.05f (Klingenberg, 2011) was used to generate a covariance matrix from the TPS input file which is a prerequisite to run Procrustes superimposition (Ferdous and

Armbruster, n.d.). This function scale images, controlling for allometry between samples, enabling comparison between images even when specimens are rotated (Rohlf and Slice, 1990). Principle component scores were then produced in Morpho J v1.05f.. For further details of this analysis please see pages 197-199 from Ford published in 2015.

Author's own:

Morphometric analysis

A morphometric PCA was plotted in R studio (R-4.3.2) using PC values produced by Ford from whole body landmarking Eigenvalues were calculated manually to see how much each principal component was contributing to the variation in morpho space. Ellipses were added to represent 95% confidence intervals for each population.

Bioinformatics pipeline for RAD sequencing

Oreochromis niloticus, a close relative of *Alcolapia* species (Ford et al., 2019), was used as the reference genome (NCBI Assembly GCF_001858045.2) following Ford et al., 2015, and this was indexed using BWA index v.0.7.17 (Li, 2013). Each forward fastq file was then aligned to the reference genome using BWA-MEM v.0.7.17 (Li, 2013) and reads were sorted using samtools v.1.17 (Li et al., 2009). Picardtools v.2.25.5 (Picard Toolkit, 2019) was used to identify PCR duplicates, and the HaplotypeCaller and GenotypeGVF functions from GATK v.4.4.0.0 were used to find variants (Van der Auwera and O'Connor 2020).The following filters were applied using bcftools (v.1.15.1; Li, 2011): SNPs only; DP = 5<100; MAF = 5%; MIS = 50%; QUAL = 20; MIS = 0.5; MAF = 0.05. Some individuals were manually removed as they of insufficient coverage. Following these filters, 38910 SNPs remained, and 113 individuals were used in the genetic PCA and fineRADstructure analysis.

Average coverage per sample was calculated using SAMtools depth (Li and Bonfield, 2023) on bam files. To combat the batch effect as there were two datasets of different coverage, the higher (mean of x23.0) and lower (mean of x6.12) datasets were intersected with BCFtools isec (v. 1.15; Danecek et al., 2021). This ensured only SNPs present in both datasets were compared.

Plink v.2 (Chen et al., 2019) was used to perform linkage pruning on the variant call format file (VCF) to remove non-independent variants as this would confound further analysis (Ravinet and Meier, 2022). A window slide of 10 SNPs and a linkage disequilibrium threshold of 0.05 were chosen based off recommendations by Ravinet and Meier (2022). Meanwhile, a window of 100 SNPs was chosen as the best fit after experimenting with different window sizes as it was not too small, which makes the data highly stochastic, while ensuring it was not too big which would lose resolution within each window, flattening potential signals in the VCF file. Plink also was used to produce a genetic PCA to investigate whether population structure exists between sites and species. The eigenvectors were plotted in R v.4.1.2 (Posit team, 2023) using ggplot2 (Wickham, 2016) and the percentage variance was calculated manually from eigenvalues.

Kanchon Dasmahapatra performed fineRADstructure analysis using the same VCF input file used in the genomic PCA (Figure 2.4B)to detect for population structure between site 14 samples when compared to other populations within Lakes Natron and Magadi. Dasmahapatra followed Malinsky's tutorial (n.d.) to calculate a coancestry matrix which is generated by comparing the pairwise relationship of multiple RAD samples (Malinsky et al., 2018). FineRadstructure then assigns clusters based upon the genetic similarity between samples and so acts independently of known population information (Malinsky, n.d.; Kanchon Dasmahapatra, personal communication, 2023). FineRadstructure also performs Bayesian clustering to predict a tree representing the shared genetic similarity between populations (Malinsky and Trucchi, 2020; Lawson, 2012). Although this tree can infer the evolutionary relationship between populations, it does not discriminate between introgression and incomplete lineage sorting so it must be treated with caution (Malinsky and Trucchi, 2020).Output heatmaps representing coancestry and relatedness between populations were plotted in R studio (R-4.3.2) using scripts from Lawson (2012).

Phylogeny on *Alcolapia* nuclear genomes

A nuclear phylogeny was generated using RAxML v.8 (Stamatakis, 2014) to investigate the evolutionary relationships between *Alcolapia* species. Due to computational demands, a subset of three-four individuals of each species/morph with the highest coverage RAD data (above x23.0 from bam files) was used. This included *A. alcalica* (site 11C), *A. latilabris (*site

11C), *A. ndalalani (*site 11C), site 14-AA, site 14-AL and *A. grahami*. *Alcolapia* sequences used were from the same batch. One RAD *O. amphimelas* individuals, sequenced by Lawson (2023), and a whole genome *O. niloticus* sample (accession number: ERR7224546) was downloaded from NCBI were also included as outgroups. After following the bioinformatics pipeline previously detailed, BCFtools view and merge were used to subset and combine VCFs containing the outgroup and the ingroups to form one VCF (v.1.15; Danecek et al., 2021). BCFtools filter was used to filter the output VCF to a GQ of 30, DP5 and SNPs only, MIS of 0.75 and a MAF of 0.05. 89983 SNPs remained after filtering. vcf2phylip was used to convert the VCF to phylip format (Oritz, 2019). RAxML uses maximum likelihood methods to calculate the best phylogeny, and this was calculated with 100 bootstrap values. The Newick file was then visualised using MEGA v.11 (Tamura et al., 2021).

mtDNA phylogeny of *Alcolapia*

As mtDNA and nuDNA evolve independently and at different rates, they may have different evolutionary trajectories (Vawter and Brown, 1986) so separate phylogenetic analysis were performed on each DNA type (Figure 2.6). mtDNA analysis was also used to search for patterns of historical hybridisation events (Rubinoff and Holland, 2005) as considerable nuclear-mtDNA discordance has been identified in *Oreochromis* (Ford et al., 2019) highlighting the role of past hybridisation events in this clade.

Qiagen Blood and Tissue DNA extraction kits were used to extract the DNA from two individuals to from each described species of *A. alcalica* (site 11C), *A. ndalalani* (site 11C), *A. latilabris* (site 11C), site 14-AA, *A. grahami*. Three mtDNA genes were investigated using two primer pairs for each gene or fragment, amplified using PCR and sent externally for Sanger sequencing by Source Bioscience. Genes NADH dehydrogenase 2 (ND2), DLoop, and 16S were chosen as they have been used in previous evolutionary studies on *Alcolapia* and African cichlids (Ford et al., 2019; Klett and Meyer, 2002). In particular, the DLoop has been shown to have considerable diversity within *Alcolapia*, with 18 haplotypes reported (Seegers et al., 2001). Additionally, there were some complete mtDNA sequences on NCBI (*A. grahami*, *O. esculentus* and *O. niloticus)* that were included in this analysis, increasing the sample within *Alcolapia* and serving as outgroups. The mtDNA sequences were inspected for ambiguous base calls and trimmed in Mega v.11 (Tamura et al., 2021), aligned by ClustalW (Thompson et al., 1994) and concatenated into a single fasta file as mtDNA is a non-recombining molecule. Maximum likelihood trees with 1,000 bs were calculated on this concatenated file in mega v.11 (Tamura et al., 2021).

Sliding windows analysis (F_{ST}) and nucleotide diversity (π)

Using SNP data, F_{ST} was performed using the pipeline from Martin (2023b) to identify regions of divergence across the genome between *Alcolapia* species that may be under selection (Martin, 2023b). The input VCF file contained 15 x 14-AL, 15 x 14-AA, 7 x site 11C *A. alcalica* and 6 x *A. latilabris* and this was filtered to a MAF of 0.05, MIS of 0.5, SNPs only, DP5-100 and a GQ of 20. 107,614 SNPs Remained after filtering. Each window contained 100 SNPs per window, overlapping by 50 SNPs. These parameters were chosen due to a trade-off between window and slide size, too small a window and noise can obscure meaningful data, while too large a window may flatten a SNP peak, losing important information. Manhattan plots were made in R v.4.1.2 to visualise results (Posit team, 2023).

This protocol was repeated with 50 SNPs overlapping by 0 SNPs so that summary histograms of F_{ST} and π across the genome could be plotted in R. It was important not to have an overlap so that each window was independent. *A. ndalalani* was not included in this analysis as 14-AL and 14-AA phenotypes most closely resembled *A. latilabris and A. alcalica* respectively, therefore it was believed that the same genes may be under selection in 14-AA and *A. alcalica* vs 14-AL and *A. latilabris*, giving rise to the similar traits in each species pair.

Quantifying morphological variation in the head region of site 14 morphs

Given that the morphometric PCA (Figure 2.4A), produced from the whole-body landmarks taken by Ford (Figure S1B), did not appear to capture variation in morphology well between 14-AL and 14-AA, linear measurements were taken to capture this variation between site 14 morphs which was most prominent in the head region. As 14-AA and 14-AL were categorised in the field based on lip thickness by Ford and Day this phenotype was investigated. It can also be observed that the upper lip of 14-AL frequently overlaps the lower lip, creating an underbite, while 14-AAs jaw are more equal in length. 14-AL also had subterminal mouths, while 14-AA was terminal. Therefore, to capture this variation, three traits were measured between site 14 morphs lip: thickness, lip protrusion and the head depth (Figure S2).

As certain key landmarks, especially the upper lip perimeter, which is vital for measuring lip thickness, were not always clear in previous specimen photographs taken by Ford, fish were pinned and re-photographed. Coordinates were then generated by landmarking photographs in ImageJ and the equation of a straight line was used to accurately calculate linear distances. Measurements were then standardised for allometry by dividing by the standard length of each fish. The scatterplot3d function in R was used to simultaneously visualise all three measurements. Two-tailed T-tests were run on each measurement to test if there was a significant difference between the means of 14-AA and 14-AL.

GWAS

GWAS was used to identify regions that may contain potential candidate genes associated with the four phenotypic traits being lip thickness, head depth, mouth protrusion and Ford and Day's discrete classification of either 14-AA or 14-AL. Multiple GWAS analysis were performed using Chehida's script (2024) which involved generating pairwise relatedness to find out the genetic similarity between two samples. The input RAD VCF contained 30 site 14 samples of both morphs, 15 x 14-AA and 15 x 14-AL, which were filtered with BCFtools (v.1.15.1; Li, 2011) to a GQ of 20, a minimum DP5, a maximum DP100, SNPs, MAF of 0.05 and a MIS of 0.5. A Bed file was generated using pLink2 (Chen et al., 2019). Filtering parameters were chosen based off recommended levels (Ravinet and Meier, undated). Too little filtering and false positives are introduced, meanwhile too stringent filtering and information is lost. Generally, a GQ score of 20 and above is trusted as reliable (Ravinet & Meier, undated). After filtering, 109216 SNPs remained, and this was used to calculate the Bonferroni threshold of 6.34. Scaffolds were arranged by size using python (Biopython,2022). It was necessary to produce a separate phenotype file for each GWAS analysis with integers used to indicate discrete phenotypes such as 14-AA or 14-AL and numbers with decimals used to designate continuous variables such as the head depth measurements. The first GWAS conducted was on the 14-AA/14-AL phenotype. A Manhattan plot was generated in R v.4.1.2 (Posit team, 2023). A zoomed in Manhattan plot of linkage group11 was produced in R to show this region in more detail where a peak was found.

Three more GWAS analysis were performed using the same input vcf.gz but filtered to a GQ of 30 to test the robustness of the results and with different phenotypes files reflecting the traits that had been measured previously; lip thickness, lip protrusion and head depth (108741 SNPs). The number of SNPs was slightly less for these three GWAS as two samples did not have measurements for these phenotypes (14-AA had 13 samples while 14-AL had 15). Therefore, in the phenotype file any samples that did not have a phenotype recorded were assigned a value of -9 which indicates to GEMMA that there is missing data.

SNP genotypes frequencies were retrieved by using grep for the significant SNP locations LG11: 6309402 and 6406194 on a geno file that had sites 2, 4, 5, 6, 9, 12, 14, 15, 17, 18, 19, 21, 24, 11A, 11B and 11C (same as the genetic PCA, Figure 2.4B). These genotype frequencies were superimposed on a map to see their distribution relative to location and species.

f_{dM} analysis

f_{dM} was used to investigate evidence for introgression at the significant SNPs found by the 14-AA/14-AL GWAS at positions LG11:6309402 and 6406194 (gibert-Fab, 2022; Ravinet and Meier, 2024; Malinsky et al., 2015). Given that there are morphological similarities between 14-AA vs *A*. *alcalica* and 14-AL and *A*. *latilabris,* it was possible that these SNPs arose from horizontal gene flow, giving rise to parallel evolution.

Scripts from Github were used to compute f_{dM} (Martin, 2024) by changing the *fd* flag to f_{dM}. f_{dM} is advantageous over the *fd* technique as *fd* can only compare between two populations at once while f_{dM} can compare multiple populations simultaneously (Martin, 2024).To test f_{dM} within the genome , the input VCF contained RAD data from site 7x 11C *A. alcalica*, 6 x site 11C *A. latilabris*, 15 x 14-AA, 15 x 14-AL and three *O. niloticus* samples downloaded from NCBI as outgroups (accession numbers: ERR7224546_1, ERR7224552_1 and ERR7224569_1). SNPs remaining after filtering = 630,350. Input VCF was filtered to a GQ30, SNPs only, DP5, DP100 and a MIS of 0.5. There were 100 SNPs per window overlapping by 0 SNPs. SNPs remaining after filtering = 630,350.

Zoomed in Manhattan plots were created in R studio (R-4.3.2) of 100,000 bp around significant SNPs identified on LG11.

Dsuite

While f_{dM} can indicate patterns of introgression for SNP positions across the genome, Dsuite can generate summary values across the genome for populations. In turn this can reveal larger genome-wide trends in introgression between multiple species (Kanchon Dasmahapatra, personal communication, 2024). Dsuite is a programme used to measure introgression between species by outputting D-statistics and the f_4-ratio across the genome (Malinsky et al., 2021). Scripts on GitHub were followed for this analysis (Malinsky, 2024).

The input VCF contained a total of 57 samples: 9 *O. niloticus* samples downloaded from NCBI to act as outgroups, 15 x 14-AA, 15 x 14-AL, 7 x 11C *A. alcalica*, 5 x 11C *A. ndalalani* and 6 x 11C A. *latilabris.* The input VCF was filtered to A GQ30, DP5, DP100, SNPs only, MIS 0.5 and a MAF of 0.05. DtriosParallel was used to calculate f4 and D statistics for all population comparisons, K (the number of jackknife blocks) was set to 100 and a Newick tree was used as an input file which was based upon previous nuclear phylogeny predictions (Figure 2.6A). Site 14 morphs were grouped together and separately to see if there were different patterns of introgression between the morphs.

Results:

Morphometric and Genetic Diversity

A PCA was generated using whole body landmarks assigned by Ford (2015) of Lake Natron species from site 11C and site 14 morphs to see how site 14 related to Lake Natron species (Figure 2.4A). Considerable overlap was seen between site 14 morphs and all three Lake Natron species, suggesting that they could be of hybrid origin. 14-AL clustered more strongly with *A. latilabris* as expected as it appears more morphologically similar, with downturned mouths, as shown by ellipses. Likewise, 14-AA overlaps more with *A. alcalica* and these two groups share more morphological similarities, including more conical, terminal shaped heads.

Considering this, the genetic similarity between *Alcolapia* populations was investigated to identify if site 14 was genetically of hybrid origin. Surprisingly, site 14 clusters distinctly from all other *Alcolapia* species and there appears to be population structuring at site 14 between 14-AL and 14-AA morphs (Figure 2.4B). Lack of overlap between site 14 and other *Alcolapia* species in PC space, suggests that reproductive barriers may be stronger than between *A. alcalica, A. grahami*, *A. latilabris* and *A. ndalanai*, although this may be due to sampling bias.

14-AA samples were closer to *A. alcalica* on PCA 2 than 14-AL, which possibly indicates more recent gene flow between 14-AA and *A. alcalica* (Figure 2.4B). There is also an outlier of 14-AA that is situated in between 14-AA and *A. alcalica*, indicating recent gene flow between these two groups (Figure 2.4B). This individual was inspected to see if it had a high level of missing data, but it was no different to other site 14 individuals and so this signal is likely true.

The batch effect was attempted to be controlled for by intersecting the data of the different coverage RAD data. Given that the same pattern occurs following this and different levels of filtering for MAF and MIS filters, this indicates that the data is robust and site 14 is indeed clustering independently of the other *Alcolapia* species (Figure 2.4B).

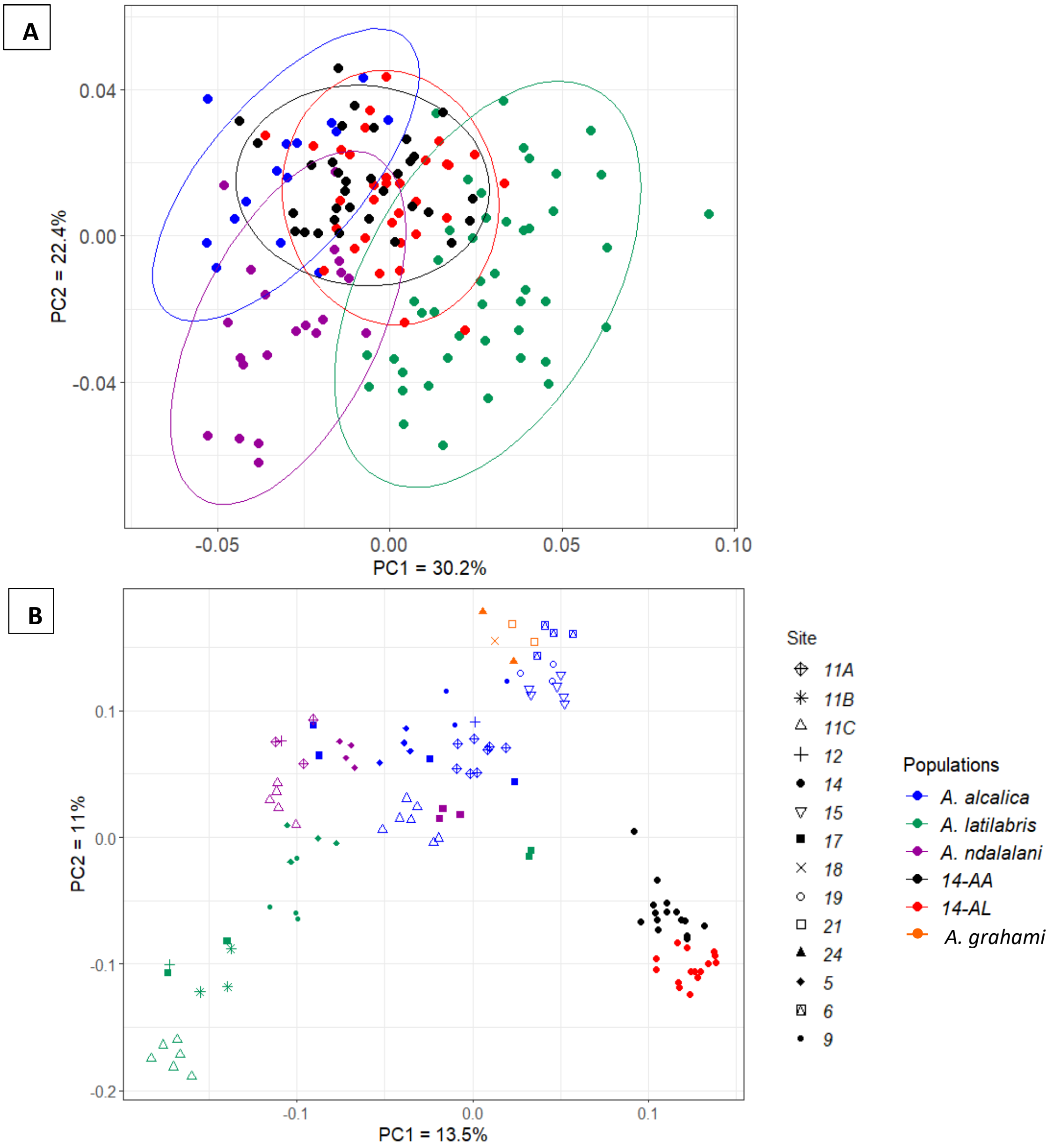


Figure 2.4: A) Morphological PCA finds little distinction between 14-AA and 14-AL and there is some overlap between *A. alcalica*, *A. ndalalani* and *A. latilabris*. Ellipses represent a 95% confidence interval. **B)** Site 14 individuals are genetically distinct from all other *Alcolapia* species. Two individuals from site 17 may be misidentified as *A. latilabris* and instead could site 14 – *A. alcalica* or *A. ndalalani* hybrids. RAD SNPs used = 38,910.

To investigate the genetic relationship between different *Alcolapia* populations further, Dasmahapatra performed fineRADstructure analysis (Figure 2.5). This found greater genetic similarity between site 14 individuals than other species comparisons.

FineRADstructure analysis divided site 14 individuals into two clusters which correspond almost exactly, bar one sample (S014_685_AD) to Ford and Days's phenotypic categorisation of 14-AA or 14-AL (Figure 2.5). S014_685_AD was identified morphologically as 14-AA by Ford and Day but groups with 14-AL in the fineRADstructure analysis. Sub clustering at site 14 indicates a genetic basis in observed traits, supporting evidence that two morphs exist at site 14. The phylogeny here must be treated with caution as the fineRADstructure analysis assumes no gene flow between lineages, so it cannot distinguish between recent introgression and incomplete lineage sorting (Malinsky and Trucchi, 2020; Dasmahapatra personal communication, 2023).

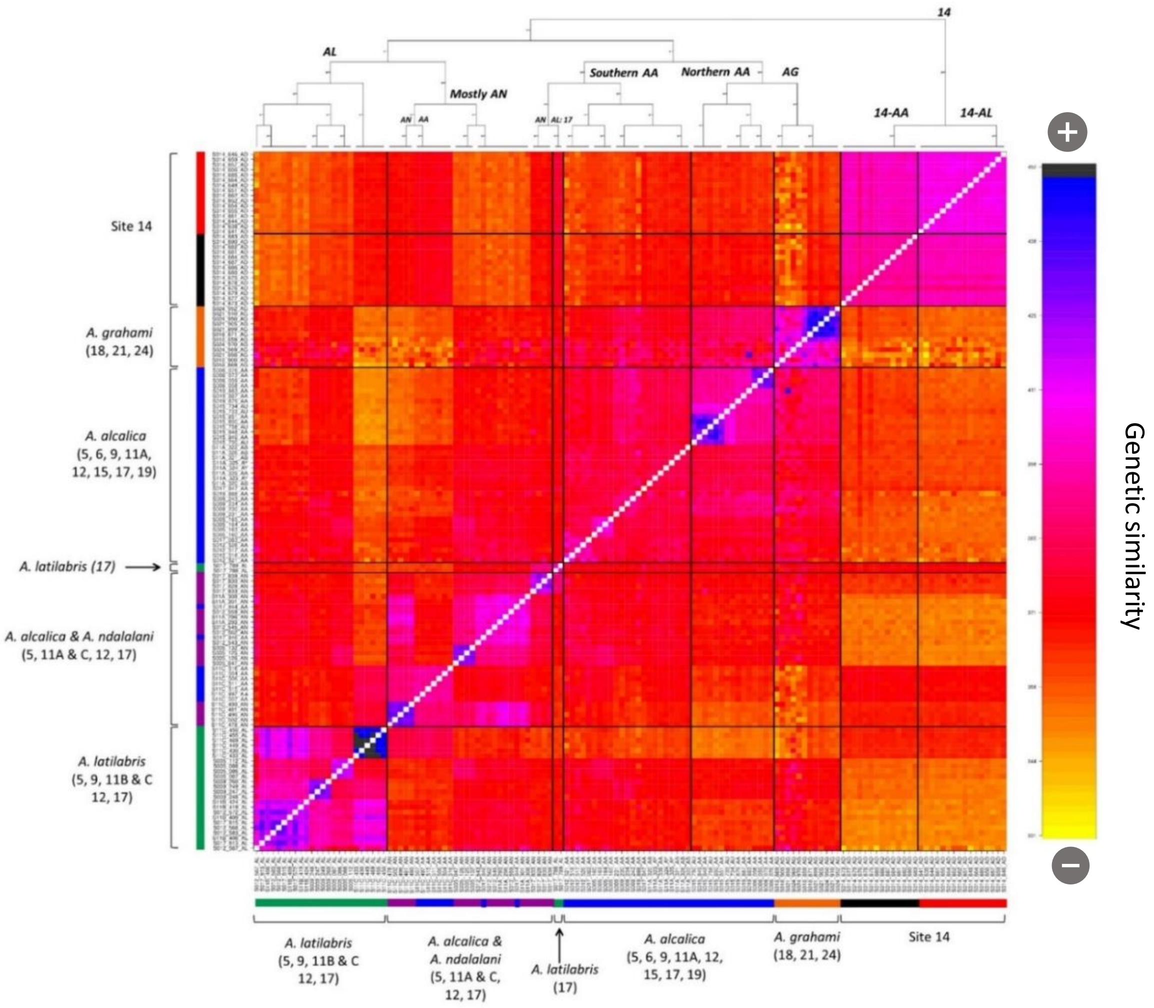


Figure 2.5: FineRADstructure analysis shows a lack of genetic similarity between site 14 and *Alcolapia* sites and species. Sub clustering is seen within site 14. Phylogeny predicts site 14 to be the sister taxon to Lake Natron and Magadi *Alcolapia* species*.* Individuals are shown on each axis, with a colour bar indicating species assigned by morphology. Black lines used for visualising groups by species/sites. Numbers in brackets are the collection sites (see Figure 1.1). AA, AN, AL and AG are abbreviations for *Alcolapia* species. RAD SNPs used = 38,910.

