Chromatin accessibility in gill tissue identifies candidate genes and loci associated with aquaculture relevant traits in tilapia

Tarang K. Mehta, Angela Man, Adam Ciezarek, Keith Ranson, David Penman, Federica Di-Palma, Wilfried Haerty

PII: S0888-7543(23)00077-0

DOI: https://doi.org/10.1016/j.ygeno.2023.110633

Reference: YGENO 110633

To appear in: *Genomics*

Received date: 17 February 2023

Revised date: 25 April 2023

Accepted date: 26 April 2023

Please cite this article as: T.K. Mehta, A. Man, A. Ciezarek, et al., Chromatin accessibility in gill tissue identifies candidate genes and loci associated with aquaculture relevant traits in tilapia, *Genomics* (2023), https://doi.org/10.1016/j.ygeno.2023.110633

Chromatin accessibility in gill tissue identifies candidate genes and loci associated with aquaculture relevant traits in tilapia

Tarang K. Mehta[1]*, Angela Man[1], Adam Ciezarek[1], Keith Ranson[2], David Penman[2], Federica Di-Palma[3,4], Wilfried Haerty[1,3]

[1] Earlham Institute (EI), Norwich, UK

[2] Institute of Aquaculture, University of Stirling, Scotland, UK

[3] School of Biological Sciences, University of East Anglia, Norwich, UK

[4] Genome British Columbia, Vancouver, Canada

* Corresponding author

Abstract

The Nile tilapia (*Oreochromis niloticus*) accounts for ~9% of global freshwater finfish production however, extreme cold weather and decreasing freshwater resources has created the need to develop resilient strains. By determining the genetic bases of aquaculture relevant traits, we can genotype and breed desirable traits into farmed strains. We generated ATAC-seq and gene expression data from *O. niloticus* gill tissues, and through the integration of SNPs from 27 tilapia species, identified 1,168 highly expressed genes (4% of all Nile tilapia genes) with highly accessible promoter regions with functional variation at transcription factor binding sites (TFBSs). Regulatory variation at these TFBSs is likely driving gene expression differences associated with tilapia gill adaptations, and differentially segregate in freshwater and euryhaline tilapia species. The generation of novel integrative data revealed candidate genes e.g., *prolactin receptor 1* and *claudin-h*, genetic relationships, and loci associated with aquaculture relevant traits like salinity and osmotic stress acclimation.

Introduction

Tilapia cichlid fish of the genus *Oreochromis*, native to Africa and the Middle East, are farmed in over 120 countries/territories [1]. The Nile tilapia, *Oreochromis niloticus*, accounted for 9% of the 49 million tonnes of freshwater finfish produced globally in 2020, the third most of any species [1]. The exponential growth in tilapia aquaculture production is largely due to their suitability for aquaculture systems and unlike most other finfish species, tilapia can grow and reproduce in many culture systems. However, climate change is leading to extreme weather events, and decreasing freshwater resources [2]. There is a pressing need to develop aquaculture systems based on strains resilient to saline waters and temperature changes. A way forward is to characterise the genetic bases responsible for such adaptive traits enabling selective breeding into farmed strains.

Such an approach involves resequencing populations to map and characterise variation, and identify signatures of selection in genomic regions associated with adaptive differences [3]. Previous population studies using whole genome resequencing of wild tilapia individuals (*O. niloticus*, *O. mossambicus*, and red tilapia) and commercial strains reported between 1.3 and 1.43 million single nucleotide polymorphisms (SNPs) [4,5], the majority of which are located within noncoding regions of the genome. Most SNPs under selection were found within those noncoding regions (45% intronic and 19% intergenic) and associated with several aquaculture relevant traits like growth, reproduction, and immunity [4,5]. Only 16 non-synonymous SNPs were identified in the three wild tilapia populations, suggesting selection on functional noncoding regions has played a prominent role during domestication/breeding [4].

We recently generated genome-wide sequencing data across 575 individuals of 27 tilapia species from across East Africa, to identify 69 million single nucleotide polymorphisms (SNPs) in the *O. niloticus* genome assessing phylogenomic patterns of recent and ancient hybridisation in tilapia (Ciezarek, A. *et al.*, bioRxiv TBC, 2023). Whilst selection in noncoding regulatory regions, including transcription factor binding sites (TFBSs) has been suggested as a key molecular mechanism contributing to the evolutionary diversification of East African cichlids, including Nile tilapia [6-8], no study has evaluated the *cis*-regulatory impact of noncoding variation towards adaptive traits in a tilapia phylogeny.

Several studies have provided evidence on the role of *cis*-regulatory elements (CREs) towards morphological diversity [9-11], including in mammals [12] and fish [13]. Whilst such studies identified CREs at particular loci, epigenetic sequencing methods can capture chromatin accessibility and transcription factor (TF) binding at the genome-wide level [14]. Only a handful of studies, all using DNA methylation, have applied epigenetic approaches to study adaptive traits in tilapia [15]; this includes the regulation of tilapia growth [16-18], sexual dimorphism [19], and sex determination [20,21]. ChiP-seq approaches have also been used to map active promoter elements associated with Nile tilapia fin development [22]. However, such techniques are only able to identify repressive or active marks at specific loci, and in the case of ChiP-seq, direct interactions between a specific protein and DNA. In contrast, the assay for transposase-accessible chromatin using sequencing (ATAC-seq) approach can robustly identify genome-wide chromatin accessibility and TF-occupancy using very few [23,24] or even single cells [25,26]. Despite the robustness of ATAC-Seq and its potential to provide a better annotation of regulatory regions, and an accurate

identification of TF-occupancy including binding sites for regulators in Nile tilapia, no such study has been performed to date on tilapia tissues and/or cells.

Here, we aim to identify genes and functional non-coding regions in Nile tilapia by characterising the open chromatin and gene expression landscape of Nile tilapia gill tissue to classify functional noncoding variation associated with environmental tolerance. For this, we characterised the open chromatin landscape of Nile tilapia gill tissue using ATAC-seq, and then identified accessible gene promoter regions with TF footprints that could account for target gene expression in the gill. We integrate this data with noncoding regulatory SNPs from a 27 species tilapia phylogeny (Ciezarek, A. *et al.*, bioRxiv TBC, 2023), and identify candidate genes with regulatory variation that could be associated with aquaculture-relevant traits, like salinity tolerance.

Results

Accessible gene promoter regions in gill tissue are functionally associated with aquaculture relevant traits

To characterise functional noncoding regions based on chromatin accessibility, we performed high-depth ATAC-seq of gill tissue from three replicate male Nile tilapia individuals, identifying 301,293 open chromatin (accessible) reproducible peaks between replicates (see '*Materials and Methods*'), that could be associated to 25,092 (average of 12 accessible peaks per gene) out of 29,552 (85%) annotated Nile tilapia genes [27]. We first annotated (Supplementary Table S1) and tested the enrichment of accessible peaks in coding and noncoding regions, finding that peaks are mostly enriched (fold enrichment >2.1, adjusted *P*-value <0.05) in both sets of conserved noncoding elements (CNEs) (Fig. 1) that we annotated in the Nile tilapia (*O. niloticus* UMD_NMBU [27]) reference genome (see '*Materials and Methods*').

We focused on gene promoter regions, taken as up to 5 kb from the transcription start site (TSS) [8], as they harbour regulatory binding sites e.g., TFBSs that have functionally diverged [3] and can be associated with gene expression differences in Nile tilapia and East African cichlids [8]. Out of the 301,293 open chromatin (accessible) peaks, 127,602 accessible peaks (average size of 464 bp) were found within gene promoters of 20,248 genes in gill tissue, with missing genes being better associated to lymphoid tissues (see '*Supplementary Information'*, Supplementary Fig. S1 – S2).

Accessible gene promoter regions could account for expression changes of genes associated with osmoregulation and salinity tolerance

To examine any correlation between gene transcription and accessible gene promoter regions, RNA-seq data was generated from the same gill tissue sample of the three replicates and gene expression measured using transcripts per million (TPM) (see '*Materials and Methods*'). Of the 29,552 Nile tilapia genes in the genome, we identified and calculated the expression (as TPM) of 20,717 (70%) genes. We found 96,784 (76%) accessible peaks could be associated to the expression of 14,842 genes.

We first tested whether these peaks are better correlated to their corresponding gene expression based on proximity to the TSS (Fig. 2a). For this, accessible peaks are categorised into four intervals of either 500 bp (12,978 peaks), 1 kb (27,975 peaks), 2kb (48,108 peaks) or up to 5 kb (127,602 peaks) from the genes TSS. We found similar weak positive correlation (R = 0.13 to 0.14, Spearman $p < 2.2e-16$) for all four intervals (TSS peak up to 500 bp, 1 kb, 2 kb and 5 kb) (Fig. 2a) and thus, we focus on all (159,708) accessible gene promoter peaks (up to 5 kb from the TSS). A higher correlation might not be observed in our datasets as accessible regions are not always associated with gene activity and can instead, be associated with repressed or poised for activation genes [28-31].

Based on an approach devised previously [31], we explore the relationship of accessible peaks and gene expression by assessing genes with combinations of high accessibility (HA) or medium–low accessibility (MA), and either highly expressed (HE) or medium–low expression (ME) (see '*Materials and Methods*').

After categorising gene peaks and expression, we were left with a total of 27,536 (22% of 127,602) accessible peaks associated to the expression of 4,509 unique genes. These could be placed into the four categories, with most genes (1,650) and peaks (10,488) in the HA-HE category, and the remaining being either HA-ME (6,961 peaks and 1,081 unique genes), MA-HE (5,990 peaks and 1,770 unique genes), or MA-ME (4,097 peaks and 1,164 unique genes). Whilst the HA-ME genes exhibit no correlation (R = -0.021, Spearman p = 0.076), and the MA-HE genes exhibit significant weak negative correlation (R = -0.034, Spearman p = 0.008), both the MA-ME (R = 0.037, Spearman p = 0.019) and HA-HE genes (R = 0.031, Spearman p = 0.0013) exhibit significant weak positive correlation (Fig. 2b). Among the 1,650 HA-HE genes, we identify genes associated with aquaculture-relevant traits, including those enriched (FDR < 0.05) for the GO terms 'cellular process' e.g., prolactin receptor 1 (*prlr1*) involved in tilapia salinity tolerance [32] and chloride voltage-gated channel 2 (*clcn2*) an osmoregulatory gene involved in adaptation to saline-alkali challenge in tilapia [33], and 'cellular respiration' e.g., cytochrome c oxidase subunit 4 (*cox4*) involved in aerobic metabolism changes under salinity challenge in tilapia gills [34] (Supplementary Fig. S3).

