Development and Validation of an Open Access Snp Array for Nile Tilapia (Oreochromis niloticus)

G3: Genes|Genomes|Genetics Early Online, published on June 12, 2020 as doi:10.1534/g3.120.401343

DEVELOPMENT AND VALIDATION OF AN OPEN ACCESS SNP ARRAY FOR NILE TILAPIA (*Oreochromis niloticus*)

Carolina Peñaloza[*], Diego Robledo[*], Agustin Barría[*], Trọng Quốc Trịnh[†], Mahirah Mahmuddin[†], Pamela Wiener[*], John A. H. Benzie[†,‡], Ross D. Houston[*]

[*] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, United Kingdom

[†] WorldFish, Penang, 10670, Malaysia

[‡] School of Biological, Earth and Environmental Sciences, University College Cork, Cork, T12 YN60, Ireland

A ~65K SNP ARRAY FOR NILE TILAPIA

Key words: GIFT, Abbassa, aquaculture, Nile tilapia, SNP array

Corresponding author: Ross D. Houston

Address: The University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG

Tel Ross D. Houston : +44 (0) 131 651 9224

Email: ross.houston@roslin.ed.ac.uk

Abstract

Tilapia are amongst the most important farmed fish species worldwide, and are fundamental for the food security of many developing countries. Several genetically improved Nile tilapia (*Oreochromis niloticus*) strains exist, such as the iconic Genetically Improved Farmed Tilapia (GIFT), and breeding programmes typically follow classical pedigree-based selection. The use of genome-wide single-nucleotide polymorphism (SNP) data can enable an understanding of the genetic architecture of economically important traits and the acceleration of genetic gain via genomic selection. Due to the global importance and diversity of Nile tilapia, an open access SNP array would be beneficial for aquaculture research and production. In the current study, a ~65K SNP array was designed based on SNPs discovered from whole-genome sequence data from a GIFT breeding nucleus population and the overlap with SNP datasets from wild fish populations and several other farmed Nile tilapia strains. The SNP array was applied to clearly distinguish between different tilapia populations across Asia and Africa, with at least ~30,000 SNPs segregating in each of the diverse population samples tested. It is anticipated that this SNP array will be an enabling tool for population genetics and tilapia breeding research, facilitating consistency and comparison of results across studies.

INTRODUCTION

Nile tilapia (*Oreochromis niloticus*) is one of the most widely farmed freshwater fish species in the world, with 4.2 million tonnes being produced in 2016 (FAO 2018). Although this species is native to Africa, Nile tilapia aquaculture has been successfully established in over fifty countries across Asia, Africa, and South America (Eknath and Hulata 2009). The popularity of tilapias stem from their overall ease of culture, which is largely based on their fast growth rate, robustness, relatively short generation interval, and ability to adapt to diverse farming systems and habitats (Ng and Romano 2013; Eknath *et al*. 1998), although see Jansen *et al*. (2019) for discussion of recent disease outbreaks. These attributes make Nile tilapia a suitable species for use in the diverse and often suboptimal farming systems of many low and middle-income countries, where it represents an important source of animal protein and social well-being (Ansah *et al*. 2014).

Several selective breeding programmes have been established for Nile tilapia (Neira 2010), among which a major success story is the development of the widely farmed Genetically Improved Farmed Tilapia (GIFT) strain. The GIFT base population was formed in the early 1990s and was composed of eight unrelated strains: four wild populations from Africa (Egypt, Ghana, Kenya and Senegal) and four widely farmed Asian strains (Israel, Singapore, Taiwan and Thailand) (Eknath *et al*. 1993). The main breeding objective of the GIFT program was to improve growth rate, but other relevant traits such as overall survival, resistance to diseases, and maturation rate were also considered (Eknath and Acosta 1998; Trọng *et al*. 2013; Komen and Trọng 2014). Breeding programs have achieved significant genetic gains for growth-related traits in this species. For instance, after five generations of artificial selection the GIFT strain showed on average a 67 % higher body weight at harvest compared

to the base population (Bentsen *et al*. 2017). Most of the genetic progress achieved to date for tilapia was obtained through traditional pedigree-based approaches. The use of genome-wide genetic markers to estimate breeding values for selection candidates via genomic selection (Meuwissen *et al*. 2001; Sonesson and Meuwissen 2009) has the potential to increase genetic gain, particularly for traits that are difficult or expensive to measure directly on the candidates. Therefore, the development and application of high density genotyping platforms would be advantageous in expediting genetic improvement in breeding programmes for Nile tilapia.

