Chromatin accessibility 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 developed and optimised ATAC-seq from *O. niloticus* gill tissues to identify regulatory regions accounting for gene expression associated with gill adaptations. We find that SNPs from 27 tilapia species are enriched in noncoding regions, with 95% of accessible gene promoter regions being SNP-containing. Regulatory variants of TF binding sites are likely driving gene expression differences associated with tilapia gill adaptations, and differentially segregate in freshwater and euryhaline tilapia species. The generation of novel open chromatin data integrated with gene expression and genetic variants revealed candidate genes, 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. 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 [2]. 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) [3,4], 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 [3,4]. 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 [3].

53 We recently generated genome-wide sequencing data across 575 individuals of 27
54 tilapia species from across East Africa, to identify 69 million single nucleotide
55 polymorphisms (SNPs) in the *O. niloticus* genome assessing phylogenomic patterns
56 of recent and ancient hybridisation in tilapia (Ciezarek, A. *et al.*, bioRxiv TBC, 2023).
57 Whilst selection in noncoding regulatory regions, including transcription factor
58 binding sites (TFBSs) has been suggested as a key molecular mechanism
59 contributing to the evolutionary diversification of East African cichlids, including Nile
60 tilapia [5-7], no study has evaluated the *cis*-regulatory impact of noncoding variation
61 towards adaptive traits in a tilapia phylogeny.
62
63 Several studies have provided evidence on the role of *cis*-regulatory elements
64 (CREs) towards morphological diversity [8-10], including in mammals [11] and fish [12].
65 Whilst such studies identified CREs at particular loci, epigenetic sequencing
66 methods can capture chromatin accessibility and TF-binding at the genome-wide
67 level [13]. Only a handful of studies, all using DNA methylation, have applied
68 epigenetic approaches to study adaptive traits in tilapia [14]; this includes the
69 regulation of tilapia growth [15-17], sexual dimorphism [18], and sex determination [19,20].
70 ChiP-seq approaches have also been used to map active promoter elements
71 associated with Nile tilapia fin development [21]. However, such techniques are only
72 able to identify repressive or active marks at specific loci, and in the case of ChiP-
73 seq, direct interactions between a specific protein and DNA. In contrast, the assay
74 for transposase-accessible chromatin using sequencing (ATAC-seq) approach can
75 robustly identify genome-wide chromatin accessibility and TF-occupancy using very
76 few [22,23] or even single cells [24,25]. Despite the robustness of ATAC-Seq and its
77 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.

91 ## Results

92

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

95 To characterise functional noncoding regions based on chromatin accessibility, we
96 performed high-depth ATAC-seq of gill tissue from three replicate male Nile tilapia
97 individuals, identifying 301,293 open chromatin (accessible) reproducible peaks
98 between replicates (see '*Materials and Methods*'), that could be associated to 25,092
99 (average of 12 accessible peaks per gene) out of 29,552 (85%) annotated Nile tilapia
100 genes [26]. We first annotated (Supplementary Table S1) and tested the enrichment of
101 accessible peaks in coding and noncoding regions, finding that peaks are mostly
102 enriched (fold enrichment >2.1, adjusted *P*-value <0.05) in both sets of conserved
103 noncoding elements (CNEs) (Fig. 1) that we annotated in the Nile tilapia (*O. niloticus*
104 UMD_NMBU [26]) reference genome (see '*Materials and Methods*').

105

106 We focused on gene promoter regions, taken as up to 5 kb from the transcription
107 start site (TSS) [7], as they harbour regulatory binding sites e.g., TFBSs that have
108 functionally diverged [6,7], and can be associated with gene expression differences in
109 Nile tilapia and East African cichlids [7]. Out of the 301,293 open chromatin
110 (accessible) peaks, 127,602 accessible peaks (average size of 464 bp) were found
111 within gene promoters of 20,248 genes in gill tissue, with missing genes being better
112 associated to lymphoid tissues (see '*Supplementary Information'*, Supplementary
113 Fig. S1 – S2).

114

115 **Accessible gene promoter regions could account for expression changes of**
116 **genes associated with osmoregulation and salinity tolerance**
117 To examine any correlation between gene transcription and accessible gene
118 promoter regions, RNA-seq data was generated from the same gill tissue sample of
119 the three replicates and gene expression measured using transcripts per million
120 (TPM) (see '*Materials and Methods*'). Of the 29,552 Nile tilapia genes in the genome,
121 we identified and calculated the expression (as TPM) of 20,717 (70%) genes. We
122 found 96,784 (76%) accessible peaks could be associated to the expression of
123 14,842 genes.
124
125 We first tested whether these peaks are better correlated to their corresponding
126 gene expression based on proximity to the transcription start site (TSS) (Fig. 2a). For
127 this, accessible peaks are categorised into four intervals of either 500 bp (12,978
128 peaks), 1 kb (27,975 peaks), 2kb (48,108 peaks) or up to 5 kb (127,602 peaks) from
129 the genes TSS. We found similar weak positive correlation (R = 0.13 to 0.14,
130 Spearman $p < 2.2e-16$) for all four intervals (TSS peak up to 500 bp, 1 kb, 2 kb and 5
131 kb) (Fig. 2a) and thus, we focus on all (159,708) accessible gene promoter peaks
132 (up to 5 kb from the TSS). A higher correlation might not be observed in our datasets
133 as accessible regions are not always associated with gene activity and can instead,
134 be associated with repressed or poised for activation genes [27-30].
135
136 Based on an approach devised previously [30], we explore the relationship of
137 accessible peaks and gene expression by assessing genes with combinations of
138 high accessibility (HA) or medium–low accessibility (MA), and either highly
139 expressed (HE) or medium–low expression (ME) (see '*Materials and Methods*').

140 After categorising gene peaks and expression, we were left with a total of 27,536
141 (22% of 127,602) accessible peaks associated to the expression of 4,509 unique
142 genes. These could be placed into the four categories, with most genes (1,650) and
143 peaks (10,488) in the HA-HE category, and the remaining being either HA-ME (6,961
144 peaks and 1,081 unique genes), MA-HE (5,990 peaks and 1,770 unique genes), or
145 MA-ME (4,097 peaks and 1,164 unique genes). Whilst the HA-ME genes exhibit no
146 correlation (R = -0.021, Spearman p = 0.076), and the MA-HE genes exhibit
147 significant weak negative correlation (R = -0.034, Spearman p = 0.008), both the
148 MA-ME (R = 0.037, Spearman p = 0.019) and HA-HE genes (R = 0.031, Spearman
149 p = 0.0013) exhibit significant weak positive correlation (Fig. 2b). Among the 1,650
150 HA-HE genes, we identify genes associated with aquaculture-relevant traits,
151 including those enriched (FDR < 0.05) for the GO terms 'cellular process' e.g., *prlr1*
152 involved in tilapia salinity tolerance [31] and *clcn2* an osmoregulatory gene involved in
153 adaptation to saline-alkali challenge in tilapia [32], and 'cellular respiration' e.g., *cox4*
154 involved in aerobic metabolism changes under salinity challenge in tilapia gills [33]
155 (Supplementary Fig. S3).
156

157 **Nile tilapia SNPs are enriched in noncoding regions**

158 Since we previously found that discrete variation in regulatory regions disrupt
159 regulatory network interactions for adaptive trait genes in Nile tilapia [6,7], we tested
160 whether functional noncoding SNPs could be associated with the regulatory activity
161 of genes involved with aquaculture-relevant traits, like environmental tolerance.
162 Building upon our previous work of identifying 69 million (69,064,774) SNPs across
163 27 tilapia species (Fig. 3a, Supplementary Table S2) mapped to the Nile tilapia
164 genome (Ciezarek, A. *et al.*, bioRxiv TBC, 2023), we identified that these SNPs are

165 significantly enriched in noncoding than coding (exonic) regions (Fig. 3b). Compared
166 to exonic regions (fold enrichment of 0.9, adjusted *P*-value <0.05), the SNPs are
167 most enriched in either intronic or up to 5 kb gene promoter (fold enrichment of 1.1,
168 adjusted *P*-value <0.05) regions (Fig. 3b).
169
170 To characterise functional noncoding variation, we find that of the 5,470,066 SNPs in
171 Nile tilapia gene promoter regions, 1,083,195 (20%) SNPs overlapping 121,404 of
172 the 127,602 (95%) accessible peaks are significantly enriched (fold enrichment of
173 1.1, adjusted *P*-value <0.05) (Fig. 3c). Of the 20,248 accessible gene promoters,
174 20,101 (99%) have SNPs in accessible regions, representing 68% of all annotated
175 Nile tilapia genes [26]. Most SNPs are rare variants (Minor Allele Frequency, MAF <=
176 0.1) with 1% more rare variants found in noncoding regions that are accessible than
177 not (Supplementary Fig. S4, see '*Supplementary Information'*). Using the 127,602
178 accessible peaks in gene promoter regions, we find that the topmost (-log10 FDR <
179 0.05 of >=10) enriched GO biological processes of 20,248 accessible gene promoter
180 regions harbouring SNPs could be functionally relevant to aquaculture traits like, for
181 example, signalling, cell communication, and cellular response to stimulus (Fig. 3d,
182 Supplementary Fig. S5). In summary, nearly three-quarters of all Nile tilapia genes
183 have putative functional noncoding variation at gene promoter regions that could be
184 associated with the regulation of aquaculture traits relevant to gill function like, for
185 example, a response to stimulus.
186
187 **Discrete variation in accessible TFBSs is likely driving gene expression**
188 **associated with tilapia gill adaptations**

189 We hypothesised that genetic variation between Nile tilapia and the 26 other tilapia
190 species (Fig. 3a) could be used to identify functional noncoding variation based
191 associated with aquaculture traits that differ between the species. For example,
192 genetic variation in gene promoter TFBSs between Nile tilapia (a freshwater species)
193 and any euryhaline species e.g., *A. grahami, O. mossambicus* or *O. urolepis* with
194 exclusive homozygous or heterozygous alternate sites could indicate positive
195 selection acting on discrete mutations, that can be functional based on chromatin
196 accessibility in Nile tilapia gill tissue, and subsequently associated to target gene
197 expression of gill-specific traits relevant to aquaculture e.g., salinity tolerance. To
198 prioritise predicted TFBSs (see '*Materials and Methods*' and '*Supplementary*
199 *Information'*), we use the best described method for ranking predictions [34] by
200 overlapping predicted TFBSs with the number of reads (tag count, TC) in the
201 footprint, and then rank TFBSs using the bit-score of the motif match within ranked
202 tag counts. Based on a mean bit-score of 11.5 and mean tag count of 111.9, we
203 identify 6,195,938 TFBSs above the means (Fig. 4a). In gene promoter regions, we
204 identify 246,296 TF footprints with 2,491,107 non-redundant TFBSs (see '*Materials*
205 *and Methods*'). Based on a mean bit-score of 11 and mean tag count of 152.6, we
206 identify 398,995 (16%) TFBSs above the means in gene promoter regions.
207
208 Using TFBSs in accessible gene promoter regions, we identified a total of 3,398,768
209 transcription factor (TF) – target gene (TG) relationships in Nile tilapia. We then
210 identify discrete mutations in TFBSs to characterise conservation or divergence and
211 therefore, whether the 3,398,768 Nile tilapia TF-TG relationships could be present or
212 absent in the other 26 species. First, we map open-chromatin peaks, footprints,
213 TFBSs and SNPs to gene promoter regions on each Nile tilapia chromosome (Fig.