Phylogenetic relationship of site 14

In view of this, it was necessary to predict the true evolutionary history of the different species within Lake Natron. RAxML predicted site 14 to have diverged from the *A. alcalica*/ *A. ndalalani*/ *A. latilabris* ancestor and that 14-AL forms a distinct clade within site 14

(Figure 2.6A). Little sequence variation was found between the three concatenated mtDNA genes, reflecting their shallow genomic divergence and possible hybridisation (Figure 2.6B). The mtDNA phylogeny remained almost identical regardless of whether Nas (which represent a failed base call) in the trace remained or were filtered out (Figure 2.6B).

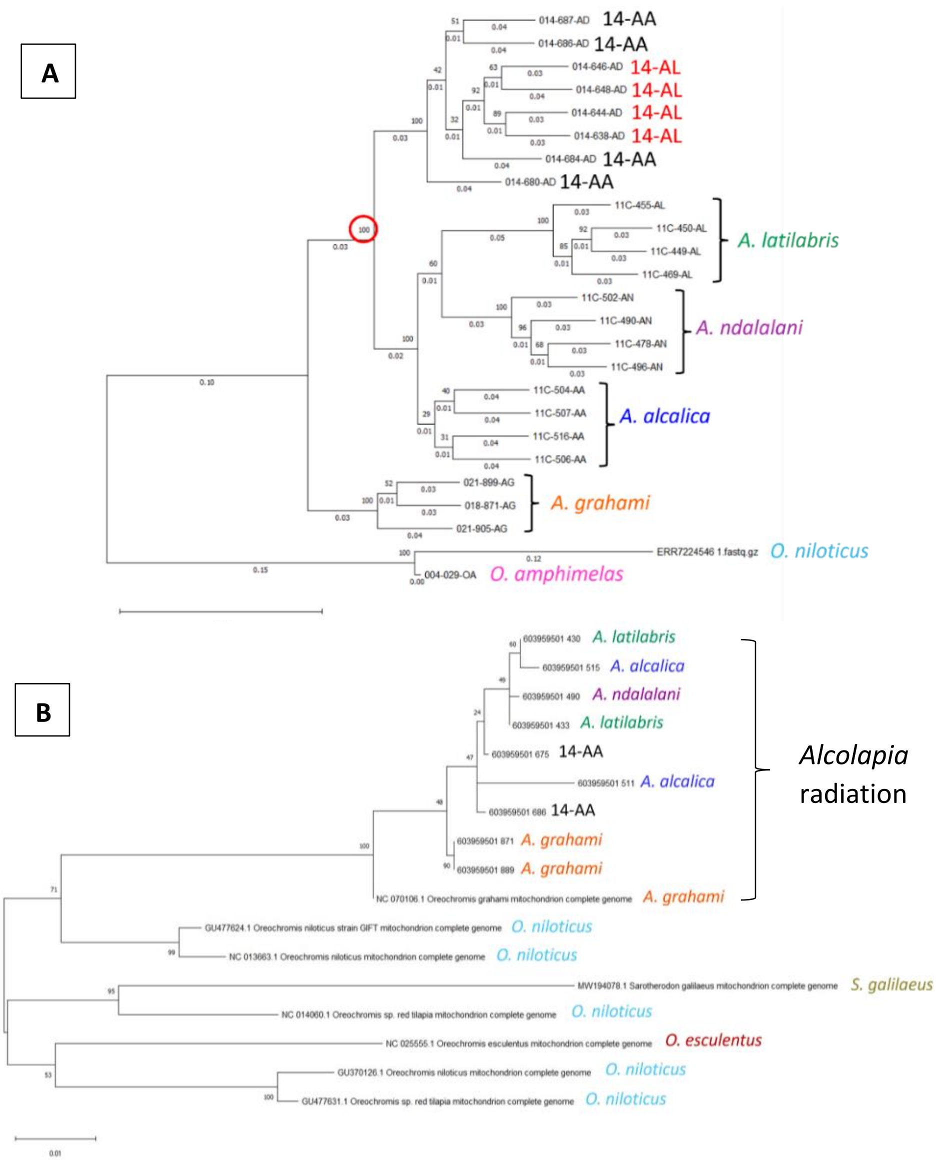


Figure 2.6: A) nuDNA ML phylogeny (83,983 SNPs) of the *Alcolapia* flock produced using RAxML identifies population 14 having diverged from other Lake Natron species with 100% bootstrap support (bs) (highlighted with a red circle). Site 14 14-AL clusters strongly as a clade (92 bs), while the relationship of 14-AA is paraphyletic, signalling more variability. Sub clustering of 14-AL provides evidence for two populations at site 14. Scale refers to % sequence divergence. **B)** The mtDNA ML tree supports the Natron species as monophyletic but does not resolve internal relationships. Total alignment size was 2,258 bases.

Quantifying phenotypic traits in the head region

As both the genetic PCA (Figure 2.4b) and the FineRADstructure analysis (Figure 2.5) found sub clustering at site 14, corresponding to field groupings, it seemed likely whole body landmarking was not capturing this variation well (Figure 2.4A). Consequently, several linear measurements capturing head depth, lip thickness and overbite were recorded to try to better capture this variation between the two site 14 morphs. Using scatterplot3D, linear measurements revealed 14-AA, the terminal morph, had thinner lips, a smaller overbite and narrower head depths, while 14-AL has a greater overbite, reflecting its more downturned mouth orientation, thicker prominent lips and a deeper head depth (Figure 2.7), incongruence with Ford (2015). It was also observed that both 14-AA and 14-AL appeared to have more of an overbite than the other Lake Natron *Alcolapia* species (Figure 1.1B). Site 14 morphs therefore segregate better than previously when whole body landmarks were used to capture morphological variation (Figure 2.4A and 2.7).

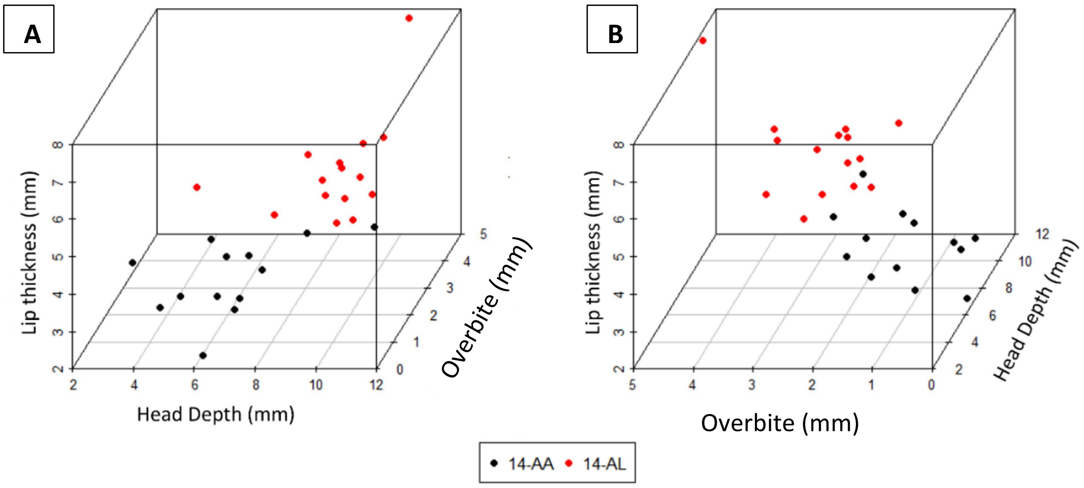


Figure 2.7: Site 14 morphs segregate in 3D space for the three measured traits. 14-AA has a smaller mean head depth (6.59mm) than 14-AL (9.07mm) (t = -4.04, df = 26, *p* = 4.19e-04). 14-AA has a smaller mean overbite (1.09mm) compared to 14-AL (2.83mm) than 14-AL (t = -5.29, df = 25.2, *p* = 1.74e-05). 14-AA had considerably thinner mean lip thickness (3.67mm) than 14-AL (5.23mm) (t = -5.61, df = 24.1, *p* = 8.77e-06). **A** = plot angle of 60, **B** = plot angle of 245.

Investigating divergence between site 14 and *Alcolapia* species

Given that there appears to be two populations at site 14 (14-AL and 14-AA) and these appear to resemble *A. alcalica* and *A. latilabris* in morphology (Figure 2.4A), it seemed feasible that they may also share genes with these species via adaptive introgression. It was therefore necessary to perform sliding windows F_{ST} to identify if an inversion was present that could be isolating site 14 from the rest of Lake Natron species and identify regions of divergence. No inversions were detected between site 14 morphs and Lake Natron species (Figure S3) and F_{ST} remained low across the genome (~>0.1) between *Alcolapia* comparisons but was elevated on LG7 for all comparisons except 14-AA/14-AL (Figure 2.8A and S3). Nucleotide diversity across the genome was also investigated to better understand the genetic diversity between *Alcolapia* populations. Surprisingly site 14 morphs had greater nucleotide diversity (Figure 2.8B), suggesting a larger population size or more hybridisation.

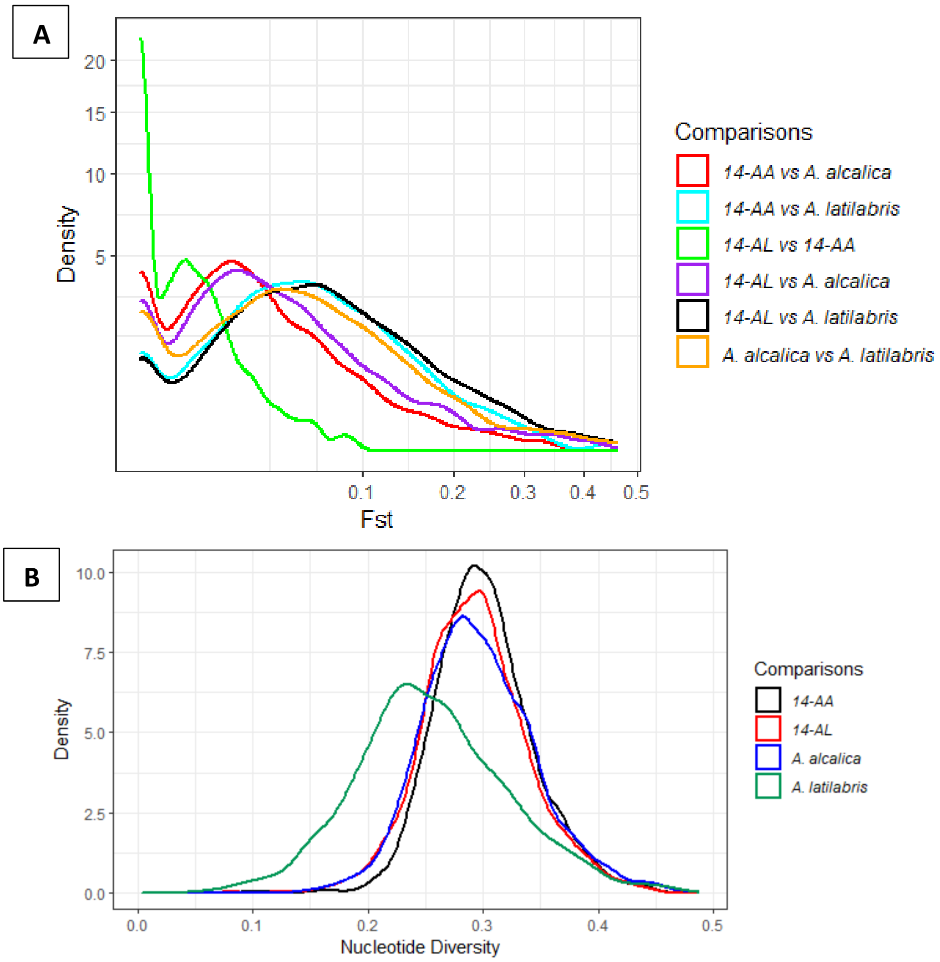


Figure 2.8: A) Density plot of sliding windows F_{ST} comparisons with 50 SNPs per window with an overlap of 0. When 14-AL and 14-AA is compared with *A. latilabris* (shown in black and blue), they have the greatest mean F_{ST} values, (0.0695 and 0.0616) indicating 14-AA and 14-AL are the most divergent from *A. latilabris*. Meanwhile 14-AL and 14-AA (shown in green) have the lowest mean F_{ST} values (0.0386) showing that they are the most closely related species across the genome**. B)** Nucleotide diversity is highest for 14-AA across the genome, followed by 14-AL. *A. latilabris* has the lowest nucleotide diversity. 50 SNP per window, overlapping by 0 with a minor allele frequency filter of 0.05, SNP data only and a missing filter of 0.5.

Investigating genomic regions associated with site 14 morphs

GWAS studies were then performed on the three measured continuous traits and Ford and Day's discrete categorisation of 14-AA/14-AL to see if genomic regions were associated with each phenotype. Only the 14-AA/14-AL category found significant SNPs (Figure 2.9 and S4). Two SNPs were found substantially above the Bonferroni threshold, so are unlikely to be false positives (Figure 2.9 A and B).

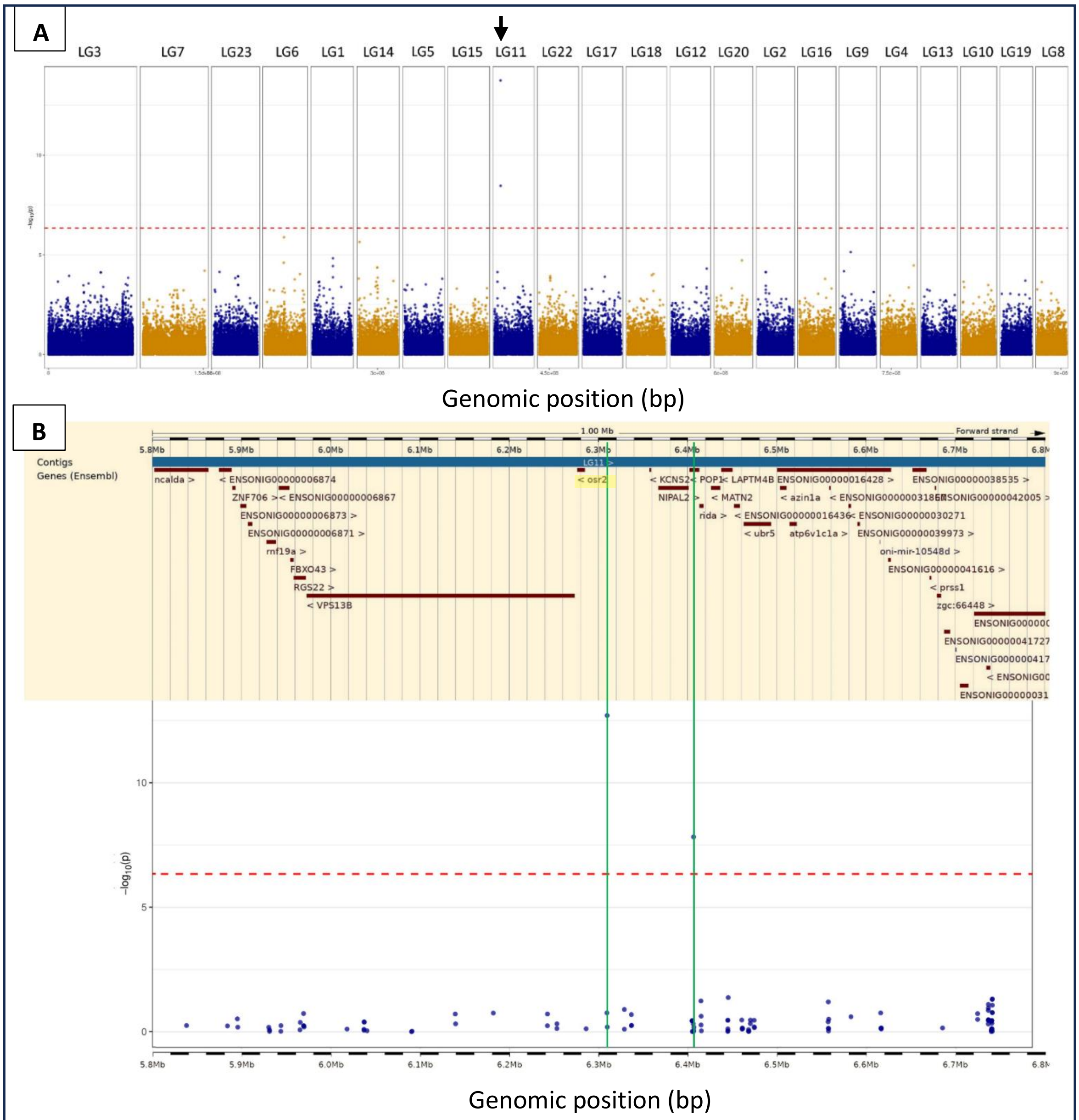


Figure 2.9: A) GWAS study on RAD data identifies a genomic peak between 14-AA and 14-AL phenotypic groups at positions LG11:6,309,402 (p = 2.013 x10^{-13}) and LG11:6,406,194 (p = 1.48 x10^{-8}) both shown with an arrow. Two SNPs are also close to the Bonferroni threshold on LG6:19,838,905 and LG6:19,949,202 which may indicate a second peak (p = 4.074236e-06 and 1.618583e-05). Filtered to GQ20. LGs are arranged by size. **B)** GWAS showing the 1MB region around significant SNPs on LG11 (shown with green lines). No strong, clear peaks beneath significant SNPs. Yellow highlights candidate gene *OSR2* ~0.025MB from most significant SNP. Built by combining author's Manhattan plot with image from the region LG11:6,309,402-6406,194 (Ensemble, 2024a) to show gene locations.

Following discovery of two SNPs differences in the region between LG11:6309402-6406194, an *O. niloticus* genome browser was used to search for candidate genes (Ensemble, 2024a). LG11:6,309,402 is located in a non-coding region while LG:6406194 is located in gene *POP1* homolog which codes for a ribonuclease P/MRP subunit (ensemble, 2024a). Just upstream of this region at position 6,276,375-6,284,679 is the *OSR2* gene which affects cartilage development in fish (Lam et al., 2013) and its homolog affects dental and cranial development in mice (Lan et al., 2001). Other nearby genes which were inspected but they did not have a clear function related to morphology.

Furthermore, for the lip protrusion GWAS, the same positions 6,309,402 (p = 2.19 x10-5) and 6,406,194 (p= 4.79x10-5) as in Figure 2.9 had lower p values, albeit not above the threshold. This indicates that the lip protrusion phenotype is associated with the SNP differences that are picked up between 14-AA and 14-AL comparisons. Therefore, it could be that the SNP is co-segregating with the alleles that cause the variation in the overbite of the lips in fish at site 14.

The genotypes of *Alcolapia* species were then investigated to see if there were patterns in their distribution at SNP locations LG11: 6309402 and 6406194. A binomial distribution exists between the genotypes of site 14 morphs with almost all of 14-AL being heterozygous for both SNPs (Figure 2.10). Interestingly these SNPs had a localised distribution being only found in 14-AL and a minority of *A. alcalica* from site 11C, suggesting it may have arisen in these populations. All 14-AA samples were homozygous at both SNP locations.

The same 14-AL individual is homozygous in each comparison. The same site 11C individual in heterozygous for both SNP positions, while the other site 11C heterozygous individual has a NA called at 6309402. Therefore, it potentially could also be heterozygous at both SNP locations, especially given that they are near one another so are likely linked.

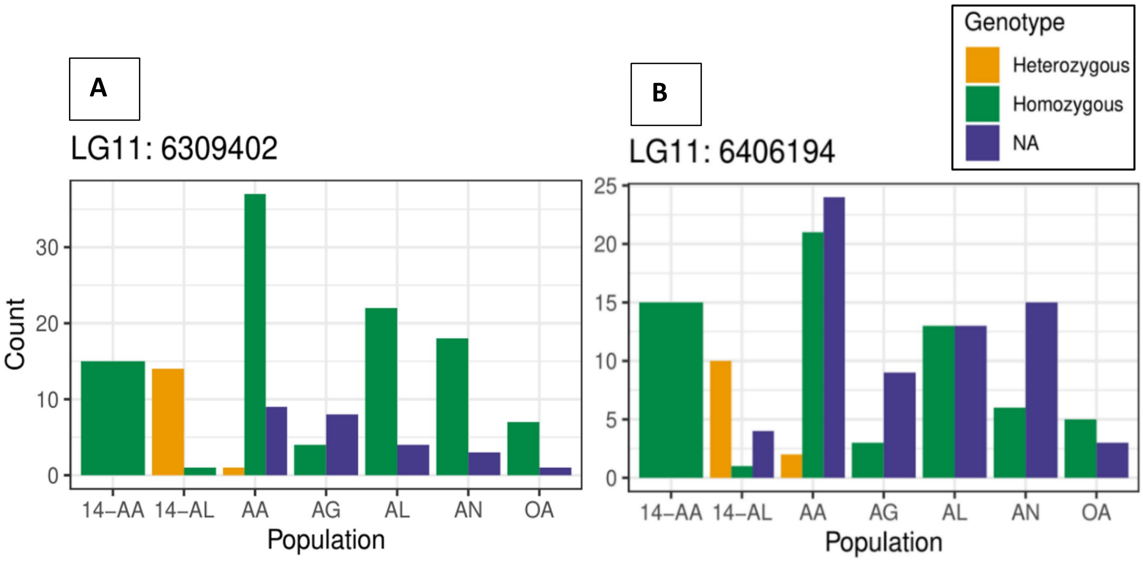


Figure 2.10: Genotype frequencies of each taxa at the significant SNPs identified by the GWAS (Figure 2.9) reveals localised distribution of alternative allele in the southeast of Lake Natron and a binomial distribution between site 14 morphs. *O. amphimelas* (OA) acts as an outgroup. **A)** At LG11:6,309,402, 14/15 individuals from 14-AL and one site 11C *A. alcalica* are heterozygous for T/G (yellow), while all other populations are homozygous for T (green) at LG11:6309402. **B)** Similarly, at LG11:6406194, 10/11 14-AL and two site 11C *A. alcalica* individuals are heterozygous for A/C while all other populations and one 14-AL individual are homozygous for A.

Gene flow is asymmetrical between *Alcolapia* species

Dsuite statistics were performed to test for the occurrence, magnitude, and direction of introgression across the genome between the Lake Natron species and site 14 morphs. This identified elevated gene flow between both site 14 morphs and *A. alcalica* and *A. ndalalani* compared with *A. latilabris* and more gene flow between 14-AL and Lake Natron species than 14-AA (Table 1).

Table 1: Dsuite results indicate that site 14 shares up to 10% more genes with *A. ndalalani* and *A. alcalica* than *A. latilabris* from nearby site 11C. There is no difference in the amount of genes that both site 14 morphs share with *A. alcalica* or *A. ndalalani.* Up to 3% more genes are shared between 14-AL and *A. alcalica*, *A. latilabris* and *A. ndalalani* when compared with 14-AA.

P1	P2	P3	Dstatistic	Z-score	p-value	f4-ratio	BBAA	ABBA	BABA
14-AA	14-AL	AA	0.005895	2.66625	0.00767	0.036818	5529.51	4308.2	4257.71
14-AA	14-AL	AL	0.013311	5.24341	1.58E-07	0.034022	5604.49	4247.6	4136
14-AA	14-AL	AN	0.006957	2.83668	0.004558	0.018991	5502.7	4275.66	4216.58
AA	AN	AL	0.007783	1.88773	0.059062	0.02385	4754.87	4483.53	4414.28
14	AL	AA	0.012317	3.01372	0.002581	0.086847	4478.04	4762.15	4646.26
AN	AA	14	0.003355	0.904314	0.365829	0.020233	4943.72	4478.51	4448.56
AL	AN	14	0.014689	3.2161	0.001299	0.081775	4766.4	4545.12	4413.52
AL	AA	14	0.018437	4.63693	3.54E-06	0.102735	4762.15	4646.26	4478.04

Given that there was phenotypic resemblance between site 14 morphs and Lake Natron species, f_{dM} statistics were performed to test for introgression. However, there was no evidence for introgression at either SNP location on LG11 as shown by the low f_{dM} values (Figure S5).

Main Findings

Linear distances, genetic PCA, fineRADstructure, RAxML and Dsuite results predict site 14 to be genetically distinct from all other *Alcolapia* species and to have diverged from the Lake Natron ancestor. Genetic and morphological sub clustering identifies two morphs at site 14 which corresponds to their field categorisation by Day and Ford. Gene flow is not equal between site 14 and *Alcolapia* species, with up to 10% more gene flow existing between site 14 and *A. alcalica and A. ndalalani* compared to *A. latilabris*. Interestingly, 3% more gene flow was found between 14-AL and A. alcalica/A. ndalalani then 14-AA which is incongruent with the proximity of genetic clusters for site 14 morphs relative to *Alcolapia* species on PC space (Table 1; Figure 2.4B). Introgression suggests speciation is occurring in the face of gene flow supporting previous findings (Ford et al., 2015; Lawson et al., 2023). GWAS analysis was run on 14-AL and 14-AA morphological groupings, which identified two candidate genes related to craniofacial and dental development, ~0.2 mb from significant

SNPs, so may be related to the different phenotypes of site 14. The heterozygous alleles have a local distribution in 14-AL and a minority of *A. alcalica* populations but not in the outgroup taxa, suggesting that these mutations arose here rather than from incomplete lineage sorting. Reassessment of the size and integrity of the *Alcolapia* flock is therefore needed based on these finding.