SNP arrays are powerful high-throughput genotyping tools that are increasingly becoming available for aquaculture species including Atlantic salmon (*Salmo salar*) (Houston *et al*. 2014; Yáñez *et al*. 2016), common carp (*Cyprinus carpio*) (Xu *et al*. 2014), rainbow trout (*Oncorhynchus mykiss*) (Palti *et al*. 2015), Pacific (*Crassostrea gigas*) and European (*Ostrea edulis*) oysters (Lapegue *et al*. 2014; Qi *et al*. 2017; Gutierrez *et al*. 2017), catfish (*Ictalurus punctatus* and *Ictalurus furcatus*) (Liu *et al*. 2014; Zeng *et al*. 2017;), Arctic charr (*Salvelinus alpinus*) (Nugent *et al*. 2019), tench (*Tinca tinca*) (Kumar *et al*. 2019), and indeed Nile tilapia (Joshi *et al*. 2018; Yáñez *et al*. 2020). Compared to other high-throughput genotyping methods, such as RAD-Seq (Baird *et al*. 2008), SNP arrays have the advantage of increased genotyping accuracy and SNP stability, as the same markers are interrogated each time (Robledo *et al*. 2018a). These platforms have been used to study the genetic architecture of diverse production traits such as growth (Tsai *et al*. 2015; Gutierrez *et al*. 2018) and disease resistance (Tsai *et al*. 2016; Bangera *et al*. 2017; Robledo *et al*. 2018b), and their utility for genomic prediction in several aquaculture species has been clearly demonstrated (for a review see Zenger *et al*. (2019)).

The two Nile tilapia SNP arrays developed to date are both focused on the broodstock strains of specific commercial breeding programmes. One of the platforms was designed based on the analysis of the GenoMar Supreme Tilapia (GST®) strain (Joshi *et al*. 2018), whereas the other platform derived from the evaluation of two strains belonging to Aquacorporación Internacional (Costa Rica) and an unspecified commercial strain from Brazil (Yáñez *et al*. 2020). These SNP arrays have been shown to be highly effective in the discovery populations, and have been used to generate high-density linkage maps and perform tests of genomic selection (Joshi *et al*. 2019; Yoshida *et al*. 2019a). However, while all of these commercial strains are related to the GIFT strain (which underpins a large proportion of global tilapia aquaculture), their utility and performance in other farmed tilapia strains, especially those inhabiting Asia and Africa, is unknown. To develop platforms that are not exclusively informative in a focal strain, ideally additional SNP panels derived from genetically diverse populations should be evaluated during the SNP selection process (Montanari *et al*. 2019). This strategy would allow mitigating ascertainment bias, and thus broadening the applicability of a SNP array.

The aim of this study was to develop a publicly available, open access ~65K SNP array for Nile tilapia based on the widely cultured GIFT strain, but that also contains informative markers in multiple tilapia strains across Asia and Africa. To achieve this, a large SNP database was generated by whole genome Illumina sequencing of pooled genomic DNA from 100 individuals from the WorldFish GIFT breeding nucleus from Malaysia. These newly discovered markers were cross-referenced with previously identified SNP panels in several populations, with the aim of prioritising markers that are informative across strains. To test the performance of the SNP array, nine Nile tilapia populations of different geographical origins and genetic

backgrounds (i.e. GIFT, GIFT-derived and non-GIFT strains / populations) were genotyped. The broad utility and open-access availability of the array is anticipated to benefit both the academic and commercial communities to advance genomic studies in this species and support ongoing and emerging breeding programmes.

MATERIALS AND METHODS

Animals, DNA extraction and sequencing

One hundred Nile tilapia broodstock samples from the 15th generation of the core GIFT Nile tilapia-breeding nucleus of WorldFish at the Aquaculture Extension Center in Jitra (Kedah, Malaysia) were used for DNA sequencing for SNP discovery. Caudal fin clips were sampled and preserved in absolute ethanol at -20˚ until shipment from Malaysia to The Roslin Institute (University of Edinburgh, UK) for DNA extraction, sequencing and genetic analysis.