214 4b). In total, we identify 7,029,307 (11% of total 69,064,774) SNPs overlapping
215 accessible gene promoter regions, TF footprints, and TFBSs with all (2042) TF
216 motifs having at least one discrete mutation in 18,899 genes (64% of the 29,552 Nile
217 tilapia genes in the genome [26]). More than half (62% - 11,790) of the 18,899 genes
218 have SNPs in 'prioritised' TFBSs where the bit-score of motif match and tag count
219 are above the means of 11 and 152.6 respectively. Of the 18,899 genes, 1,562 (8%)
220 have high accessibility and high expression (HA-HE), which makes up the majority
221 (95%) of all 1,650 HA-HE genes. This indicates that discrete variation in accessible
222 gene promoter TFBSs could be a core factor driving high levels of gene expression
223 associated with environmental tolerances of the gill. Using both measures, we
224 narrowed this down to 1,168 genes (71% of all HE-HE genes or 6% of all genes with
225 a SNP in an accessible gene promoter TFBS) that are categorised as 1) having a
226 SNP in a 'prioritised TFBSs' where the bit-score of motif match and tag count are
227 above the mean of all gene promoter TFBSs; and 2) being a 'HA-HE gene' based on
228 peak-expression correlations. To further narrow down the 1,168 genes, we focus on
229 genes 3) with 'genetic variation' that exists between species exhibiting a gradient of
230 environmental tolerance e.g., freshwater versus saline water; and 4) that are
231 enriched for biological process GO terms that are associated with gill adaptations of
232 traits with aquaculture relevance e.g., response to stimulus (Fig. 3d, Supplementary
233 Fig. S5) or cellular process (Supplementary Fig. S3). We use these four described
234 criteria for identifying examples of candidate genes that are likely involved in gill
235 function associated with environmental tolerance in tilapia.
236
237 **Discrete mutations at accessible regulatory sites of adaptive trait genes are**
238 **associated with environmental tolerance networks in tilapia species**

239 Using the four described criteria (prioritised TFBSs, HA-HE gene, genetic variation
240 between diverse species, and aquaculture relevant GO terms) on the set of 1,168
241 genes, we first identified STAT1-*prlr1* (Fig. 5) as a candidate TF-TG relationship that
242 could be involved in gill adaptations associated with tilapia environmental tolerance.
243 The STAT1 TFBS is proximal (<400 bp) to the *prlr1* TSS (Fig. 5a-b) and with a
244 significant signal of activity (Fig. 5c), has a few standard deviations higher bit-score
245 of motif match (11.04 ± 2.61 SD) and tag count (152.6 ± 0.12 SD) than the means of
246 all predicted gene promoter TFBSs. Differences in the regulation of prolactin (PRL)
247 and its receptors (PRLRs) have been shown to underly variation in salinity tolerance
248 of tilapias [31], and are differentially expressed between *A. grahami* and the freshwater
249 species, *O. leucostictus* gill transcriptomes [35]. For this reason, genetic variation at
250 the TFBS for STAT1 could account for *prlr1* gene expression changes associated
251 with salinity response in different tilapia species. We identified a rare variant (LG7
252 position 18107947; MAF=0.03) at the third position in the predicted STAT1 TFBS
253 (Fig. 5b and Fig. 5d). The rare variant is homozygous alternate G/G (n=14) and
254 heterozygous alternate T/G (n=9) in the highly adapted extremophile species, *A.*
255 *grahami,* but homozygous reference T/T in Nile tilapia and the other 25 tilapia
256 species (Fig. 5d). STAT1 is a transcriptional activator that can be phosphorylated via
257 osmotic stress [36]; this suggests that STAT1 could be a key regulator of *prlr1* gene
258 expression in most tilapia species gills in response to osmotic stress however, the
259 loss of STAT1 regulation of *prlr1* could necessitate gill adaptations to extreme
260 alkaline and saline waters that are natural to *A. grahami* [37,38].
261
262 In another example, we identify variation in the gene promoter region of *cldnh* (also
263 referred to as *cldn3-like* and *cldn28a*), associated with expression differences in

264 stickleback ecotypes [39] and euryhaline tilapia [40] under salinity challenge, as well as
265 involved in permeability changes associated with salinity acclimation in freshwater
266 tilapia species too [41]. Specifically, we identify SREBF2-*cldnh* (Fig. 6) as another
267 candidate TF-TG relationship that could be involved in gill adaptations associated
268 with salinity acclimation, especially of freshwater species. The SREBF2 TFBS is
269 proximal (<330 bp) to the *cldnh* TSS (Fig. 6a-b) with a significant signal of activity
270 (Fig. 6c), and a few standard deviations higher bit-score of motif match (11.04 ± 0.4
271 SD) and tag count (152.6 ± 0.3 SD) than the means of all predicted gene promoter
272 TFBSs. We identified a rare variant (LG10 position 17991869; MAF=0.07) at the 4th
273 position in the predicted SREBF2 TFBS (Fig. 6b and Fig. 6d). We identify the rare
274 variant is mostly homozygous alternate G/G (n=19) and heterozygous alternate A/G
275 (n=14) in the freshwater species, *O. esculentus,* but homozygous reference A/A in
276 Nile tilapia, *O. esculentus* (n=4) and most individuals, except *O. "mtera"* , of the other
277 25 tilapia species (Fig. 6d). Despite being a freshwater species, *O. esculentus*
278 individuals with heterozygous and homozygous alternate genotypes were collected
279 from saline/alkaline water bodies, namely Hombolo Dam [42], Lake Malimbe [43], and
280 Lake Rukwa [44,45]. Such genetic variation at the SREBF2 TFBS could therefore
281 account for differential expression of the *cldnh* gene, which has been previously
282 shown to reorganise gill tight junctions in response to salinity acclimation of
283 euryhaline [40] and freshwater [41] tilapia. Discrete mutations in regulatory sites of
284 aquaculture relevant genes could therefore be driving gene regulatory evolution
285 associated with salinity acclimation of the gills.

286 ## Discussion

287 Most (~90%) tilapia aquaculture production is based on Nile tilapia [46] owing to its
288 high growth and reproduction rate in several culture systems. However, with climate
289 change leading to extreme weather and human competition decreasing freshwater
290 resources, there is a crucial need to breed tilapia strains that are resilient to broad
291 environmental conditions. One method to determine the genetic bases responsible
292 for environmental traits e.g., salinity tolerance, involves firstly characterising genetic
293 variation and signatures of selection associated with adaptive traits in resequenced
294 populations [2] and then secondly, using epigenetic sequencing e.g., ATAC-Seq to
295 annotate functional genomic regions and RNA sequencing to characterise whether
296 these loci can drive gene expression change. Despite the impact of noncoding
297 regulatory evolution towards the diversification of cichlids, including Nile tilapia [5-7]
298 and breeding wild tilapia populations [3,4], no previous work has applied ATAC-Seq to
299 accurately identify transcriptional regulators in tilapia tissues and/or cells, let alone
300 classify functional noncoding variation associated with adaptive traits.
301
302 By performing high-depth ATAC-seq of Nile tilapia gill tissues, we identify 301,293
303 accessible (open chromatin) regions in three replicates that can be associated to
304 85% of all annotated Nile tilapia genes [26]. No previous work has carried out ATAC-
305 seq in gill tissue however, we identify more peaks in our study compared to 11 other
306 zebrafish tissues (66k-180k peaks) but fewer than the 436k merged non-redundant
307 peaks across all tissues [47]. The identification of more single tissue peaks than
308 zebrafish is possibly due to tissue-specific differences and sampling of more
309 replicates. More than 40% (127,602) of the accessible regions can be found within

310 gene promoter regions of 20,248 genes, with genes without a reproducible peak
311 being better associated to functions of lymphoid tissues not sampled here.
312
313 We were able to associate 96,784 (76%) accessible gene promoter regions to the
314 expression of 14,842 genes. After determining no difference in weak positive
315 correlation of accessible peaks to gene expression based on proximity to the TSS,
316 we found that most genes (1,650 out of 4,509) could be categorised as highly
317 accessible and highly expressed (HA-HE) based on signals and counts above the
318 70th percentiles. The HA-HE genes are enriched for genes involved in salinity
319 acclimation however, the overall set exhibits weak positive correlation of accessibility
320 and gene expression. A higher correlation might not be observed in our datasets as
321 accessible regions are not always associated with gene activity and can instead, be
322 associated with repressed or poised for activation genes [27-30]. Like previous work in
323 *C. elegans* [27], this would need to confirmed using relevant (H3K27me3 and H3K4
324 methylation) ChiP-Seq marks.
325
326 Whilst our previous work focused on regulatory variation in a phylogeny of five cichlid
327 species, including Nile tilapia [6,7], no previous work has examined functional
328 noncoding variation in the tilapia phylogeny. Building upon our recent work of
329 identifying 69 million SNPs in the Nile tilapia genome based on resequencing
330 populations of 27 tilapia species (Ciezarek, A. *et al.*, bioRxiv TBC, 2023), we found
331 SNPs are enriched in noncoding regions and more specifically, gene promoter
332 regions. We identified that gene promoter regions with accessible SNPs are enriched
333 for processes associated with gill function e.g., response to stimulus, and therefore
334 conclude that much like the adaptive radiations of East African cichlids [6,7], discrete