Discussion

Adaptive radiations are exceptionally useful for increasing our understanding of speciation and the genetic basis of ecologically relevant traits. This study reveals another lineage of *Alcolapia*, site 14, presenting further support of speciation in the face of gene flow. Previously it was suggested that site 14 may be of hybrid origin (Ford, 2015), however there is little evidence to support this as site 14 is predicted to be sister taxa to Lake Natron species, there is little overlap in genetic PCA space between site 14 and Alcolapia and there is a lack of shared coancestry and clustering between site 14 and Alcolapia (Figure 2.5). However, elevated nucleotide diversity across site 14 genomes may be indicative of hybridisation or alternatively, due to larger population size of this lineage.

Previous research found 14-AL to be significantly smaller than 14-AA with a mean standard length of 37.1mm compared to 50.7mm (Ford, 2015).

The lack of genetic distinction between species on the PCA space and low F_{ST} (mode ~0.05, range 0-~0.46) between *Alcolapia* comparisons further supports how this species flock are extremely young but it questions the integrity of the taxonomic status of Lake Natron and Magadi species. Despite this, each species is phenotypically distinct within this species flock with evidence for fine scale ecological niche partitioning (Ford et al., 2016; Lawson, 2023). Given that site 14 appears more genetically distinct than *A. grahami* (Figures 2.4B and 2.5) the predicted outgroup, this provides good support for site 14 being a species. Some taxonomists use morphology, meristics and colour patterning to classify species as seen in eight newly described cichlids species from Lake Malawi (Pauers, 2016; Pauers and Phiri, 2023). Quantifying the phenotypic attributes and producing lectotypes are therefore required to better describe and distinguish site 14 and its two morphs. Given that prezygotic mating isolation exists between *Alcolapia* species (Lawson et al., 2023), there likely to be assortative mating between site 14 and Lake Natron species, being strongest between site 14 and *A. latilabris* due to lower rates of gene flow (Table 1).

These findings are congruent with Ford et al., (2015) which identifies extensive introgression within *Alcolapia* species. However, this study contradicts Lawson (2023) who predicts *A. grahami* and *A. alcalica* to be sister species but supports Ford et al., 2015 who model *A. alcalica* as the sister species to *A. ndalalani and A. latilabris*. Perhaps this lack of cohesion between studies was due to incomplete knowledge of the *Alcolapia* species flock.

As Figure 2.6A predicts that population 14 diverged from Lake Natron taxa subsequently to *A. grahami*, therefore population 14 cannot be older than 7Ka-1.5Mya which is the estimated divergence time of *A. grahami* from the Lake Natron taxa (Ford et al., 2019). It is not possible to gain an accurate estimation of the age of site 14 species using molecular clocks without a fossilised ancestor of site 14 - *A. alcalica/A. latilabris/A. ndalalani*. In contrary, there have been several fossils found in the Tungan Hills dated 10-13 Mya in Kenya, north of Lake Natron, which are identified as being morphologically closest to extant *Alcolapia* and show adaptions to soda lake conditions (Kevrekidis et al., 2020). This suggest that tolerance to alkaline conditions originated much older than the emergence of the *Alcolapia* clade, ~10,000 years ago (Williamson *et al*. 1993; Tichy and Seegers 1999; Roberts et al., 1993). Therefore genetic analysis between these fossils, closely related *Oreochromis species* and the revised *Alcolapia* flock (including site 14) is required to identify if this is indeed an ancestor or close relative of *Alcolapia* Given the date of these fossils (10-13 Mya, Kevrekidis et al., 2020) it is unlikely to be useful for calibrating molecular clocks as they are older than the *Alcolapia* flock (~10,000 years old, Williamson *et al*. 1993; Tichy and Seegers 1999; Roberts et al., 1993). Therefore, further searching for younger fossils around Lake Natron could elucidate the ages of Lake Natron species.

Given that body shape predicts ecological niche in cichlids (Clabaut et al., 2007) it would be interesting to investigate whether site 14 morphs fill different ecological guilds. 14-AL may be more of a benthic feeder, able to grasp algae from rocks like *A. latilabris*, while 14-AA may be better suited to catching and chasing prey, like *A. alcalica* with its more conical shaped head. Head width is different between site 14 morphs (Ford, 2015) and smaller widths have been observed with the ability to feed in between cervices in *A. latilabris*, (Lawson, 2023), so perhaps this has a similar function in 14-AL. Furthermore, in some cichlid species jaw protrusion is important for creating a vacuum aiding in engulfing prey

(Wainwright et al., 2001). Perhaps as 14-AL has a large lip protrusion, it feeds via this same mechanism, but this remains to be investigated.

Consequently, one question that arose is whether parallel evolution occurs within *Alcolapia* as both *A. latilabris* and 14-AL possess enlarged lips (Figure 1.1B; Ford, 2015). This trait has been shown to convergently evolve in Lake Malawi cichlids (Masonick et al., 2022). However, sliding windows F_{ST} did not reveal any clear genomic regions that were shared exclusively between *Alcolapia* comparisons (Figure S3). Furthermore, there was little evidence for introgression at LG11:6309402 and 6406194 between 14-AA, 14-AL and *A. alcalica* or *A. latilabris*, shown by low f_{dM} values at this region (Figure S5). However, given that this is RAD data, there are large gaps in the genomic sequence, so repeated analysis with whole genome data of each morph would be required to rule out parallel and collateral evolution.

Given that GWAS analysis identified SNP LG11:6406194 to be in a protein coding gene called *POP1* (ensemble 2024a; Figure 2.9B), in situ hybridisation analysis of RNAs produced by this gene would be informative to establish the exact cellular and tissue location of gene transcripts produced by *POP1* (Chu et al., 2019). This would help to elucidate the functional role of *POP1*, regarding the tissues and cells that it is expressed for site 14 morphs. Following this, RNAseq analysis could be used to quantify any differences in protein expression of *POP1* between site 14 morphs and *Alcolapia* species (Wang et al., 2009). This same procedure could be repeated for nearby genes *OSR2* to confirm or reject its role in cranial and dental morphology within site 14 individuals. Given that the other SNP, LG11:6309402 is located in a non-coding region, its direct impact may have to be assessed using alternative methods such as snpSTARRseq. This technique identifies the functional role of non-coding regions to identify whether they affect transcription (Tunc Morova et al., 2022) as it is possible SNP LG:6309402 may have cis or trans effects on other genes such as nearby *OSR2*.

As the two SNPs identified at LG11:6309402 and 6406194 were the only significant peaks identified (Figure 2.9), they may be having a large role in the phenotypic differences observed between 14-AA and 14-AL. It would therefore be useful to identify the heritability of each SNP to find out how much each contributes to phenotypic variation (Zhu and Zhou, 2020). The dominance of these SNPs is also unknown and could be achieved by breeding heterozygous 14-AL individuals, genotyping and trait measuring.

The genes surrounding the significant SNPs on LG11 were investigated to see if they had any relatable function to the different morphs identified at site 14. Most genes, such as *NIPAL2*, a magnesium ion transporter (Goytain et al, 2008), have specific cellular or molecular function, so it is difficult to see how they are linked to the morphological differences seen between site 14 morphs. However, a gene called *OSR2* (Odd-Skipped Related Transcription Factor 2), located ~25,000bp away from the region LG11:6309402-6406194 (Figure 2.10), is a promising gene candidate. The homolog of *OSR2* is necessary for normal dental and craniofacial development in mice (Lan et al., 2001). Lan et al., (2001) found *OSR2* to be expressed in the maxillary process and mandibular which are facial regions where differences in structure have been noticed between site 14 *Alcolapia* morphs. Furthermore, *OSR2* is involved in dental organogenesis (Jia et al., 2016) and controls the number of rows of teeth in mice (Zhang et al., 2009). When *OSR2* is knocked out, mice show prolific tooth development (Zhang et al., 2009). This is especially interesting as Seegers, Sonnenburg and Tichy (2001) found *Alcolapia* morphs that look like 14-AL just 600m away from site 14 and describes them to have “unique dentition” and less curved teeth than *A. latilabris*. Therefore, it would be interesting to reinspect the size, quantity, shape, and number of rows of the teeth between site 14 morphs, perform GWAS on these traits, followed by reverse genetics by KO of *OSR2* should differences be found.

By contrast, in a study involving *Danio rerio* (zebrafish) embryos, a closer relative to *Alcolapia*, *OSR2* was found to be crucial for fin chondrogenesis development and is expressed in the pronephric (early kidney organ) (Lam et al., 2013). However, full development of zebrafish teeth normally occurs by 26 days (Van der Heyden & Huysseune, 2000), therefore it may have been too premature to observe *OSR2* gene expression in the mouth and head region during the study by Lam et al., (2013) as they present results up to 55 hours post fertilization. Given that *OSR2* is pleiotropic and highly conserved (Lam et al., 2013; Zhang et al., 2009; Jia et al., 2016), this gene could still be having a role in dental formation and craniofacial development in site 14 morphs and remains the most probably gene candidate.

Smaller studies often have less power for detecting polygenic traits as each loci have a small effect on the trait of interest (Beavis, 1995). Furthermore, additional whole genome sequencing may reveal additional peaks along the genome that might also be under selection and supporting SNPs that have been removed by filtering. In *Amphilophus*, a genus of cichlid, body shape and tooth size are controlled by 5-7 small effect loci (Kautt et al., 2020). It is therefore not surprising that no significant SNPs were identified in the GWAS analysis when the phenotypes were either lip protrusion or head depth. Meanwhile lip thickness is believed to be oligogenic with a SNP on CHR8 contributing 77% of the phenotypic variation between Midas cichlids (Kautt et al., 2020). Interestingly, when the phenotype was lip protrusion, p-values were lower (albeit significant) on CHR11 position 6309402 and 6406194 (Figure S4), which are the same SNPs that were identified as being significant on the 14-AL/14-AA GWAS (Figure 2.9 and 12). This indicates that the lip protrusion trait is somewhat capturing the genetic variation found between site 14 morphs. Further phenotyping of fixed skeletal landmarks by X-raying fish may also reveal craniofacial structure better than measuring from soft tissue that may have some unreliability. Landmarks taken from different viewpoints other than the lateral view are also required to capture additional variation.

F_{ST} was very low across most of the genome for species comparisons (Figure 2.8A), which reflects the youth of this radiation and extensive hybridisation between taxa. Site 14-AA and 14-AL are more closely related to *A. alcalica* than *A. latilabris* as shown by the sliding windows F_{ST} density plot (Figure 2.8). This is supported by the genetic PCA (Figure 2.4B), whereby population 14 are closest to *A*. *alcalica* on PC1 (13.5% of the variance). This idea is also reinforced as there are individuals that cluster in between *A. alcalica/A. ndalalani* and population 14 in genetic PCA space, implying gene flow between these species (two site 17 individuals originally classified as morphologically *A. latilabris,* Figure 2.4B). These site 17 samples are unlikely F1 hybrids as they are closer to *A. alcalica* and *A*. *ndalalani* on PCA space than site 14. These same site 17 individuals cluster with *A. alcalica* on the Fine RAD plot but show more relatedness to site 14 individuals than any other interspecific pairwise comparison (Figure 2.5). This hints that population 14 has more northern distribution and warrants future investigation and sampling to the north and south of site 14 in the east of Lake Natron.

No inversion was detected by the sliding windows F_{ST} plot (Figure S3). However as RAD techniques were used, only a small proportion of the genome was sampled (Miller et al., 2007), so it's possible that an inversion (or other SNPs and indels) could still be present that has been undetected in this analysis as the sliding windows could be too large, masking the effect of an inversion. Furthermore, as *Alcolapia* is a very young clade, even if an inversion exists within population 14, it may be difficult to detect as there has been limited time for divergence to emerge, illustrated by relatively low F_{ST} across the *Alcolapia* genomes comparisons (Figure 2.8A and S3). Given that inversions are important in reproductively isolating species (Wellenreuther and Bernatchez, 2018), further whole genome sequencing is needed to fully rule out this possibility.

There is a lack of alignment between nuDNA and mtDNA phylogenies seen in this study which supports Ford et al., (2019), who find widespread incongruence within *Oreochromis*. A lack in mtDNA phylogeny resolution within *Alcolapia* (Figure 2.6B), is likely due to the youth of this species flock and as it readily hybridises (Ford et al., 2015), allowing for mtDNA to be exchanged between lineages. This could also indicate that these mtDNA genes are evolving more slowly than nuDNA as the rate of evolution varies between DNA types (Allio et al., 2017). As there are a limited number of non-coding regions within mtDNA, there are few alternative genes that could have been investigated when constructing mtDNA trees for *Alcolapia*. Given the high bootstrap supports and resolution within the nuclear phylogeny, this seems more rigorous than the mtDNA for reconstructing the evolutionary relationship of *Alcolapia*.

It is surprising that site 14's nucleotide diversity is the highest within *Alcolapia* (Figure 2.8B), considering that it has only been found in one region within Lake Natron. One might have expected *A. alcalica* to have the highest nucleotide diversity given that it has the greatest known distribution around Lake Natron (Ford, 2015), so likely has the largest population size (Figure 3a). High nucleotide diversity of site 14 could reflect more extensive hybridisation occurring and introgression into this population. Indeed, Dsuite results indicated that site 14 had up to 10% more gene flow with *A. alcalica* and *A. ndalalani* rather than *A. latilabris* shown by the F4 ratios (Table 1). Given that 14-AL has 3% more gene flow with *A. alcalica* and *A. ndalalani* than *A. latilabris*, it is surprising that 14-AL doesn't have higher nucleotide diversity than 14-AA (Table 1 & Figure 2.8B). Perhaps 14-AA has a greater population size than 14-AL and so harbours more diversity across the genome. Furthermore, as site 14 has relatively high nucleotide diversity compared to other *Alcolapia* species, it may inhabit a more stable habitat and so is less strongly affected by water level change that may cause genetic bottlenecks and corrode genetic diversity at other sites. Elevated rates of nucleotide diversity in site 14 may be explained by how site 14 is predicted to be older than other Lake Natron species (Figures 2.6A), as molecular clock theory predicts mutations accumulate with time (Kumar, 2005).

The genotype frequencies across samples were then investigated at SNPs LG11: 6309402 and 6406194 to see if they varied by location and between species. Given that the heterozygous genotype frequencies have a local distribution and were not found in the outgroup (Figure 2.10), *O. amphimelas*, it seems unlikely that they have persisted through incomplete lineage sorting. Instead, as the heterozygous alleles have the highest frequency in 14-AL, it seems probable that these two mutations arose within this population. A couple of nearby sites 11C *A. alcalica* individuals also exhibit heterozygous genotypes at SNPs LG11:6309402 and 6406194, which is likely due to gene flow with nearby 14-AL. This is supported by elevated rates of genome wide introgression between 14-AL and site 11C *A. alcalica* (Table 1). Moreover, the binomial distribution of the genotype frequencies between site 14 morphs, may indicate that these polymorphisms are under divergent selection. Perhaps heterozygous alleles provide adaptive function and are also being selected back into *A*. *alcalica* populations. As no homozygous individuals were detected for the less common nucleotides at either SNP location, this could either indicate that G/G and/or C/C is deleterious, or that these SNPs are of such low frequency in the population, that homozygosity for the alternative allele is rare.

Efforts were made to see if there were any regional differences that could explain site 14s genetic isolation from other *Alcolapia*, but there were no obvious geological differences from maps between site 14 and nearby sites that contain all three species. However, within Lake Natron, there is likely to be vast heterogeneity in abiotic stressors such as natrocarbonate concentration, salinity and temperature, especially where freshwater mixes with the lake. It is known that population 14 inhabits an inlet (Figure 2.3) which is supported by streams. Together, this could be placing local selection pressures on population 14 but remains to be investigated. Site 14 appears to be connected to the southerly lagoon within Lake Natron even when water levels are low (Google Earth Historical Maps, 2013), enabling migration between nearby southerly sites. It is therefore not clear how environmental factors may have contributed to the speciation at site 14. Perhaps instead assortative mating, which has been observed in *Alcolapia* (Lawson et al., 2023), could be contributing more strongly to the genetic isolation of population 14. No study has been done on the colouration and patterning of *Alcolapia*, which could be important for providing courtship cues within each species.

Given that many island adaptive radiation are endemic, they are vulnerable and subject to high extinction rates (Fernández-Palacios et al., 2021). The IUCN Red List will need to assess the conservation status of *Alcolapia* site 14 which inhabits an insular system. Currently *A. alcalica* is listed as endangered (Bayona, 2006a) while *A. latilabris* and *A. grahami* are categorised as vulnerable. This is due to their limited distribution and habitat fragmentation (Hanssens, 2006; Bayona, 2006 b). Given this, it is likely that site 14 will also be categorised as vulnerable or assigned a more extreme conservation status of critically endangered, especially since it is only documented at one region which is less than 10km^2 (Seegers et al., 2001; IUCN, n.d.). Therefore, more extensive sampling around the Gelai region is needed to pinpoint the extent of site 14's distribution. Additionally, further sequencing of site 14 individuals could be used to generate population estimates for site 14, as sample sizes were too small to make reliable predictions during this project. Standardised photographs of living site 14 are also required to document its coloration as *Alcolapia* specimens lose their colour upon storage (Figure 2.3). This information will be important for species recognition, its classification status of this new species and for decisions regarding its conservation.

Chapter 3: Genetic analysis of mouth shape variation in *Alcolapia alcalica*

Abstract

Interspecific phenotypic diversity can be indicative of early divergence. Within cichlids, there are numerous adaptive radiations which are often large and complex. *Oreochromis Alcolapia*, a small radiation with substantial variability in head morphology, proves a simple model for understanding the evolution of new traits. The genetic basis of upturned and terminal mouth phenotypes remains unknown in *A. alcalica* which are segregating between two subsites in the east of Lake Natron, Tanzania. Variation in mouth phenotype was quantified in 50 specimens using linear measurements in the head region. All measurements appeared continuous rather than discrete, suggesting multiple loci contribute to mouth orientation. Using whole genome sequence data from these specimens, a GWAS identified 3 significant SNPs associated with mouth morphology. One of these is located ~0.2 Mb away from major craniofacial developmental genes *ANKMY2a* and *SOSTDC1a* which are involved in *sonic hedgehog (SHH)* and *wnt* regulation. These candidates need validation using gene expression and knock-out experiments. This research provides further evidence for the genetic basis of craniofacial morphology.

Introduction

Adaptive radiations are characterised by the rapid formation of many species from a common ancestor and high levels of phenotypic variation, which become specialised to a particular ecological niche to avoid competitive exclusion (Schulter, 2000). The presence of phenotypic diversity within a species can therefore indicate early divergence as a trait may be under selection as it is better suited to a specific environment (Kusche et al., 2015). Cichlids are a model family for understanding the emergence of biodiversity (Burress, 2014) as certain clade readily speciate and present exceptional phenotypic diversity in morphology, trophic level and male nuptial colouration (Seehausen and Schluter, 2004). Diversity within this family is driven by multiple factors including sexual selection

(Seehausen and Schluter, 2004), capacity to hybridise (Meier et al., 2023), the evolution of new traits (Conith and Albertson, 2021) and ecological opportunity (Ngoepe et al., 2023).

The propensity for cichlids to radiate is related to the evolution of a second set of jaws, enabling decoupling of food acquisition and processing, freeing the oral jaws to specialise to different food resources (Conith and Albertson, 2021). In species with one set of jaws, adaptations to different food resources are confined to the mechanical trade-off between power and speed, limiting functional diversity in mouth morphology (Burress et al., 2020). Pharyngognathy is therefore widely regarded as a key innovation, as it enables cichlids inhabiting the same region to divergently evolve and avoid competitive exclusion (Burress et al., 2020). The oral jaws function by capturing prey, while the pharyngeal jaws, which have evolved from gill arches and possess teeth, grind down food (Conith and Albertson, 2021). Functional decoupling within cichlids has facilitated their trophic diversity (Hulsey et al., 2006), as seen in the Lake Victoria cichlids, where up to four trophic levels exist (Meier et al., 2023).

Within the *Oreochromis Alcolapia*, a young subgenus of five species, there is considerable phenotypic diversity in mouth shape and orientation, making this a good, simplified system to study the emergence of novel phenotypic traits (Figure 1.1B and 2.6A). Despite no obvious difference being found between the pharyngeal jaws of *Alcolapia*, there is considerable variation in the oral jaw morphology (Ford et al., 2016). *A. latilabris* has distinct hypertrophied downturned lips with large, exposed teeth that it uses to rasp algae from rocks (Lawson, 2023). *A. ndalalani* also has a downturned mouths orientation but it has a narrow head. The vast majority of *A. alcalica* and *A. grahami* have terminal mouths except for several populations of genetically distinct upturned *A. grahami* found in Little Magadi (Kavembe et al., 2016). There is also an isolated pool of water in the northeast of Lake Natron (site 15, Figures 3.1, 3.2 and 3.3), where upturned *A. alcalica* are found (Ford et al., 2015; Seegers et al., 2001). Terminal mouthed *A. alcalica*, which are the more widespread phenotype, inhabit the lake margins at a stream just a few meters from this pool (Julia Day field notes 2012 and 2017). The pond may connect to the main lagoon during periods of flooding facilitating gene flow between the morphs, however they were not connected in July 2017 indicating segregation of phenotypes between the sites (Julia Day personal communication, 2024). However, there is variation in severity of the angle of the mouth orientation (Figure 3.1), suggesting that this is a continuous trait which will be investigated and confirmed in this thesis.

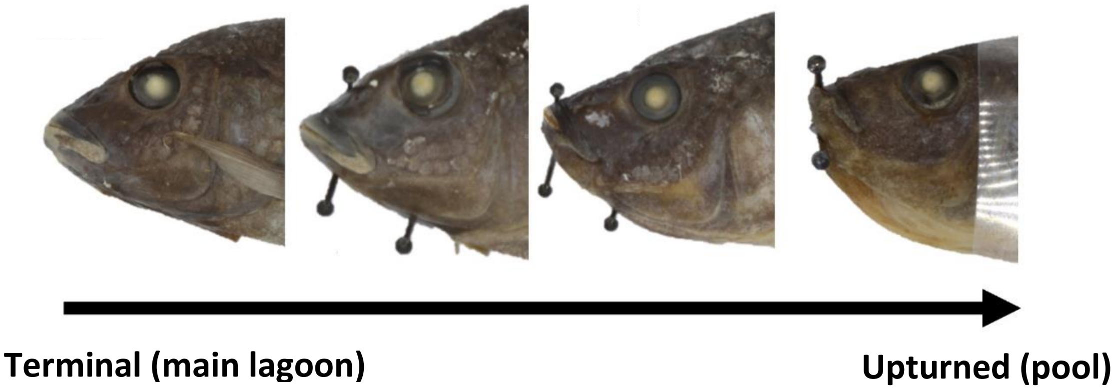


Figure 3.1: Phenotypic diversity in mouth orientation of preserved *A. alcalica* from site 15.

The upturned mouth orientation of *A. alcalica* could provide clues to its dietary habits and trophic level that it occupies. Keppeler et al., (2020) finds that fish with upturned mouths are more likely to be higher in the food chain than those exhibiting terminal and downturned mouth apparatus. This adaptation allows upturned fish, such as Freshwater Hatchet fish (*Gasteropelecidae*) or weavers (*Echiichthys vipera*), to predate on fish swimming above them. Additionally, some cichlid species with upturned mouths have been shown to predate on the eggs and juveniles of other cichlids which is known as paedophagy (McKaye et al., 1983). *Cyrtocara orthognathus*, a Lake Malawi cichlid, use their heads to ram mouthbrooding females, which causes the females to expel their young into the water column which the paedophageous cichlid then predate upon. Similar behaviours have also been reported in Lake Vicotria fish and paedophagy has been reported in *Alcolapia* species (Seegers and Tichy 1999; Ford et al, 2016). Alternatively, upturned *A. alcalica* may feed more on algal film and insects found at the surface as the bulk of *Alcolapia's* diet at other sites is composed of cyanobacteria, albeit some insect species (Lawson, 2023).

The genetic basis of mouth shape is unknown in *A. alcalica*. This study aims to identify whether mouth orientation is a complex or discrete trait by capturing variation in the positioning of the oral jaws. Following phenotyping, several GWAS will be conducted to identify if there are any loci associated with mouth orientation.

Figure 3.2: Small, isolated pool of water circled in red which is just a few meters wide and is nicknamed “the old mine”, taken in July 2017 by Julia Day at coordinates -2.433361 S, 36.10175 E. This site has been observed to be inhabited by exclusively upturned *A. alcalica* (Julia Day fieldnotes, 2017). Julia Day noted that this pool had an extremely high density of *A. alcalica* and was deeper than the surrounding bodies of water connected to Lake Natron.