Genomic DNA was isolated from the tilapia fin clips using a salt-based extraction method (Aljanabi and Martinez 1997). The integrity of the DNA samples was assessed by performing an agarose gel electrophoresis. DNA quality was also evaluated by estimating the 280/260 and 230/280 ratios on a NanoDrop 1000 UV spectrophotometer. The concentration of the DNA extractions was measured with the Qubit dsDNA BR assay kit (Invitrogen, Life technologies). Samples were diluted to 50 ng/ul and then combined in equimolar concentrations to generate two pools of 50 (different) individuals each. Library preparation and sequencing services were provided by Edinburgh Genomics (University of Edinburgh, UK). DNA pools were prepared for sequencing using a TruSeq PCR-free kit (Illumina, San Diego). The two

pools were then sequenced at a minimum 90X depth of coverage on an Illumina HiSeq X platform with a 2x150 bp read length.

SNP discovery in the GIFT strain

The quality of the sequencing output was assessed using FastQC v.0.11.5 (Andrews 2010). Quality filtering and removal of residual adaptor sequences was conducted on read pairs using Trimmomatic v.0.38 (Bolger *et al*. 2014). Specifically, Illumina specific adaptors were trimmed from the reads, leading and trailing bases with a Phred score less than 20 were removed, and reads were trimmed if the average Phred score over four consecutive bases was less than 20. Only read pairs that had a post-filtering-length longer than 36 bp were retained. Cleaned paired-end reads were aligned to the *Oreochromis niloticus* genome assembly published by Conte *et al*. (2017) (Genbank accession GCF_001858045.2) using BWA v0.7.17 (Li and Durbin 2009). To minimise biased estimates of allele frequencies, PCR duplicates were removed from the dataset using SAMtools v1.6 (Li *et al*. 2009). Variants were called from the pools with the software Freebayes v1.0.2 (Garrison and Marth 2012 *preprint*) if (i) at least three reads supported the alternate allele or (ii) the SNP allele frequency in the pool was above 0.02, whichever condition was met first. As a first filtering step, only SNPs that had no interfering variants within less than 40 bp on either side were retained. The resulting vcf file was then filtered to obtain a list of high quality variants with vcffilter v1.0.0 (https://github.com/vcflib/vcflib); bi-allelic SNPs meeting the following criteria were kept for further evaluation: (i) a minimum coverage of 50X and maximum coverage of 150X, (ii) presence of supporting reads on both strands, (iii) at least two reads balanced to each side of the site and (iv)

more than 90% of the observed alternate and reference alleles are supported by properly paired reads. To enrich the platform for variants located on or nearby genes, polymorphisms were annotated and classified using the software SnpEff v4.3 (Cingolani *et al*. 2012). This list of candidate SNPs were sent as 71-mer nucleotide sequences to ThermoFisher for *in silico* probe scoring.

Overlap between GIFT SNPs and other datasets

In order to reduce ascertainment bias and increase the utility of the platform across multiple strains, we prioritised markers that also segregated in other strains / populations. The candidate GIFT SNP discovery panel was compared with four other lists of variants. The first panel of variants used for comparison were identified in an inter-generational sample of individuals of the Abbassa strain, a selectively bred Nile tilapia strain from Egypt (Abbassa breeding panel: 6,163 SNPs) (Lind *et al*. 2017). The second SNP panel corresponds to variants discovered in wild fish populations from the region of Abbassa, Egypt (Abbassa wild panel: 6,749 SNPs). The third SNP panel was obtained from a Nile tilapia stock that had been selected for growth for over ten years in Kenya, and that was initially founded by individuals from several populations from East Africa (Kenya breeding panel: 33,085 SNPs). The fourth panel of variants derived from the joint analysis of farmed and wild fish populations from Tanzania (Tanzania panel: 2,182 SNPs). In addition, and as a quality control check, the candidate list of GIFT SNPs was cross-referenced against a panel of markers identified in a sub-sample of the WorldFish GIFT population at Jitra, Malaysia (Wageningen panel: 7,298 SNPs) (Van Bers *et al*. 2012).