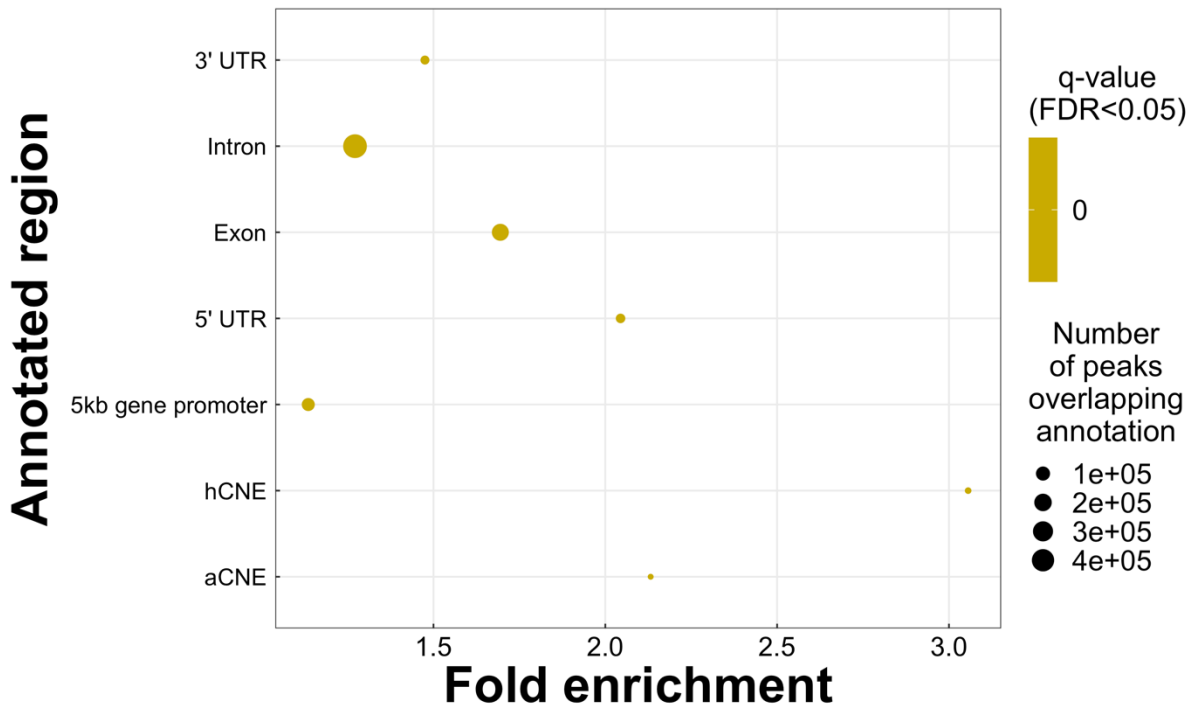
335 variation in regulatory regions can also drive gene regulatory rewiring of tilapia
336 adaptive trait genes. This is supported by the fact that nearly all (99%) accessible
337 gene promoter regions have a SNP with most (90%) being a rare variant. Gene
338 promoter regions are enriched for TFBSs [7], as are accessible regions [47,48], and thus,
339 variation, including rare variation [49], at TFBSs is a major target of genetic selection [50]
340 and a key contributor to adaptive phenotypes [6,7,51]. Using TFBSs in accessible gene
341 promoter regions, we identified a total of 3,398,768 transcription factor (TF) – target
342 gene (TG) relationships in Nile tilapia. As we use experimental data, this is less than
343 the 5,900,174 TF-TG edges we *in silico* predicted in Nile tilapia in our previous work
344 [7], but similar to the 3,505,491 TF-TG relationships found in human based on *in silico*
345 predictions overlapping ENCODE footprints [52].
346
347 Using a stringent criterion to identify candidate TF-TG relationships, involving
348 prioritising statistically significant TFBSs with genetic variation between ecologically
349 divergent species and genes that exhibit high accessibility and high expression
350 compared to all genes, we identified 1,168 genes that could be associated to
351 aquaculture relevant traits, like salinity acclimation. Some tilapia species can thrive in
352 adverse environmental conditions, including varying levels of salinity and alkaline
353 water. Nile tilapia is intolerant to high salinity (optimal is up to 16 parts per thousand,
354 ppt) [53] and requires freshwater (pH 6-9) [54,55] whereas euryhaline species like
355 *Alcolapia grahami*, nested within the *Oreochromis* genus [56], have adapted to
356 extreme environments of pH 9-11.5 and high salt concentrations of >20 ppt [37,38]. In
357 *prlr1*, a gene associated with tilapia salinity response [31], a rare genetic variant
358 between Nile tilapia and *A. grahami* in the STAT1 TFBSs could account for osmotic
359 stress response in most tilapia species, but the loss of this site could enable gill

360 adaptation in extremophiles like *A. grahami.* This is further supported by the role of
361 STAT1 as an osmotic stress induced transcriptional activator [36] and differential
362 expression of prolactin receptors in gills of *A. grahami* and freshwater tilapia species
363 [35]. On the other hand, the introduction and acclimation of freshwater species, like *O.*
364 *esculentus*, to saline/alkaline waters e.g., Lake Rukwa in Tanzania [44,45], could be
365 driven by genetic variation at the SREBF2 TFBS, and account for differential
366 expression of *cldnh,* a gene responsible for reorganising gill tight junctions in
367 response to salinity exposure of freshwater tilapia [41]. Salinity acclimation first
368 requires the detection of osmotic changes and then secondly, a physiological
369 response to restore osmotic homeostasis. Accordingly, excessive ions are secreted
370 by specialised cells in the gill, ionocytes, containing ion pumps and transporters [57,58].
371 As a result, there are differential gill-specific transcriptional responses to salinity
372 exposure in different tilapia species, triggering turnover of ionocytes as well as
373 immune and cell stress response [59]. We therefore suggest that the interactions of
374 both STAT1-*prlr1* and SREBF2-*cldnh* are likely involved in salinity acclimation along
375 the tilapia phylogeny. Discrete rewiring of existing ancestral variation at these
376 important functional loci could be a mechanism to allow populations to rapidly adapt
377 to new ecological niches e.g., *O. esculentus* introductions to Lake Rukwa [60], and
378 also form the basis for breeding genotypes of desirable traits into farmed strains.
379
380 This novel epigenome of not only Nile tilapia gill, but the first recognised assessment
381 of open chromatin in any fish gill tissue, will enable a better functional annotation of
382 noncoding regions to identify causative variants associated with aquaculture relevant
383 traits. Now that the ATAC-seq protocol has been optimised for tilapia fish, we can
384 characterise the open chromatin landscape in the other 26 tilapia species to confirm

and prioritise variants identified in this study, and define trait loci for other environmental traits e.g., temperature tolerance. 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.

390 **Figures**

391



392

393 **Fig. 1 – Annotation of open chromatin peaks in Nile tilapia gill.** Fold enrichment

394 of open chromatin (accessible) peaks overlapping coding and noncoding regions in

395 the Nile tilapia genome. Circles show enriched annotated region (y-axis) of

396 significance (FDR <0.05, to right) and fold enrichment (x-axis) values of all annotated

397 peaks. Number of peaks overlapping each annotation shown by size of each circle.

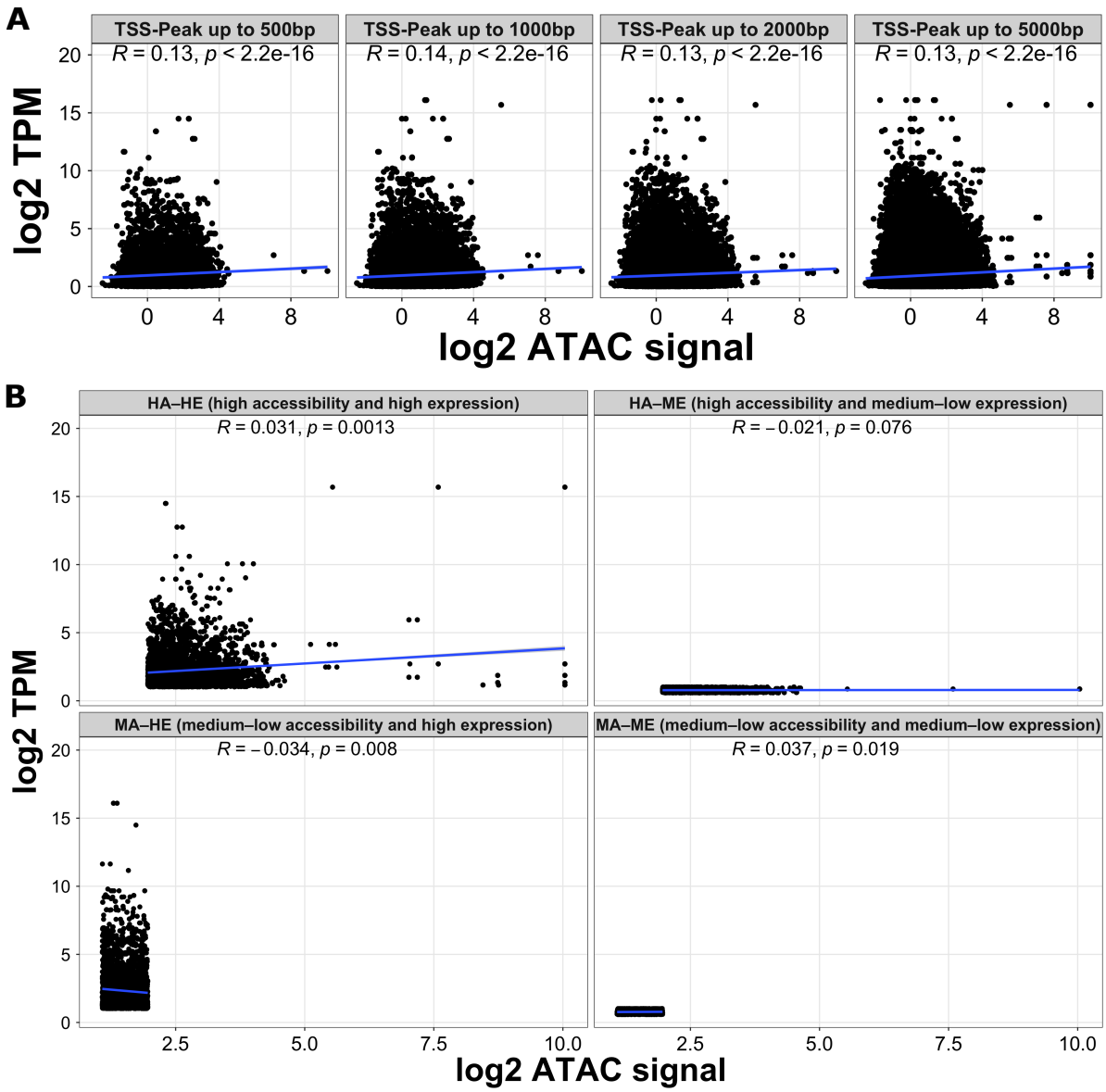
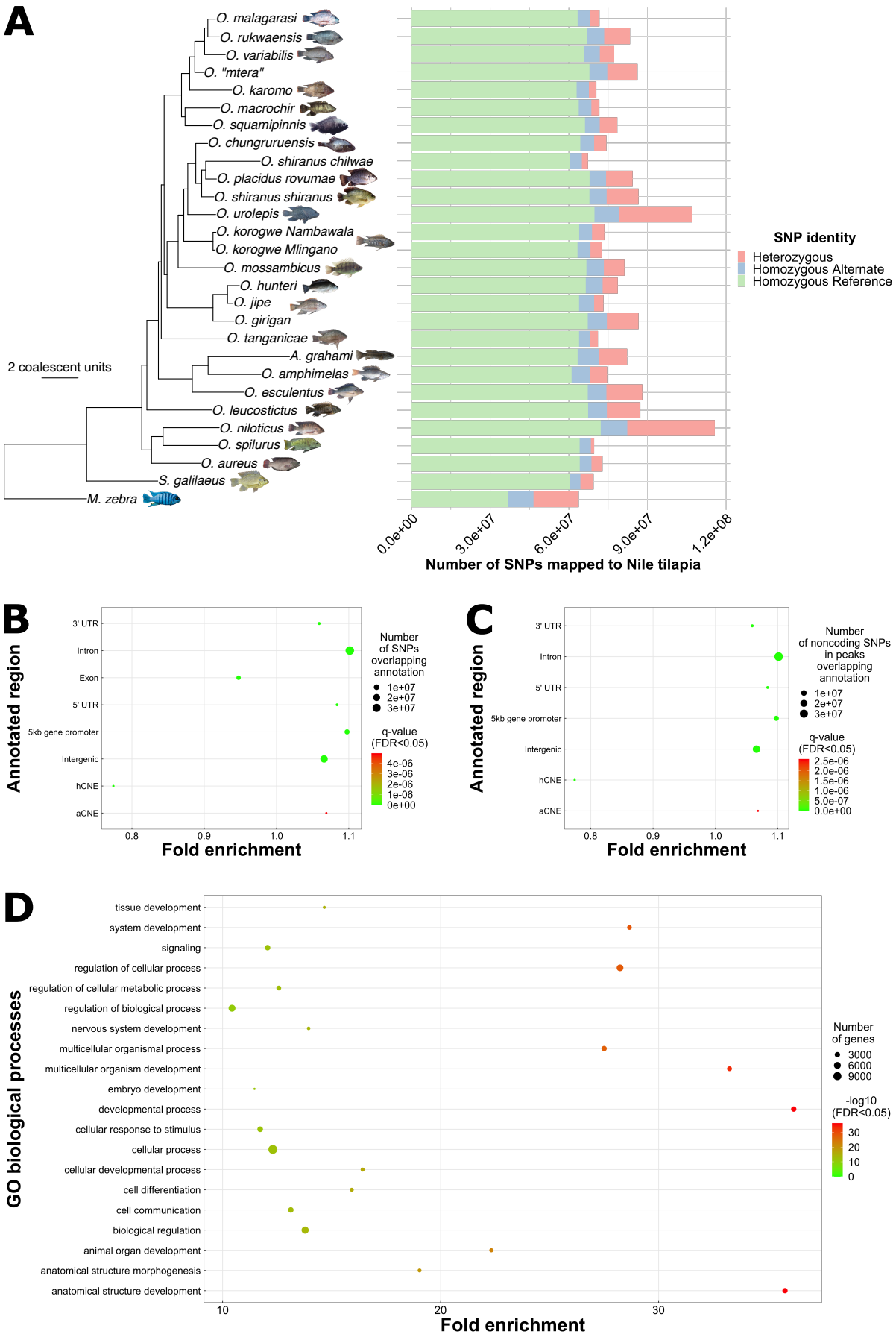