Nile tilapia SNPs are enriched in noncoding regions

Since we previously found that discrete variation in regulatory regions disrupt regulatory network interactions for adaptive trait genes in Nile tilapia [7,8], we tested whether functional noncoding SNPs could be associated with the regulatory activity of genes involved with aquaculture-relevant traits, like environmental tolerance. Building upon our previous work of identifying 69 million (69,064,774) SNPs across 27 tilapia species (Fig. 3a, Supplementary Table S2) mapped to the Nile tilapia

genome (Ciezarek, A. *et al.*, bioRxiv TBC, 2023), we identified that these SNPs are significantly enriched in noncoding than coding (exonic) regions (Fig. 3b). Compared to exonic regions (fold enrichment of 0.9, adjusted *P*-value <0.05), the SNPs are most enriched in either intronic or up to 5 kb gene promoter (fold enrichment of 1.1, adjusted *P*-value <0.05) regions (Fig. 3b).

To characterise functional noncoding variation, we find that of the 5,470,066 SNPs in Nile tilapia gene promoter regions, 1,083,195 (20%) SNPs overlapping 121,404 of the 127,602 (95%) accessible peaks are significantly enriched (fold enrichment of 1.1, adjusted *P*-value <0.05) (Fig. 3c). Of the 20,248 accessible gene promoters, 20,101 (99%) have SNPs in accessible regions, representing 68% of all annotated Nile tilapia genes [27]. Most SNPs are rare variants (Minor Allele Frequency, MAF <= 0.1) with 1% more rare variants found in noncoding regions that are accessible than not (Supplementary Fig. S4, see '*Supplementary Information*'). Using the 127,602 accessible peaks in gene promoter regions, we find that the topmost (-log10 FDR < 0.05 of >=10) enriched GO biological processes of 20,248 accessible gene promoter regions harbouring SNPs could be functionally relevant to aquaculture traits like, for example, signalling, cell communication, and cellular response to stimulus (Fig. 3d, Supplementary Fig. S5). In summary, nearly three-quarters of all Nile tilapia genes have putative functional noncoding variation at gene promoter regions that could be associated with the regulation of aquaculture traits relevant to gill function like, for example, a response to stimulus.

Discrete variation in accessible TFBSs is likely driving gene expression associated with tilapia gill adaptations

We hypothesise that genetic variation between Nile tilapia and the 26 other tilapia species (Fig. 3a) could be used to identify functional noncoding variation in TFBSs of orthologous genes associated with aquaculture traits that differ between the species. We first prioritise predicted TFBSs (see '*Materials and Methods*' and '*Supplementary Information*') using the best described method for ranking predictions [35] by overlapping predicted TFBSs with the number of reads (tag count, TC) in the footprint, and then rank TFBSs using the bit-score of the motif match within ranked tag counts. Based on a mean bit-score of 11.5 and mean tag count of 111.9, we identify 6,195,938 TFBSs above the means (Fig. 4a). In gene promoter regions, we identify 246,296 TF footprints with 2,491,107 non-redundant TFBSs (see '*Materials and Methods*'). Based on a mean bit-score of 11 and mean tag count of 152.6, we identify 398,995 (16%) TFBSs above the means in gene promoter regions.

Using TFBSs in accessible gene promoter regions, we identified a total of 3,398,768 transcription factor (TF) – target gene (TG) relationships in Nile tilapia. We then identify discrete mutations in TFBSs to characterise conservation or divergence and therefore, whether the 3,398,768 Nile tilapia TF-TG relationships could be present or absent in the other 26 species. First, we map open-chromatin peaks, footprints, TFBSs and SNPs to gene promoter regions on each Nile tilapia chromosome (Fig. 4b). In total, we identify 7,029,307 (11% of total 69,064,774) SNPs overlapping accessible gene promoter regions, TF footprints, and TFBSs with all (2042) TF motifs having at least one discrete mutation in 18,899 genes (64% of the 29,552 Nile tilapia genes in the genome [27]). More than half (62% - 11,790) of the 18,899 genes have SNPs in 'prioritised' TFBSs where the bit-score of motif match and tag count are above the means of 11 and 152.6 respectively. Of the 18,899 genes, 1,562 (8%)

have high accessibility and high expression (HA-HE), which makes up the majority (95%) of all 1,650 HA-HE genes. Using both of the above measures, we narrowed this down to 1,168 genes (71% of all HE-HE genes or 6% of all genes with a SNP in an accessible gene promoter TFBS) that are categorised as 1) having a SNP in a 'prioritised TFBSs' where the bit-score of motif match and tag count are above the mean of all gene promoter TFBSs; and 2) being a 'HA-HE gene' based on peak-expression correlations. To further narrow down the 1,168 genes, we focus on genes 3) with 'genetic variation' that exists between species exhibiting a gradient of environmental tolerance e.g., freshwater versus saline water; and 4) that are enriched for biological process GO terms that are associated with gill adaptations of traits with aquaculture relevance e.g., response to stimulus (Fig. 3d, Supplementary Fig. S5) or cellular process (Supplementary Fig. S3). We use these four described criteria for identifying examples of candidate genes that are likely involved in gill function associated with environmental tolerance in tilapia.

Discrete mutations at accessible regulatory sites of adaptive trait genes are associated with environmental tolerance networks in tilapia species

Using the four described criteria (prioritised TFBSs, HA-HE gene, genetic variation between diverse species, and aquaculture relevant GO terms) on the set of 1,168 genes, we first identified signal transducer and activator of transcription 1 (STAT1) - *prlr1* (Fig. 5) as a candidate TF-TG relationship that could be involved in gill adaptations associated with tilapia environmental tolerance. The STAT1 TFBS is proximal (<400 bp) to the *prlr1* TSS (Fig. 5a-b) and with a significant signal of activity (Fig. 5c), has a few standard deviations higher bit-score of motif match (11.04 ± 2.61 SD) and tag count (152.6 ± 0.12 SD) than the means of all predicted gene promoter

TFBSs. We identified a rare variant (LG7 position 18107947; MAF=0.03) at the third position in the predicted STAT1 TFBS (Fig. 5b and Fig. 5d). The rare variant is homozygous alternate G/G (n=14) and heterozygous alternate T/G (n=9) in the highly adapted extremophile species, *A. grahami,* but homozygous reference T/T in Nile tilapia and the other 25 tilapia species (Fig. 5d). STAT1 is a transcriptional activator that can be phosphorylated via osmotic stress [36] and prolactin receptors underly variation in salinity tolerance of tilapias [32]; this suggests that STAT1 could be a key regulator of *prlr1* gene expression in most tilapia species gills in response to osmotic stress however, the loss of STAT1 regulation of *prlr1* could necessitate gill adaptations to extreme alkaline and saline waters that are natural to *A. grahami* [37,38].

In another example, we identify variation in the gene promoter region of claudin-h (*cldnh* - also referred to as *cldn3-like* and *cldn28a*), associated with expression differences in stickleback ecotypes [39] and euryhaline tilapia [40] under salinity challenge, as well as involved in permeability changes associated with salinity acclimation in freshwater tilapia species too [41]. Specifically, we identify sterol regulatory element binding transcription factor 2 (SREBF2) - *cldnh* (Fig. 6) as another candidate TF-TG relationship that could be involved in gill adaptations associated with salinity acclimation, especially of freshwater species. The SREBF2 TFBS is proximal (<330 bp) to the *cldnh* TSS (Fig. 6a-b) with a significant signal of activity (Fig. 6c), and a few standard deviations higher bit-score of motif match (11.04 ± 0.4 SD) and tag count (152.6 ± 0.3 SD) than the means of all predicted gene promoter TFBSs. We identified a rare variant (LG10 position 17991869; MAF=0.07) at the 4th position in the predicted SREBF2 TFBS (Fig. 6b and Fig. 6d). We identify the rare variant is mostly homozygous alternate G/G (n=19) and heterozygous

alternate A/G (n=14) in the freshwater species, *O. esculentus,* but homozygous reference A/A in Nile tilapia, *O. esculentus* (n=4) and most individuals, except *O. "mtera"* , of the other 25 tilapia species (Fig. 6d). Despite being a freshwater species, *O. esculentus* individuals with heterozygous and homozygous alternate genotypes were collected from saline/alkaline water bodies, namely Hombolo Dam [42], Lake Malimbe [43], and Lake Rukwa [44,45]. Such genetic variation at the SREBF2 TFBS could therefore account for differential expression of the *cldnh* gene, which has been previously shown to reorganise gill tight junctions in response to salinity acclimation of euryhaline [40] and freshwater [41] tilapia. Discrete mutations in regulatory sites of aquaculture relevant genes could therefore be driving gene regulatory evolution associated with salinity acclimation of the gills.

Discussion

Most (~90%) tilapia aquaculture production is based on Nile tilapia [46] owing to its high growth and reproduction rate in several culture systems. However, with climate change leading to extreme weather and human competition decreasing freshwater resources, there is a crucial need to breed tilapia strains that are resilient to broad environmental conditions. One method to determine the genetic bases responsible for environmental traits e.g., salinity tolerance, involves firstly characterising genetic variation and signatures of selection associated with adaptive traits in resequenced populations [3] and then secondly, using epigenetic sequencing e.g., ATAC-Seq to annotate functional genomic regions and RNA sequencing to characterise whether these loci can drive gene expression change. Despite the impact of noncoding regulatory evolution towards the diversification of cichlids, including Nile tilapia [6-8] and breeding wild tilapia populations [4,5], no previous work has applied ATAC-Seq to accurately identify transcriptional regulators in tilapia tissues and/or cells, let alone classify functional noncoding variation associated with adaptive traits.

By performing high-depth ATAC-seq of Nile tilapia gill tissues, we identify 301,293 accessible (open chromatin) regions in three replicates that can be associated to 85% of all annotated Nile tilapia genes [27]. No previous work has carried out ATAC-seq in gill tissue however, we identify more peaks in our study compared to 11 other zebrafish tissues (66k-180k peaks) but fewer than the 436k merged non-redundant peaks across all tissues [47]. The identification of more single tissue peaks than zebrafish is possibly due to tissue-specific differences and sampling of more replicates. More than 40% (127,602) of the accessible regions can be found within

gene promoter regions of 20,248 genes, with genes without a reproducible peak being better associated to functions of lymphoid tissues not sampled here.

We were able to associate 96,784 (76%) accessible gene promoter regions to the expression of 14,842 genes. After determining no difference in weak positive correlation of accessible peaks to gene expression based on proximity to the TSS, we found that most genes (1,650 out of 4,509) could be categorised as highly accessible and highly expressed (HA-HE) based on signals and counts above the 70th percentiles. The HA-HE genes are enriched for genes involved in salinity acclimation however, the overall set exhibits weak positive correlation of accessibility and gene expression. A higher correlation might not be observed in our datasets as accessible regions are not always associated with gene activity and can instead, be associated with repressed or poised for activation genes [28-31]. Like previous work in *C. elegans* [28], this would need to confirmed using relevant (H3K27me3 and H3K4 methylation) ChiP-Seq marks.