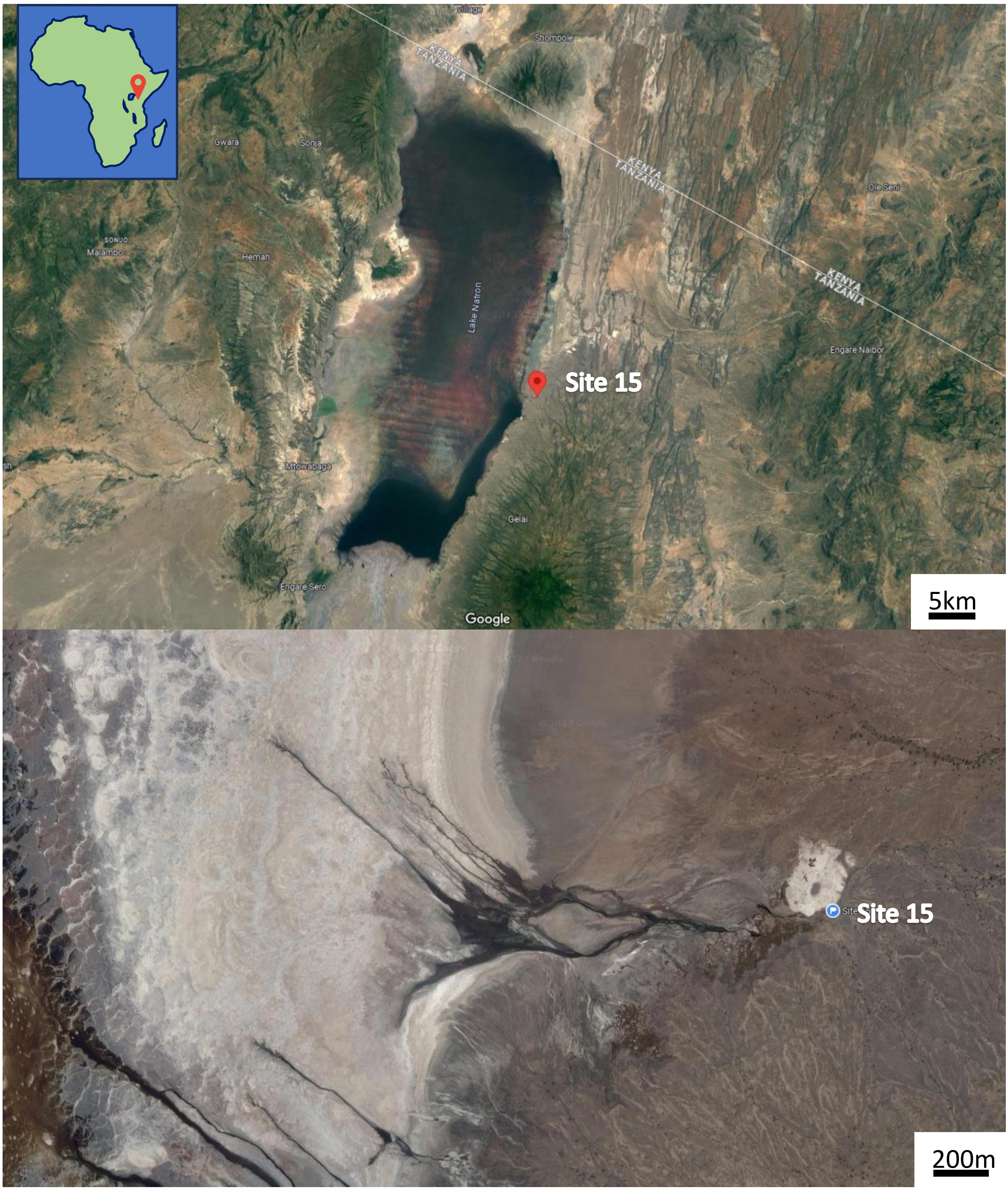


Figure 3.3: Arial photography of site 15 situated in the east of Lake Natron by streams (Google Maps, 2023).

Methods

Specimen collection by Day, Ford, Shechonge, Turner and White, 2017

In July 2017, two subsites were visited in the East of Lake Natron, west of the Gelai Volcano which are known to foster *A. alcalica* with mouth orientation polymorphisms (Figures 3.1-3.3; Ford, 2015). No other locations have this polymorphism documented for this species. The first site is a stream connected to the main lagoon and contained *A. alcalica* with terminal mouth morphologies. Meanwhile an adjacent small pool of water contained *A. alcalica* with distinctively upturned mouth orientations (Figure 3.1, Julia Day personal communication), indicating phenotype segregation. These sites are referred to by Lawson (2023) as site 15A (terminal) and 15B (upturned), and both have the coordinates: -2.433361 S, 36.10175 E due to their proximity. Site 15 is therefore a collective term for these two sites. Fish collected were categorised as either upturned (pool) or terminal (lagoon) by Ford and Day and a total of 58 specimens were collected from the pool (35 x upturned) and lagoon (23 x terminal).The pool was observed to be just a few meters wide but much deeper than the nearby lagoon and both sites remained isolated during the dry season when visited in January 2012 and July 2017 by Day and Ford (Figure 3.2, Julia Day personal communication, 2023). Fish were euthanised using clove oil and pectoral fin clips were taken from these and preserved in 100% ethanol.

Authors Own:

Phenotyping variation in mouth position

On closer inspection, there appeared to be fish with intermediate mouth morphologies, indicating that mouth orientation was continuous rather than categorical. To investigate this, preserved fish were photographed on their left-hand side, pinned externally with their mouths closed. Some previously documented landmarks were used (landmarks, 2, 5 12, 13, 16 see Fig. 6.2 from Ford, 2015). However landmark 14 called “posterior point of mouth cleavage” previously used by Ford (2015) was inconsistent with upturned and terminal fish at site 15, so a different landmark was chosen to capture mouth angle variation. An additional landmark was chosen within the head region by drawing a tangent to the upper lip along the axis of the mouth opening (landmark 4, Figure 3.4). This was used to measure mouth angle. TpsDig2 version 2.32 was used to plot the landmarks (Rohlf, 2023).

Three traits within the head region were measured to capture mouth shape variation. Distance 1 is the horizontal distance between the tip of the snout (landmark 3) and the bottom of the anterior most point of the top lip (landmark 5). Meanwhile distance 2 captures variation in the vertical length between the centre of the eye (landmark 6) and the bottom of the anterior most point of the lips (landmark 5). Finally, mouth angle was captured by measuring the degrees at which a tangent to the mouth axis (4 and 5) intercepts a perpendicular line to landmarks 1 and 2 (see Figure 3.4).

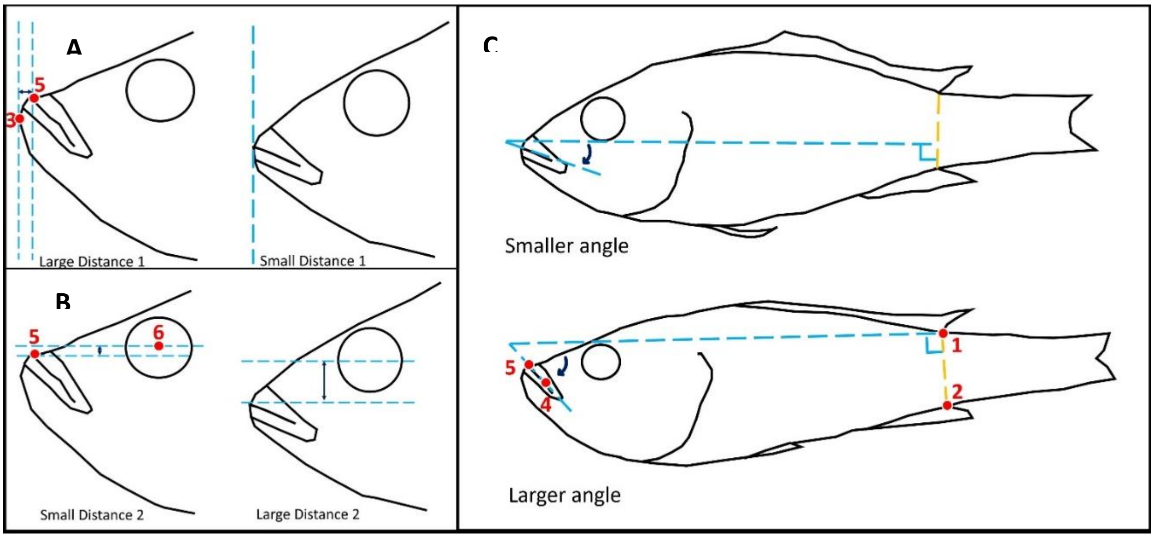


Figure 3.4: Diagrams of three measured traits to quantify mouth orientation. More upturned fish are predicted to have larger distance 1 **(A)**, smaller distance 2 **(B)** and larger mouth angles **(C)**. Landmarks: 1) Insertion of the dorsal fin, 2) insertion of the anal fin, 3) tip of the snout, 4) axis of the mouth opening, 5) the bottom of the anterior most point on the upper lips, 6) centre of the eye. Orange line used to form a reference for measuring angles and distances. Distance between landmarks 1-3 was used to control for allometry.

Tpsdig2 version 2.31 generated a series of coordinates enabling trigonometry to be used to measure the distances and angles. A total of four phenotypic traits were generated (three continuous and one discrete).

DNA extraction

DNA was extracted from 50 fin clippings (22 males; 28 females) using a Qiagen DNeasy Blood & Tissue Kit and quantified using the Qubit protocol. Fish with more extreme phenotypic measurements were prioritised to be sequenced. All DNA extracts were run on an agarose gel to check for RNA contamination and DNA fragmentation by comparing to a 1kb DNA ladder. Samples of low DNA concentration were concentrated using ethanol precipitation. Samples were then tested for purity using a nanodrop spectrophotometer and sent to NEOF in Liverpool for whole genome re-sequencing.

Bioinformatics pipeline

BWA mem was used to map and index raw whole genome files against the *O. niloticus* reference genome (Li, 2013). Samtools depth was used to calculate an average coverage of x14.8 on site 15 bam files across the genome (Li and Bonfield, 2023). Picard tools was used to remove PCR duplicates and haplotype calling and genotyping was performed using GATK 4.3.0 (Van der Auwera and O'Connor, 2020). BCFtools 1.15 (Danecek et al., 2021) was used to filter the VCF to contain only SNPs, depth of 5-100, GQ of 30, Missing of 0.5 and a MAF of 0.05. 2,413,661 SNPs remained after filtering. 3,527,226 SNPs remained after filtering when all the previous filters are kept the same apart from GQ, which was reduced to 20. Plink was used to perform linkage pruning and to generate a PCA to check for population structure (PLINK/2.00a2.3-GCC-10.3.0). Eigenvectors were plotted using R studio (R-4.3.2) and eigenvalues were transformed manually into PC values.

Scatterplot3D was used to visualise the traits measured and Mann-Whittney U tests were performed to test for significance on each trait (as the data was not normally distributed).

GWAS

GWAS enables identification of SNPs that are associated with the trait of interest which can be used to identify potential candidate genes (Rathan et al., 2022). Despite being few sequenced samples for site 15, GWAS analysis can still be performed to produce good predictions as low population structure exists between extremely upturned and terminal

A. alcalica at this site (Ford, 2015; Lawson, 2023), therefore there will be fewer false positives making it a good model to study genes related to mouth orientation.

Bed files were generated using plink2 and univariate linear mixed models in GEMMA were used to perform GWAS (Zhou and Stephens, 2012). The four different GWAS phenotypes were run to test if genetic differences could be observed between specimens with different mouth morphologies. This included, distance1, distance 2, angle1 and Ford and Day's categorisation (discrete either upturned or terminal).

The following modules were used to run GWAS in a Unix terminal: plink, math/Eigen/3.4.0-GCCcore-11.3.0 and numlib/OpenBLAS/0.3.20-GCC-11.3.0.Python was also used to sort the scaffolds in ascending order (Biopython/1.79-foss-2022a).

GWAS results were plotted in R studio (R-4.3.2, Posit Team, 2023) following a script provided by Yacine Ben Chehida that contained the following packages: ggfortify, qqman, tidyverse, ggtext, normentR, grid, gridExtra, dplyr, sjlabelled, ggplot2, cowplot, quotidieR and ggthemes.

Results

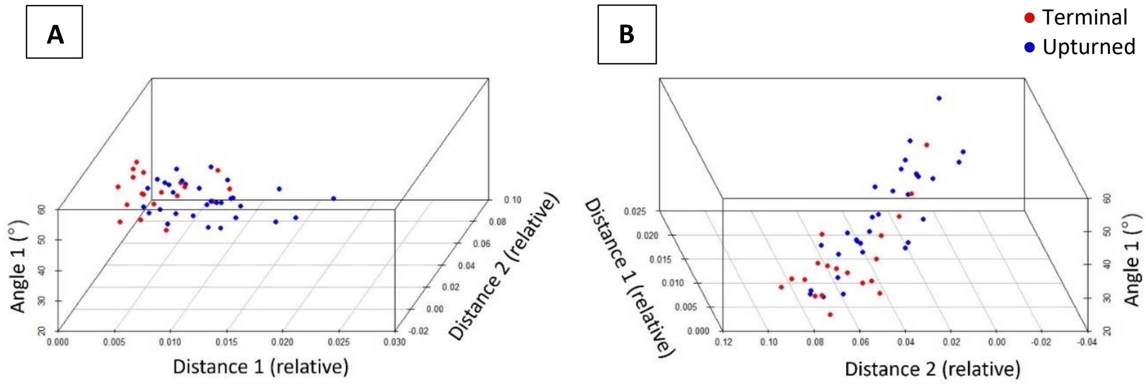


Figure 3.5: 3D scatterplots show variation in head morphology between upturned (31) and terminal (19) morphs. Distances are normalised to control for allometry. **A** = Plot angle 65 °**, B =** plot angle = 295 °.

Linear measurements identified mouth orientation as a continuous trait (Figure 3.5). Upturned fish have larger distance 1, smaller distance 2, and greater angle in mouth as expected. All three continuous traits have significantly different distributions when grouped by Ford and Day's categorisation of upturned and terminal. Distances are all relative to the size of the fish so are in arbitrary units. The median distance1 for upturned fish (0.00774 a.u.) is significantly greater than terminal phenotypes (0.00178 a.u.) ($W = 100$, p = 4.75e-05). Furthermore, the median distance 2 for terminal fish (0.0662 a.u.) is significantly greater than upturned fish (0.0411p a.u.) ($W = 449$, $p = 0.00162$). The median angle1 for terminal fish (34.9°) is significantly lower than upturned fish (42.4 °) ($W = 185$, p = 0.0284).

Next, a genetic PCA on whole genome data grouped by upturned/terminal category was performed to see if population structure existed. No population structure was identified between these morphs from site 15 (Figure 3.6), indicating a GWAS would be possible with a small sample size.

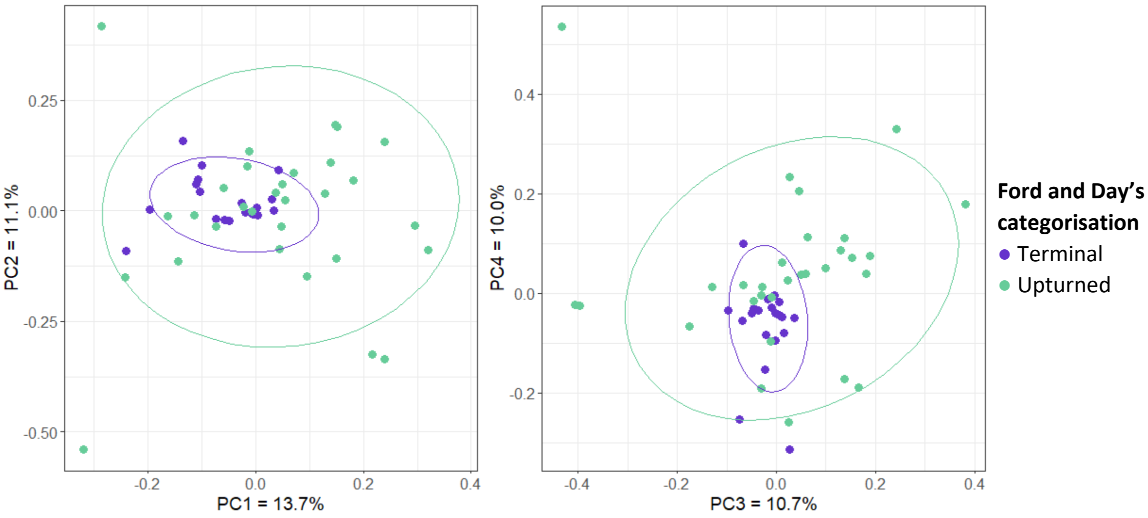


Figure 3.6: Genetic PCA of LD pruned SNPs (2,413,661 SNPs) does not show any clustering by morphological categorisation. Extensive overlap between upturned and terminal *A. alcalica* DNA from site 15 when grouped by Ford and Day's categorisation, indicating high genetic similarity between samples. Looser clustering on PC 1-4 between upturned morphs indicates higher genetic diversity than terminal morphs. 2,413,661 SNPs remaining after filtering; GQ30, DP5, DP100, MIS 0.5, MAF 0.05, SNP only.

Several GWAS were performed to identify any candidate genes that may be associated with mouth orientation and to see if there was evidence for selection or inversions. GWAS analysis can be used to search for inversions as it can show patterns between specific phenotypes and sections of the genome which have very small p-values corresponding to samples where base sequences have been reversed. Likewise, very small p-values at base pair positions flanking a SNP can be used to indicate regions and genes under selection in a GWAS analysis. This is as nearby alleles are linked to a SNP under selection Although GWAS were performed on all four phenotypic traits (angle of the mouth, Ford and Day's

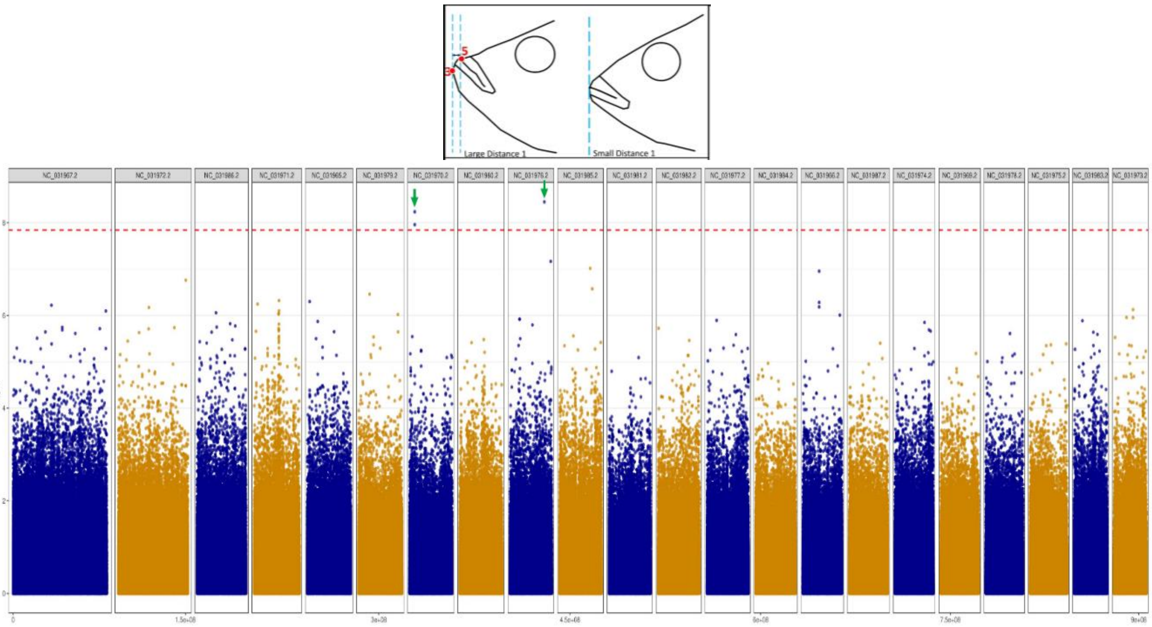


Figure 3.7: GWAS identifies three SNPs (shown with green arrows) above the Bonferroni threshold of 7.84 when distance 1 is the phenotype. SNPs are located on CHR 5 (NC_031970.2) and 11 (NC_031976.2). GQ = 20. Chromosomes are arranged by size from left to right being the largest to smallest with x-axis = accumulative base pair position across the genome.

discrete ranking, distance 1 and 2), only distance 1 found significant SNPs associated with mouth orientation (see Figure 3.7 and S8). Three SNPs were found, one on LG11, two on LG5 (Figure 3.7).

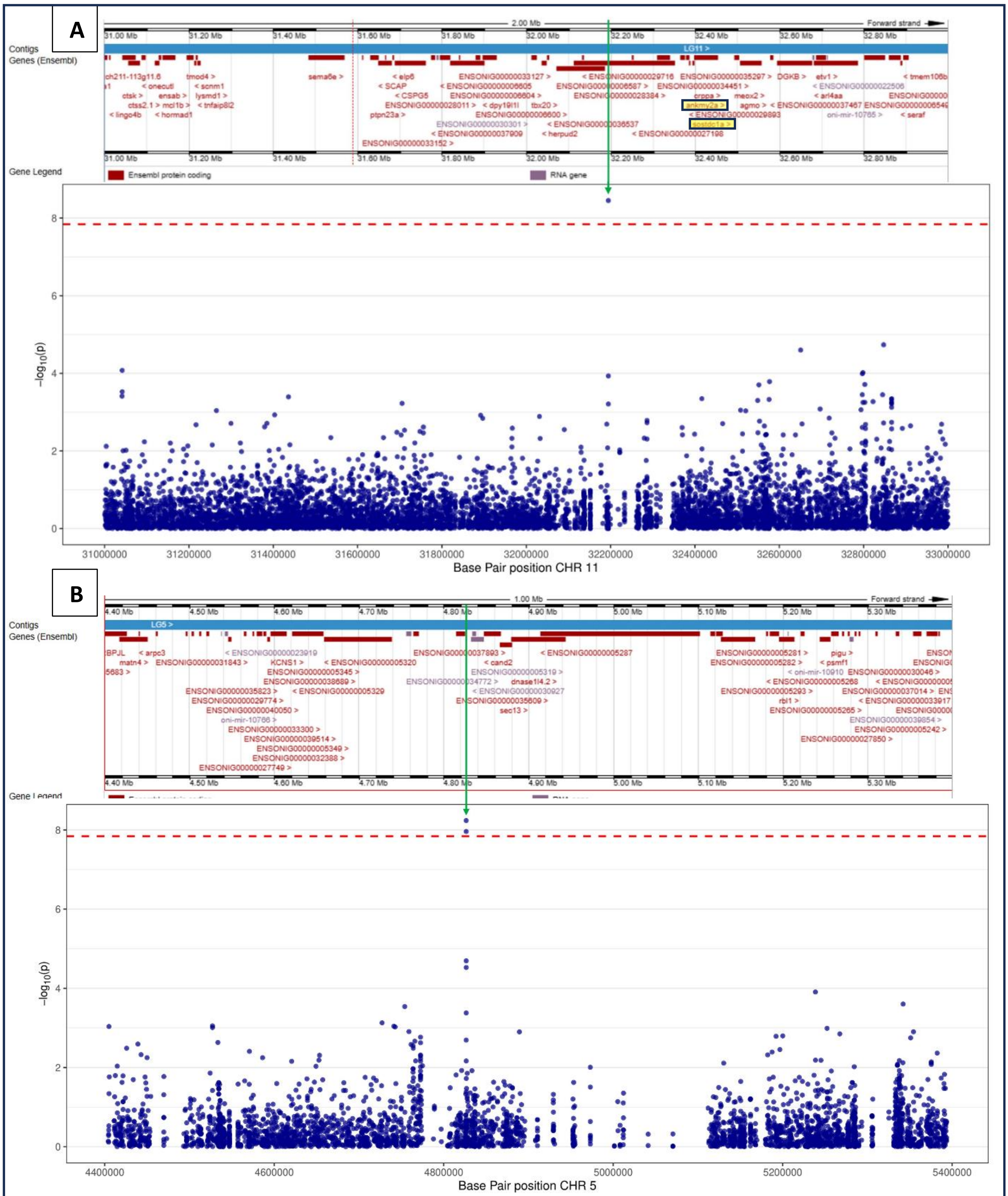


Figure 3.8: Zoomed in GWAS of distance 1 around significant SNPs identified in Figure 3.7. **A)** SNP LG11: 32,194,615 (shown by a green arrow) has a *p* = 8.67x10-9 and is located in an unknown protein coding gene ENSONIG00000028384 (Ensemble, 2024b). Yellow boxes = genes of interest **B)** Two significant SNPs are in non-coding regions at CHR5:4826575 and 4826582 (shown by green arrow) with *p* = 2.52x10-8 and 1.37x10-8 respectively. Screenshots taken of gene regions from *O. niloticus* genome browser (Ensemble, 2024b and c).

The region around the significant SNPs found in Fig 3.7 was inspected more closely to see if there was evidence for selection which would be illustrated by smaller p values close to a SNP in a GWAS analysis (Figure 3.8). All three SNPs were not well supported by nearby linked SNPs. However, the gaps around peak LG11: 32, 194,615 reflect filtered regions of low mapping quality, and so more SNPs may exist flanking this SNP.

The SNP located on LG11: 32, 194,615 is ~0.2Mb from *SOSTDC1a* and *ANKMY2a* (referred to in the discussion as possible gene candidates). No promising gene candidates were found in the 1MB region around the two SNPs on LG5, although many of these were of unknown function. At more stringent filtering, GQ30, only LG11:32, 194,615 is significant (Figure S7), suggesting SNPs on LG5 are below GQ30.