Fig. 2 – Correlation of accessible peaks accounting for gene expression change in Nile tilapia gill. (A) Average *log2* gene expression (as transcripts per million - TPM, y-axis) and average *log2* ATAC signal (as gene promoter peak counts, x-axis) for each gene across the three replicates, split into 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

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


A
O. malagarasi
O. rukwaensis
O. variabilis
O. "mtera"
O. karomo
O. macrochir
O. squamipinnis
O. chungruruensis
O. shiranus chilwae
O. placidus rovumae
O. shiranus shiranus
O. urolepis
O. korogwe Nambawala
O. korogwe Mlingano
O. mossambicus
O. hunteri
O. jipe
O. girigan
O. tanganicae
A. grahami
O. amphimelas
O. esculentus
O. leucostictus
O. niloticus
O. spilurus
O. aureus
S. galilaeus
M. zebra
2 coalescent units
SNP identity
Heterozygous
Homozygous Alternate
Homozygous Reference
0.0e+00
3.0e+07
6.0e+07
9.0e+07
1.2e+08
Number of SNPs mapped to Nile tilapia
B
Annotated region
3' UTR
Intron
Exon
5' UTR
5kb gene promoter
Intergenic
hCNE
aCNE
0.8
0.9
1.0
1.1
Fold enrichment
Number of SNPs overlapping annotation
1e+07
2e+07
3e+07
q-value (FDR<0.05)
4e-06
3e-06
2e-06
1e-06
0e+00
C
Annotated region
3' UTR
Intron
5' UTR
5kb gene promoter
Intergenic
hCNE
aCNE
0.8
0.9
1.0
1.1
Fold enrichment
Number of noncoding SNPs in peaks overlapping annotation
1e+07
2e+07
3e+07
q-value (FDR<0.05)
2.5e-06
2.0e-06
1.5e-06
1.0e-06
5.0e-07
0.0e+00
D
GO biological processes
tissue development
system development
signaling
regulation of cellular process
regulation of cellular metabolic process
regulation of biological process
nervous system development
multicellular organismal process
multicellular organism development
embryo development
developmental process
cellular response to stimulus
cellular process
cellular developmental process
cell differentiation
cell communication
biological regulation
animal organ development
anatomical structure morphogenesis
anatomical structure development
10
20
30
Fold enrichment
Number of genes
3000
6000
9000
-log10 (FDR<0.05)
30
20
10
0

414 **Fig. 3 – Single nucleotide polymorphisms (SNPs) in the Nile tilapia genome. (A)**
415 Coalescent phylogeny of 27 tilapia species and 2 outgroups inferred using ASTRAL
416 (Ciezarek, A. *et al.*, bioRxiv TBC, 2023) with number and identity of SNPs in each
417 species based on mapping to the Nile tilapia genome. Phylogenies inferred based on
418 14,535 10 kb windows across the genome. Units = coalescent units; All posterior
419 probabilities = 1.0. Species photo credits: George Turner, Martin Genner, Antonia
420 Ford, Benjamin Ngatunga, Geraldine Kavembe, and Fishbase (Luc de Vos, Brain
421 Gatwicke, Andrew Nightingale, and Royal Museum for Central Africa). **(B)** Fold
422 enrichment of SNPs overlapping coding and noncoding regions in the Nile tilapia
423 genome. Circles show enriched annotated region (y-axis) of significance (FDR
424 <0.05, heatmap to right) and fold enrichment (x-axis) values of all SNPs mapped to
425 the Nile tilapia genome. Number of SNPs overlapping each annotation shown by
426 size of each circle. **(C)** Fold enrichment of SNPs overlapping open chromatin
427 (accessible) peaks in noncoding regions of the Nile tilapia genome. Circles show
428 enriched annotated region (y-axis) of significance (FDR <0.05, heatmap to right) and
429 fold enrichment (x-axis) values of all SNPs in annotated noncoding peaks. Number
430 of noncoding SNPs in annotated noncoding peaks shown by size of each circle. **(D)**
431 Gene ontology (GO) biological process enrichment of accessible gene promoter
432 regions containing SNPs in the Nile tilapia genome. GO terms are subset for
433 enrichment of -log10 (FDR < 0.05) of more than 10, with the full figure provided as
434 Supplementary Fig. S5. Circles show enriched term (y-axis) of significance (FDR
435 <0.05, heatmap to right) and fold enrichment (x-axis) values of GO terms. Number of
436 genes with promoter regions containing SNPs shown by size of each circle.