Whilst our previous work focused on regulatory variation in a phylogeny of five cichlid species, including Nile tilapia [7,8], no previous work has examined functional noncoding variation in the tilapia phylogeny. Building upon our recent work of identifying 69 million SNPs in the Nile tilapia genome based on resequencing populations of 27 tilapia species (Ciezarek, A. *et al.*, bioRxiv TBC, 2023), we found SNPs are enriched in noncoding regions and more specifically, gene promoter regions. We identified that gene promoter regions with accessible SNPs are enriched for processes associated with gill function e.g., response to stimulus, and therefore conclude that much like the adaptive radiations of East African cichlids [7,8], discrete

variation in regulatory regions can also drive gene regulatory rewiring of tilapia adaptive trait genes. This is supported by the fact that nearly all (99%) accessible gene promoter regions have a SNP with most (90%) being a rare variant. Gene promoter regions are enriched for TFBSs [8], as are accessible regions [47,48], and thus, variation, including rare variation [49], at TFBSs is a major target of genetic selection [50] and a key contributor to adaptive phenotypes [7,8,51]. Using TFBSs in accessible gene promoter regions, we identified a total of 3,398,768 transcription factor (TF) – target gene (TG) relationships in Nile tilapia. As we use experimental data, this is less than the 5,900,174 TF-TG edges we *in silico* predicted in Nile tilapia in our previous work [8], but similar to the 3,505,491 TF-TG relationships found in human based on *in silico* predictions overlapping ENCODE footprints [52].

Using a stringent criterion to identify candidate TF-TG relationships, involving prioritising statistically significant TFBSs with genetic variation between ecologically divergent species and genes that exhibit high accessibility and high expression (HA-HE) compared to all genes, we identified that the majority (71% - 1,168 genes) of all 1,650 HA-HE genes could be associated to aquaculture relevant traits, like salinity acclimation. This indicates that discrete variation in accessible gene promoter TFBSs could be a core factor driving high levels of gene expression associated with environmental tolerances of the gill. Some tilapia species can thrive in adverse environmental conditions, including varying levels of salinity and alkaline water. Nile tilapia is intolerant to high salinity (optimal is up to 16 parts per thousand, ppt) [53] and requires freshwater (pH 6-9) [54,55] whereas euryhaline species like *Alcolapia grahami*, nested within the *Oreochromis* genus [56], have adapted to extreme environments of pH 9-11.5 and high salt concentrations of >20 ppt [37,38]. Genetic variation in gene

promoter TFBSs between Nile tilapia (a freshwater species) and euryhaline species e.g., *A. grahami* with exclusive homozygous or heterozygous alternate sites could indicate positive selection acting on discrete mutations, that can be functional based on chromatin accessibility in Nile tilapia gill tissue, and subsequently associated to target gene expression of gill-specific traits relevant to aquaculture e.g., salinity tolerance. In *prlr1*, a gene associated with tilapia salinity response [32], a rare genetic variant between Nile tilapia and *A. grahami* in the STAT1 TFBSs could account for osmotic stress response in most tilapia species, but the loss of this site could enable gill adaptation in extremophiles like *A. grahami.* This is supported by the role of STAT1 as an osmotic stress induced transcriptional activator [36] and differential expression of prolactin receptors in the gills of *A. grahami* and freshwater tilapia species [57]. Furthermore, the functional impact of regulatory variation has been highlighted for the associated prolactin (*prl*) gene, whereby a TFBS polymorphism in the gene promoter has been associated with differential *prl* gene expression as well as growth in salt water of a single [58] and three out of nine [59] *O. mossambicus* × *O. niloticus* F_2 hybrid families. Prolactin is an osmoregulatory hormone that mediates water and ion permeability of the gills in euryhaline teleost fish, that is activated once it combines with the prolactin receptor, *prlr1*. We therefore suggest that discrete variation at functionally associated TFBSs e.g., STAT1 in the *prlr1* gene promoter, can ultimately regulate the activity of the freshwater-adapting prolactin hormone in euryhaline tilapia. Relatedly, the introduction and acclimation of freshwater species, like *O. esculentus*, to saline/alkaline waters e.g., Lake Rukwa in Tanzania [44,45], could be driven by genetic variation at the SREBF2 TFBS, and account for differential expression of *cldnh*. Since *cldnh* is responsible for reorganising gill tight junctions in response to salinity exposure of freshwater tilapia [41], important permeability changes

associated with salinity exposure of primarily freshwater tilapia species is likely controlled by the regulation of *cldnh* by SREBF2. Salinity acclimation first requires the detection of osmotic changes and then secondly, a physiological response to restore osmotic homeostasis. Accordingly, excessive ions are secreted by specialised cells in the gill, ionocytes, containing ion pumps and transporters [60,61]. As a result, there are differential gill-specific transcriptional responses to salinity exposure in different tilapia species, triggering turnover of ionocytes as well as immune and cell stress response [62]. Given the associated functional roles, we therefore suggest that the interactions of both STAT1-*pnr1* and SREBF2-*cldnh* are likely involved in salinity acclimation in species along the tilapia phylogeny. Discrete rewiring of existing ancestral variation at these important functional loci could be a mechanism to allow populations to rapidly adapt to new ecological niches e.g., *O. esculentus* introductions to Lake Rukwa [63]. The mechanistic effect of polymorphisms in gene promoter regions and functionally associated TFBSs has been previously shown for growth genes, ultimately affecting growth traits in largemouth bass [64] and Nile tilapia [65]. Our findings can therefore form the basis for breeding genotypes of desirable traits like salinity acclimation into farmed strains.

To conclude, by characterising the open chromatin landscape of Nile tilapia gill tissue using ATAC-seq, we identified accessible gene promoter regions with active TF footprints accounting for target gene expression in the gill. We integrate this data with noncoding regulatory SNPs from a 27 species tilapia phylogeny (Ciezarek, A. *et al.*, bioRxiv TBC, 2023), and as a case study, identify candidate genes with regulatory variation that could be associated with aquaculture relevant traits, like salinity tolerance. Using genetic variation along the 27 species tilapia phylogeny, we investigated the plausibility of

candidate regulators and regulatory regions having an association with the evolution of salinity acclimation in freshwater and euryhaline species. Overall, our integrative approach identified over 1000 candidate genes and associated regulators, and we find that discrete mutations in transcriptional regulatory sites of these genes could be driving gene regulatory evolution associated with salinity and osmotic stress acclimation of the gills along the tilapia phylogeny. This novel epigenome of not only Nile tilapia gill, but the first recognised assessment of open chromatin in any fish gill tissue, will enable a better functional annotation of noncoding regions to identify causative variants associated with aquaculture relevant traits. Now that the ATAC-seq protocol has been optimised for tilapia fish, we can characterise the open chromatin landscape in the other 26 tilapia species to confirm and prioritise variants identified in this study and define trait loci and genes for several economically beneficial traits e.g., growth, temperature tolerance, disease resistance, and nutrient metabolism. The novel methods and resources generated here can be ultimately used to guide genomic selection programmes aimed at the genetic improvement of species to acclimate to adverse environmental conditions.

Figures

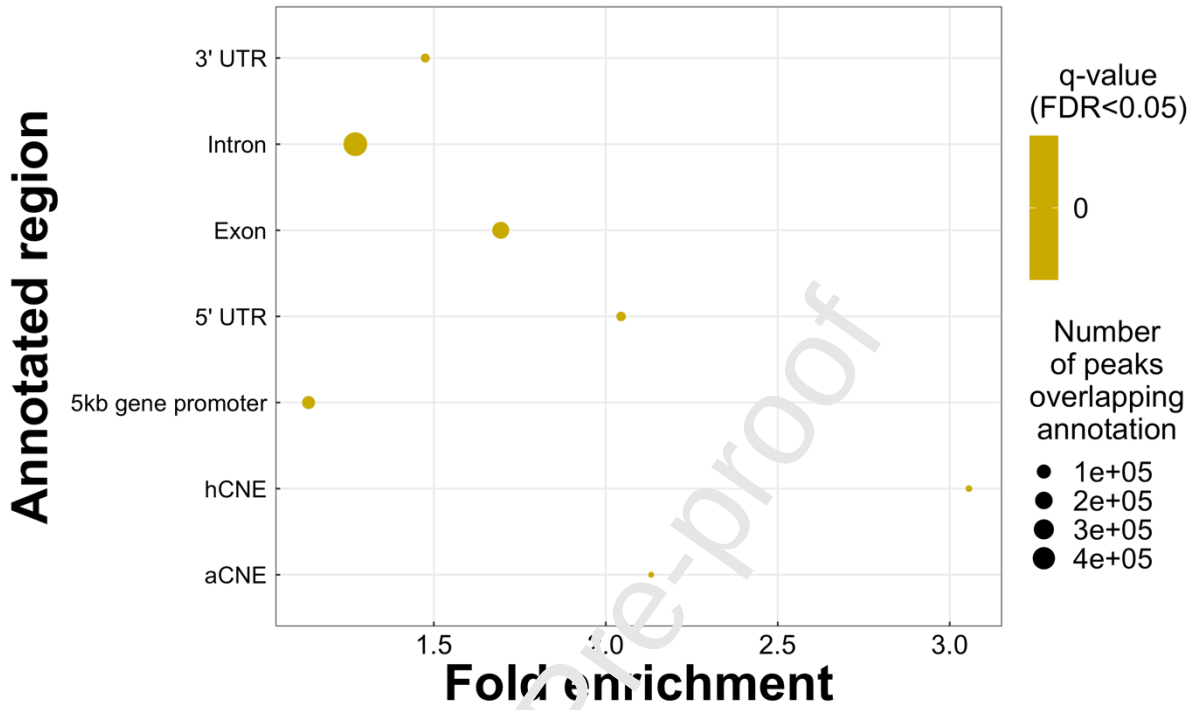

Fig. 1 – Annotation of open chromatin peaks in Nile tilapia gill. Fold enrichment of open chromatin (accessible) peaks overlapping coding and noncoding regions in the Nile tilapia genome. Circles show enriched annotated region (y-axis) of significance (FDR <0.05, to right) and fold enrichment (x-axis) values of all annotated peaks. Number of peaks overlapping each annotation shown by size of each circle.