Lastly, the genotype frequencies of the three different SNPs were investigated to identify if there were patterns between sex, genotype and the trait of interest (distance 1). The four most extremely upturned females had the same genotype so mouth orientation may be sex linked (Figure 3.9). The alternative genotype has the greatest distance 1, indicating more upturned fish (Figure 3.9). On inspection of the genotype file prior to filtering, no C/C individuals existed for LG11:32,194,615.

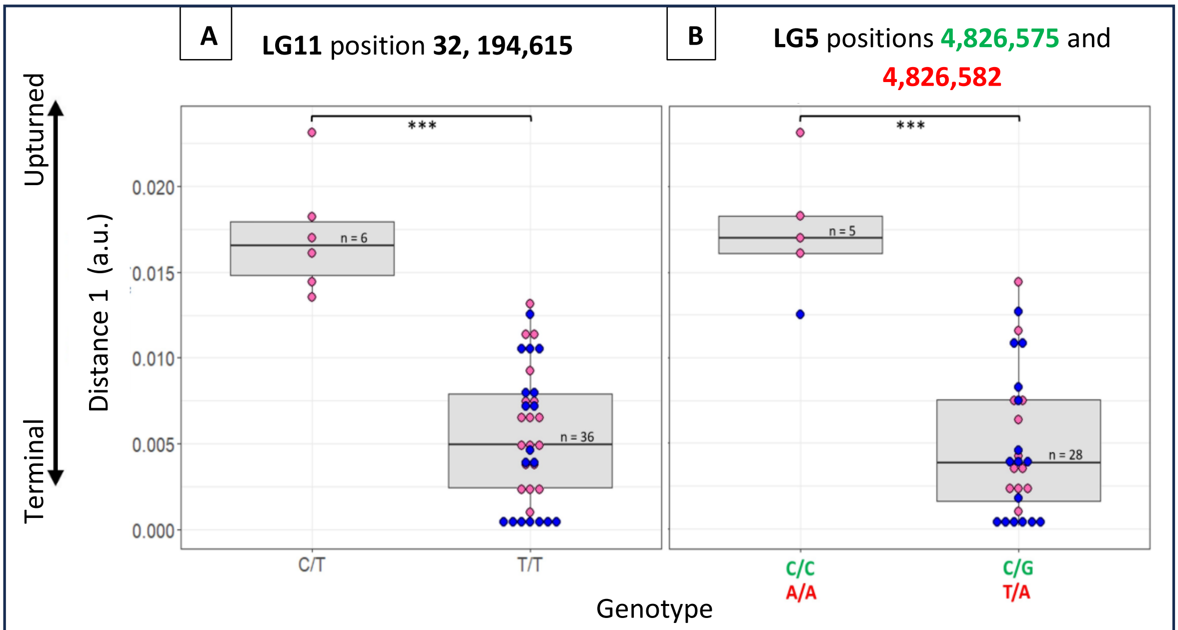


Figure 3.9: A) Heterozygotes have greater distance 1, so are more upturned, than homozygote T/T individuals. All C/T individuals are female, while T/T and N/N are composed of both sexes. T/T are slightly skewed towards 0. Wilcox test: *W* = 216, *p* = 0.000112 **B)** Genotype frequency patterns were identical for CHR5: 4,826,575 and 4,826,582 likely due to linkage. Homozygous genotypes have greater distance 1. GQ = 20. Wilcox test: *W* = 138, *p* = 0.000701. Blue dots are males while pink ones are females.

Main findings

Linear and angular measurements captured continuous variation within the mouth region for site 15 *A. alcalica* between upturned and terminal morphs (Figure 3.5), indicating mouth orientation is a complex trait. No population structure was seen at site 15 *A. alcalica*, indicating gene flow between morphs which is concordant with previous findings (Lawson, 2023). However, the terminal morphs clustered more closely together on genetic PC space (Figure 3.6), indicating they have lower genetic diversity than upturned morphs. The GWAS on distance 1, which captures the relationship between the mouth opening and the tip of the snout, found three significant SNPs. The SNP located at LG11: 32194615 is near two genes *SOSTDC1a* and *ANKMY2a* ~0.2Mb away (Ensemble, 2024b and c) that are promising gene candidates as they are related to craniofacial development (Cho et al., 2015; Saita et al., 2014). More than one region is associated with mouth orientation polymorphism (this study; Ford, 2015; Kavembe et al., 2016), indicating mouth orientation is polygenic. No SNPs were found to be associated with the other phenotypes. Only the most extremely upturned females possessed the alternative genotype.

Discussion

Three SNPs were identified as being significantly associated with distance 1 when GQ=20. However, all SNPs are not well clear of the Bonferroni threshold, therefore there is always the possibility that they are false positives. Furthermore, when GQ is increased to 30, both SNPs on LG5 are no longer significant and may be being removed due to the filtering (Figure S7), which questions the integrity of these findings. There may be more support for the peak on LG11 as bordering SNPs were filtered out due to low mapping quality (below a GQ of 20). Furthermore, only 50 samples were used in this GWAS, therefore a larger study is needed as craniofacial morphology is a complex trait (Powder and Albertson, 2016) and so small effect loci may not be identified. In support of this, four of the most severely upturned individuals in this study possess all three of the less common genotypes (C/T, C/C or A/A), suggesting mouth orientation is polygenic and that each locus is contributing additively to the upturned trait (Figure 3.9). Additionally, the genetic architecture body shape, a similar type of trait which captures variation in the head region, has been found to be polygenic in Midas cichlid species (Kautt et al., 2020). It would therefore be insightful to know how much of the variation in mouth orientation is explained by each SNP and future work could estimate the heritability of these significant SNPs using GEMMA (Zhu and Zhou, 2020).

Interestingly, all the heterozygote genotypes at LG11 position 32,194,615 were female and had the greatest distance 1s (Figure 3.9a), indicating that the most severely upturned fish are more likely to be female. More replicas are needed to increase certainty of this pattern. However, as the upturned phenotype is continuous (Figure 3.5), it cannot definitely be used to predict the sex of the specimen. . Mouth orientation is less informative of the sex at LG5 at positions 4,826,575 and 4,826,582, as a male grouped with the less common homozygous genotypes, representing the upturned group (Figure 3.9B). Perhaps the SNP on LG11 is of larger effect size, as LG11 has no overlap in linear measurements when categorised by sex or genotype unlike the two SNPs on LG5, which have slightly bigger p-values compared to LG11 (Figure 3.9). These findings support previous theory that this trait is sex-linked which Ford (2015) theorised may be due to disparities in the occupation within the water column between morphs which is implied by differences in dietary niches between the sexes. Interestingly, mouth angle is strongly sexually dimorphic in *A. burtoni,* a cichlid species native to Lake Tanganika*,* with females exhibiting a tenfold difference in head morphology than males, having more upturned mouth orientation, leaner and longer heads (Figure S9; Theis et al. 2014).

Just ~0.2Mb away from the SNP located on LG11 in the large protein coding gene ENSONIG00000028384, is a gene called *ANKMY2a* (see Figure 3.8A; Ensemble, 2024b). Expression of *ANKMY2a* influences the sonic hedgehog (*SHH*) pathway, which is a critical gene in cell development that is conserved between mammals and fish (Saita et al., 2014). *ANKMY2a* upregulation positively affects the *SHH* pathway and likewise downregulation of *ANKMY2a*, negatively regulates *SHH* activity (Saita et al., 2014). *SHH* is expressed in the teeth, operculum and gill filaments of cichlids and inhibition of this pathway leads to abnormalities in the formation of premaxilla, teeth, gill filaments and oral jaws of *Haplochromis piceatus*, a Lake Victoria cichlid in a different order to *A. alcalica* (Cho et al., 2015). Upregulation of *SHH* leads to greater jaw size in *Astyanax mexicanus,* a species of fish from the same class (Actinopterygii) as *Alcolapia* species (Yamamoto et al., 2009). Furthermore, inhibition of *SHH* affected the position of the mouth opening in *Astyanax mexicanus,* causing a more downturned mouth (Figure 3E and F in Yamamoto et al., 2009). *SHH* is expressed in the head region of developing *A. alcalica* embryos (White, 2021). Therefore, it's feasible that the SNP difference detected in the unknown gene ENSONIG00000028384 (Figure 3.7 and 3.8A), could be affecting *ANKMY2a* expression, perhaps as a transcription factor, influencing the regulation of *SHH* during development. In turn variations in *SHH* could be generating the range of mouth orientation phenotypes observed in *A. alcalica* at site 15, Lake Natron. To test this further, a series of Transcriptome-wide association studies, protein assays and Western blots could be conducted in the tissues known to express *SHH*, comparing between developing terminal and upturned fish embryos. It would also be insightful to quantify the expression patterns of ENSONIG00000028384 between the tissues of upturned and terminal *A. alcalica* morphs, focusing on developing cranial bones, connective tissues and muscles. RNAseq could be used to quantify potential differences between tissues and/or morphs, improving our understanding of any variation in the gene expression profiles of fish with different mouth orientations and the role of this unknown gene (Wang et al., 2009). Likewise in situ hybridisation could be used to visualise identify the exact location of ENSONIG00000028384 expression in cells and tissues within the head region of developing *A. alcalica* upturned and terminal embryos (Chu et al., 2019).

Although the SNP associated with mouth orientation (LG11:32194615) is located in an unknown gene, SNPs can alter DNA methylation, affecting the transcription of nearby genes (Vohra et al., 2020). Another gene called *SOSTDC1A* is located ~0.2Mb from the SNP on LG11:32194615 and is known to be expressed in the teeth, nasal bone, and gill filaments of *H. piceatus* (Cho et al., 2015). *SOSTDC1* expression inhibits Wnt and bmp pathways which are important in bone development (Krause et al., 2010). Variation in WNT concentration has been shown to affect pre-orbital morphology as it influences bone ossification within the head region of Lake Malawi cichlids (Parsons et al., 2014). It is therefore believed to be an important contributor in generating phenotypic diversity within Lake Malawi cichlid species (Parsons et al., 2014). To test whether the SNP at LG11:32194615 is affecting DNA methylation, thereby influencing *SOSTDC1A* expression and corresponding downstream pathways, CpGenie could be used to predict the functional role of this SNP located within a non-coding region (Zeng and Gifford, 2017).

Furthermore, *bmp4* expression affects cichlid jaw shape, influencing craniofacial diversity (Albertson and Kocher, 2006). Considering this, SNP LG11:32194615 located within an unknown protein coding gene (ENSONIG00000028384) at could be affecting the transcription of nearby of nearby *SOSTDC1* (for genomic distances see Ensemble, 2024b). *SOSTDC1*, which negatively regulates Wnt and bmp expression, could be affecting the bone structure within the pre-orbital region of *A. alcalica*, leading to the diversity in head orientation seen at site 15. Likewise, monitoring of the expression of Wnt and bmp is needed between morphs to test this theory.

On the contrary, *O. niloticus* is a terminal tilapia species from the same genus as *A. alcalica* and possess a cytosine residue at LG11:32194615 (Ensemble, 2024b), like upturned *A. alcalica* (Figure 3.9). The genotype of *O. niloticus* is unknown at this position as it is haploid but occurrence of the cytoside SNP substitution in the ENSONIG00000028384 gene contradicts theory that is causing phenotypic change in mouth orientation. Therefore, KO experiments of this gene are necessary in to identify its function, widening our knowledge of the role of this unknown gene.

Conflicting results between the diets of upturned and terminal *A. alcalica* are found (Ford, 2015), while others remain inconclusive (Lawson, 2023). Ford (2015) finds a significant difference between stable carbon isotope levels between upturned and terminal *A. alcalica* but considerable dietary overlap for Schoener's index. Given that isotype analysis represents long term dietary trends (Pearson et al., 2015), compared to stomach contents analysis which fluctuates daily, Ford's (2015) findings imply that site 15 morphs have differences in diets between upturned and terminal fish and are partitioning the niche at site 15. Furthermore, upturned *A. alcalica* at site 15 have been shown to consume fish scales (Figure 6.6D, Ford, 2015), providing evidence that the upturned morph may be related to cannibalistic behaviour, which may indicate another trophic level in *Alcolapia* but this remains to be tested. This is interesting given that fish with upturned mouths are more likely to be predators (Keppler et al., 2020 cited in Kopf et al., 2020). Therefore, further long-term sampling of stomach contents analysis is needed so that the species richness and diversity of taxa consumed by site 15 morphs can be identified and monitored. This may infer whether the different morphs have fine scale niche partitioning like other described *Alcolapia* (Ford et al.., 2016) and are evolving to avoid competitive exclusion.

Despite no signals of selective sweeps being identified in this study, Ford found broad elevation in F_{ST} between upturned and terminal *A. alcalica* around a peak on LG11 and another peak on LG7 (Figure 4.7, Ford, 2015), indicating recent divergent selection. However, this peak is at different genomic window positions LG11:9,100,000-11,400,000 compared to the one found in this study (LG11: 32,194,615), therefore represent different SNPs. Given that mouth orientation appears sex related (Ford, 2015), and *Alcolapia* are mouthbrooders (Sutton et al., 2020), perhaps a more upturned mouth could provide a fitness benefit for raising offspring. Divergence in mouth orientation axis could therefore be under sexual selection.

Deeper lakes are associated with higher diversification rates in cichlids as they provide greater variation in ecological niches (Wagner et al., 2012). Therefore, as the pool is deeper than the main lagoon (Julia Day, field notes, 2017), there is likely to be more habitat disparity, providing subtle differences in selection pressures with depth. Ford (2015) also noted that the pool was stagnant which is unique as the other sites sampled around Lake Natron (also used in this study) had flowing water (Ford, 2015). Therefore, greater habitat range at site 15 may provide ecological release, enabling higher genetic and phenotypic diversity at this site as seen in Lake Victoria cichlids (Ngoepe et al., 2023).

Upon close inspection of the region near LG5: 4,826,575 and 4,826,582, there were no obvious genes that may be related to mouth orientation (Ensemble, 2024c). Most of the genes around this SNP are of unknown function. As the SNP is in a non-coding region it could be an enhancer for another gene as these do not need to be located next to the transcription start site to affect its expression (Smemo et al., 2014). In fact, an enhancer can be up to 1mb away from the TSS to regulate expression (Yang and Adli, 2019). Considering this, all known genes were inspected 1mb either side of LG5 at positions LG5: 4,826,575 and 4,826,582 but there was no strong evidence for any relevant genes within this region. One gene, sp7 is cited to be involved in the diversification of cichlid jaw phenotypes, however evidence for this in the primary source could not be found (Gunther et al.,2013 as cited in Irisarri et al., 2018). To test whether SNPs at positions LG5: 4,826,575 and 4,826,582 are transcription factors, snpSTARRseq could be performed to reveal if they have functional roles (Tunc Morova et al., 2022).

A previous study investigated the genetic basis of upturned mouth phenotypes in Little Magadi compared to terminal *A. grahami* which inhabit nearby geographically isolated Lake Magadi. Kavembe et al (2016) found one locus under selection on Chr 6 but found no relevant candidate genes that relate to mouth orientation (Kavembe et al., 2016). As this SNP is on a different chromosome to the SNPs identified in this study, there is no evidence to suggest that the same alleles for mouth orientation are evolving collaterally. Furthermore, upturned phenotypes from Little Magadi look less severe as those found at site 15 *A. alcalica* in Lake Natron (Figure S6). Despite morphological similarities between upturned *A. alcalica* and *A. grahami*, the two species cluster less strongly genetically and so Ford (2015) suggested that the ancestor of the *Alcolapia* flock may have possessed terminal to slightly upturned mouth phenotypes based on morphological reconstruction.

Perhaps there needs to be improved phenotyping of the fish to capture more variation between the upturned and terminal phenotypes. One paper on Malawi cichlids measured mouth angle and the oral jaw length but this relies on a landmark used that is very hard to place which captures the pivot of the jaw (Figure 3, landmark 3, Feller et al., 2020). This could therefore introduce error into the measurements. Alternatively, it may be useful to remove the top layer of skin as seen in Parsons et al., 2014 (Figure 1) or x-ray the fish to see anatomical features and skeletal structures clearly to aid in measuring mouth angles and jaw lengths. Furthermore, it may also be useful to investigate the number and cusps of teeth and see if this relates to different niches of site 15 morphs as different feeding behaviours specialised for biting or suction convergently evolve tooth phenotypes associated with them (Albertson and Kocher, 2006). For instance, biting phenotypes that scrape algae from rocks have many short tricuspid teeth, while Malawi cichlids with longer jaws are specialised for suction feeding on plankton and prey and have fewer, larger, and less cusped teeth (Albertson and Kocher, 2006). It therefore would be interesting to see if teeth formation changes in relation to an upturned phenotype in *Alcolapia*.

Chapter 4: Conclusions and future suggestions

This study not only describe the genetic basis of morphological traits including mouth orientation which may be sex linked, but it provides strong evidence for a new species of *Alcolapia* emerging in the face of gene flow. This study supports theory that craniofacial morphology is controlled by multiple loci and draws attention to*OSR2*, which has been linked with cranial facial structure in other animal systems. This needs to be treated with caution however as there are many genes nearby the SNPs that have unknown function so will need further assessment to determine the gene or genes with functional importance in regard to mouth orientation. . KO experiments of genes *ANKMY2a, SOSTDC1A*,*OSR2* and unknown genes near or surrounding identified SNPs are therefore needed to see how they impact the expression of other genes including *SHH* and embryonic development in the cranial region of developing site 14 and *A. alcalica* individuals. Further phylogenetic modelling of the *Alcolapia* species flock is needed and performing a Multispecies Coalescent species tree such as SNAPP to capture coalescent events, estimations of species divergence and population sizes would be useful for conservational purposes (Bryant et al., 2012). In addition to this, lectotypes as seen in Seegers and Tichy (1999) are required for further morphological classification of this newly described lineage and its two morphs.

Further to this investigation, it would be interesting to see if ecological divergence exists between the two morphs at site 14. To understand this, gut contents analysis could be performed to see if ecological partitioning is occurring, with the expectation that 14-AL possess more benthic role as it resembled *A. latilabris* more closely. Meanwhile, 14-AA may have a diet most similar to *A. alcalica* which eats more vascular plants and has a terminal mouth. The percentage of grit in the stomach may also indicate niche divergence as seen in Ford et al., (2016). As the gut contents analysis of upturned and terminal Alcolapia was inconclusive, future investigation is needed to see if these morphs vary in their diets and therefore ecological niches. Likewise, as gut content analysis of upturned and terminal *A. alcalica* morphs was inconclusive (Lawson, 2023), similar experiments could be repeated on these morphs to see if upturned phenotypes are more cannibalistic, have higher predation rates on surface insects or consume more vascular plant material.

Given that extremely upturned *A. alcalica* were all females in this research (samples were collected in 2015) and from a previous study at the same location collected in January 2012 (Ford, 2015), it implies mouth orientation may be under sexual selection. Perhaps the angle of the mouth affects life history traits, such as greater survival rate and number of offspring raised as *Alcolapia* are known to be mouth brooders. Perhaps upturned mouth is a defence strategy to terminal A. alcalica which have been shown to have increased rates of paedophagy when compared to other *Alcolapia* species (Figure 2D, Ford et al., 2016).

The genetic basis of yellow-blue colour polymorphism and presence of melanism in *A. latilabris* and *A. alcalica* needs further investigating (Ford, 2015). Previous research has identified regions of divergence between blue-yellow colour morphs on LG1, LG5 and LG17 (Ford, 2015, Figure 4.7 plot O) shown by raised F_{ST} at these locations. As blue-yellow polymorphism exists in two species within the radiation, it's possible it may have arisen by recent introgression. It is also not known whether this trait is continuous or discrete. Given that sexual dichromatism is positively linked with species richness in cichlids (Wagner et al., 2012), it is likely that this is a contributing factor for reproductive isolation between Lake Natron taxa and site 14 morphs which are known to be darker (Ford, 2015). Assortative mating experiments between site 14 morphs and *Alcolapia* would therefore provide a measure of the strength of reproductive isolation as seen in Lawson et al., (2023).

A better survey of Lake Natron is required, particularly on the Western shores where springs are numerous (Dawson, 2008) and unexplored, but these are difficult to visit due to their remoteness and lack of infrastructure (Julia Day personal communication, 2023). Further sampling is also needed in the northeast which may reveal isolation by distance between northern and southern *A. alcalica*, which are known to have population structure (Ford et al., 2015). Sampling around site 14 would help to understand the distribution and abundancy of population 14.

References

Abzhanov, A. (2010). Darwin's Galápagos finches in modern biology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, [online] 365(1543), pp.1001–1007. doi:https://doi.org/10.1098/rstb.2009.0321.

Aguillon, S.M., Dodge, T.O., Preising, G.A. and Schumer, M. (2022). Introgression. *Current Biology*, 32(16), pp.R865–R868. doi:https://doi.org/10.1016/j.cub.2022.07.004.

Albertson, R.C. and Kocher, T.D. (2006). Genetic and developmental basis of cichlid trophic diversity. *Heredity*, 97(3), pp.211–221. doi:https://doi.org/10.1038/sj.hdy.6800864.

Allio, R., Donega, S., Galtier, N. and Nabholz, B. (2017). Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Molecular Biology and Evolution*, 34(11), pp.2762–2772. doi:https://doi.org/10.1093/molbev/msx197.

Amador, L.I., Simmons, N.B. and Giannini, N.P. (2019). Aerodynamic reconstruction of the primitive fossil bat Onychonycteris finneyi (Mammalia: Chiroptera). *Biology Letters*, 15(3), p.20180857. doi:https://doi.org/10.1098/rsbl.2018.0857.

Balentine, C.M. and Bolnick, D.A. (2022). Parallel evolution in human populations: A biocultural perspective. *Evolutionary Anthropology: Issues, News, and Reviews*. doi:https://doi.org/10.1002/evan.21956.

Barrera-Guzmán, A.O., Aleixo, A., Shawkey, M.D. and Weir, J.T. (2017). Hybrid speciation leads to novel male secondary sexual ornamentation of an Amazonian bird. *Proceedings of the National Academy of Sciences of the United States of America*, 115(2). doi:https://doi.org/10.1073/pnas.1717319115.

Bayona, J.D.R. (2006a). *The IUCN Red List of Threatened Species*, *IUCN Red List of Threatened Species*. Available at: https://www.iucnredlist.org/species/60452/12368336 (Accessed: 27 October 2023).

Bayona, J.& A. (2006b) *The IUCN Red List of Threatened Species*, *IUCN Red List of Threatened Species*. Available at: https://www.iucnredlist.org/species/60453/12368415 (Accessed: 27 October 2023).

Bayona, J.D.R. (2006c). *IUCN Red List of Threatened Species: Oreochromis amphimelas*. [online] IUCN Red List of Threatened Species. Available at: https://www.iucnredlist.org/species/60629/12388607.

Beavis, W.D. (1995) The power and deceit of QTL experiments: Lessons from comparative QTL studies. Proceedings of the Forty-ninth Annual Corn and Sorghum Industry Research Conference ASTA, Washington, 252-268.

Benton, M.J., Wilf, P. and Sauquet, H. (2021). The Angiosperm Terrestrial Revolution and the origins of modern biodiversity. *New Phytologist*, 233(5). doi:https://doi.org/10.1111/nph.17822.

Biopython. (2022). Biopython 1.79-foss-2022a [Software]. University of Birmingham. https://bear-apps.bham.ac.uk/applications/2022a/Biopython/1.79-foss-2022a/

Burgarella, C., Barnaud, A., Kane, N.A., Jankowski, F., Scarcelli, N., Billot, C., Vigouroux, Y. and Berthouly-Salazar, C. (2019). Adaptive Introgression: An Untapped Evolutionary Mechanism for Crop Adaptation. *Frontiers in Plant Science*, 10. doi:https://doi.org/10.3389/fpls.2019.00004.

Butlin, R.K., Galindo, J. and Grahame, J.W. (2008). Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1506), pp.2997–3007. doi:https://doi.org/10.1098/rstb.2008.0076.

Castric V, Bechsgaard J, Schierup MH, Vekemans X (2008). Repeated adaptive introgression at a gene under multiallelic balancing selection. PLoS Genet. 4. e1000168 p. https://doi.org/1000110.1001371/journal.pgen.1000168

Chehida, Y.B. (2024). *Ithominii_convergence/GWAS/GWAS_Single_SP at main · yacinebenchehida/Ithominii_convergence*. [online] GitHub. Available at: https://www.google.com/url?q=https://github.com/yacinebenchehida/Ithominii_convergence/tree/main/GWAS/GWAS_Single_SP&source=gmail&ust=1718445306196000&usg=AOvVaw3I1SkZ3qUSoBk1rwNU4W3S [Accessed 14 Jun. 2024].

Chen, Z.-L., Meng, J.-M., Cao, Y., Yin, J.-L., Fang, R.-Q., Fan, S.-B., Liu, C., Zeng, W.-F., Ding, Y.-H., Tan, D., Wu, L., Zhou, W.-J., Chi, H., Sun, R.-X., Dong, M.-Q. and He, S.-M. (2019). A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nature Communications*, 10(1). doi:https://doi.org/10.1038/s41467-019-11337-z.

Chu, Y.-H., Hardin, H., Zhang, R., Guo, Z. and Lloyd, R.V. (2019). In situ hybridization: Introduction to techniques, applications and pitfalls in the performance and interpretation of assays. *Seminars in Diagnostic Pathology*, 36(5), pp.336–341. doi:https://doi.org/10.1053/j.semdp.2019.06.004.

Ciezarek, A.G. *et al.* (2023). *Ancient and ongoing hybridization in the Oreochromis cichlid fishes* [Preprint]. doi:10.1101/2023.05.19.541459.