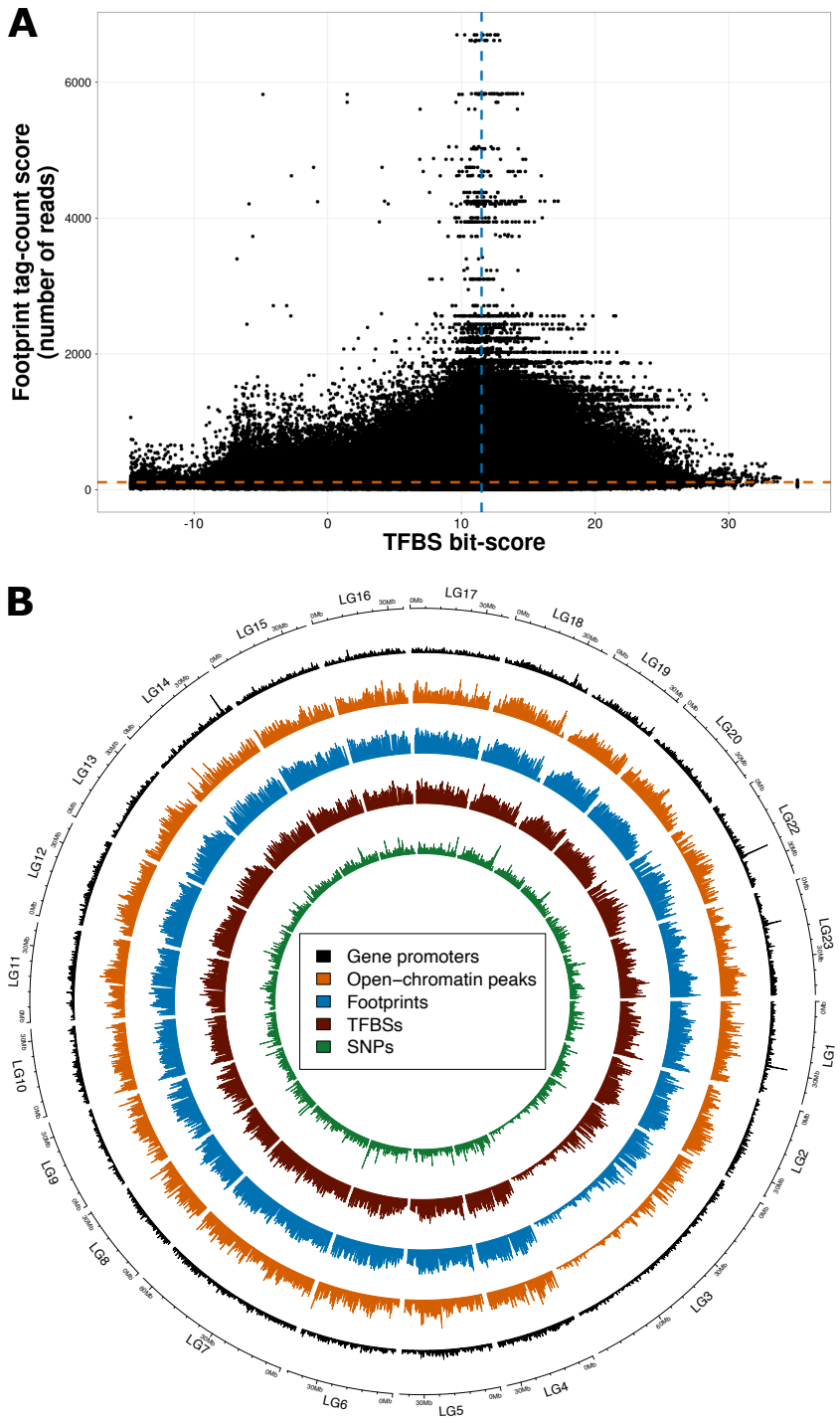
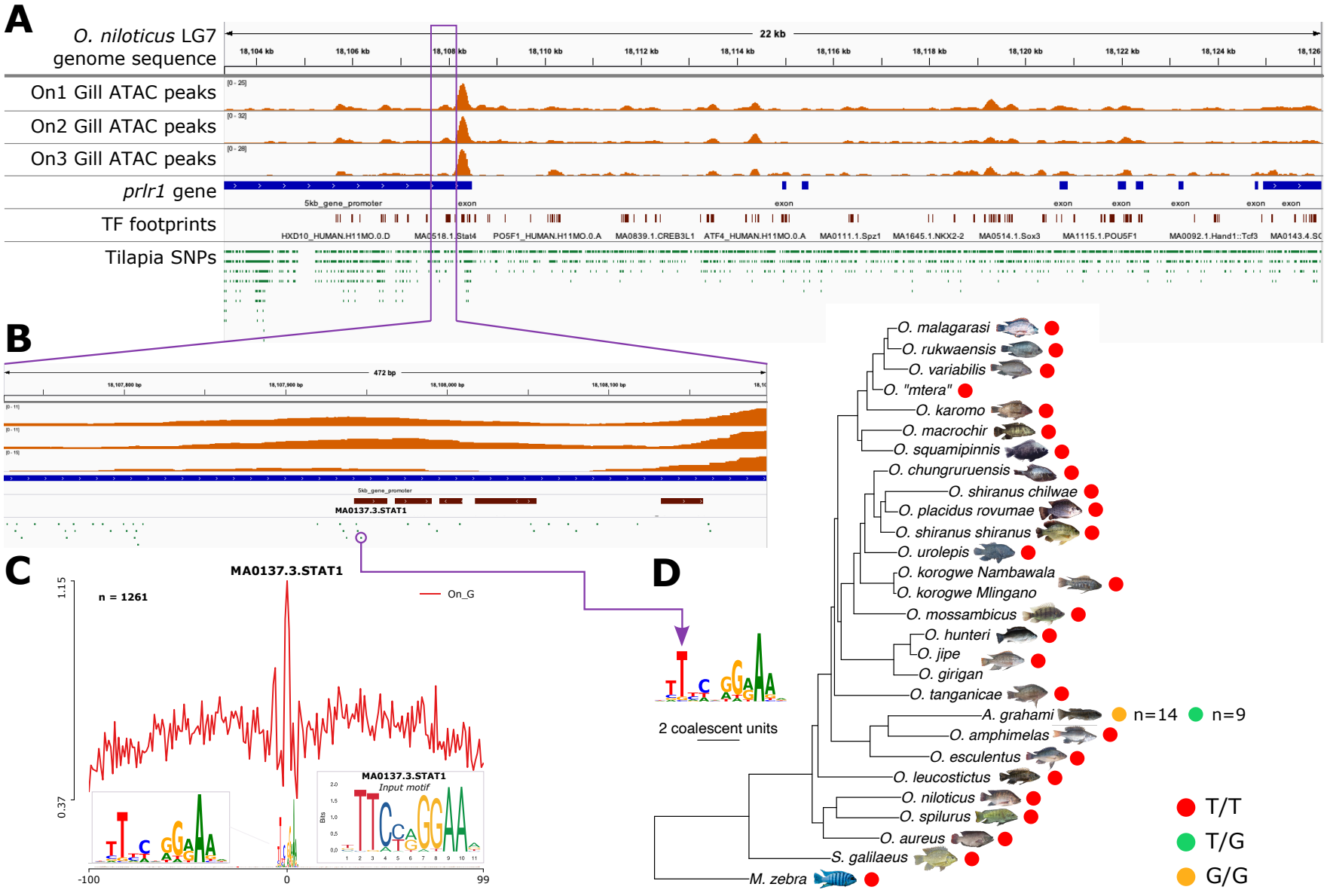
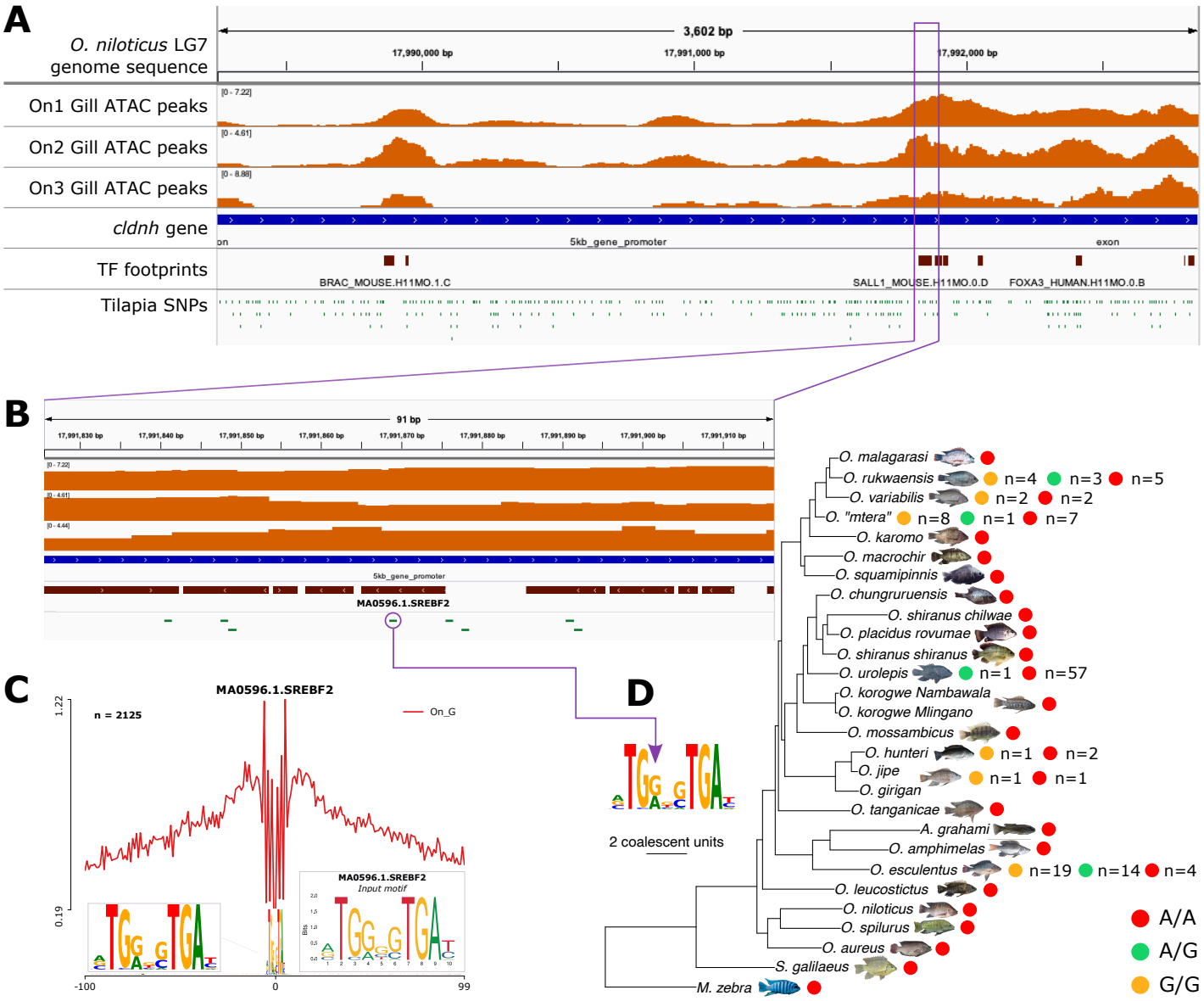


Fig. 4 – Prioritising genetic variation of TFBSs in accessible peaks accounting for gene expression change in Nile tilapia gill. (A) Tag-count (TC) score of TF footprints (y-axis) and TFBS bit-score of motif match (x-axis) of genome-wide predicted TFBSs based on footprints in accessible gill peaks. **(B)** 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).


A
O. niloticus LG7 genome sequence
22 kb
On1 Gill ATAC peaks
On2 Gill ATAC peaks
On3 Gill ATAC peaks
prlr1 gene
5kb_gene_promoter
exon
TF footprints
HXD10_HUMAN.H11MO.0.D
MA0518.1.Stat4
PO5F1_HUMAN.H11MO.0.A
MA0839.1.CREB3L1
ATF4_HUMAN.H11MO.0.A
MA0111.1.Spz1
MA1645.1.NKX2-2
MA0514.1.Sox3
MA1115.1.POU5F1
MA0092.1.Hand1::Tcf3
Tilapia SNPs
B
472 bp
MA0137.3.STAT1
C
MA0137.3.STAT1
n = 1261
On_G
1.15
0.37
-100
0
99
MA0137.3.STAT1
Input motif
Bits
D
2 coalescent units
O. malagarasi
O. rukwaensis
O. variabilis
O. "mtera"
O. karomo
O. macrochir
O. squamipinnis
O. chungruruensis
O. shiranus chilwae
O. placidus rovumae
O. shiranus shiranus
O. urolepis
O. korogwe Nambawala
O. korogwe Mlingano
O. mossambicus
O. hunteri
O. jipe
O. girigan
O. tanganicae
A. grahami
O. amphimelas
O. esculentus
O. leucostictus
O. niloticus
O. spilurus
O. aureus
S. galilaeus
M. zebra
n=14
n=9
T/T
T/G
G/G

445

446 **Fig. 5 – Accessible peaks in the Nile tilapia *prlr1* gene promoter with variation at the STAT1 TF footprint. (A)** Nile tilapia
447 track of the *prlr1* gene sequence (blue annotations) in LG7 with accessible peaks across three biological gill replicates (orange
448 peaks), TF footprints (maroon marks) and tilapia SNPs (green marks). **(B)** Expanded *prlr1* gene promoter region (blue annotation)
449 of accessible peaks (in orange) across three gill replicates above the STAT1 TF footprint (maroon mark) and containing a SNP
450 (green mark with purple circle). **(C)** Line plot profile of the STAT1 TF footprint (n=1261) across Nile tilapia gene promoter regions
451 showing positional (x-axis) signal of activity (red line, y-axis) centred on the footprint (at position 0) with flanking sequence.
452 Expanded STAT1 TF footprint shown to left and original input motif used for predicting the site shown to right. **(D)** Variation from the
453 SNP in Fig. 5b demarcated in the STAT1 TF footprint with homozygous reference (red dots), homozygous alternate (green dots)
454 and heterozygous (yellow dots) alleles for the SNP shown for each species in the tilapia phylogeny (based on Fig. 3a).