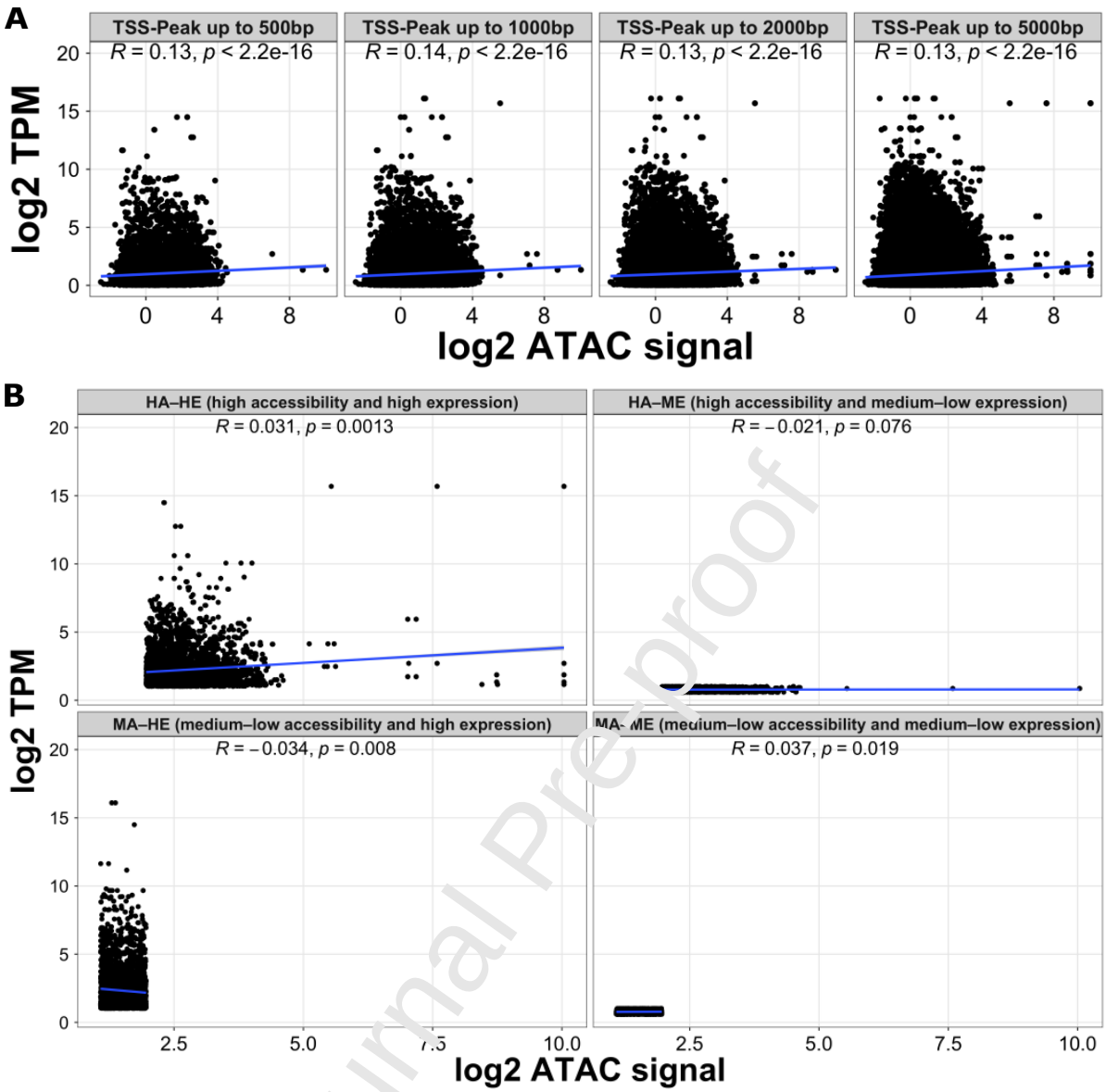

Fig. 2 – Correlation of accessible peaks accounting for gene expression change in Nile tilapia gill. (A) Average *log2* gene expression (as transcripts per million - TPM, y-axis) and average *log2* ATAC signal (as gene promoter peak counts, x-axis) for each gene across the three replicates, split into four categories based on peak summit distance to transcription start site (TSS). Correlation co-efficient (*R,* blue trendline) and *p-value* shown for each category. **(B)** Average *log2* gene expression (as transcripts per million - TPM, y-axis) and average *log2* ATAC signal (as all gene promoter peak counts i.e., TSS-peak up to 5000 bp category in Fig. 2a, x-axis) for each gene across the three replicates, split by high accessibility and high

expression to left (HA-HE - both average TPM and ATAC signal values are above the 70th percentiles) and to right, medium-low accessibility and high expression (MA-ME - average ATAC signal count less than the 50th percentile and TPM higher than the 70th percentile). Correlation co-efficient (*R,* blue trendline) and *p-value* shown for each category.

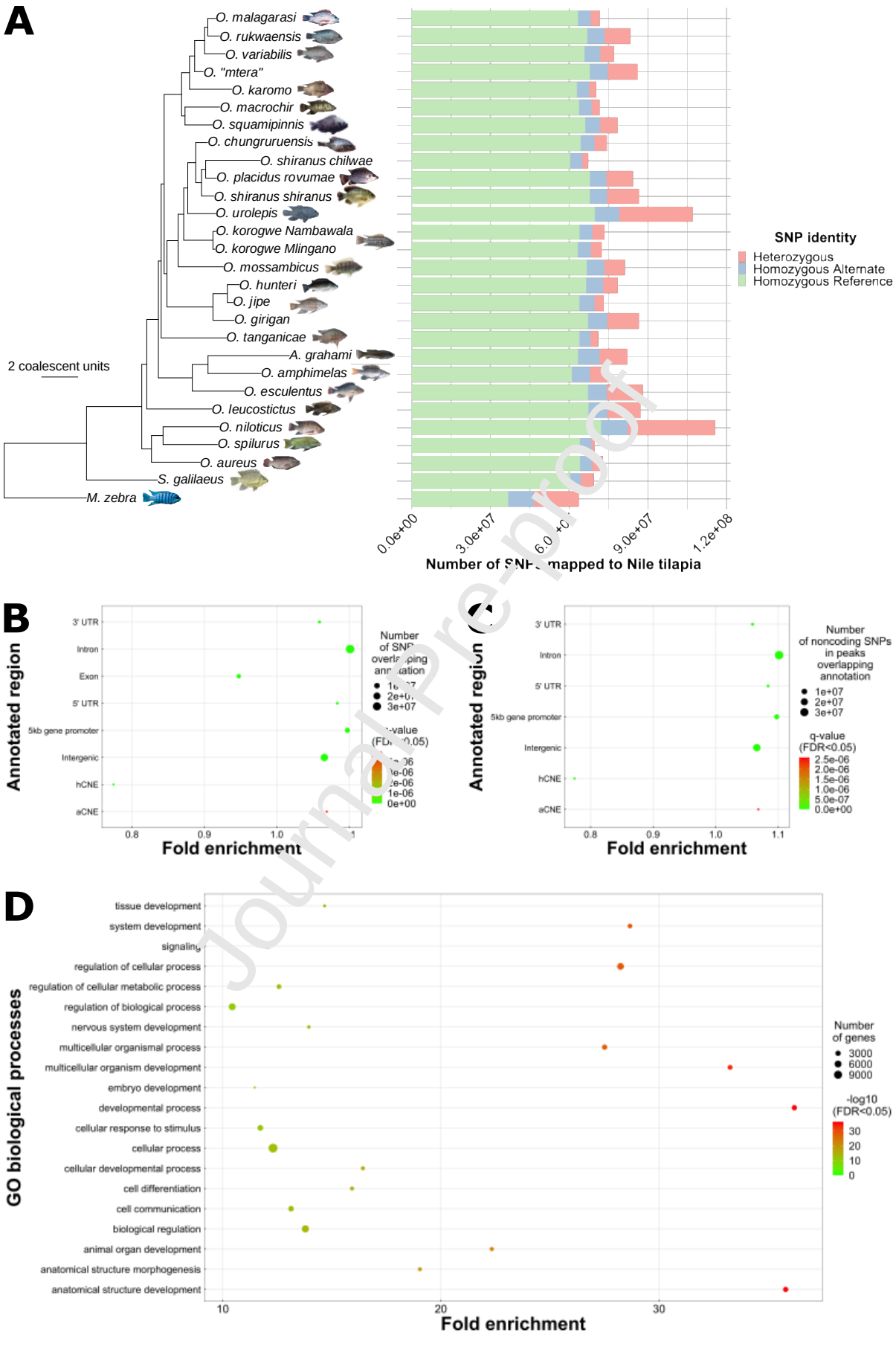

Fig. 3 – Single nucleotide polymorphisms (SNPs) in the Nile tilapia genome. (A) Coalescent phylogeny of 27 tilapia species and 2 outgroups inferred using ASTRAL (Ciezarek, A. *et al.*, bioRxiv TBC, 2023) with number and identity of SNPs in each species based on mapping to the Nile tilapia genome. Phylogenies inferred based on 14,535 10 kb windows across the genome. Units = coalescent units; All posterior probabilities = 1.0. Species photo credits: George Turner, Martin Genner, Antonia Ford, Benjamin Ngatunga, Geraldine Kavembe, and Fishbase (Luc de Vos, Brain Gatwicke, Andrew Nightingale, and Royal Museum for Central Africa). **(B)** Fold enrichment of SNPs overlapping coding and noncoding regions in the Nile tilapia genome. Circles show enriched annotated region (y-axis) of significance (FDR <0.05, heatmap to right) and fold enrichment (x-axis) values of all SNPs mapped to the Nile tilapia genome. Number of SNPs overlapping each annotation shown by size of each circle. **(C)** Fold enrichment of SNPs overlapping open chromatin (accessible) peaks in noncoding regions of the Nile tilapia genome. Circles show enriched annotated region (y-axis) of significance (FDR <0.05, heatmap to right) and fold enrichment (x-axis) values of all SNPs in annotated noncoding peaks. Number of noncoding SNPs in annotated noncoding peaks shown by size of each circle. **(D)** Gene ontology (GO) biological process enrichment of accessible gene promoter regions containing SNPs in the Nile tilapia genome. GO terms are subset for enrichment of -log10 (FDR < 0.05) of more than 10, with the full figure provided as Supplementary Fig. S5. Circles show enriched term (y-axis) of significance (FDR <0.05, heatmap to right) and fold enrichment (x-axis) values of GO terms. Number of genes with promoter regions containing SNPs shown by size of each circle.