Clabaut, C., Bunje, P.M.E., Salzburger, W. and Meyer, A. (2007). GEOMETRIC MORPHOMETRIC ANALYSES PROVIDE EVIDENCE FOR THE ADAPTIVE CHARACTER

OF THE TANGANYIKAN CICHLID FISH RADIATIONS. *Evolution*, 61(3), pp.560–578. doi:https://doi.org/10.1111/j.1558-5646.2007.00045.x.

Clarisse, L., Van Damme, M., Gardner, W., Coheur, P.-F. ., Clerbaux, C., Whitburn, S., Hadji-Lazaro, J. and Hurtmans, D. (2019). Atmospheric ammonia (NH3) emanations from Lake Natron's saline mudflats. *Scientific Reports*, 9(1). doi:https://doi.org/10.1038/s41598-019-39935-3.

Coe, M.J. (1966). *The biology of tilapia grahami boulenger in Lake Magadi, Kenya*. Basel: Verlag für Recht und Gesellschaft.

Coyne, J.A. (1985). The genetic basis of Haldane's rule. *Nature*, 314(6013), pp.736–738. doi:https://doi.org/10.1038/314736a0.

Coyne, J.A. and Orr, H.A. (1989). Patterns of Speciation in Drosophila. *Evolution*, 43(2), p.362. doi:https://doi.org/10.2307/2409213.

Schluter, D. (2000). *The Ecology of Adaptive Radiation*. OUP Oxford.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M. and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). doi:https://doi.org/10.1093/gigascience/giab008.

Dawson, J.B. (2008). Chapter 6 Rift-associated sedimentary basins. *Geological Society, London, Memoirs*, 33(1), pp.33–38. doi:https://doi.org/10.1144/m33.6.

Elmer, K.R., Fan, S., Kusche, H., Luise Spreitzer, M., Kautt, A.F., Franchini, P. and Meyer, A. (2014). Parallel evolution of Nicaraguan crater lake cichlid fishes via non-parallel routes. *Nature Communications*, 5(1). doi:https://doi.org/10.1038/ncomms6168.

Elmer, K.R., Reggio, C., Wirth, T., Verheyen, E., Salzburger, W. and Meyer, A. (2009). 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*, 106(32), pp.13404–13409. doi:https://doi.org/10.1073/pnas.0902299106.

Ensemble (2024a). *Region in Detail*. [online] uswest.ensembl.org. Available at: https://www.ensembl.org/Oreochromis_niloticus/Location/View?db=core;r=LG11:5800000-6800000 [Accessed 10 Jan. 2024]. *Oreochromis niloticus* genome browser.

Ensemble (2024b). *Region in Detail*. [online] uswest.ensembl.org. Available at: https://www.ensembl.org/Oreochromis_niloticus/Location/View?db=core;r=LG11:31000000-33000000 [Accessed 10 Jan. 2024]. *Oreochromis niloticus* genome browser.

Ensemble (2024c). *Region in Detail.* [online] uswest.ensemble.org. Available at: https://www.ensembl.org/Oreochromis_niloticus/Location/View?db=core;r=LG5:4400000-5400000 [Accessed 10 Jan. 2024. *Oreochromis niloticus* genome browser

Feller, A.F. and Seehausen, O. (2022). Genetic architecture of adaptive radiation across two trophic levels. *Bern Open Repository and Information System (University of Bern)*, 289(1974). doi:https://doi.org/10.1098/rspb.2022.0377.

Ferdous, S. and Armbruster, J. (n.d.). *ACSI II Guide to Geometric Morphometrics in MorphoJ ACSI II Guide to Geometric Morphometrics in MorphoJ*. [online] Available at: https://webhome.auburn.edu/~armbrjw/gmguide.pdf [Accessed 18 Jun. 2024].

Fernández-Palacios, J.M., Kreft, H., Irl, S.D.H., Norder, S., Ah-Peng, C., Borges, P.A.V., Burns, K.C., de Nascimento, L., Meyer, J.-Y., Montes, E. and Drake, D.R. (2021). Scientists' warning – The outstanding biodiversity of islands is in peril. *Global Ecology and Conservation*, 31, p.e01847. doi:https://doi.org/10.1016/j.gecco.2021.e01847.

Ford, A.G.P. (2015). *Evolutionary relationships of East African soda lake cichlid fish*. [Doctoral Thesis] pp.1–351. Available at: https://discovery.ucl.ac.uk/id/eprint/1471875/ [Accessed 24 Jan. 2024].

Ford, A.G.P., Dasmahapatra, K.K., Rüber, L., Gharbi, K., Cezard, T. and Day, J.J. (2015). High levels of interspecific gene flow in an endemic cichlid fish adaptive radiation from an extreme lake environment. *Molecular Ecology*, 24(13), pp.3421–3440. doi:https://doi.org/10.1111/mec.13247.

Friedman, M., Keck, B.P., Dornburg, A., Eytan, R.I., Martin, C.H., Hulsey, C.D., Wainwright, P.C. and Near, T.J. (2013). Molecular and fossil evidence place the origin of cichlid fishes long after Gondwanan rifting. *Proceedings of the Royal Society B: Biological Sciences*, 280(1770), p.20131733. doi:https://doi.org/10.1098/rspb.2013.1733.

Garcin, Y., Vincens, A., Williamson, D., Buchet, G. and Guiot, J. (2007). Abrupt resumption of the African Monsoon at the Younger Dryas—Holocene climatic transition. *Quaternary Science Reviews*, 26(5-6), pp.690–704. doi:https://doi.org/10.1016/j.quascirev.2006.10.014.

Getenet, M., OtáloraF., Emmerling, F., Al-Sabbagh, D. and García-Ruiz, J.M. (2023). Mineral precipitation and hydrochemical evolution through evaporitic processes in soda brines (East African Rift Valley). *Chemical Geology*, 616, pp.121222–121222. doi:https://doi.org/10.1016/j.chemgeo.2022.121222.

Gibert, J. and Deharveng, L. (2002). Subterranean Ecosystems: A Truncated Functional Biodiversity. *BioScience*, 52(6), p.473. doi:https://doi.org/10.1641/0006-3568(2002)052[0473:seatfb]2.0.co;2.

gibert-Fab (2022). *gibert-Fab/ABBA-BABA*. [online] GitHub. Available at: https://github.com/gibert-Fab/ABBA-BABA?tab=readme-ov-file [Accessed 31 Jan. 2024].

Google Earth Pro, Historical Maps. (1995). Lake Natron. 2°20’36.80" S 36°02’45.04" E. 603m elevation.

Google Earth Pro, Historical Maps. (2016). Lake Natron. 2°20’36.80" S 36°02’45.04" E. 603m elevation.

Google Earth Pro, Historical Maps. (2017). Lake Natron. 2°20’36.80" S 36°02’45.04" E. 603m elevation.

Google Maps. (2023). *Lake Natron*. [online] Available at: https://www.google.com/maps/place/Lake+Natron/@-2.0407699 [Accessed 16 Sep. 2023].

Google Maps. (2024). *Lake Natron*. [online] Available at: https://www.google.com/maps/place/Lake+Natron/@-2.0407699 [Accessed 16 Jan. 2024].

Goytain, A., Hines, R.M. and Quamme, G.A. (2008). Functional characterization of NIPA2, a selective Mg^{2+} transporter. *American Journal of Physiology-Cell Physiology*, 295(4), pp.C944–C953. doi:https://doi.org/10.1152/ajpcell.00091.2008.

Grant, P.R. and Grant, B.R. (2015). Introgressive hybridization and natural selection in Darwin’s finches. *Biological Journal of the Linnean Society*, 117(4), pp.812–822. doi:https://doi.org/10.1111/bij.12702.

Hanssens, M.& S. (2006) *The IUCN Red List of Threatened Species*, *IUCN Red List of Threatened Species*. Available at: https://www.iucnredlist.org/species/60319/12348298#assessment-information (Accessed: 27 October 2023).

Hulsey, C.D., León, F.J.G. and Rodiles-Hernández, R. (2006). MICRO- AND MACROEVOLUTIONARY DECOUPLING OF CICHLID JAWS: A TEST OF LIEM’S KEY INNOVATION HYPOTHESIS. *Evolution*, 60(10), pp.2096–2109. doi:https://doi.org/10.1111/j.0014-3820.2006.tb01847.x.

Illumina HiSeq 2500 paired end sequencing; Oreochromis SNP Panel - Sra - NCBI (no date) *National Center for Biotechnology Information*. Available at: https://www.ncbi.nlm.nih.gov/sra/ERR7224546 (Accessed: 25 October 2023).

Irisarri, I., Singh, P., Koblmüller, S., Torres-Dowdall, J., Henning, F., Franchini, P., Fischer, C., Lemmon, A.R., Lemmon, E.M., Thallinger, G.G., Sturmbauer, C. and Meyer, A. (2018). Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. *Nature Communications*, [online] 9(1), p.3159. doi:https://doi.org/10.1038/s41467-018-05479-9.

IUCN (n.d.). *SUMMARY OF THE FIVE CRITERIA (A-E) USED TO EVALUATE IF A TAXON BELONGS IN AN IUCN RED LIST THREATENED CATEGORY (CRITICALLY ENDANGERED, ENDANGERED OR VULNERABLE).1*. [online] Iucnredlist.org. Available at:

https://nc.iucnredlist.org/redlist/content/attachment_files/summary_sheet_en_web.pdf [Accessed 24 Jan. 2024].

Iucnredlist.org. (2013). *SUMMARY OF THE FIVE CRITERIA (A-E) USED TO EVALUATE IF A TAXON BELONGS IN AN IUCN RED LIST THREATENED CATEGORY (CRITICALLY ENDANGERED, ENDANGERED OR VULNERABLE)*. [online] Available at: https://nc.iucnredlist.org/redlist/content/attachment_files/summary_sheet_en_web.pdf.

Jae Young Choi, Dai, X., Alam, O., Peng, J., Priyesh Rughani, Hickey, S.E., Harrington, E., Juul, S., Ayroles, J.F., Purugganan, M.D. and Stacy, E.A. (2021). Ancestral polymorphisms shape the adaptive radiation of*Metrosideros*across the Hawaiian Islands. *Proceedings of the National Academy of Sciences of the United States of America*, 118(37). doi:https://doi.org/10.1073/pnas.2023801118.

Jia, S., Kwon, H.-J.E., Lan, Y., Zhou, J., Liu, H. and Jiang, R. (2016). Bmp4-Msx1 signaling and Osr2 control tooth organogenesis through antagonistic regulation of secreted Wnt antagonists. *Developmental Biology*, [online] 420(1), pp.110–119. doi:https://doi.org/10.1016/j.ydbio.2016.10.001.

Johannsson, O.E., Bergman, H.L., Wood, C.M., Laurent, P., Kavembe, D.G., Bianchini, A., Maina, J.N., Chevalier, C., Bianchini, L.F., Papah, M.B. and Ojoo, R.O. (2014). Air breathing in Magadi tilapiaAlcolapia grahami, under normoxic and hyperoxic conditions, and the association with sunlight and reactive oxygen species. *Journal of Fish Biology*, 84(3), pp.844–863. doi:https://doi.org/10.1111/jfb.12289.

Johnson, T.C., Kelts, K. and Odada, E. (2000). The Holocene History of Lake Victoria. *AMBIO: A Journal of the Human Environment*, 29(1), pp.2–11. doi:https://doi.org/10.1579/0044-7447-29.1.2.

Joint Research Centre (2013) *Africa Soil Atlas - Part 2: Soil Information*. Available at: https://esdac.jrc.ec.europa.eu/Library/Maps/Africa_Atlas/Documents/JRC_africa_soil_atlas_part2.pdf (Accessed: 11 January 2024).

Jønsson, K.A., Fabre, P.-H., Fritz, S.A., Etienne, R.S., Ricklefs, R.E., Jørgensen, T.B., Fjeldså, J., Rahbek, C., Ericson, P.G.P., Woog, F., Pasquet, E. and Irestedt, M. (2012). Ecological and evolutionary determinants for the adaptive radiation of the Madagascan vangas. *Proceedings of the National Academy of Sciences*, [online] 109(17), pp.6620–6625. doi:https://doi.org/10.1073/pnas.1115835109.

Kautt, A.F., Kratochwil, C.F., Nater, A., Machado-Schiaffino, G., Olave, M., Henning, F., Torres-Dowdall, J., Härer, A., Hulsey, C.D., Franchini, P., Pippel, M., Myers, E.W. and Meyer, A. (2020). Contrasting signatures of genomic divergence during sympatric speciation. *Nature*, [online] pp.1–6. doi:https://doi.org/10.1038/s41586-020-2845-0.

Kavembe, G.D., Kautt, A.F., Machado-Schiaffino, G. and Meyer, A. (2016). Eco-morphological differentiation in Lake Magadi tilapia, an extremophile cichlid fish living in hot, alkaline and hypersaline lakes in East Africa. *Molecular Ecology*, 25(7), pp.1610–1625. doi:https://doi.org/10.1111/mec.13461.

Keppeler, F.W., Montaña, C.G. and Winemiller, K.O. (2020). The relationship between trophic level and body size in fishes depends on functional traits. *Ecological Monographs*, 90(4). doi:https://doi.org/10.1002/ecm.1415.

Kevrekidis, C. *et al.* (2020) 'New cichlid fossils from the middle-late miocene alkaline lakes of Africa', *Journal of Vertebrate Paleontology*, 40(4). doi:10.1080/02724634.2020.1805621.

Kevrekidis, C., Ruthensteiner, B., Cerwenka, A.F., Penk, S. and Reichenbacher, B. (2020). New Cichlid Fossils from the Middle-Late Miocene Alkaline Lakes of Africa. *Journal of Vertebrate Paleontology*, 40(4), pp.e1805621–e1805621. doi:https://doi.org/10.1080/02724634.2020.1805621.

Kimura, M. (1991). The neutral theory of molecular evolution: A review of recent evidence. *The Japanese Journal of Genetics*, 66(4), pp.367–386. doi:https://doi.org/10.1266/jjg.66.367.

Klett, V. and Meyer, A. (2002). What, if Anything, is a Tilapia?—Mitochondrial ND2 Phylogeny of Tilapiines and the Evolution of Parental Care Systems in the African Cichlid Fishes. *Molecular Biology and Evolution*, 19(6), pp.865–883. doi:https://doi.org/10.1093/oxfordjournals.molbev.a004144.

Klingenberg, C.P., (2011). MorphoJ: an integrated software package for geometric morphometrics. Molecular Ecology Resources, 11, pp.353-357.

Kocher, T.D. (2004). Adaptive evolution and explosive speciation: the cichlid fish model. *Nature Reviews Genetics*, 5(4), pp.288–298. doi:https://doi.org/10.1038/nrg1316.

Kopf, R.K., Yen, J.D.L., Nimmo, D.G., Brosse, S. and Villéger, S. (2020). Global patterns and predictors of trophic position, body size and jaw size in fishes. *Global Ecology and Biogeography*, 30(2), pp.414–428. doi:https://doi.org/10.1111/geb.13227.

Krause, C., Korchynskyi, O., de Rooij, K., Weidauer, S.E., de Gorter, D., van Bezooijen, R., Hatsell, S.J., Economides, A.N., Müller, T., Löwik, C.W.G.M. and ten Dijke, P. (2010). Distinct Modes of Inhibition by Sclerostin on Bone Morphogenetic Protein and Wnt Signaling Pathways. *Journal of Biological Chemistry*, 285(53), pp.41614–41626. doi:https://doi.org/10.1074/jbc.m110.153890.

Kumar, S. (2005). Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, 6(8), pp.654–662. doi:https://doi.org/10.1038/nrg1659.

Kusche, H., Elmer, K.R. and Meyer, A. (2015). Sympatric ecological divergence associated with a color polymorphism. *BMC Biology*, 13(1). doi:https://doi.org/10.1186/s12915-015-0192-7.

Lam, P.-Y., Kamei, C.N., Mangos, S., Mudumana, S., Liu, Y. and Drummond, I.A. (2013). odd-skipped related 2 is required for fin chondrogenesis in zebrafish. *Developmental Dynamics*, 242(11), pp.1284–1292. doi:https://doi.org/10.1002/dvdy.24026.

Lameck, A.S., Skutai, J. and Boros, E. (2023). Review of chemical properties of inland soda and saline waters in East Africa (rift valley region). *Journal of Hydrology: Regional Studies*, 46, pp.101323–101323. doi:https://doi.org/10.1016/j.ejrh.2023.101323.

Lan, Y., Kingsley, P.D., Cho, E.-S. and Jiang, R. (2001). Osr2, a new mouse gene related to Drosophila odd-skipped, exhibits dynamic expression patterns during craniofacial, limb, and kidney development. *Mechanisms of Development*, 107(1-2), pp.175–179. doi:https://doi.org/10.1016/s0925-4773(01)00457-9.

Lawson, D. (2012). *The world's most powerful ancestry inference software*. [online] www.paintmychromosomes.com. Available at: http://www.paintmychromosomes.com/ [Accessed 18 Jun. 2024]. Code for building the fineRADstructure plot is available to download on this hyperlink in the folder "FineSTRUCTURE R scripts" followed by "FinestructureRcode.zip" which contains the necessary file 'FinestructureLibrary' produced in 2016.

Lawson, M., (2023). *An integrative approach to studying speciation in a recent adaptive radiation.* PhD Thesis, University of York.

Lawson, M., Hayle, M., Shechonge, A.H., Nyingi, W.D., Ford, A., Hoffman, J.I., Day, J.J., Turner, G.F. and Dasmahapatra, K.K. (2023). Sympatric and allopatric Alcolapia soda lake cichlid species show similar levels of assortative mating. *Frontiers in Ecology and Evolution*, 11. doi:https://doi.org/10.3389/fevo.2023.1150083.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

Li, H. (2011). Tabix: Fast retrieval of sequence features from generic TAB-delimited files.

Li, H. and Bonfield, J. (2023). *samtools-depth(1) manual page*. [online] www.htslib.org. Available at: https://www.htslib.org/doc/samtools-depth.html [Accessed 24 Jan. 2024].

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Liem, K.F. (1973). Evolutionary Strategies and Morphological Innovations: Cichlid Pharyngeal Jaws. *Systematic Zoology*, 22(4), p.425. doi:https://doi.org/10.2307/2412950.

Losos, J. B. (2011). Lizards in an evolutionary tree: ecology and adaptive radiation of anoles, Vol. 10. California: Univ of California Press.

MacFadden BJ. (2005). Fossil horses—evidence for evolution. Science 307(5716):1728–30

Malinsky, M. (n.d.). *fineRADstructure*. [online] Milan Malinsky. Available at: https://www.milan-malinsky.org/fineradstructure [Accessed 18 Jun. 2024].

Malinsky, M. (2024). *Dsuite/README.md at master · millanek/Dsuite*. [online] GitHub. Available at: https://github.com/millanek/Dsuite/blob/master/README.md [Accessed 12 Jan. 2024].

Malinsky, M. and Trucchi, E. (2020). *fineradstructure-activity*. [online] Evolution and Genomics. Available at: https://evomics.org/learning/population-and-speciation-genomics/2020-population-and-speciation-genomics/fineradstructure-activity/ [Accessed 6 Nov. 2023].

Malinsky, M., Emiliano Trucchi, Lawson, D. and Falush, D. (2018). RADpainter and fineRADstructure: Population Inference from RADseq Data. *Molecular Biology and Evolution*, 35(5), pp.1284–1290. doi:https://doi.org/10.1093/molbev/msy023.

Malinsky, M., Matschiner, M. and Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, [online] 21(2), pp.584–595. doi:https://doi.org/10.1111/1755-0998.13265.

Malinsky, M., Challis, R.J., Tyers, A.M., Schiffels, S., Terai, Y., Ngatunga, B.P., Miska, E.A., Durbin, R., Genner, M.J. and Turner, G.F. (2015). Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science (New York, N.Y.)*, [online] 350(6267), pp.1493–1498. doi:https://doi.org/10.1126/science.aac9927.

Marques, D.A., Meier, J.I. and Seehausen, O. (2019). A Combinatorial View on Speciation and Adaptive Radiation. *Trends in Ecology & Evolution*, 34(6), pp.531–544. doi:https://doi.org/10.1016/j.tree.2019.02.008.

Martin, S., (2023a). s*imonhmartin/genomics_general.* [online] GitHub. Available at:https://github.com/simonhmartin/genomics_general#trees-for-sliding-windows. [Accessed 27 Apr. 2023].

Martin, S. (2023b). *simonhmartin/genomics_general*. [online] GitHub. Available at: https://github.com/simonhmartin/genomics_general#diversity-and-divergence-analyses-in-sliding-windows [Accessed 4 Jan. 2024].

Martin, S. (2024). *simonhmartin/genomics_general*. [online] GitHub. Available at: https://github.com/simonhmartin/genomics_general/tree/master?tab=readme-ov-file#abba-baba-statistics-in-sliding-windows Note: See the section title ‘ABBA-BABA statistics in sliding windows’ and change the *fd* flag to f_{dM} to run f_{dM} analysis.

Masonick, P., Meyer, A. and Hulsey, C.D. (2022). Phylogenomic Analyses Show Repeated Evolution of Hypertrophied Lips Among Lake Malawi Cichlid Fishes. *Genome Biology and Evolution*, 14(4). doi:https://doi.org/10.1093/gbe/evac051.

Matschiner, M., Böhne, A., Ronco, F. and Salzburger, W. (2020). The genomic timeline of cichlid fish diversification across continents. *Nature Communications*, 11(1). doi:https://doi.org/10.1038/s41467-020-17827-9.

McGee, M.D. *et al.* (2020) ‘The ecological and genomic basis of explosive adaptive radiation’, *Nature*, 586(7827), pp. 75–79. doi:10.1038/s41586-020-2652-7.

McKaye, K.R. and Kocher, T. (1983). Head ramming behaviour by three paedophagous cichlids in Lake Malawi, Africa. *Animal Behaviour*, 31(1), pp.206–210. doi:https://doi.org/10.1016/s0003-3472(83)80190-0.

Meier, J.I. *et al.* (2023). Cycles of fusion and fission enabled rapid parallel adaptive radiations in African cichlids, *Science*, 381(6665). doi:10.1126/science.ade2833.

Meier, J.I., Marques, D.A., Mwaiko, S., Wagner, C.E., Excoffier, L. and Seehausen, O. (2017). Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, 8(1).

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A. and Johnson, E.A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2), pp.240–248. doi:https://doi.org/10.1101/gr.5681207.

Mishmar, D. (2020) ‘MtDNA in the crossroads of evolution and disease’, *Nature Reviews Molecular Cell Biology*, 21(4), pp. 181–181. doi:10.1038/s41580-020-0213-4.

Nevado, B. *et al.* (2019) ‘Adaptive evolution is common in rapid evolutionary radiations’, *Current* no. 13, June 2015, pp. 3421–40, doi:https://doi.org/10.1111/mec.13247.

Ngoepe, N., Muschick, M, Kishe, M.A., Mwaiko, S., Temoltzin-Loranca,Y., King, L., Mustaphi, C.C., Heiri, O., Wienhues G., Vogel, H., Cuenca-Cambronero, M., Tinner, W., Grosjean, M., Matthews, B. and Seehausen, O. (2023). A continuous fish fossil record reveals key insights into adaptive radiation. *Nature*, 622(7982), pp.315–320. doi:https://doi.org/10.1038/s41586-023-06603-6.

Ngoepe, N., Muschick, M., Kishe, M.A., Mwaiko, S., Temoltzin-Loranca, Y., King, L., Mustaphi, C.C., Heiri, O., Wienhues, G., Vogel, H., Cambronero, M.C., Tinner, W., Grosjean, M., Matthews, B. and Seehausen, O. (2023). A continuous fish fossil record reveals key insights into adaptive radiation. *Nature*, 622(7982), pp.315–320. doi:https://doi.org/10.1038/s41586-023-06603-6.

Nie, Y., Fu, Y.-T., Wang, W., Li, R., Tang, W.-Q. and Li, G. (2022). Comparative analyses of the fragmented mitochondrial genomes of wild pig louse Haematopinus apri from China and Japan. *International Journal for Parasitology: Parasites and Wildlife*, 18, pp.25–29. doi:https://doi.org/10.1016/j.ijppaw.2022.03.013.

Noor, M.A.F., Grams, K.L., Bertucci, L.A. and Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, [online] 98(21), pp.12084–12088. doi:https://doi.org/10.1073/pnas.221274498.

Oreochromis Niloticus Genome Assembly o_niloticus_umd_nmbu - NCBI - NLM (no date) *National Center for Biotechnology Information*. Available at: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_001858045.2/ (Accessed: 25 October 2023).

Oreochromis niloticus genome assembly O_niloticus_UMD_NMBU[Internet]. NCBI. 2018 [cited 2022 Dec 22]. Available from: https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_001858045.2/

Orr, H.A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics*, 139(4), pp.1805–1813. doi:https://doi.org/10.1093/genetics/139.4.1805.

Ortiz, E.M. (2019). vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. DOI:10.5281/zenodo.2540861

Ottenburghs, J. (2022). Avian introgression patterns are consistent with Haldane's Rule. *Journal of Heredity*, 113(4). doi:https://doi.org/10.1093/jhered/esac005.

Pardo-Diaz, C., Salazar, C., Baxter, S.W., Mérot, C., Wilsea M.B. Figueiredo-Ready, Mathieu Joron, W. Owen McMillan and Jiggins, C.D. (2012). Adaptive Introgression across Species Boundaries in Heliconius Butterflies. *Plos Genetics*, 8(6), pp.e1002752–e1002752. doi:https://doi.org/10.1371/journal.pgen.1002752.