A
O. niloticus LG7 genome sequence
3,602 bp
17,990,000 bp
17,991,000 bp
17,992,000 bp
On1 Gill ATAC peaks
On2 Gill ATAC peaks
On3 Gill ATAC peaks
cldnh gene
5kb_gene_promoter
exon
TF footprints
BRAC_MOUSE.H11MO.1.C
SALL1_MOUSE.H11MO.0.D
FOXA3_HUMAN.H11MO.0.B
Tilapia SNPs
B
91 bp
MA0596.1.SREBF2
C
MA0596.1.SREBF2
n = 2125
On_G
1.22
0.19
-100
0
99
Input motif
D
2 coalescent units
O. malagarasi
O. rukwaensis
n=4
n=3
n=5
O. variabilis
n=2
n=2
O. "mtera"
n=8
n=1
n=7
O. karomo
O. macrochir
O. squamipinnis
O. chungruruensis
O. shiranus chilwae
O. placidus rovumae
O. shiranus shiranus
O. urolepis
n=1
n=57
O. korogwe Nambawala
O. korogwe Mlingano
O. mossambicus
O. hunteri
n=1
n=2
O. jipe
n=1
n=1
O. girigan
O. tanganicae
A. grahami
O. amphimelas
O. esculentus
n=19
n=14
n=4
O. leucostictus
O. niloticus
O. spilurus
O. aureus
S. galilaeus
M. zebra
A/A
A/G
G/G

456

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

466 ## Materials and Methods

467

468 **Tissue dissection**

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

476

477 **Cell preparation**

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

484

485 **Transposition reaction and purification**

486 Pelleted nuclei were gently resuspended in a 50 µL transposition reaction mix
487 composed of 25 µL 2× TD Buffer (Illumina), 2.5 µL Tn5 Transposes (Illumina) and
488 22.5 µL Nuclease-Free H_2O. Transposition reaction was incubated at 37°C for 30
489 min and purified using the PCR purification MinElute Kit (QIAGEN), according to

490 manufacturer's protocol. Transposed DNA was eluted in 10 µL Elution Buffer (10mM
491 Tris buffer, pH 8) and stored at -20°C.
492
493 **PCR amplification**
494 Transposed DNA was amplified in a final volume of 50 µL composed of 10 µL
495 Transposed DNA, 10 µL Nuclease Free H_2O, 2.5 µL 25µM P7 adapter, 2.5 µL 25µM
496 P5 adapter and 25 µL NEBNext High-Fidelity 2x PCR Master Mix (NEB).
497 Transposed DNA was amplified for 5 min at 72°C, 30 secs at 98°C, and 11 cycles of
498 10 secs at 98°C, 30 secs at 63°C and 1 min at 72°C in a PCR thermocycler.
499 Amplified transposed DNA was purified using the PCR purification MinElute Kit
500 (QIAGEN) and eluted in 20 µL Elution Buffer (10mM Tris buffer, pH 8), according to
501 manufacturer's protocol. To remove excess primers for final ATAC libaries, an
502 additional 1× Agencourt AMPure XP bead (Beckman Coulter*)* clean-up was
503 performed and eluted in 0.1× filtered TE. ATAC libraries were quantified on the Qubit
504 4 fluorometer (Invitrogen) and size distribution assessed on Agilent Tapestation
505 and/or Bioanalyser.
506
507 **DNA extraction**
508 DNA was purified from gill filaments using the DNeasy Blood and Tissue kit
509 (QIAGEN) according to manufacturer's protocol. DNA was quantified on the
510 Nanodrop 2000 (Thermo Scientific) and Qubit 4 fluorometer (Invitrogen), and used 1
511 ng for DNA library preparation using the Nextera XT DNA Library Preparation kit
512 (Illumina), according to manufacturer's protocol. Library size distribution was
513 assessed on Agilent Tapestation and/or Bioanalyser. DNA libraries, obtained from

514 naked DNA, were used as internal controls to determine background levels of
515 genomic DNA accessibility and Tn5 transposase sequence cleavage bias.
516
517 **ATAC and control DNA sequencing**
518 Three ATAC and three corresponding naked DNA control libraries were equimolar
519 pooled and 50 bp paired-end sequenced at Earlham Institute on the Illumina
520 NovaSeq 6000 platform using an S2 flow cell, generating an average of 90 million
521 (ATAC-seq) and 16 million (naked DNA control) reads per library.
522
523 **ATAC-seq and control DNA processing**
524 Sequence adaptors were removed and trimmed for quality from raw paired-end
525 reads using Trim Galore! (v 0.6.5) (https://github.com/FelixKrueger/TrimGalore) and
526 FastQC (v 0.11.9) [61] using default parameters. Read alignment, post alignment
527 filtering and ATAC peak calling were performed according to the ENCODE projects
528 'ATAC-seq Data Standards and Processing Pipeline' for replicated data
529 (https://www.encodeproject.org/atac-seq/). Briefly, trimmed reads were mapped to
530 the Nile tilapia reference genome (*O. niloticus* UMD_NMBU) [26] using bowtie2
531 (v2.2.6) [62], with parameters '`–k 4 –X2000 –mm`' and outputted in BAM format using
532 SAMtools (v1.9) [63]. Since ATAC-seq generates high proportions (15-50% in a typical
533 experiment) of mitochondrial mapped reads, any reads mapping to the mitochondrial
534 genome were identified using BLAST (v2.3.0) [64] and removed from the BAM file
535 using SAMtools (v1.9) [63]. The resulting BAM files were sorted, and duplicated reads
536 were marked using Sambamba v0.6.5 [65]. Duplicated, unmapped, non-primary
537 alignment, and failing platform QC reads were filtered out using SAMtools (v1.9) [63],
538 retaining reads mapped as proper pairs, and fragment length distributions were

539 plotted using Picard (v1.140) (https://github.com/broadinstitute/picard/). At each step,
540 the recommended parameters from the ENCODE pipeline were applied. BAM files
541 were converted to *tagalign* files using Bedtools (v2.30.0) [66] and Tn5 shifting of ATAC
542 mappings carried out prior to peak calling. Peaks were identified using macs2
543 (v2.1.1) [67,68] with the shifted tag as test and corresponding control DNA as input with
544 parameters '`-f BED -p 0.05 --nomodel --shift -75 --extsize 150`'. Narrow
545 peaks were used to create coverage tracks using *bedClip* and *bedToBigBed* in the
546 UCSC-tools package (v333) (http://hgdownload.cse.ucsc.edu/admin/exe/). Following
547 the ENCODE pipeline, Irreproducible Discovery Rate (IDR) peaks of true replicates
548 were flagged as either true (<0.1) or false (≥0.1) using idr v2.0.4
549 (https://github.com/kundajelab/idr), taking reproducible (true) peaks between
550 replicates. The fraction of reads in peaks (FRiP) were calculated using Bedtools
551 (v2.30.0) [66] and demarcated as pass (>0.3) or acceptable (>0.2) according to
552 ENCODE guidelines. FRiP was not used as a QC measure and instead, transcription
553 start site (TSS) enrichment was calculated using ATACseqQC (v1.18.0) [69], with the
554 TSS enrichment QC requirement of a significant signal value at the centre of the
555 distribution being applied. All samples passed this criterion for further analysis. A
556 union of peaks from all three biological replicates was created and used for
557 subsequent analyses.
558

559 **Identifying conserved noncoding elements (CNEs)**

560 We use a similar approach as applied previously [5] to call CNEs in Nile tilapia based
561 on evolutionary constraint in a five cichlid species phylogeny. A multiple genome
562 alignment (MGA) of four cichlid genomes (*M. zebra* UMD2a [70]; *P. nyererei* v1 [5]; *A.*
563 *burtoni* v1 [5]; *N. brichardi* v1 [5]) and Nile tilapia (*O. niloticus* UMD_NMBU [26]) was

564 created using Cactus (v2.0.3) [71] and outputted in Multiple Alignment Format (MAF).
565 A neutral substitution model was created using the *phyloFit* function of PHAST (v1.5)
566 [72] by fitting a time reversible substitution 'REV' model and parameters '–tree
567 "((((Metriaclima_zebra,Pundamilia_nyererei),Astatotilapia_burtoni),N
568 eolamprologus_brichardi),Oreochromis_niloticus)" --subst-mod REV'.
569 The five cichlid MGA was split by *O. niloticus* chromosomes/scaffolds using the
570 *mafSplit* function in the UCSC-tools package (v333)
571 (http://hgdownload.cse.ucsc.edu/admin/exe/).
572 The neutral substitution model and MGAs of each *O. niloticus* reference
573 chromosome/scaffold were used as input to predict conserved noncoding elements
574 (CNEs) using the *phastCons* function of PHAST (v1.5) [72] with parameters '--
575 target-coverage 0.3 --expected-length 30 --most-conserved --
576 estimate-trees --msa-format MAF'. Conservation score outputs were used to
577 define highly conserved CNEs (hCNEs) with sequence identity ≥90% over ≥30bp
578 and pseudo accelerate/diverged CNEs (pseudo aCNEs) with sequence identity
579 <90% over ≥30bp. The evolutionary conservation-acceleration (CONACC) score of
580 each pseudo aCNE was predicted by running *phyloP* [73] of the PHAST (v1.5)
581 package [72] using the likelihood ratio test (LRT) '--method LRT' on the CNE --
582 features with their corresponding neutral substitution model in '--mode CONACC'.
583 Pseudo aCNEs with a negative CONACC score, likelihood ratio <0.05, and
584 significant divergence (altsubscale >1), were defined as significantly deviating from
585 the neutral model, and therefore as true aCNEs.
586
587 **Peak annotation and enrichment**

588 Up to 5 kb gene promoter regions were annotated in the *O. niloticus* UMD1 [26]
589 genome according to the method used in our previous study [7]. Narrow peaks either
590 overlapping or most proximal to all annotated features [26] in the genome were
591 mapped using the *intersect* function of Bedtools (v2.30.0) [66], and each peak
592 assigned to a feature and/or gene accordingly. The enrichment of accessible peaks
593 in coding and noncoding regions was tested using the Genome Association Tester
594 (GAT) tool [74]. The accessible peaks were provided as segments of interest to test, a
595 bed file of annotations to test against, and workspace as the length of each
596 scaffold/chromosome. GAT was ran using the following parameters: `--verbose=5,`
597 `--counter=segment-overlap, --ignore-segment-tracks, --qvalue-`
598 `method=BH --pvalue-method=norm`. We use the Benjamini-Hochberg [75] false
599 discovery rate (FDR) to assess enrichment of peaks in annotated regions, with a
600 statistical cut-off of FDR < 0.05.
601
602 **Gene Ontology (GO) enrichment**
603 GO enrichment analyses of genes was conducted using the 'g:GOst' module of
604 g:Profiler (https://biit.cs.ut.ee/gprofiler/gost) [76], version e105_eg52_p16_e84549f
605 (February 2022), using the *O. niloticus* database. We use the false discovery rate
606 (FDR) corrected hypergeometric *p*-value to assess enrichment of GO terms, with a
607 statistical cut-off of FDR < 0.05.
608
609 **Transcription factor (TF) footprinting**
610 TF footprints were characterised using HINT-ATAC in the Regulatory Genomic
611 Toolbox (v0.13.0) [77] using a stringent false positive rate (FPR) of 0.0001, with both
612 Nile tilapia specific and cichlid-wide position weight matrices (PWMs) as defined in

613 our previous study [7], as well as vertebrate PWMs from JASPAR (v9.0) [78],
614 HOCOMOCO [79], GTRD [80], and UniPROBE [81]. TF footprint line plots were generated
615 with the 'differential analysis' module of HINT-ATAC in the Regulatory Genomic
616 Toolbox (v0.13.0) package [77] using bias-corrected signals. Redundant TFBSs in the
617 same, or overlapping positions were filtered based on selecting the highest bit-score
618 of motif match for each overlapping TF.
619
620 **RNA extraction and sequencing**
621 RNA was purified from each tissue using the RNeasy Plus Mini kit (QIAGEN)
622 according to manufacturer's protocol. RNA and DNA content were quantified on the
623 Qubit 4 fluorometer (Invitrogen) and integrity assessed on Agilent Tapestation and/or
624 Bioanalyser, taking samples with RIN≥7 and <15% genomic DNA. A total of 45/55
625 (82%) samples passed these criteria for selection (Supplementary Table S2). A total
626 of 45 stranded RNA libraries were prepared using the NEBNext Ultra II Direction
627 RNA-seq kit according to manufacturer's protocol. All stranded RNA-seq libraries
628 were equimolar pooled and 150 bp paired-end sequenced at Earlham Institute on 1
629 lane of the Illumina NovaSeq 6000 platform using an S4 v1.5 flow cell, generating an
630 average of 70 million reads per library.
631
632 **RNA-seq processing**
633 Read quality was assessed using FastQC (v 0.11.9) [61] and Trim Galore! (v 0.6.5)
634 (https://github.com/FelixKrueger/TrimGalore) was used to remove adapters and trim
635 for low-quality from the raw paired-end reads using default settings. All reads were
636 then mapped to the *O. niloticus* UMD1 genome [26] using HISAT2 (v 2.2.1) [82] with
637 default parameters. Mapping QC was carried out using QualiMap (v 2.2.1) [83] with