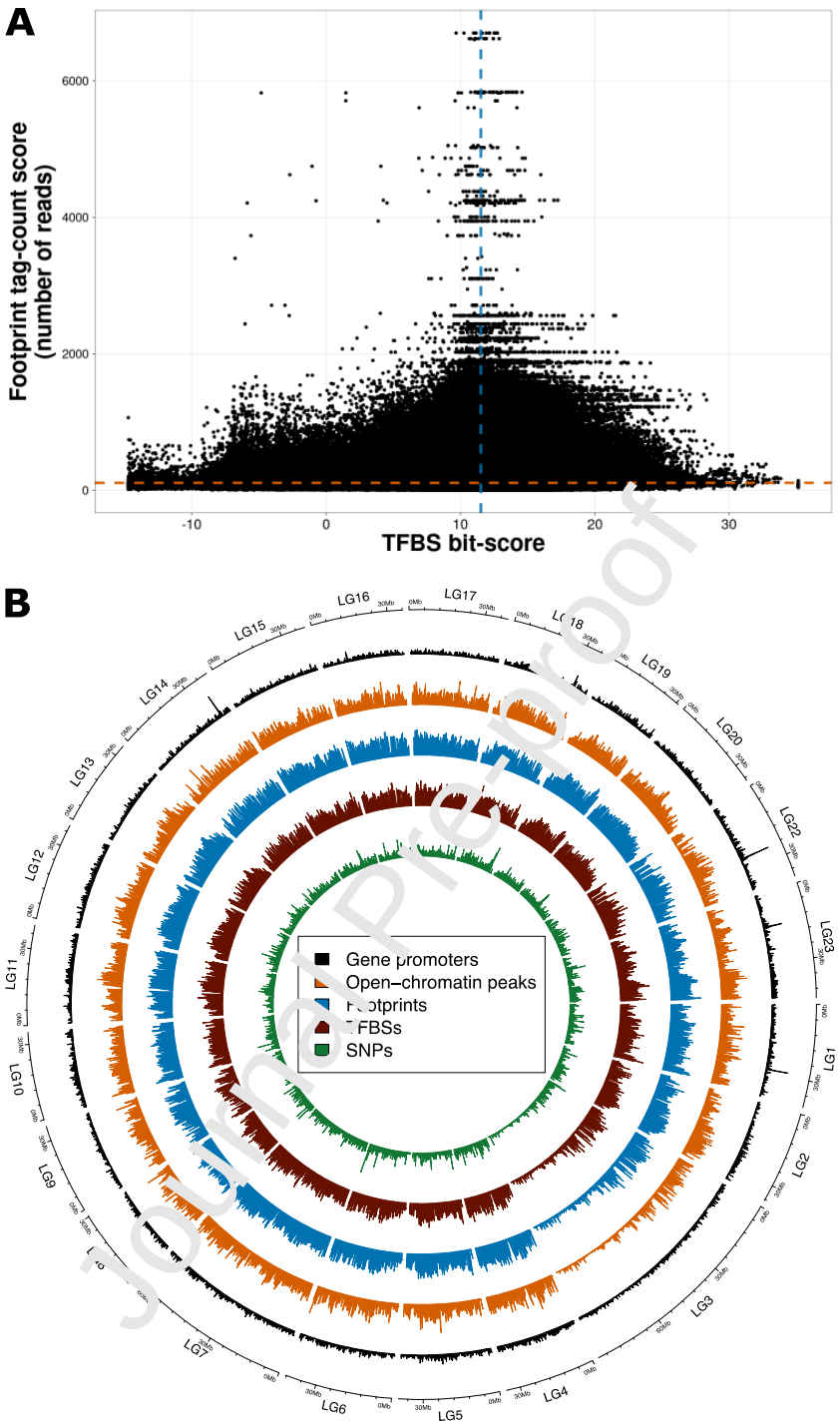

Fig. 4 – Prioritising genetic variation of TFBSs in accessible peaks accounting for gene expression change in Nile tilapia gill. (A) Tag-count (TC) score of TF footprints (y-axis) and TFBS bit-score of motif match (x-axis) of genome-wide predicted TFBSs based on footprints in accessible gill peaks. **(B)** Circos plot of open-chromatin peaks (orange), footprints (blue), TFBSs (maroon), SNPs (green), and gene promoter regions (black) on each Nile tilapia chromosome (outer track).

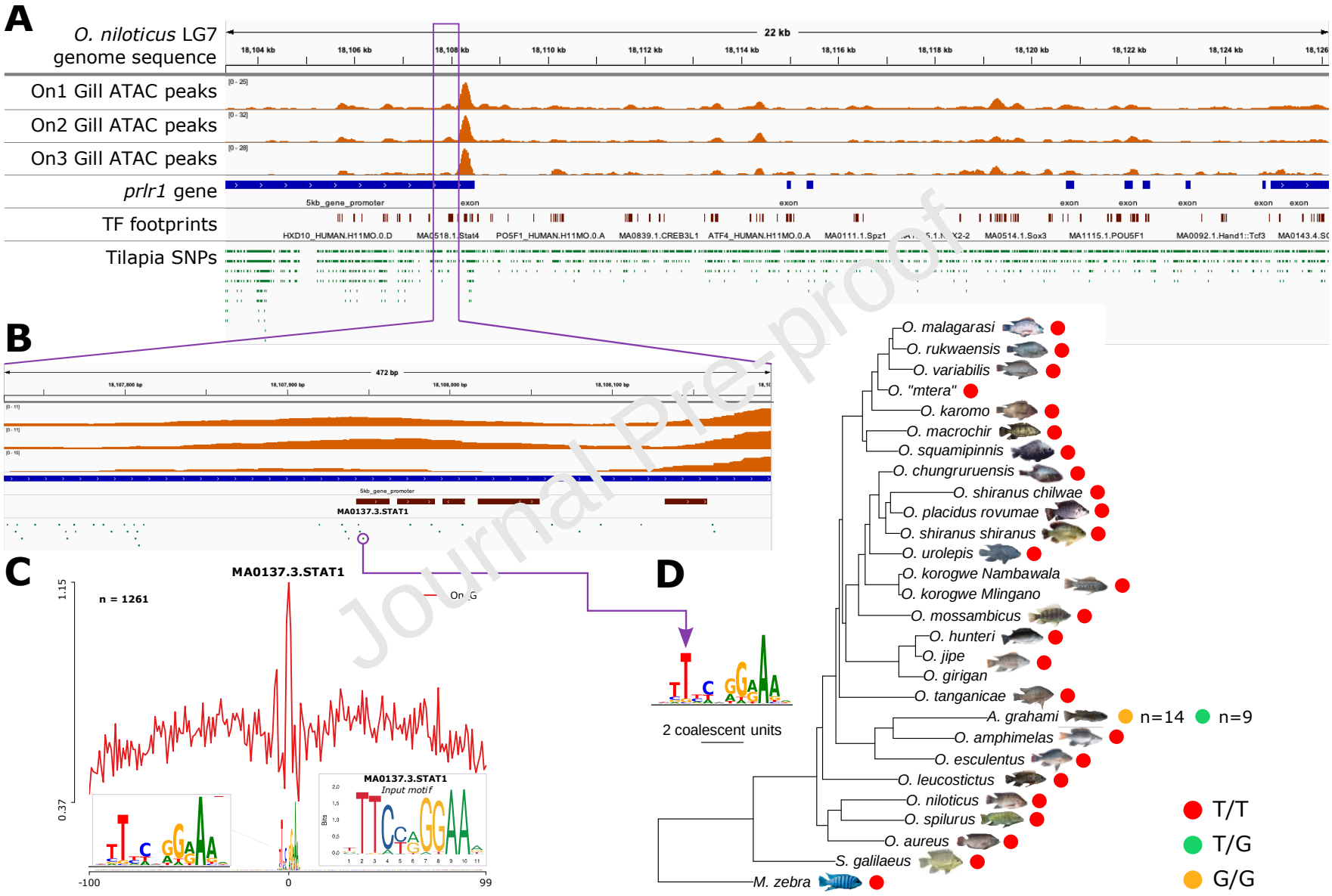

Fig. 5 – Accessible peaks in the Nile tilapia *prlr1* gene promoter with variation at the STAT1 TF footprint. (A) Nile tilapia track of the *prlr1* gene sequence (blue annotations) in LG7 with accessible peaks across three biological gill replicates (orange peaks), TF footprints (maroon marks) and tilapia SNPs (green marks). **(B)** Expanded *prlr1* gene promoter region (blue annotation) of accessible peaks (in orange) across three gill replicates above the STAT1 TF footprint (maroon mark) and containing a SNP (green mark with purple circle). **(C)** Line plot profile of the STAT1 TF footprint (n=1231) across Nile tilapia gene promoter regions showing positional (x-axis) signal of activity (red line, y-axis) centred on the footprint (at position 0) with flanking sequence. Expanded STAT1 TF footprint shown to left and original input motif used for predicting the site shown to right. **(D)** Variation from the SNP in Fig. 5b demarcated in the STAT1 TF footprint with homozygous reference (red dots), homozygous alternate (green dots) and heterozygous (yellow dots) alleles for the SNP shown for each species in the tilapia phylogeny (based on Fig. 3a).