Parsons, K.J., Trent Taylor, A., Powder, K.E. and Albertson, R.C. (2014). Wnt signalling underlies the evolution of new phenotypes and craniofacial variability in Lake Malawi cichlids. *Nature Communications*, 5(1). doi:https://doi.org/10.1038/ncomms4629.

Patton, A.H., Margres, M.J., Epstein, B., Eastman, J., Harmon, L.J. and Storfer, A. (2020). Hybridizing salamanders experience accelerated diversification. *Scientific Reports*, 10(1). doi:https://doi.org/10.1038/s41598-020-63378-w.

Pauers, M.J. (2016). Two New and Remarkably Similarly Colored Species of*Labeotropheus*(Perciformes: Cichlidae) from Lake Malaŵi, Africa. *Copeia*, 104(3), pp.628–638. doi:https://doi.org/10.1643/ci-15-360.

Pauers, M.J. and Phiri, T.B. (2023). Six New Species of Labeotropheus (Cichliformes: Cichlidae) from the Malaŵian Shore of Lake Malaŵi, Africa. *Ichthyology & herpetology*, 111(2). doi:https://doi.org/10.1643/i2021055.

Pearson, J.A., Haddow, S.D., Hillson, S.W., Knüsel, C.J., Larsen, C.S. and Sadvari, J.W. (2015). Stable carbon and nitrogen isotope analysis and dietary reconstruction through the life course at Neolithic Çatalhöyük, Turkey. *Journal of Social Archaeology*, 15(2), pp.210–232. doi:https://doi.org/10.1177/1469605315582983.

Picard Toolkit., (2019). Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/; Broad Institute

Portik, D.M., Bell, R.C., Blackburn, D.C., Bauer, A.M., Barratt, C.D., Branch, W.R., Burger, M., Channing, A., Colston, T.J., Conradie, W., Dehling, J.M., Drewes, R.C., Ernst, R., Greenbaum, E., Gvoždík, V., Harvey, J., Hillers, A., Hirschfeld, M., Jongsma, G.F.M. and Kielgast, J. (2019). Sexual Dichromatism Drives Diversification within a Major Radiation of African Amphibians. *Systematic Biology*, 68(6), pp.859–875. doi:https://doi.org/10.1093/sysbio/syz023.

Posit team (2023). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

Powder, K.E. and Albertson, R.C. (2016). Cichlid fishes as a model to understand normal and clinical craniofacial variation. *Developmental Biology*, 415(2), pp.338–346. doi:https://doi.org/10.1016/j.ydbio.2015.12.018.

Rathan, N.D., Krishna, H., Ellur, R.K., Sehgal, D., Govindan, V., Ahlawat, A.K., Krishnappa, G., Jaiswal, J.P., Singh, J.B., SV, S., Ambati, D., Singh, S.K., Bajpai, K. and Mahendru-Singh, A. (2022). Genome-wide association study identifies loci and candidate genes for grain micronutrients and quality traits in wheat (Triticum aestivum L.). *Scientific Reports*, 12(1). doi:https://doi.org/10.1038/s41598-022-10618-w.

Ravinet, M. and Meier, J. (no date) *Filtering and handling vcfs*, *Speciation & Population Genomics: a how-to-guide*. Available at: https://speciationgenomics.github.io/filtering_vcfs/ (Accessed: 07 June 2024).

Ravinet, M. and Meier, J. (2024). *Sliding window differentiation, variance and introgression*. [online] Speciation & Population Genomics: a how-to-guide. Available at: https://speciationgenomics.github.io/sliding_windows/.

Ravinet, M., and Meier, J. (2022) “Population Structure: PCA.” *Speciation & Population Genomics: A How To-Guide*, speciationgenomics.github.io/pca/. Accessed 28 Apr. 2023.

Reite, O.B., Maloiy, G.MO and Aasenhaug, B. (1974). pH, Salinity and Temperature Tolerance of Lake Magadi Tilapia. *Nature*, 247(5439), pp.315–315. doi:https://doi.org/10.1038/247315a0.

Rieseberg, L.H. (1997) ‘Hybrid origins of plant species’, *Annual Review of Ecology and Systematics*, 28(1), pp. 359–389. doi:10.1146/annurev.ecolsys.28.1.359.

Rieseberg, L.H., Archer, M.A. and Wayne, R.K. (1999). Transgressive segregation, adaptation and speciation. *Heredity*, 83(4), pp.363–372. doi:https://doi.org/10.1038/sj.hdy.6886170.

Rieseberg, L.H., Widmer, A., Arntz, A.M. and Burke, B. (2003). The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1434), pp.1141–1147. doi:https://doi.org/10.1098/rstb.2003.1283.

Roberts, N., Taieb, M., Barker, P., Damnati, B., Icole, M. and Williamson, D., 1993. Timing of the Younger Dryas event in East Africa from lake-level changes. Nature, 366(6451), pp.146-148.

Rohlf, F.J. (2023). *Data acquisition*. [online] www.sbmorphometrics.org. Available at: https://www.sbmorphometrics.org/soft-dataacq.html [Accessed 24 Jan. 2024].

Rohlf, F.J. and Slice, D. (1990). Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. *Systematic Zoology*, 39(1), p.40. doi:https://doi.org/10.2307/2992207.

Rubin, C.-J., Enbody, E.D., Dobreva, M.P., Arhat Abzhanov, Davis, B.W., Sangeet Lamichhaney, Pettersson, M.E., C. Grace Sprehn, Valle, C.A., Vasco, K., Wallerman, O., B. Rosemary Grant, Grant, P.R. and Andersson, L. (2021). Darwin's finches - an adaptive radiation constructed from ancestral genetic modules. *bioRxiv (Cold Spring Harbor Laboratory)*. doi:https://doi.org/10.1101/2021.09.17.460815.

Rohlf, F.J., (2013). tpsDig, version 2.17. Department of Ecology and Evolution, State University of New York at Stony Brook, New York.

Rubinoff, D. and Holland, B.S. (2005). Between Two Extremes: Mitochondrial DNA is neither the Panacea nor the Nemesis of Phylogenetic and Taxonomic Inference. *Systematic Biology*, [online] 54(6), pp.952–961. doi:https://doi.org/10.1080/10635150500234674.

Sadier, A., Urban, D.J., Anthwal, N., Howenstine, A.O., Sinha, I. and Sears, K.E. (2021). Making a bat: The developmental basis of bat evolution. *Genetics and Molecular Biology*, [online] 43(1 Suppl 2). doi:https://doi.org/10.1590/1678-4685-GMB-2019-0146.

Saita, S., Shirane, M., Tohru Ishitani, Shimizu, N. and Nakayama, K.I. (2014). Role of the ANKMY2-FKBP38 Axis in Regulation of the Sonic Hedgehog (Shh) Signaling Pathway. *Journal of Biological Chemistry*, 289(37), pp.25639–25654. doi:https://doi.org/10.1074/jbc.m114.558635.

Sánchez-Guillén, R.A., Córdoba-Aguilar, A., Cordero-Rivera, A. and Wellenreuther, M. (2014). Rapid evolution of prezygotic barriers in non-territorial damselflies. *Biological Journal of the Linnean Society*, 113(2), pp.485–496. doi:https://doi.org/10.1111/bij.12347.

Sardell, J.M. and Uy, J.A.C. (2016). Hybridization following recent secondary contact results in asymmetric genotypic and phenotypic introgression between island species ofMyzomelahoneyeaters. *Evolution*, 70(2), pp.257–269. doi:https://doi.org/10.1111/evo.12864.

Scientific Names where Genus Equals Oreochromis (no date) *Fishbase.mnhn.fr*. Available at: https://fishbase.mnhn.fr/Nomenclature/ScientificNameSearchList.php (Accessed: 26 June 2023).

Seegers, L., Sonnenberg, R., &; Tichy, H. (2001). The Alcolapia group, a remarkable species flock*, Journal of aquaculture and aquatic sciences*, 289, pp.335-364

Seegers, L., Sonnenberg, R., and Yamamoto, R. (1999). Molecular analysis of the Alcolapia flock from lakes Natron and Magadi, Tanzania and Kenya (Teleostei: Cichlidae), and implications for their systematics and evolution. Ichthyol. Explor. Freshw. 10, 175–199.

Seegers,L., Tichy, H. (1999). Exploration of Freshwaters. Vols. 10, Nos. 2. pp 97-200.

Seehausen, O. (2004). 'Hybridization and adaptive radiation', *Trends in Ecology & Evolution*, 19(4), pp. 198–207. doi:10.1016/j.tree.2004.01.003.

Seehausen, O. (2004). Hybridization and adaptive radiation. *Trends in Ecology & Evolution*, 19(4), pp.198–207. doi:https://doi.org/10.1016/j.tree.2004.01.003.

Seehausen, O. (2013) 'Conditions when hybridization might predispose populations for adaptive radiation', *Journal of Evolutionary Biology*, 26(2), pp. 279–281. doi:10.1111/jeb.12026.

Seehausen, O. and Schluter, D. (2004). Male–male competition and nuptial–colour displacement as a diversifying force in Lake Victoria cichlid fishes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(1546), pp.1345–1353. doi:https://doi.org/10.1098/rspb.2004.2737.

Seehausen, O., van Alphen, J.J.M. and Witte, F. (1997). Cichlid Fish Diversity Threatened by Eutrophication That Curbs Sexual Selection. *Science*, 277(5333), pp.1808–1811. doi:https://doi.org/10.1126/science.277.5333.1808.

Seifert, B., Kulmuni, J. and Pamilo, P. (2010). Independent hybrid populations of Formica polyctena X rufa wood ants (Hymenoptera: Formicidae) abound under conditions of forest fragmentation. *Evolutionary Ecology*, 24(5), pp.1219–1237. doi:https://doi.org/10.1007/s10682-010-9371-8.

Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., Lee, J.H., Puviindran, V., Tam, D., Shen, M., Son, J.E., Vakili, N.A., Sung, H.-K., Naranjo, S., Acemel, R.D. and Manzanares, M. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, [online] 507(7492), pp.371–375. doi:https://doi.org/10.1038/nature13138.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312–1313. doi:https://doi.org/10.1093/bioinformatics/btu033.

Stiassny, M.L.J. and Meyer, A. (1999). Cichlids of the Rift Lakes. *Scientific American*, 280(2), pp.64–69. doi:https://doi.org/10.1038/scientificamerican0299-64.

Stroud, J.T. and Losos, J.B. (2016). Ecological Opportunity and Adaptive Radiation. *Annual Review of Ecology, Evolution, and Systematics*, 47(1), pp.507–532. doi:https://doi.org/10.1146/annurev-ecolsys-121415-032254.

Su, G., Villéger, S. and Brosse, S. (2018). Morphological diversity of freshwater fishes differs between realms, but morphologically extreme species are widespread. *Global Ecology and Biogeography*, 28(2), pp.211–221. doi:https://doi.org/10.1111/geb.12843.

Sung Won Cho, J.C. van Rijssel, Witte, F., de, G. and Richardson, M.K. (2015). The sonic hedgehog signaling pathway and the development of pharyngeal arch Derivatives in Haplochromis piceatus, a Lake Victoria cichlid. *Journal of Oral Biosciences*, 57(3), pp.148–156. doi:https://doi.org/10.1016/j.job.2015.04.001.

Sung Won Cho, J.C. van Rijssel, Witte, F., de, G. and Richardson, M.K. (2015). The sonic hedgehog signaling pathway and the development of pharyngeal arch Derivatives in Haplochromis piceatus, a Lake Victoria cichlid. *Journal of Oral Biosciences*, 57(3), pp.148–156. doi:https://doi.org/10.1016/j.job.2015.04.001.

Sutton, G., White, L.J., Ford, A.G.P., Shechonge, A., Day, J.J., Dasmahapatra, K.K. and Pownall, M.E. (2020). Exploring the Expression of Cardiac Regulators in a Vertebrate Extremophile: The Cichlid Fish Oreochromis (Alcolapia) alcalica. *MDPI (MDPI AG)*, 8(4), pp.22–22. doi:https://doi.org/10.3390/jdb8040022.

Tamura, K., Stecher, G. and Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology And Evolution*, 38(7). doi:https://doi.org/10.1093/molbev/msab120.

Tao, R., Sack, L. and Rosindell, J. (2021). Biogeographic Drivers of Evolutionary Radiations. *Frontiers in Ecology and Evolution*, 9. doi:https://doi.org/10.3389/fevo.2021.644328.

Taylor, E.B., Boughman, J.W., Groenenboom, M., Sniatynski, M., Schluter, D. and Gow, J.L. (2005). Speciation in reverse: morphological and genetic evidence of the collapse of a three-spined stickleback (Gasterosteus aculeatus) species pair. *Molecular Ecology*, 15(2), pp.343–355. doi:https://doi.org/10.1111/j.1365-294x.2005.02794.x.

Thawornwattana, Y., Seixas, F., Yang, Z. and Mallet, J. (2023). Major patterns in the introgression history of Heliconius butterflies. *eLife*, 12. doi:https://doi.org/10.7554/elife.90656.3.

Theis, A., Ronco, F., Indermaur, A., Salzburger, W. and Egger, B. (2014). Adaptive divergence between lake and stream populations of an East African cichlid fish. *Molecular Ecology*, 23(21), pp.5304–5322. doi:https://doi.org/10.1111/mec.12939.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22), pp.4673–4680. doi:https://doi.org/10.1093/nar/22.22.4673.

Turissini, D.A., McGirr, J.A., Patel, S.S., David, J.R. and Matute, D.R. (2017). The Rate of Evolution of Postmating-Prezygotic Reproductive Isolation in Drosophila. *Molecular Biology and Evolution*, 35(2), pp.312–334. doi:https://doi.org/10.1093/molbev/msx271.

Urban, S., Nater, A., Meyer, A. and Kratochwil, C.F. (2020). Different Sources of Allelic Variation Drove Repeated Color Pattern Divergence in Cichlid Fishes. *Molecular Biology and Evolution*, 38(2), pp.465–477. doi:https://doi.org/10.1093/molbev/msaa237.

Van der Auwera GA & O'Connor BD. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition).* O'Reilly Media.

Van der heyden, C. and Huysseune, A. (2000) ‘Dynamics of Tooth Formation and replacement in the zebrafish (Danio rerio) (Teleostei, Cyprinidae)’, *Developmental Dynamics*, 219(4), pp. 486–496. doi:10.1002/1097-0177(2000)9999:9999<::aid-dvdy1069>3.0.co;2-z.

Vawter, L. and Brown, W. (1986). Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. *Science*, 234(4773), pp.194–196. doi:https://doi.org/10.1126/science.3018931.

Verheyen, E., Salzburger, W., Snoeks, J. and Meyer, A. (2003). Origin of the Superflock of Cichlid Fishes from Lake Victoria, East Africa. *Science*, 300(5617), pp.325–329. doi:https://doi.org/10.1126/science.1080699.

Vohra, M., Sharma, A.R., Prabhu B, N. and Rai, P.S. (2020). SNPs in Sites for DNA Methylation, Transcription Factor Binding, and miRNA Targets Leading to Allele-Specific Gene Expression and Contributing to Complex Disease Risk: A Systematic Review. *Public Health Genomics*, 23(5-6), pp.155–170. doi:https://doi.org/10.1159/000510253.

Wagner, C.E., Harmon, L.J. and Seehausen, O. (2012). Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, 487(7407), pp.366–369. doi:https://doi.org/10.1038/nature11144.

Wainwright, P.C., Ferry-Graham, L.A., Waltzek, T.B., Carroll, A.M., Hulsey, C.D. and Grubich, J.R. (2001). Evaluating the use of ram and suction during prey capture by cichlid fishes. *Journal of Experimental Biology*, 204(17), pp.3039–3051. doi:https://doi.org/10.1242/jeb.204.17.3039.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), pp.57–63. doi:https://doi.org/10.1038/nrg2484.

Wellenreuther, M. and Bernatchez, L. (2018). Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology & Evolution*, 33(6), pp.427–440. doi:https://doi.org/10.1016/j.tree.2018.04.002.

White, L. (2021). *Life in extreme environments: adaptation and evolution of a soda lake cichlid; Alcolapia alcalica*. [PDF] pp.3–208. Available at: https://etheses.whiterose.ac.uk/28729/1/White_203010828_CorrectedThesisClean.pdf [Accessed 16 Nov. 2023].

White, L.J., Rose, M., Lawson, M., Joyce, D., Smith, A.M., Thomas, G.H., Dasmahapatra, K.K. and Pownall, M.E. (2022). Two closely related ureotelic fish species of the genus *Alcolapia* express different levels of ammonium transporters in gills. *Biology Open*, 11(11). doi:https://doi.org/10.1242/bio.059575.

Wickham, H. (2016) 'ggplot2: Elegant Graphics for Data Analysis', Springer-Verlag New York.

Widmer, A., Lexer, C. and Cozzolino, S. (2008). Evolution of reproductive isolation in plants. *Heredity*, [online] 102(1), pp.31–38. doi:https://doi.org/10.1038/hdy.2008.69.

Williamson, D. *et al.* (1993) ‘Equatorial extension of the younger Dryas event: Rock magnetic evidence from Lake Magadi (Kenya)’, *Global and Planetary Change*, 7(1–3), pp. 235–242. doi:10.1016/0921-8181(93)90053-q.

Wilson, G.P., Evans, A.R., Corfe, I.J., Smits, P.D., Fortelius, M. and Jernvall, J. (2012). Adaptive radiation of multituberculate mammals before the extinction of dinosaurs. *Nature*, 483(7390), pp.457–460. doi:https://doi.org/10.1038/nature10880.

Wood, C., Perry, S., Wright, P., Bergman, H. and Randall, D., 1989. Ammonia and urea dynamics in the Lake Magadi tilapia, a ureotelic teleost fish adapted to an extremely alkaline environment. Respiration Physiology, 77(1), pp.1-20.

Wood, C.M., Brix, K.V., De Boeck, G., Bergman, H.L., Bianchini, A., Bianchini, L.F., Maina, J.N., Johannsson, O.E., Kavembe, G.D., Papah, M.B., Letura, K.M. and Ojoo, R.O. (2016). Mammalian metabolic rates in the hottest fish on earth. *Scientific Reports*, 6(1). doi:https://doi.org/10.1038/srep26990.

Wright, P.A., Perry, S.F., Randall, D.J., Wood, C.M. and Bergman, H.L. (1990). The Effects of Reducing Water pH and Total CO2 on a Teleost Fish Adapted to an Extremely Alkaline Environment. *The Journal of Experimental Biology*, 151(1), pp.361–369. doi:https://doi.org/10.1242/jeb.151.1.361.

Yang, J. and Adli, M. (2019). Mapping and Making Sense of Noncoding Mutations in the Genome. *Cancer Research*, 79(17), pp.4309–4314. doi:https://doi.org/10.1158/0008-5472.can-19-0905.

Yang, L., Tan, Z., Wang, D., Xue, L., Guan, M., Huang, T. and Li, R. (2014). Species identification through mitochondrial rRNA genetic analysis. *Scientific Reports*, [online] 4(1). doi:https://doi.org/10.1038/srep04089.

Zaccara, S., Crosa, G., Vanetti, I., Binelli, G., Harper, D., Mavuti, K.M., Balarin, J.D. and Britton, J. (2014). Genetic and morphological analyses indicate high population mixing in the endangered cichlid Alcolapia flock of East Africa. *Conservation Genetics*, 15(2), pp.429–440. doi:https://doi.org/10.1007/s10592-013-0551-2.

Zeng, H. and Gifford, D.K. (2017). Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Research*, 45(11), pp.e99–e99. doi:https://doi.org/10.1093/nar/gkx177.

Zhang, Z., Lan, Y., Chai, Y. and Jiang, R. (2009). Antagonistic Actions of Msx1 and Osr2 Pattern Mammalian Teeth into a Single Row. *Science*, [online] 323(5918), pp.1232–1234. doi:https://doi.org/10.1126/science.1167418.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7), pp.821–824. doi:https://doi.org/10.1038/ng.2310.

Zhu, H. and Zhou, X. (2020). Statistical methods for SNP heritability estimation and partition: A review. *Computational and Structural Biotechnology Journal*, 18, pp.1557–1568. doi:https://doi.org/10.1016/j.csbj.2020.06.011.

Supplementary

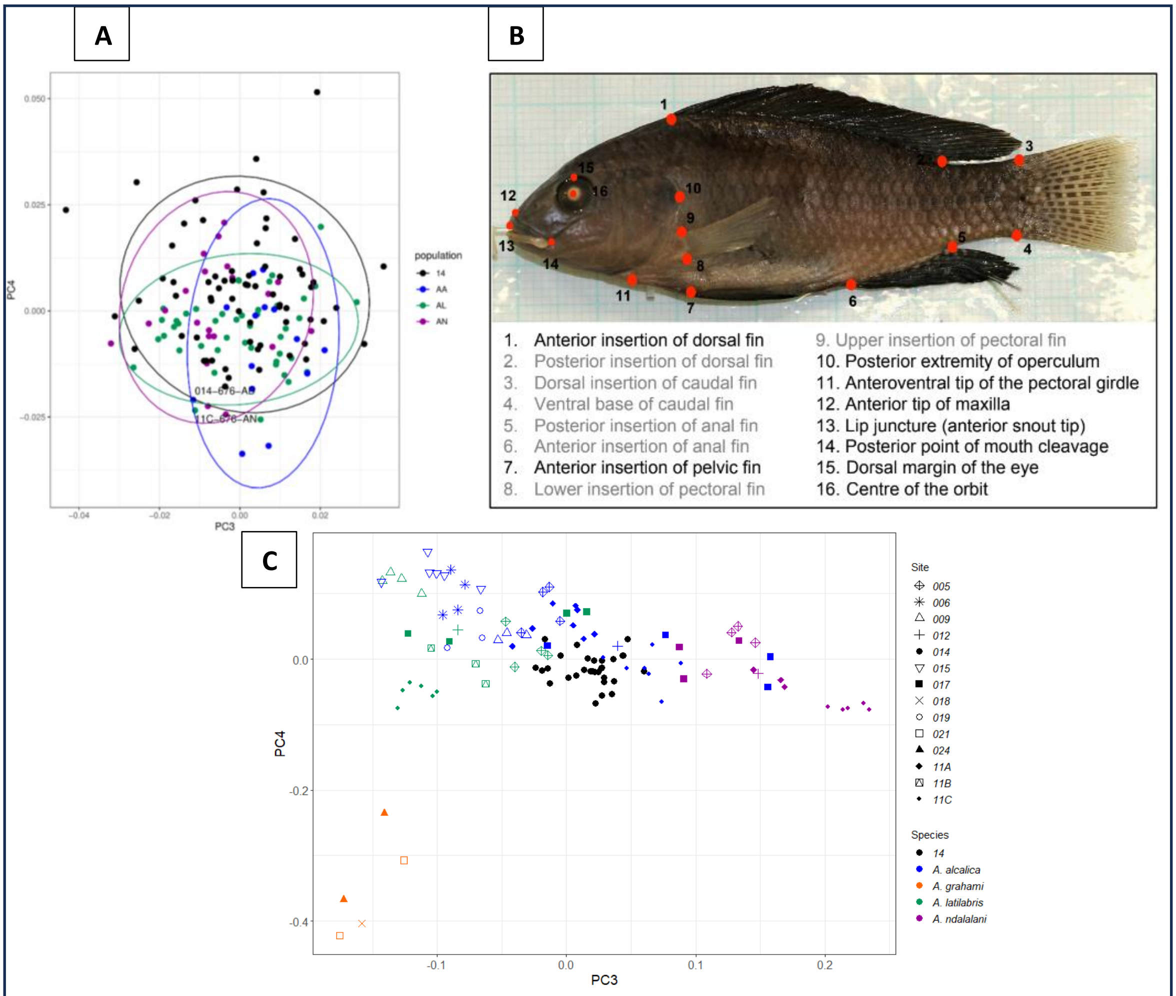


Figure S1: A) Morphometric PCA showing considerable overlap in morphospace between Lake Natron species and site 14. **B)** Landmarks 1-16 are those used to generate eigenvectors for the morphometric PCA of site 14 compared to 11C *A. alcalica*, *A. ndalalani* and *A. latilabris. (Figure 6.2,* Ford, 2015). **C)** Genetic PCA illustrating *A. grahami* is the most genetically distinct on PC 3 and 4, rather than site 14. Filtering and SNPS are the same as Figure 2.4B. PC3 = 8.77, PC4 = 5.78.