638 default ‘rnaseq’ parameters. The final BAM file was sorted using samtools (v1.16.1)
639 [63] and transcript abundance was calculated using ‘htseq-count’ in the HTSeq (v
640 2.0.2) [84] package.
641
642 **Identifying ATAC peak and gene expression association**
643 Given the large number of accessible peaks and overlapping genes (127,602 peaks
644 and 20,248 genes), we devised an approach to reduce and identify the peaks that
645 could regulate their gene targets expression. For gene expression, we use an
646 approach applied previously [85] to calculate transcript per million (TPM) values by first
647 normalizing the transcript count by the gene length, as calculated using GTFtools [86],
648 followed by the library size, as calculated using QualiMap (v 2.2.1) [83], and then
649 carrying out a *log2*(x +1) transformation of 1) each genes TPM in each replicate; and
650 2) the mean TPM for each gene across biological replicates. Each genes mean TPM
651 across the three replicates is used to assess correlations with ATAC signal whereas
652 replicate specific TPM are used to study any differences between replicates. The
653 ATAC signal was processed using an approach applied previously [87] to obtain the
654 number of independent Tn5 insertions in collated gene promoter peaks; the number
655 of insertion sites (peak counts) was counted using collated narrow peaks against a
656 merged BAM file of all three replicates using Bedtools (v2.30.0) *coverage* [66]. After
657 creating a peak count matrix, the counts matrix was normalised using edgeR [88]
658 counts per million (CPM) ‘`log=TRUE, prior.count=5`’ followed by a quantile
659 normalization using the preprocessCore
660 (https://github.com/bmbolstad/preprocessCore) *normalize.quantiles* module in R (v
661 4.4.2). After merging the three replicates ATAC signal using the *log2* average from

662 the normalized counts matrices, the average *log2*-TPM and *log2*-ATACSignal for
663 each gene was plotted with the correlation co-efficient (r) calculated in R (v 4.4.2).
664
665 Based on an approach devised previously [30] to categorise putative activated,
666 repressed or poised genes, peak-gene relationships are assigned to one of four
667 groups according to lower than 50th (medium-low) and higher than 70th (high)
668 percentiles of gene promoter accessibility (*log2*-ATACSignal) and gene expression
669 (*log2*-TPM). The four groups are (1) MA–ME (medium–low accessibility and
670 medium–low expression); (2) HA–HE (high accessibility and high expression); (3)
671 HA–ME (high accessibility and medium–low expression); (4) MA–HE (medium–low
672 accessibility and high expression). Any genes that did not fall into one of the four
673 groups were not used in the analysis to maintain stringent gene groups.
674
675 **Identifying SNPs with exclusive variant types between Nile tilapia and other**
676 **species**
677 The VCF file of the 69 million SNPs mapped to the Nile tilapia (*O. niloticus*
678 UMD_NMBU [26]) reference genome in our previous study (Ciezarek, A. *et al.*, bioRxiv
679 TBC, 2023) was first normalised to split multiallelic sites into multiple rows using
680 bcftools (v 1.12) [63] with the parameters '`norm -m-any -Ov`'. After outputting the first
681 four columns from the resulting file, we use bcftools '`query`' to tabulate the number
682 of samples having each variant type, and then bcftools '`view`' to report each sample
683 with each type of variant. SNPs that are present in prioritised TFBSs of candidate
684 genes are then filtered for variant types (heterozygous and/or homozygous alternate)
685 that are exclusive to Nile tilapia and any other species.

686 ## Declarations

687 **Data Availability**

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

692

693 **Competing interests**

694 The authors declare that they have no competing interests.

695

696 **Acknowledgments**

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

705

706 **Funding**

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

711 BBS/E/T/000PR9819). Part of this work was delivered via the BBSRC National
712 Capability in Genomics and Single Cell Analysis (BBS/E/T/000PR9816) at Earlham
713 Institute by members of the Genomics Pipelines Group.
714

715 **Author contributions**

716 KR and DP sacrificed and dissected fresh fish tissues; TKM prepared tissue-specific
717 cells to perform transposition reactions and purifications for ATAC libraries; TKM and
718 AM performed tagmentation of genomic DNA controls, library amplification and QC;
719 AC identified variants in the tilapia phylogeny; TKM processed ATAC-seq and RNA-
720 seq data, identified CNEs, annotated peaks with enrichment analysis, ran GO
721 enrichment, identified TF footprints, and identified ATAC peak and gene expression
722 associations; TKM and WH wrote the manuscript with input from AM, AC, KR, DP
723 and FDP.
724

725 **Corresponding authors**

726 Correspondence to Tarang.Mehta@earlham.ac.uk

727 ## References

728 1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**,
729 68-74 (2015).
730 2. Vitti, J.J., Grossman, S.R. & Sabeti, P.C. Detecting natural selection in
731 genomic data. *Annual review of genetics* **47**, 97-120 (2013).
732 3. Hong Xia, J. *et al.* Signatures of selection in tilapia revealed by whole genome
733 resequencing. *Scientific Reports* **5**, 14168-14168 (2015).
734 4. Cadiz, M.I. *et al.* Whole genome re-sequencing reveals recent signatures of
735 selection in three strains of farmed Nile tilapia (Oreochromis niloticus). *Sci*
736 *Rep* **10**, 11514 (2020).
737 5. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African
738 cichlid fish. *Nature* **2**, 17-19 (2014).
739 6. Mehta, T.K. *et al.* Evolution of miRNA-Binding Sites and Regulatory Networks
740 in Cichlids. *Mol Biol Evol* **39**, msac146 (2022).
741 7. Mehta, T.K. *et al.* Evolution of regulatory networks associated with traits under
742 selection in cichlids. *Genome Biology* **22**, 25-25 (2021).
743 8. Gompel, N., Prud'Homme, B., Wittkopp, P.J., Kassner, V.A. & Carroll, S.B.
744 Chance caught on the wing: cis-regulatory evolution and the origin of pigment
745 patterns in Drosophila. *Nature* (2005).
746 9. Prud'homme, B. *et al.* Repeated morphological evolution through cis-
747 regulatory changes in a pleiotropic gene. *Nature* (2006).
748 10. Jeong, S., Rokas, A. & Carroll, S.B. Regulation of Body Pigmentation by the
749 Abdominal-B Hox Protein and Its Gain and Loss in Drosophila Evolution. *Cell*
750 (2006).
751 11. Cretekos, C.J. *et al.* Regulatory divergence modifies limb length between
752 mammals. *Genes and Development* (2008).
753 12. Chan, Y.F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by
754 recurrent deletion of a pitxl enhancer. *Science (New York, N.Y.)* **327**, 302-305
755 (2010).
756 13. Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the
757 regulatory epigenome. in *Nature Reviews Genetics* (2019).

758 14. Liu, Z., Zhou, T. & Gao, D. Genetic and epigenetic regulation of growth,
759 reproduction, disease resistance and stress responses in aquaculture. *Front*
760 *Genet* **13**, 994471 (2022).
761 15. Zhong, H. *et al.* DNA methylation of pituitary growth hormone is involved in
762 male growth superiority of Nile tilapia (Oreochromis niloticus). *Comparative*
763 *Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **171**,
764 42-48 (2014).
765 16. Podgorniak, T., Brockmann, S., Konstantinidis, I. & Fernandes, J.M.O.
766 Differences in the fast muscle methylome provide insight into sex-specific
767 epigenetic regulation of growth in Nile tilapia during early stages of
768 domestication. *Epigenetics* **14**, 818-836 (2019).
769 17. Konstantinidis, I. *et al.* Epigenetic mapping of the somatotropic axis in Nile
770 tilapia reveals differential DNA hydroxymethylation marks associated with
771 growth. *Genomics* **113**, 2953-2964 (2021).
772 18. Wan, Z.Y. *et al.* Genome-wide methylation analysis identified sexually
773 dimorphic methylated regions in hybrid tilapia. *Sci Rep* **6**, 35903 (2016).
774 19. Sun, L.X. *et al.* Global DNA Methylation Changes in Nile Tilapia Gonads
775 during High Temperature-Induced Masculinization. *PLoS One* **11**, e0158483
776 (2016).
777 20. Wang, Y.Y. *et al.* Epigenetic control of cyp19a1a expression is critical for high
778 temperature induced Nile tilapia masculinization. *Journal of Thermal Biology*
779 **69**, 76-84 (2017).
780 21. Kratochwil, C.F. & Meyer, A. Mapping active promoters by ChIP-seq profiling
781 of H3K4me3 in cichlid fish - a first step to uncover cis-regulatory elements in
782 ecological model teleosts. *Molecular Ecology Resources*, n/a-n/a
783 (2014).
784 22. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J.
785 Transposition of native chromatin for fast and sensitive epigenomic profiling of
786 open chromatin, DNA-binding proteins and nucleosome position. *Nature*
787 *methods* **10**, 1213-8 (2013).
788 23. Corces, M.R. *et al.* An improved ATAC-seq protocol reduces background and
789 enables interrogation of frozen tissues. *Nature Methods* (2017).
790 24. Buenrostro, J.D. *et al.* Single-cell chromatin accessibility reveals principles of
791 regulatory variation. *Nature* **523**, 486-490 (2015).