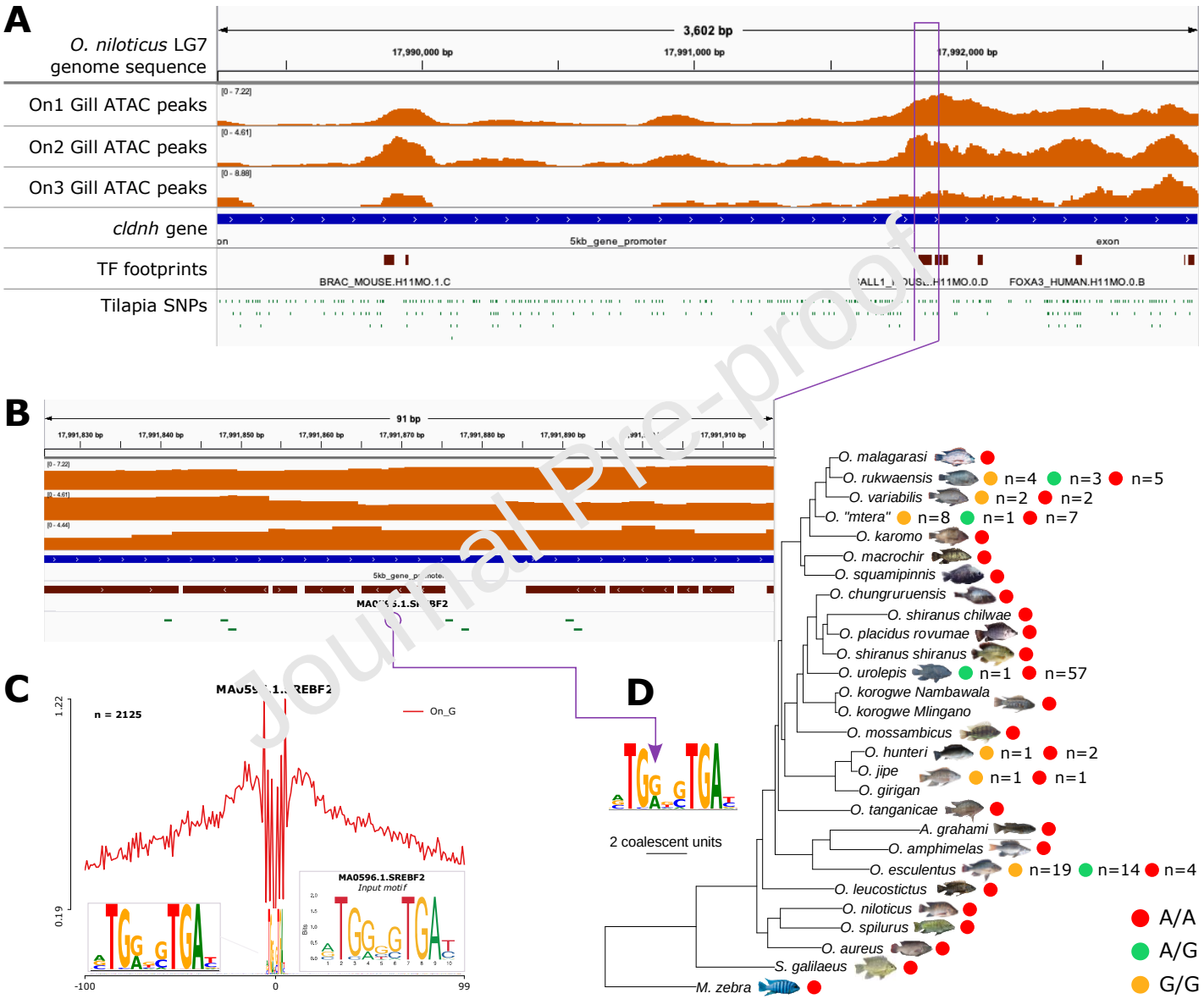

Fig. 6 – Accessible peaks in the Nile tilapia *cldnh* gene promoter with variation at the SREBF2 TF footprint. (A) Nile tilapia track of the *cldnh* gene sequence (blue annotations) in LG10 with accessible peaks across three biological gill replicates (orange peaks), TF footprints (maroon marks) and tilapia SNPs (green marks). **(B)** Expanded *cldnh* gene promoter region (blue annotation) of accessible peaks (in orange) across three gill replicates above the SREBF2 TF footprint (maroon mark) and containing a SNP (green mark with purple circle). **(C)** Line plot profile of the SREBF2 TF footprint (n=2125) across Nile tilapia gene promoter regions showing positional (x-axis) signal of activity (red line, y-axis) centred on the footprint (at position 0) with flanking sequence. Expanded SREBF2 TF footprint shown to left and original input motif used for predicting the site shown to right. **(D)** Variation from the SNP in Fig. 6b demarcated in the SREBF2 TF footprint with homozygous reference (red dots), homozygous alternate (green dots) and heterozygous (yellow dots) alleles for the SNP shown for each species in the tilapia phylogeny (based on Fig. 3a).

Materials and Methods

Tissue dissection

All animal procedures were approved by University of Stirling Animal Welfare and Ethical Review Body (AWERB) and carried out in accordance with approved guidelines. Three male *O. niloticus* individuals were sacrificed according to Home Office schedule 1 killing using overdose of MS-222 (tricaine) at University of Stirling, UK. The gill raker, arch and filaments were dissected from each individual, and one half was stored in RNAlater (1:5 ratio) for naked DNA and RNA extraction, and the other half immediately used for ATAC cell preparation.

Cell preparation

Around 50,000 cells were harvested from the gill filaments by taking a 1mm biopsy punch, counting nuclei and then spun at 500×g for 5 min at 4°C. Cells were washed once with 50µl of ice-cold 1× PBS buffer and pelleted at 500×g for 5 min at 4°C. Cell pellets were gently resuspended in 50µl of 0.05% cold lysis buffer (10mM Tris-HCl pH7.4, 10mM NaCl, 3mM $MgCl_2$, 0.05% IGEPAL CA-630), and immediately pelleted at 500×g for 10 min at 4°C and then stored on ice.

Transposition reaction and purification

Pelleted nuclei were gently resuspended in a 50 µL transposition reaction mix composed of 25 µL 2× TD Buffer (Illumina), 2.5 µL Tn5 Transposes (Illumina) and 22.5 µL Nuclease-Free H_2O. Transposition reaction was incubated at 37°C for 30 min and purified using the PCR purification MinElute Kit (QIAGEN), according to

manufacturer's protocol. Transposed DNA was eluted in 10 µL Elution Buffer (10mM Tris buffer, pH 8) and stored at -20°C.

PCR amplification

Transposed DNA was amplified in a final volume of 50 µL composed of 10 µL Transposed DNA, 10 µL Nuclease Free H_2O, 2.5 µL 25µM P7 adapter, 2.5 µL 25µM P5 adapter and 25 µL NEBNext High-Fidelity 2x PCR Master Mix (NEB). Transposed DNA was amplified for 5 min at 72°C, 30 secs at 98°C, and 11 cycles of 10 secs at 98°C, 30 secs at 63°C and 1 min at 72°C in a PCR thermocycler. Amplified transposed DNA was purified using the PCR purification MinElute Kit (QIAGEN) and eluted in 20 µL Elution Buffer (10mM Tris buffer, pH 8), according to manufacturer's protocol. To remove excess primers for final ATAC libaries, an additional 1× Agencourt AMPure XP bead (Beckman Coulter) clean-up was performed and eluted in 0.1× filtered TE. ATAC libraries were quantified on the Qubit 4 fluorometer (Invitrogen) and size distribution assessed on Agilent Tapestation and/or Bioanalyser.

DNA extraction

DNA was purified from gill filaments using the DNeasy Blood and Tissue kit (QIAGEN) according to manufacturer's protocol. DNA was quantified on the Nanodrop 2000 (Thermo Scientific) and Qubit 4 fluorometer (Invitrogen), and used 1 ng for DNA library preparation using the Nextera XT DNA Library Preparation kit (Illumina), according to manufacturer's protocol. Library size distribution was assessed on Agilent Tapestation and/or Bioanalyser. DNA libraries, obtained from

naked DNA, were used as internal controls to determine background levels of genomic DNA accessibility and Tn5 transposase sequence cleavage bias.

ATAC and control DNA sequencing

Three ATAC and three corresponding naked DNA control libraries were equimolar pooled and 50 bp paired-end sequenced at Earlham Institute on the Illumina NovaSeq 6000 platform using an S2 flow cell, generating an average of 90 million (ATAC-seq) and 16 million (naked DNA control) reads per library.

ATAC-seq and control DNA processing

Sequence adaptors were removed and trimmed for quality from raw paired-end reads using Trim Galore! (v 0.6.5) (https://github.com/FelixKrueger/TrimGalore) and FastQC (v 0.11.9) [66] using default parameters. Read alignment, post alignment filtering and ATAC peak calling were performed according to the ENCODE projects 'ATAC-seq Data Standards and Processing Pipeline' for replicated data (https://www.encodeproject.org/atac-seq/). Briefly, trimmed reads were mapped to the Nile tilapia reference genome (*O. niloticus* UMD_NMBU) [27] using bowtie2 (v2.2.6) [67], with parameters '–k 4 –X2000 –mm' and outputted in BAM format using SAMtools (v1.9) [68]. Since ATAC-seq generates high proportions (15-50% in a typical experiment) of mitochondrial mapped reads, any reads mapping to the mitochondrial genome were identified using BLAST (v2.3.0) [69] and removed from the BAM file using SAMtools (v1.9) [68]. The resulting BAM files were sorted, and duplicated reads were marked using Sambamba v0.6.5 [70]. Duplicated, unmapped, non-primary alignment, and failing platform QC reads were filtered out using SAMtools (v1.9) [68], retaining reads mapped as proper pairs, and fragment length distributions were

plotted using Picard (v1.140) (https://github.com/broadinstitute/picard/). At each step, the recommended parameters from the ENCODE pipeline were applied. BAM files were converted to *tagalign* files using Bedtools (v2.30.0) [71] and Tn5 shifting of ATAC mappings carried out prior to peak calling. Peaks were identified using macs2 (v2.1.1) [72,73] with the shifted tag as test and corresponding control DNA as input with parameters '-f BED -p 0.05 --nomodel --shift -75 --extsize 150'. Narrow peaks were used to create coverage tracks using *bedClip* and *bedToBigBed* in the UCSC-tools package (v333) (http://hgdownload.cse.ucsc.edu/admin/exe/). Following the ENCODE pipeline, Irreproducible Discovery Rate (IDR) peaks of true replicates were flagged as either true (<0.1) or false (≥0.1) using idr v2.0.4 (https://github.com/kundajelab/idr), taking reproducible (true) peaks between replicates. The fraction of reads in peaks (FRiP) were calculated using Bedtools (v2.30.0) [71] and demarcated as pass (>0.3) or acceptable (>0.2) according to ENCODE guidelines. FRiP was not used as a QC measure and instead, transcription start site (TSS) enrichment was calculated using ATACseqQC (v1.18.0) [74], with the TSS enrichment QC requirement of a significant signal value at the centre of the distribution being applied. All samples passed this criterion for further analysis. A union of peaks from all three biological replicates was created and used for subsequent analyses.

Identifying conserved noncoding elements (CNEs)

We use a similar approach as applied previously [6] to call CNEs in Nile tilapia based on evolutionary constraint in a five cichlid species phylogeny. A multiple genome alignment (MGA) of four cichlid genomes (*M. zebra* UMD2a [75]; *P. nyererei* v1 [6]; *A. burtoni* v1 [6]; *N. brichardi* v1 [6]) and Nile tilapia (*O. niloticus* UMD_NMBU [27]) was

created using Cactus (v2.0.3) [76] and outputted in Multiple Alignment Format (MAF). A neutral substitution model was created using the *phyloFit* function of PHAST (v1.5) [77] by fitting a time reversible substitution 'REV' model and parameters '—tree "((((Metriaclima_zebra,Pundamilia_nyererei),Astatotilapia_burtoni),Neolamprologus_brichardi),Oreochromis_niloticus)" --subst-mod REV'. The five cichlid MGA was split by *O. niloticus* chromosomes/scaffolds using the *mafSplit* function in the UCSC-tools package (v333) (http://hgdownload.cse.ucsc.edu/admin/exe/).

The neutral substitution model and MGAs of each *O. niloticus* reference chromosome/scaffold were used as input to predict conserved noncoding elements (CNEs) using the *phastCons* function of PHAST (v1.5) [77] with parameters '--target-coverage 0.3 --expected-length 30 --most-conserved --estimate-trees --msa-format MAF'. Conservation score outputs were used to define highly conserved CNEs (hCNEs) with sequence identity ≥90% over ≥30bp and pseudo accelerate/diverged CNEs (pseudo aCNEs) with sequence identity <90% over ≥30bp. The evolutionary conservation-acceleration (CONACC) score of each pseudo aCNE was predicted by running *phyloP* [78] of the PHAST (v1.5) package [77] using the likelihood ratio test (LRT) '--method LRT' on the CNE --features with their corresponding neutral substitution model in '--mode CONACC'. Pseudo aCNEs with a negative CONACC score, likelihood ratio <0.05, and significant divergence (altsubscale >1), were defined as significantly deviating from the neutral model, and therefore as true aCNEs.

Peak annotation and enrichment

Up to 5 kb gene promoter regions were annotated in the *O. niloticus* UMD1 [27] genome according to the method used in our previous study [8]. Narrow peaks either

overlapping or most proximal to all annotated features [27] in the genome were mapped using the *intersect* function of Bedtools (v2.30.0) [71], and each peak assigned to a feature and/or gene accordingly. The enrichment of accessible peaks in coding and noncoding regions was tested using the Genome Association Tester (GAT) tool [79]. The accessible peaks were provided as segments of interest to test, a bed file of annotations to test against, and workspace as the length of each scaffold/chromosome. GAT was ran using the following parameters: --verbose=5, --counter=segment-overlap, --ignore-segment-tracks, --qvalue-method=BH --pvalue-method=norm. We use the Benjamini-Hochberg [80] false discovery rate (FDR) to assess enrichment of peaks in annotated regions with a statistical cut-off of FDR < 0.05.