Table S1: Location of *Alcolapia* sites

Site Number	South coordinate	East coordinate
5	-2.597583	35.918417
6	-2.430397	35.895392
9	-2.479278	35.888223
11A	-2.591	36.009445
11B	-2.591667	36.012056
11C	-2.592667	36.016056
12	-2.618972	35.999806
14	-2.527028	36.046139
15	-2.433361	36.10175
17	-2.456278	36.087806
18	-2.00113888889	36.23197222
19	-2.145833333	36.05575
21	-1.84444444444	36.22425
24	-0.396028	36.107583

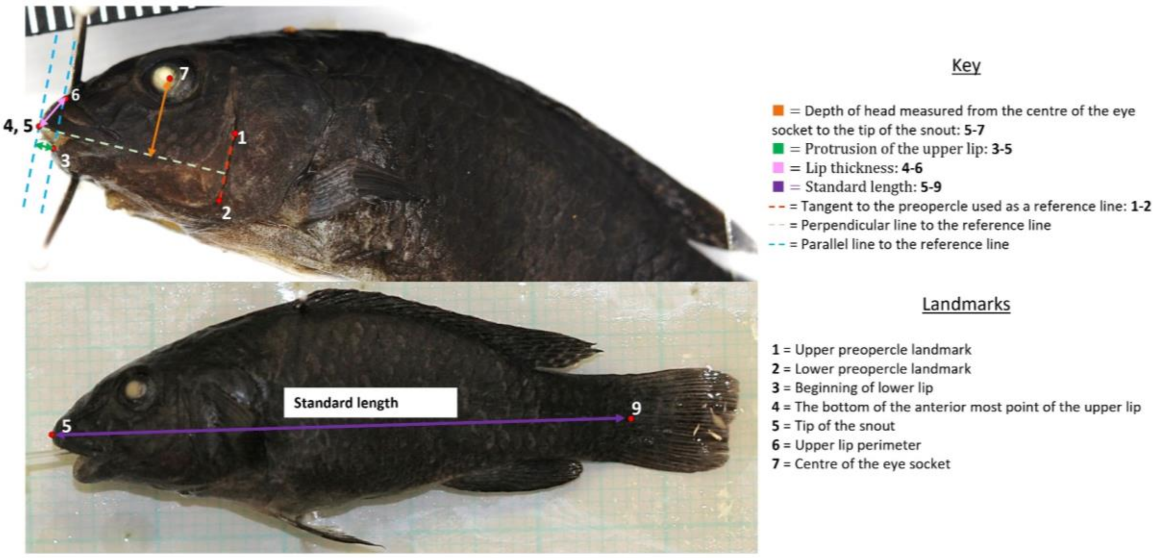


Figure S2: Diagram of the three different morphological traits. Fish were standardised using the standard lengths.

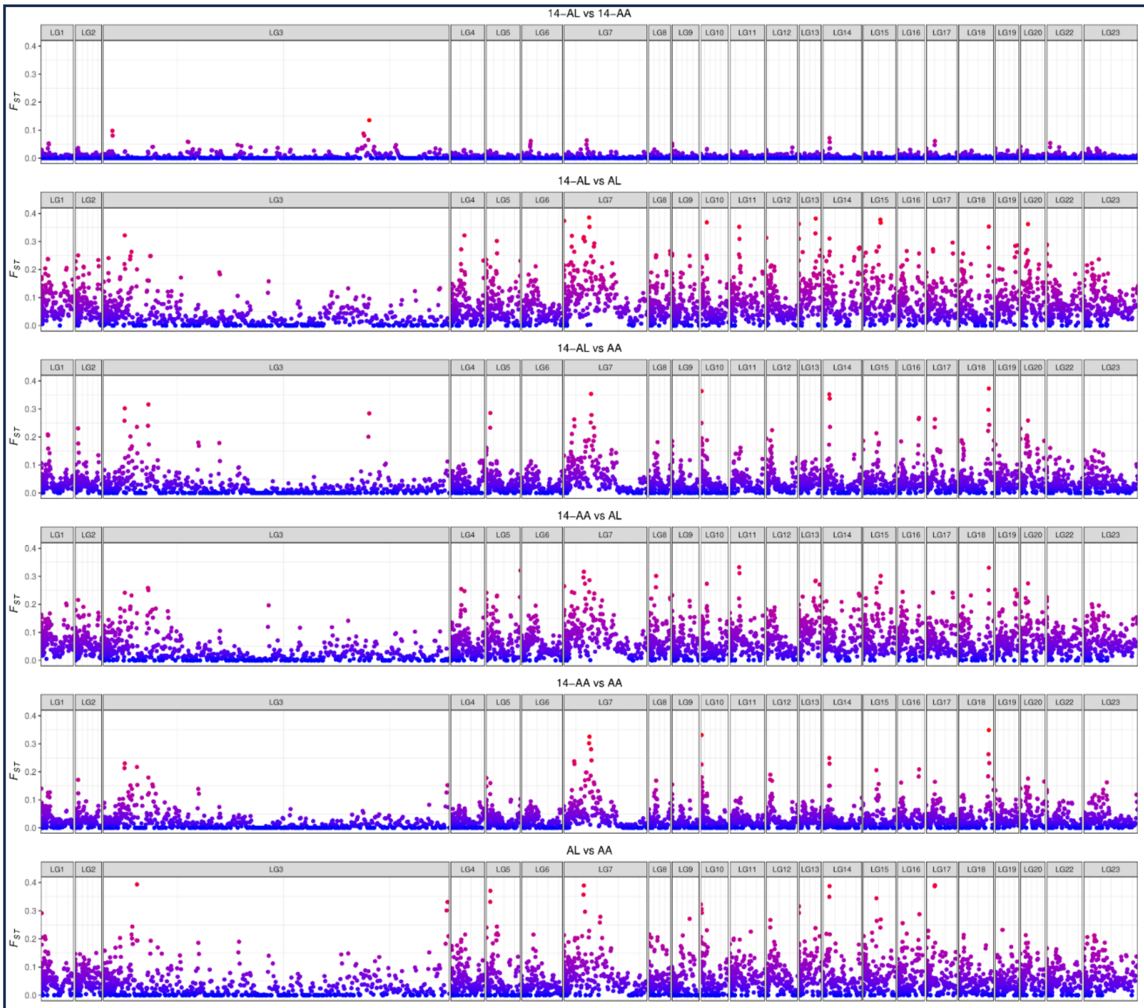


Figure S3: Sliding windows F_{ST} of RAD data from site 14 morphs compared with *A. alcalica* (AA) and *A. latilabris* (AL) shows there is no apparent inversion across all comparisons. F_{ST} is low across most windows for *Alcolapia* comparisons. 100 SNPs per window overlapping by 50 SNPs. 107,614 SNPs. *x* axis = genome position.

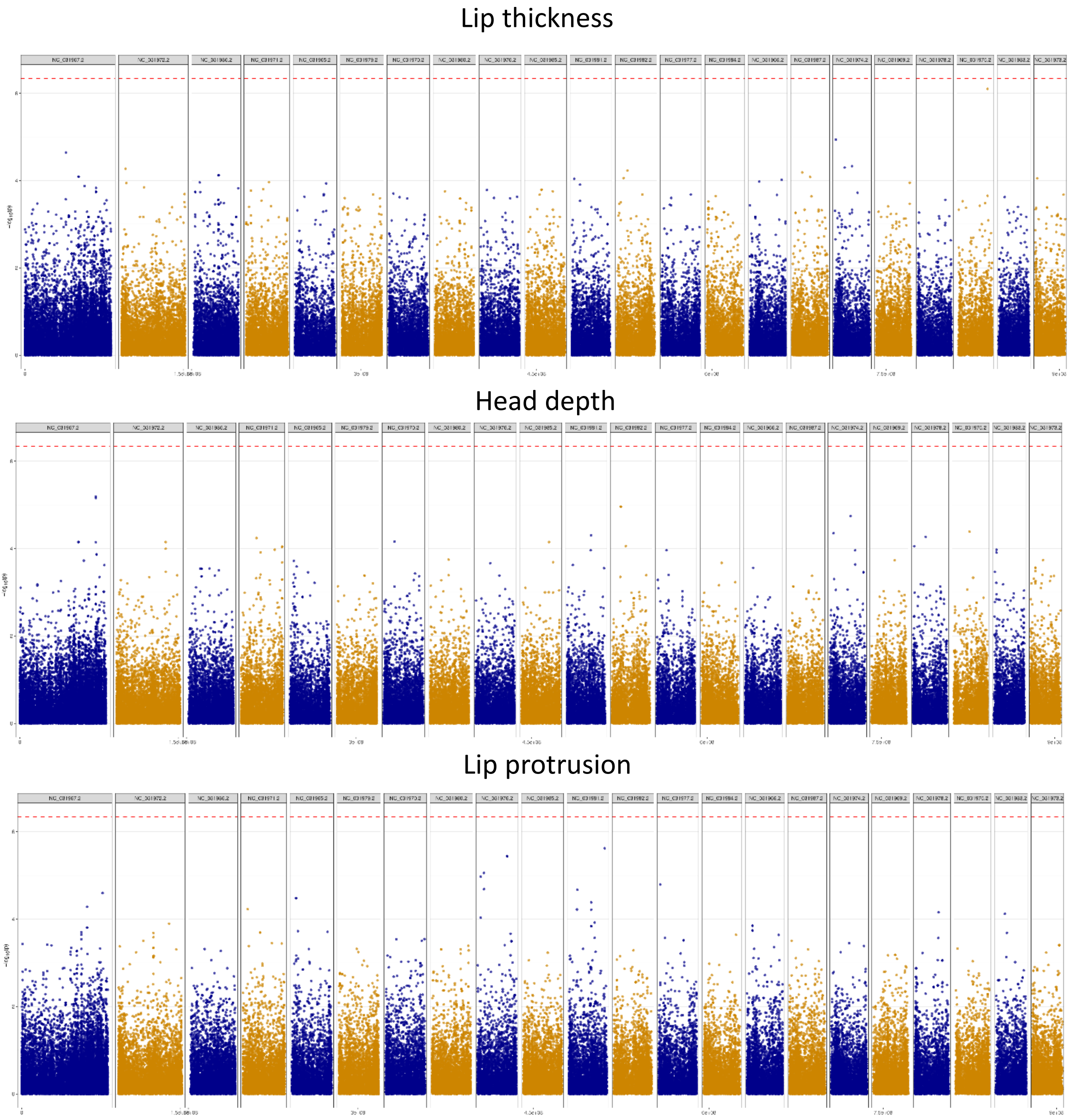


Figure S4: GWAS on site 14 using different phenotypes identified no significant SNPs. Filters used included GQ30, MIS 0.5 and MAF 0.05.

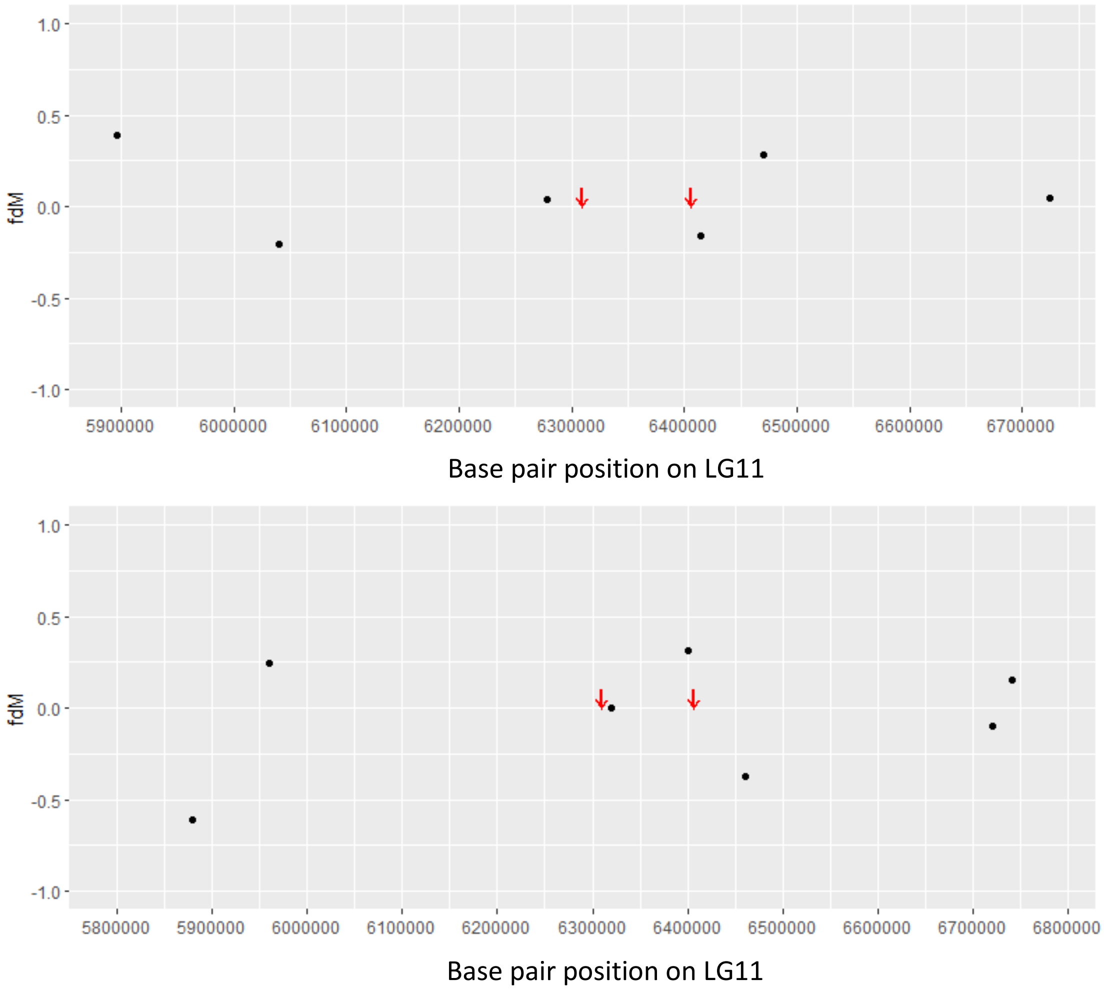


Figure S5: There is no evidence for introgression between either 14-AL or 14-AA and *A. alcalica* or *A. latilabris* at positions LG11: 6309402 and 6406194 (shown here by red arrows) as f_{dM} is close to 0. These are the two SNPs with very low P. values previously identified in Figure 2.9. Both plots have 100 SNPs per window, overlap of 0, a depth of 5-100, missingness of 0.5, only SNPs and a quality score of 30. A positive f_{dM} would indicate P2 and P3 are more closely related than P1 and P3 (Martin, 2024) while a negative f_{dM} indicates the inverse relationship. **Top:** P1 = 14-AA, P2 = 14-AL, P3 = *A. latilabris*. **Bottom:** P1 = 14-AA, P2 = 14-AL, P3 = *A. alcalica*. X- axis = bp. SNPs remaining after filtering = 630,350.

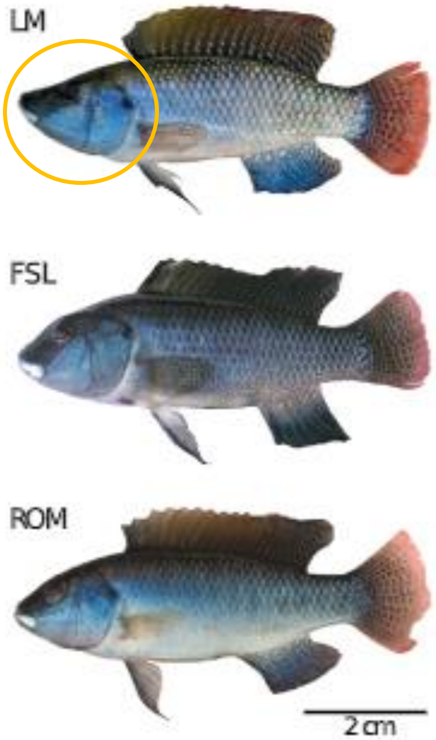


Figure S6: *A. grahami* from LM (Little Magadi), FSL (Fish Springs Lagoon) and ROM (rest of Lake Magadi). LM is **upturned**, while FLS and ROM are **terminal**. Annotated from Figure 1 Kavembe et al., (2016).

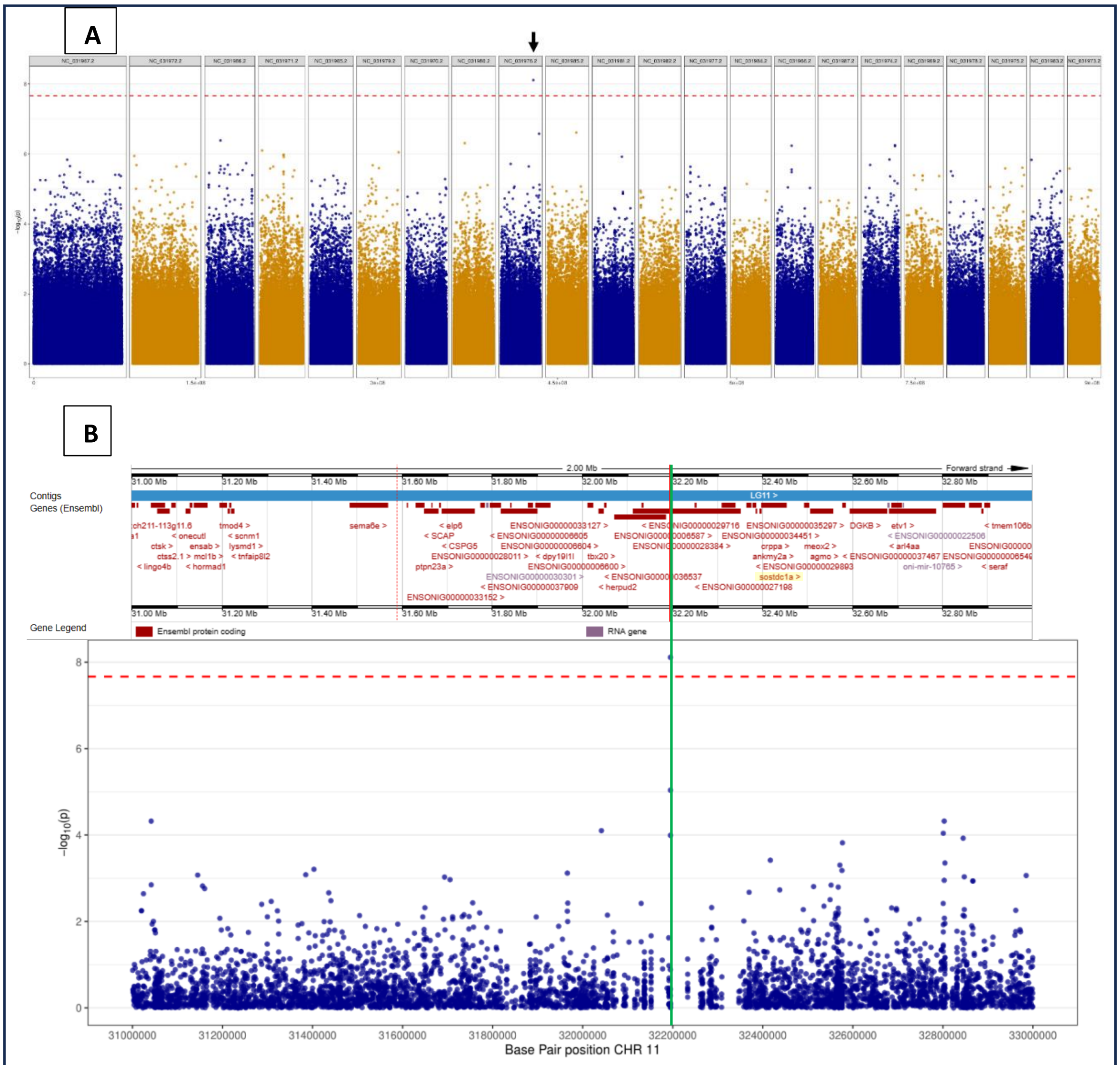
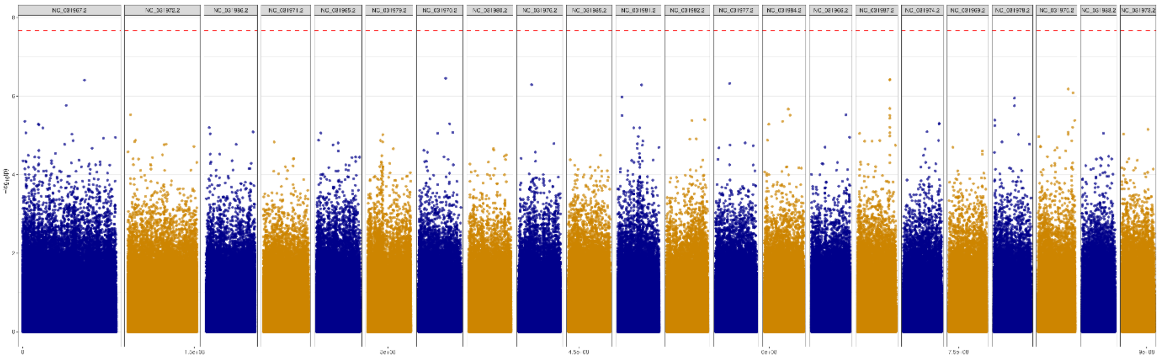
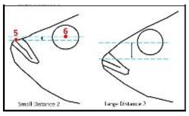


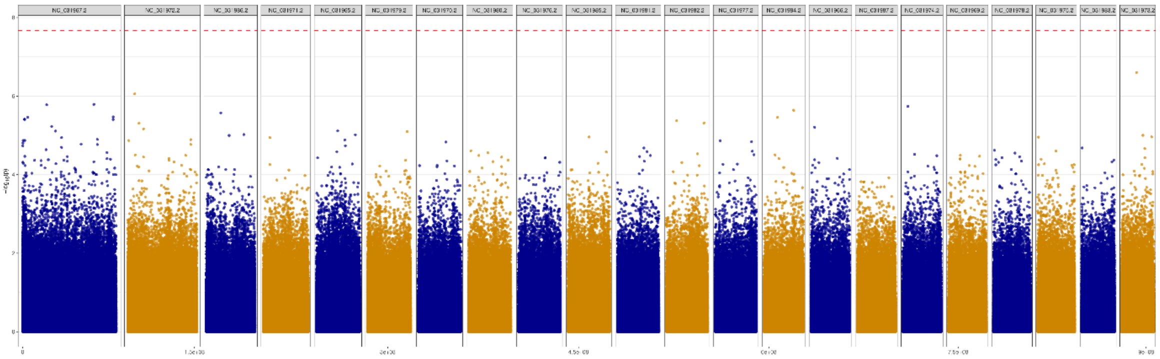
Figure S7: A) GWAS on site 15 comparisons when the phenotype is distance1. One SNP is above the Bonferroni threshold of 7.666028 on scaffold NC_031976.2 (LG11) at base pair position 32194615 shown with an arrow. Filtering parameters = **GQ30**, DP5, DP100, MIS 0.5, MAF 0.05, SNPs only. **B**) Zoomed in section on CHR11 with significant SNP show with a green line.

A) Ford and Day's discrete rank. **Terminal or Upturned**

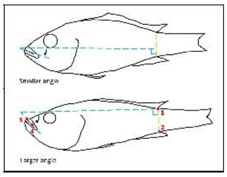


B) Distance 2





C) Mouth angle



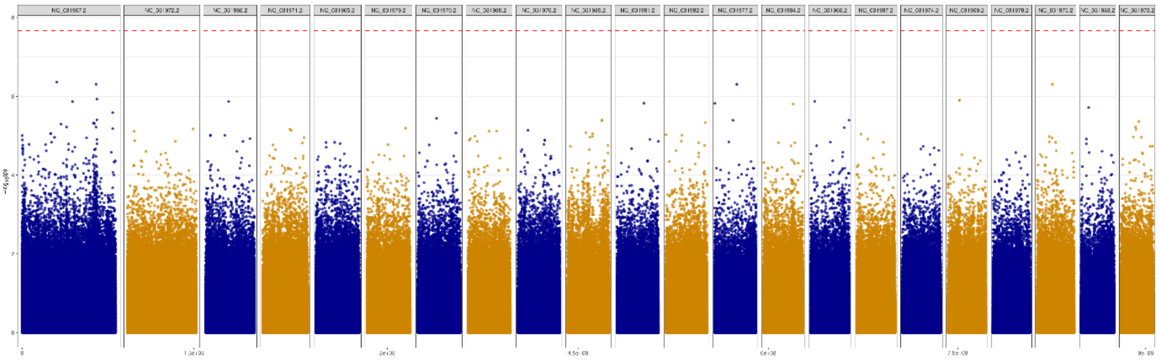


Figure S8: GWAS on whole genome data using phenotypic traits characterising mouth orientation **(A-C)** finds no significant peaks for *A. alcalica* at site 15. *x*-axis = chromosomes ordered by size (largest to smallest) with accumulative base pair position from left to right. GQ30.

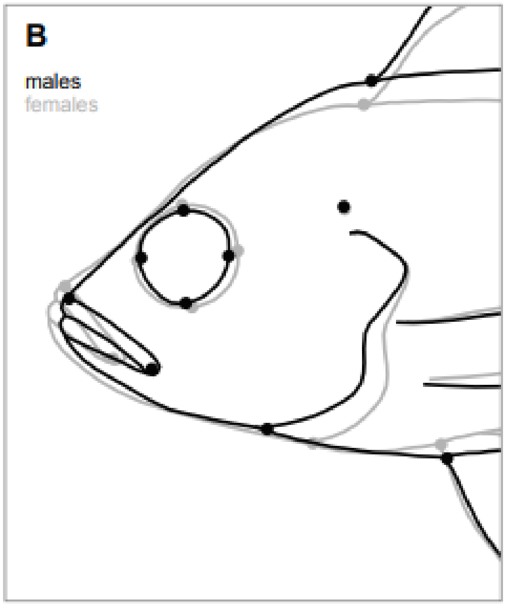


Figure S9: Females have more upturned mouth morphologies than male *A. burtoni* from Lake Tanganika (screenshot from S1B, Theis et al., 2014). This is a similar phenotype seen between site 15 *A. alcalica* upturned and terminal morphs.