792 25. A, C.D. *et al.* Multiplex single-cell profiling of chromatin accessibility by
793 combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
794 26. Conte, M.A., Gammerdinger, W.J., Bartie, K.L., Penman, D.J. & Kocher, T.D.
795 A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome
796 reveals the structure of two sex determination regions. *BMC Genomics* **18**,
797 341-341 (2017).
798 27. Daugherty, A.C. *et al.* Chromatin accessibility dynamics reveal novel
799 functional enhancers in C. elegans. *Genome Res* **27**, 2096-2107 (2017).
800 28. Ackermann, A.M., Wang, Z., Schug, J., Naji, A. & Kaestner, K.H. Integration
801 of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature
802 genes. *Mol Metab* **5**, 233-244 (2016).
803 29. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from
804 properties to genome-wide predictions. *Nat Rev Genet* **15**, 272-86 (2014).
805 30. Starks, R.R., Biswas, A., Jain, A. & Tuteja, G. Combined analysis of dissimilar
806 promoter accessibility and gene expression profiles identifies tissue-specific
807 genes and actively repressed networks. *Epigenetics Chromatin* **12**, 16 (2019).
808 31. Yamaguchi, Y. *et al.* Acute salinity tolerance and the control of two prolactins
809 and their receptors in the Nile tilapia (Oreochromis niloticus) and Mozambique
810 tilapia (O. mossambicus): A comparative study. *General and comparative*
811 *endocrinology* **257**, 168-176 (2018).
812 32. Su, H., Ma, D., Zhu, H., Liu, Z. & Gao, F. Transcriptomic response to three
813 osmotic stresses in gills of hybrid tilapia (Oreochromis mossambicus female x
814 O. urolepis hornorum male). *BMC Genomics* **21**, 110 (2020).
815 33. Hu, Y.C., Chung, M.H. & Lee, T.H. An assay of optimal cytochrome c oxidase
816 activity in fish gills. *Anal Biochem* **553**, 38-45 (2018).
817 34. Gusmao, E.G., Allhoff, M., Zenke, M. & Costa, I.G. Analysis of computational
818 footprinting methods for DNase sequencing experiments. *Nat Methods* **13**,
819 303-9 (2016).
820 35. Kavembe, G.D., Franchini, P., Irisarri, I., Machado-Schiaffino, G. & Meyer, A.
821 Genomics of Adaptation to Multiple Concurrent Stresses: Insights from
822 Comparative Transcriptomics of a Cichlid Fish from One of Earth's Most
823 Extreme Environments, the Hypersaline Soda Lake Magadi in Kenya, East
824 Africa. *Journal of Molecular Evolution* (2015).

825 36. Li, Y. *et al.* Role of p38α Map kinase in Type I interferon signaling. *J Biol*
826 *Chem* **279**, 970-9 (2004).
827 37. Johansen, K., Maloiy, G.M. & Lykkeboe, G. A fish in extreme alkalinity. *Respir*
828 *Physiol* **24**, 159-62 (1975).
829 38. Wood, C.M. *et al.* Mammalian metabolic rates in the hottest fish on earth. *Sci*
830 *Rep* **6**, 26990 (2016).
831 39. Gibbons, T.C., Metzger, D.C.H., Healy, T.M. & Schulte, P.M. Gene expression
832 plasticity in response to salinity acclimation in threespine stickleback ecotypes
833 from different salinity habitats. *Mol Ecol* **26**, 2711-2725 (2017).
834 40. Tipsmark, C.K. *et al.* Regulation of gill claudin paralogs by salinity, cortisol
835 and prolactin in Mozambique tilapia (Oreochromis mossambicus). *Comp*
836 *Biochem Physiol A Mol Integr Physiol* **199**, 78-86 (2016).
837 41. Tipsmark, C.K., Baltzegar, D.A., Ozden, O., Grubb, B.J. & Borski, R.J. Salinity
838 regulates claudin mRNA and protein expression in the teleost gill. *Am J*
839 *Physiol Regul Integr Comp Physiol* **294**, R1004-14 (2008).
840 42. Shemsanga, C. *et al.* Origin and mechanisms of high salinity in Hombolo Dam
841 and groundwater in Dodoma municipality Tanzania, revealed. *Applied Water*
842 *Science* **7**, 2883-2905 (2017).
843 43. Katunzi, E.F.B. Conservation of satellite lakes as refugia for the endangered
844 species of Lake Victoria: a case study of Lake Malimbe. *National Fisheries*
845 *Resources Research Institute Publications* (1995).
846 44. Mapenzi, L.L., Shimba, M.J., Moto, E.A., Maghembe, R.S. & Mmochi, A.J.
847 Heavy metals bio-accumulation in tilapia and catfish species in Lake Rukwa
848 ecosystem Tanzania. *Journal of Geochemical Exploration* **208**, 106413
849 (2020).
850 45. Barker, P., Telford, R., Gasse, F. & Thevenon, F. Late Pleistocene and
851 Holocene palaeohydrology of Lake Rukwa, Tanzania, inferred from diatom
852 analysis. *Palaeogeography, Palaeoclimatology, Palaeoecology* **187**, 295-305
853 (2002).
854 46. FAO. The State of World Fisheries and Aquaculture 2022. Towards Blue
855 Transformation. *Rome, FAO* (2022).
856 47. Yang, H. *et al.* A map of cis-regulatory elements and 3D genome structures in
857 zebrafish. *Nature* **588**, 337-343 (2020).

858 48. Thurman, R.E. *et al.* The accessible chromatin landscape of the human
859 genome. *Nature* **489**, 75-82 (2012).
860 49. Martin-Trujillo, A. *et al.* Rare genetic variation at transcription factor binding
861 sites modulates local DNA methylation profiles. *PLoS Genet* **16**, e1009189
862 (2020).
863 50. Arbiza, L. *et al.* Genome-wide inference of natural selection on human
864 transcription factor binding sites. *Nat Genet* **45**, 723-9 (2013).
865 51. Tseng, C.C. *et al.* Genetic Variants in Transcription Factor Binding Sites in
866 Humans: Triggered by Natural Selection and Triggers of Diseases. *Int J Mol*
867 *Sci* **22**(2021).
868 52. Plaisier, C.L. *et al.* Causal Mechanistic Regulatory Network for Glioblastoma
869 Deciphered Using Systems Genetics Network Analysis. *Cell Systems* **3**, 172-
870 186 (2016).
871 53. El-Leithy, A.A.A. *et al.* Optimum salinity for Nile tilapia (Oreochromis niloticus)
872 growth and mRNA transcripts of ion-regulation, inflammatory, stress- and
873 immune-related genes. *Fish Physiol Biochem* **45**, 1217-1232 (2019).
874 54. USEPA. Quality criteria for water. *Washington, D.C.: Office of Water*
875 *Regulations and Standards.* (1986).
876 55. Mustapha, M.K. & Atolagbe, S.D. Tolerance level of different life stages of
877 Nile tilapia Oreochromis niloticus (Linnaeus, 1758) to low pH and acidified
878 waters. *The Journal of Basic and Applied Zoology* **79**, 46 (2018).
879 56. Ford, A.G.P. *et al.* Molecular phylogeny of Oreochromis (Cichlidae:
880 Oreochromini) reveals mito-nuclear discordance and multiple colonisation of
881 adverse aquatic environments. *Mol Phylogenet Evol* **136**, 215-226 (2019).
882 57. Edwards, S.L. & Marshall, W.S. 1 - Principles and Patterns of Osmoregulation
883 and Euryhalinity in Fishes. in *Fish Physiology*, Vol. 32 (eds. McCormick, S.D.,
884 Farrell, A.P. & Brauner, C.J.) 1-44 (Academic Press, 2012).
885 58. Zadunaisky, J.A. Chloride cells and osmoregulation. *Kidney Int* **49**, 1563-7
886 (1996).
887 59. Campo, A. *et al.* Different transcriptomic architecture of the gill epithelia in Nile
888 and Mozambique tilapia after salinity challenge. *Comp Biochem Physiol Part*
889 *D Genomics Proteomics* **41**, 100927 (2022).

890 60. Shechonge, A. *et al.* Widespread colonisation of Tanzanian catchments by
891 introduced Oreochromis tilapia fishes: the legacy from decades of deliberate
892 introduction. *Hydrobiologia* **832**, 235-253 (2019).
893 61. Andrews, S. FastQC: A quality control tool for high throughput sequence data
894 [Online]. *Available online at:*
895 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
896 62. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. in
897 *Nature Methods* Vol. 9 357-359 (2012).
898 63. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**,
899 giab008-giab008 (2021).
900 64. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC*
901 *Bioinformatics* **10**, 421-421 (2009).
902 65. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. Sambamba: fast
903 processing of NGS alignment formats. *Bioinformatics* **31**, 2032-2034 (2015).
904 66. Quinlan, A.R. & Hall, I.M. BEDTools: A flexible suite of utilities for comparing
905 genomic features. *Bioinformatics* **26**, 841-842 (2010).
906 67. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X.S. Identifying ChIP-seq
907 enrichment using MACS. *Nature Protocols* **7**, 1728-1740 (2012).
908 68. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology*
909 **9**, R137-R137 (2008).
910 69. Ou, J. *et al.* ATACseqQC: a Bioconductor package for post-alignment quality
911 assessment of ATAC-seq data. *BMC Genomics* **19**, 169-169 (2018).
912 70. Conte, M.a. & Kocher, T.D. An improved genome reference for the African
913 cichlid, Metriaclima zebra. *BMC Genomics* **16**, 724-724 (2015).
914 71. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the
915 thousand-genome era. *Nature* **587**, 246-251 (2020).
916 72. Hubisz, M.J., Pollard, K.S. & Siepel, A. PHAST and RPHAST: phylogenetic
917 analysis with space/time models. *Briefings in Bioinformatics* **12**, 41-51 (2011).
918 73. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of
919 nonneutral substitution rates on mammalian phylogenies. *Genome research*
920 **20**, 110-21 (2010).
921 74. Heger, A., Webber, C., Goodson, M., Ponting, C.P. & Lunter, G. GAT: A
922 simulation framework for testing the association of genomic intervals.
923 *Bioinformatics* **29**, 2046-2048 (2013).

924 75. Benjamini, Y. & Hochberg, Y. Benjamini Y, Hochberg Y. Controlling the false
925 discovery rate: a practical and powerful approach to multiple testing. *Journal*
926 *of the Royal Statistical Society B* **57**, 289-300 (1995).
927 76. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis
928 and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**,
929 W191-W198 (2019).
930 77. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq.
931 *Genome Biology* **20**, 45-45 (2019).
932 78. Castro-Mondragon, J.A. *et al.* JASPAR 2022: the 9th release of the open-
933 access database of transcription factor binding profiles. *Nucleic Acids*
934 *Research* **50**, D165-D173 (2022).
935 79. Kulakovskiy, I.V. *et al.* HOCOMOCO: A comprehensive collection of human
936 transcription factor binding sites models. *Nucleic Acids Research* **41**, D195-
937 202 (2013).
938 80. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: A
939 database of transcription factor binding sites identified by ChIP-seq
940 experiments. *Nucleic Acids Research* **45**, D61-D67 (2017).
941 81. Hume, M.A., Barrera, L.A., Gisselbrecht, S.S. & Bulyk, M.L. UniPROBE,
942 update 2015: New tools and content for the online database of protein-binding
943 microarray data on protein-DNA interactions. *Nucleic Acids Research* **43**,
944 D117-D122 (2015).
945 82. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based
946 genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature*
947 *Biotechnology* **37**, 907-915 (2019).
948 83. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced
949 multi-sample quality control for high-throughput sequencing data.
950 *Bioinformatics (Oxford, England)* **32**, 292-294 (2016).
951 84. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with
952 high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166-
953 169 (2015).
954 85. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq
955 data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
956 86. Li, H.-D., Lin, C.-X. & Zheng, J. GTFtools: a software package for analyzing
957 various features of gene models. *Bioinformatics* **38**, 4806-4808 (2022).

87. Sanghi, A. *et al.* Chromatin accessibility associates with protein-RNA correlation in human cancer. *Nat Commun* **12**, 5732 (2021).
88. 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).