Gene Ontology (GO) enrichment

GO enrichment analyses of genes was conducted using the 'g:GOst' module of g:Profiler (https://biit.cs.ut.ee/gprofiler/gost) [81], version e105_eg52_p16_e84549f (February 2022), using the *O. niloticus* database. We use the false discovery rate (FDR) corrected hypergeometric *p*-value to assess enrichment of GO terms, with a statistical cut-off of FDR < 0.05.

Transcription factor (TF) footprinting

TF footprints were characterised using HINT-ATAC in the Regulatory Genomic Toolbox (v0.13.0) [82] using a stringent false positive rate (FPR) of 0.0001, with both Nile tilapia specific and cichlid-wide position weight matrices (PWMs) as defined in our previous study [8], as well as vertebrate PWMs from JASPAR (v9.0) [83], HOCOMOCO [84], GTRD [85], and UniPROBE [86]. TF footprint line plots were generated

with the 'differential analysis' module of HINT-ATAC in the Regulatory Genomic Toolbox (v0.13.0) package [82] using bias-corrected signals. Redundant TFBSs in the same, or overlapping positions were filtered based on selecting the highest bit-score of motif match for each overlapping TF.

RNA extraction and sequencing

RNA was purified from each tissue using the RNeasy Plus Mini kit (QIAGEN) according to manufacturer's protocol. RNA and DNA content were quantified on the Qubit 4 fluorometer (Invitrogen) and integrity assessed on Agilent Tapestation and/or Bioanalyser, taking samples with RIN≥7 and <15% genomic DNA. A total of 45/55 (82%) samples passed these criteria for selection (Supplementary Table S2). A total of 45 stranded RNA libraries were prepared using the NEBNext Ultra II Direction RNA-seq kit according to manufacturer's protocol. All stranded RNA-seq libraries were equimolar pooled and 150 bp paired-end sequenced at Earlham Institute on 1 lane of the Illumina NovaSeq 6000 platform using an S4 v1.5 flow cell, generating an average of 70 million reads per library.

RNA-seq processing

Read quality was assessed using FastQC (v 0.11.9) [66] and Trim Galore! (v 0.6.5) (https://github.com/FelixKrueger/TrimGalore) was used to remove adapters and trim for low-quality from the raw paired-end reads using default settings. All reads were then mapped to the *O. niloticus* UMD1 genome [27] using HISAT2 (v 2.2.1) [87] with default parameters. Mapping QC was carried out using QualiMap (v 2.2.1) [88] with default 'rnaseq' parameters. The final BAM file was sorted using samtools (v1.16.1)

[68] and transcript abundance was calculated using 'htseq-count' in the HTSeq (v 2.0.2) [89] package.

Identifying ATAC peak and gene expression association

Given the large number of accessible peaks and overlapping genes (127,602 peaks and 20,248 genes), we devised an approach to reduce and identify the peaks that could regulate their gene targets expression. For gene expression, we use an approach applied previously [90] to calculate transcript per million (TPM) values by first normalizing the transcript count by the gene length, as calculated using GTFtools [91], followed by the library size, as calculated using QualiMap (v 2.2.1) [88], and then carrying out a *log2*(x +1) transformation of 1) each genes TPM in each replicate; and 2) the mean TPM for each gene across biological replicates. Each genes mean TPM across the three replicates is used to assess correlations with ATAC signal whereas replicate specific TPM are used to study any differences between replicates. The ATAC signal was processed using an approach applied previously [92] to obtain the number of independent Tn5 insertions in collated gene promoter peaks; the number of insertion sites (peak counts) was counted using collated narrow peaks against a merged BAM file of all three replicates using Bedtools (v2.30.0) *coverage* [71]. After creating a peak count matrix, the counts matrix was normalised using edgeR [93] counts per million (CPM) 'log=TRUE, prior.count=5' followed by a quantile normalization using the preprocessCore (https://github.com/bmbolstad/preprocessCore) *normalize.quantiles* module in R (v 4.4.2). After merging the three replicates ATAC signal using the *log2* average from the normalized counts matrices, the average *log2*-TPM and *log2*-ATACSignal for each gene was plotted with the correlation co-efficient (r) calculated in R (v 4.4.2).

Based on an approach devised previously [31] to categorise putative activated, repressed or poised genes, peak-gene relationships are assigned to one of four groups according to lower than 50th (medium-low) and higher than 70th (high) percentiles of gene promoter accessibility (*log2*-ATACSignal) and gene expression (*log2*-TPM). The four groups are (1) MA–ME (medium–low accessibility and medium–low expression); (2) HA–HE (high accessibility and high expression); (3) HA–ME (high accessibility and medium–low expression); (4) MA–HE (medium–low accessibility and high expression). Any genes that did not fall into one of the four groups were not used in the analysis to maintain stringent gene groups.

Identifying SNPs with exclusive variant types between Nile tilapia and other species

The VCF file of the 69 million SNPs mapped to the Nile tilapia (*O. niloticus* UMD_NMBU [27]) reference genome in our previous study (Ciezarek, A. *et al.*, bioRxiv TBC, 2023) was first normalised to split multiallelic sites into multiple rows using bcftools (v 1.12) [68] with the parameters 'norm -m-any -Ov'. After outputting the first four columns from the resulting file, we use bcftools 'query' to tabulate the number of samples having each variant type, and then bcftools 'view' to report each sample with each type of variant. SNPs that are present in prioritised TFBSs of candidate genes are then filtered for variant types (heterozygous and/or homozygous alternate) that are exclusive to Nile tilapia and any other species.

Declarations

Data Availability

All sequencing data generated for this article is available under the ENA study accession PRJEB59919. All other data underlying this article is available in the article, its supplementary material or in the following repository: https://doi.org/10.6084/m9.figshare.22100825.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

The authors acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation. The authors would like to acknowledge Tom Barker, Vanda Knitlhoffer, Leah Catchpole, and Suzanne Henderson of the Genomics Pipelines Group at Earlham Institute for data generation including preparation of RNA Seq libraries, pooling, and sequencing. The authors would also like to acknowledge the Scientific Computing group, as well as support for the physical HPC infrastructure and data centre delivered via the NBI Research Computing group.

Funding

The investigations were supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation; this research was funded by the BBSRC Core Strategic Programme Grant BB/CSP1720/1 (TKM, AM, AC, FDP and WH) and its constituent work packages (BBS/E/T/000PR9818 and

BBS/E/T/000PR9819). Part of this work was delivered via the BBSRC National Capability in Genomics and Single Cell Analysis (BBS/E/T/000PR9816) at Earlham Institute by members of the Genomics Pipelines Group.

Author contributions

KR and DP sacrificed and dissected fresh fish tissues; TKM prepared tissue-specific cells to perform transposition reactions and purifications for ATAC libraries; TKM and AM performed tagmentation of genomic DNA controls, library amplification and QC; AC identified variants in the tilapia phylogeny; TKM processed ATAC-seq and RNA-seq data, identified CNEs, annotated peaks with enrichment analysis, ran GO enrichment, identified TF footprints, and identified ATAC peak and gene expression associations; TKM and WH wrote the manuscript with input from AM, AC, KR, DP and FDP.

Corresponding authors

Correspondence to Tarang.Mehta@earlham.ac.uk

References

1. FAO. The State of World Fisheries and Aquaculture 2022. Towards Blue Transformation. *Rome, FAO* (2022).
2. Intergovernmental Panel on Climate Change. Working, G., II. *Climate change 2022 : impacts, adaptation and vulnerability : the Working Group II contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, (IPCC, Intergovernmental Panel on Climate Change, Geneva, Switzerland, 2022).
3. Vitti, J.J., Grossman, S.R. & Sabeti, P.C. Detecting natural selection in genomic data. *Annual review of genetics* **47**, 97-120 (2013).
4. Hong Xia, J. *et al.* Signatures of selection in tilapia revealed by whole genome resequencing. *Scientific Reports* **5**, 14168-14168 (2015).
5. Cadiz, M.I. *et al.* Whole genome re-sequencing reveals recent signatures of selection in three strains of farmed Nile tilapia (Oreochromis niloticus). *Sci Rep* **10**, 11514 (2020).
6. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **2**, 17-19 (2014).
7. Mehta, T.K. *et al.* Evolution of miRNA-Binding Sites and Regulatory Networks in Cichlids. *Mol Biol Evol* **39**, msac146 (2022).
8. Mehta, T.K. *et al.* Evolution of regulatory networks associated with traits under selection in cichlids. *Genome Biology* **22**, 25-25 (2021).
9. Gompel, N., Prud'Homme, B., Wittkopp, P.J., Kassner, V.A. & Carroll, S.B. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* (2005).
10. Prud'homme, B. *et al.* Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* (2006).
11. Jeong, S., Rokas, A. & Carroll, S.B. Regulation of Body Pigmentation by the Abdominal-B Hox Protein and Its Gain and Loss in Drosophila Evolution. *Cell* (2006).
12. Cretekos, C.J. *et al.* Regulatory divergence modifies limb length between mammals. *Genes and Development* (2008).

13. Chan, Y.F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitxl enhancer. *Science (New York, N.Y.)* **327**, 302-305 (2010).
14. Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. in *Nature Reviews Genetics* (2019).
15. Liu, Z., Zhou, T. & Gao, D. Genetic and epigenetic regulation of growth, reproduction, disease resistance and stress responses in aquaculture. *Front Genet* **13**, 994471 (2022).
16. Zhong, H. *et al.* DNA methylation of pituitary growth hormone is involved in male growth superiority of Nile tilapia (Oreochromis niloticus). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **171**, 42-48 (2014).
17. Podgorniak, T., Brockmann, S., Konstantinidis, I. & Fernandes, J.M.O. Differences in the fast muscle methylome provide insight into sex-specific epigenetic regulation of growth in Nile tilapia during early stages of domestication. *Epigenetics* **14**, 818-836 (2019).
18. Konstantinidis, I. *et al.* Epigenetic mapping of the somatotropic axis in Nile tilapia reveals differential DNA hydroxymethylation marks associated with growth. *Genomics* **113**, 2951-2964 (2021).
19. Wan, Z.Y. *et al.* Genome-wide methylation analysis identified sexually dimorphic methylated regions in hybrid tilapia. *Sci Rep* **6**, 35903 (2016).
20. Sun, L.X. *et al.* Global DNA Methylation Changes in Nile Tilapia Gonads during High Temperature-Induced Masculinization. *PLoS One* **11**, e0158483 (2016).
21. Wang, Y.Y. *et al.* Epigenetic control of cyp19a1a expression is critical for high temperature induced Nile tilapia masculinization. *Journal of Thermal Biology* **69**, 76-84 (2017).
22. Kratochwil, C.F. & Meyer, A. Mapping active promoters by ChIP-seq profiling of H3K4me3 in cichlid fish - a first step to uncover cis-regulatory elements in ecological model teleosts. *Molecular Ecology Resources*, n/a-n/a (2014).
23. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of

open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-8 (2013).

24. Corces, M.R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods* (2017).
25. Buenrostro, J.D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).
26. A, C.D. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
27. Conte, M.A., Gammerdinger, W.J., Bartie, K.L., Penman, D.J. & Kocher, T.D. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. *BMC Genomics* **18**, 341-341 (2017).
28. Daugherty, A.C. *et al.* Chromatin accessibility dynamics reveal novel functional enhancers in C. elegans. *Genome Res* **27**, 2096-2107 (2017).
29. Ackermann, A.M., Wang, Z., Schug, J., Naji, A. & Kaestner, K.H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol Metab* **5**, 233-244 (2016).
30. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-86 (2014).
31. Starks, R.R., Biswas, A., Jain, A. & Tuteja, G. Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenetics Chromatin* **12**, 16 (2019).
32. Yamaguchi, Y. *et al.* Acute salinity tolerance and the control of two prolactins and their receptors in the Nile tilapia (Oreochromis niloticus) and Mozambique tilapia (O. mossambicus): A comparative study. *General and comparative endocrinology* **257**, 168-176 (2018).
33. Su, H., Ma, D., Zhu, H., Liu, Z. & Gao, F. Transcriptomic response to three osmotic stresses in gills of hybrid tilapia (Oreochromis mossambicus female x O. urolepis hornorum male). *BMC Genomics* **21**, 110 (2020).
34. Hu, Y.C., Chung, M.H. & Lee, T.H. An assay of optimal cytochrome c oxidase activity in fish gills. *Anal Biochem* **553**, 38-45 (2018).
35. Gusmao, E.G., Allhoff, M., Zenke, M. & Costa, I.G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat Methods* **13**, 303-9 (2016).

36. Li, Y. *et al.* Role of p38α Map kinase in Type I interferon signaling. *J Biol Chem* **279**, 970-9 (2004).
37. Johansen, K., Maloiy, G.M. & Lykkeboe, G. A fish in extreme alkalinity. *Respir Physiol* **24**, 159-62 (1975).
38. Wood, C.M. *et al.* Mammalian metabolic rates in the hottest fish on earth. *Sci Rep* **6**, 26990 (2016).
39. Gibbons, T.C., Metzger, D.C.H., Healy, T.M. & Schulte, P.M. Gene expression plasticity in response to salinity acclimation in threespine stickleback ecotypes from different salinity habitats. *Mol Ecol* **26**, 2711-2725 (2017).
40. Tipsmark, C.K. *et al.* Regulation of gill claudin paralogs by salinity, cortisol and prolactin in Mozambique tilapia (Oreochromis mossambicus). *Comp Biochem Physiol A Mol Integr Physiol* **199**, 78-86 (2016).
41. Tipsmark, C.K., Baltzegar, D.A., Ozden, O., Grubb, B.J. & Borski, R.J. Salinity regulates claudin mRNA and protein expression in the teleost gill. *Am J Physiol Regul Integr Comp Physiol* **294**, R1004-14 (2008).
42. Shemsanga, C. *et al.* Origin and mechanisms of high salinity in Hombolo Dam and groundwater in Dodoma municipality Tanzania, revealed. *Applied Water Science* **7**, 2883-2905 (2017).
43. Katunzi, E.F.B. Conservation of satellite lakes as refugia for the endangered species of Lake Victoria: a case study of Lake Malimbe. *National Fisheries Resources Research Institute Publications* (1995).
44. Mapenzi, L.L., Shimba, M.J., Moto, E.A., Maghembe, R.S. & Mmochi, A.J. Heavy metals bio-accumulation in tilapia and catfish species in Lake Rukwa ecosystem Tanzania. *Journal of Geochemical Exploration* **208**, 106413 (2020).
45. Barker, P., Telford, R., Gasse, F. & Thevenon, F. Late Pleistocene and Holocene palaeohydrology of Lake Rukwa, Tanzania, inferred from diatom analysis. *Palaeogeography, Palaeoclimatology, Palaeoecology* **187**, 295-305 (2002).
46. FAO. The State of World Fisheries and Aquaculture 2022. Towards Blue Transformation. *Rome, FAO* (2022).
47. Yang, H. *et al.* A map of cis-regulatory elements and 3D genome structures in zebrafish. *Nature* **588**, 337-343 (2020).

48. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
49. Martin-Trujillo, A. *et al.* Rare genetic variation at transcription factor binding sites modulates local DNA methylation profiles. *PLoS Genet* **16**, e1009189 (2020).
50. Arbiza, L. *et al.* Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* **45**, 723-9 (2013).
51. Tseng, C.C. *et al.* Genetic Variants in Transcription Factor Binding Sites in Humans: Triggered by Natural Selection and Triggers of Diseases. *Int J Mol Sci* **22**(2021).
52. Plaisier, C.L. *et al.* Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis. *Cell Systems* **3**, 172-186 (2016).
53. El-Leithy, A.A.A. *et al.* Optimum salinity for Nile tilapia (Oreochromis niloticus) growth and mRNA transcripts of ion-regulation, inflammatory, stress- and immune-related genes. *Fish Physiol Biochem* **45**, 1217-1232 (2019).
54. USEPA. Quality criteria for water. *Washington, D.C.: Office of Water Regulations and Standards.* (1986).
55. Mustapha, M.K. & Atolaghe, S.D. Tolerance level of different life stages of Nile tilapia Oreochromis niloticus (Linnaeus, 1758) to low pH and acidified waters. *The Journal of Basic and Applied Zoology* **79**, 46 (2018).
56. Ford, A.G.P. *et al.* Molecular phylogeny of Oreochromis (Cichlidae: Oreochromini) reveals mito-nuclear discordance and multiple colonisation of adverse aquatic environments. *Mol Phylogenet Evol* **136**, 215-226 (2019).
57. Kavembe, G.D., Franchini, P., Irisarri, I., Machado-Schiaffino, G. & Meyer, A. Genomics of Adaptation to Multiple Concurrent Stresses: Insights from Comparative Transcriptomics of a Cichlid Fish from One of Earth's Most Extreme Environments, the Hypersaline Soda Lake Magadi in Kenya, East Africa. *Journal of Molecular Evolution* (2015).
58. Streelman, J.T. & Kocher, T.D. Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. *Physiol Genomics* **9**, 1-4 (2002).
59. Velan, A. *et al.* Association between polymorphism in the Prolactin I promoter and growth of tilapia in saline-water. *Aquaculture Reports* **1**, 5-9 (2015).

60. Edwards, S.L. & Marshall, W.S. 1 - Principles and Patterns of Osmoregulation and Euryhalinity in Fishes. in *Fish Physiology*, Vol. 32 (eds. McCormick, S.D., Farrell, A.P. & Brauner, C.J.) 1-44 (Academic Press, 2012).
61. Zadunaisky, J.A. Chloride cells and osmoregulation. *Kidney Int* **49**, 1563-7 (1996).
62. Campo, A. *et al.* Different transcriptomic architecture of the gill epithelia in Nile and Mozambique tilapia after salinity challenge. *Comp Biochem Physiol Part D Genomics Proteomics* **41**, 100927 (2022).
63. Shechonge, A. *et al.* Widespread colonisation of Tanzanian catchments by introduced Oreochromis tilapia fishes: the legacy from decades of deliberate introduction. *Hydrobiologia* **832**, 235-253 (2019).
64. Li, X.H. *et al.* Polymorphisms in the 5′ flanking region of the insulin-like growth factor I gene are associated with growth traits in largemouth bass Micropterus salmoides. *Fisheries Science* **75**, 351-358 (2009).
65. Jaser, S.K.K., Dias, M.A.D., Lago, A.A., Reis Neto, R.V. & Hilsdorf, A.W.S. Single nucleotide polymorphisms in the growth hormone gene of Oreochromis niloticus and their association with growth performance. *Aquaculture Research* **48**, 5835-5845 (2017).
66. Andrews, S. FastQC: A quality control tool for high throughput sequence data [Online]. *Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2010).
67. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. in *Nature Methods*, Vol. 9 357-359 (2012).
68. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008-giab008 (2021).
69. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421-421 (2009).
70. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034 (2015).
71. Quinlan, A.R. & Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
72. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq enrichment using MACS. *Nature Protocols* **7**, 1728-1740 (2012).

73. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**, R137-R137 (2008).
74. Ou, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC Genomics* **19**, 169-169 (2018).
75. Conte, M.a. & Kocher, T.D. An improved genome reference for the African cichlid, Metriaclima zebra. *BMC Genomics* **16**, 724-724 (2015).
76. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246-251 (2020).
77. Hubisz, M.J., Pollard, K.S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* **12**, 41-51 (2011).
78. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20**, 110-21 (2010).
79. Heger, A., Webber, C., Goodson, M., Ponting, C.P. & Lunter, G. GAT: A simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046-2048 (2013).
80. Benjamini, Y. & Hochberg, Y. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289-300 (1995).
81. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191-W198 (2019).
82. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biology* **20**, 45-45 (2019).
83. Castro-Mondragon, J.A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **50**, D165-D173 (2022).
84. Kulakovskiy, I.V. *et al.* HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research* **41**, D195-202 (2013).
85. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: A database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Research* **45**, D61-D67 (2017).

86. Hume, M.A., Barrera, L.A., Gisselbrecht, S.S. & Bulyk, M.L. UniPROBE, update 2015: New tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Research* **43**, D117-D122 (2015).
87. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907-915 (2019).
88. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics (Oxford, England)* **32**, 292-294 (2016).
89. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166-169 (2015).
90. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
91. Li, H.-D., Lin, C.-X. & Zheng, J. GTFtools: a software package for analyzing various features of gene models. *Bioinformatics* **38**, 4806-4808 (2022).
92. Sanghi, A. *et al.* Chromatin accessibility associates with protein-RNA correlation in human cancer. *Nat Commun* **12**, 5732 (2021).
93. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

CRediT author statement

Tarang K. Mehta: Conceptualization, Methodology, Software, Formal analysis, Validation, Investigation, Writing- Original draft preparation, Project administration, Writing- Reviewing and Editing. **Angela Man:** Validation, Investigation, Writing- Reviewing and Editing. **Adam Ciezarek:** Software, Formal analysis, Writing- Reviewing and Editing. **Keith Ranson:** Resources, Investigation, Writing- Reviewing and Editing. **David Penman:** Resources, Investigation, Writing- Reviewing and Editing. **Federica Di-Palma:** Funding acquisition, Writing- Reviewing and Editing. **Wilfried Haerty:** Funding acquisition, Investigation, Supervision, Writing- Reviewing and Editing.

Highlights

- 85% of all Nile tilapia genes have associated accessible regions in gill tissue.
- SNPs from 27 tilapia species are enriched in accessible gene promoter regions.
- Accessible gene promoter regions associated to the expression of ~15k genes.
- Regulatory variation identified >1k genes associated with environmental tolerance.
- Discrete regulatory mutations drive salinity acclimation of tilapia gills.