SNP selection

The process of selecting the final panel of SNPs for inclusion on the Applied Biosystems Axiom Tilapia Genotyping Array was as follows. First, SNPs that were previously identified as being associated with phenotypic sex were included (Palaiokostas *et al*. 2013, 2015) (Supplementary Table S1). Second, all SNPs that were shared with at least one other SNP panel – either Abbassa breeding, Abbassa wild, Kenya breeding, Tanzania or Wageningen – were considered as high priority markers and included directly on the array. In addition, for each SNP that was submitted for evaluation, ThermoFisher assigns a design score (p-convert value) to both 35 bp probes flanking the variant. Probes with a high p-convert value indicate an assay with a higher probability of SNP conversion. Based on their p-convert value, probes can be classified as either 'recommended', 'neutral', 'not recommended' or 'not possible'. For downstream analysis, SNPs that had at least one probe that was either 'recommended' or 'neutral' were retained. Next, SNPs were filtered according to their minor allele frequency (MAF) by removing markers with an average MAF (estimated from the two sequenced pools) < 0.05 or >0.45. The latter MAF threshold was applied to avoid spurious SNPs resulting from sequence differences between paralogues. Additional criteria for SNP selection included filtering out *A/T* and *G/C* variants, as compared to other polymorphisms they require twice as many assays on a ThermoFisher Axiom platform. From the remaining list of high confidence SNPs identified in the discovery population, polymorphisms located in exons were prioritized. To fill the remaining target of ~65K, SNPs were selected from those located either within a gene or at most at a 1 kb distance. The strategy of enriching for SNPs on genes was followed because they are more likely to alter protein function, and therefore may have a larger effect on

phenotypes compared to variants occurring outside genes (Jorgenson & Witte 2006). To obtain a uniform physical distribution across the Nile tilapia genome, all chromosomes and 130 of the longest scaffolds were divided into 10-kb non-overlapping windows, and the SNP with the highest MAF within each interval was selected for inclusion in the platform. Finally, for 1-Mb regions exhibiting the lowest number of markers, the SNP with the highest MAF within the region was included manually.

SNP array validation

The ThermoFisher Axiom ~65K Nile tilapia SNP array designed in this study was tested by genotyping nine Nile tilapia populations of different geographical locations and genetic backgrounds (Table 1). The tested fish belonged to one wild population from Egypt (Abbassa wild) and six genetically improved strains. The evaluated strains were the (i) Genetically Improved Farmed Tilapia (GIFT) (Eknath and Acosta 1998; Eknath *et al*. 1993), (ii) Genetically Enhanced Tilapia-Excellent (GET-EXCEL) (Tayamen 2004), (iii) Brackish water Enhanced Saline Tilapia (BEST) (Tayamen *et al*. 2004), (iv) Freshwater Aquaculture Centre (FAC) selected Tilapia (FaST) (Bolivar 1998), and improved strains from (v) Kenya and (vi) Abbassa (Egypt). For each representative strain, a single population was sampled, with the exception of the GIFT strain, for which three populations from different countries were evaluated, Malaysia (discovery population), Bangladesh and Philippines.

In total, 135 individuals, comprising 15 fish of balanced sex ratios in each population, were genotyped by IndentiGEN (Ireland) using the Nile tilapia ~65K SNP array. To perform a principal component analysis (PCA) on the genome-wide SNP data the

following SNPs and samples were retained using PLINK v1.9 (Chang *et al*. 2015): (i) SNPs of the Poly High Resolution class (i.e. high quality markers with three well-resolved genotype clusters) (ii) markers with a call rate > 0.95, (iii) individuals with a call rate > 0.90, and (iv) one SNP of a pair showing high linkage disequilibrium (r^2 > 0.7). In addition, for individuals with greater than 80 % identity-by-state (IBS) with another individual, only one was retained for further analysis. The structure of the 135 individuals genotyped with the SNP array was investigated using the R package LEA (Frichot and François 2015), with the significance of the identified components evaluated with Tracy-Widom statistics (Tracy and Widom 1994).

Summary statistics of SNPs

The levels of observed and expected heterozygosity (Ho, He) for each Nile tilapia strain / population were calculated, and 95 % confidence intervals of Ho estimated based on 1,000 bootstrap replicates. To evaluate the informativeness of the SNPs on the array, the average MAF values per strain / population were calculated and classified into five different categories: Common (MAF > 0.3); Intermediate (0.3 > MAF > 0.1); Low (0.1 > MAF > 0.05); Rare (MAF <0.05); and Fixed (MAF = 0).

Linkage disequilibrium magnitude and decay

To estimate linkage disequilibrium (LD) we used a version of the SNP dataset to which all individual and SNP QC filters were applied (see SNP array validation section), except the removal of markers based on pairwise LD. As a pairwise measure of LD, r^2 (Hill and Robertson 1968) was chosen because it is most frequently used in the context of association mapping (Ardlie *et al*. 2002). Moreover,

other LD metrics such as D' are highly affected by sample size (McRae *et al*. 2002) and its use is not recommended when sample sizes are small. LD was estimated separately for each strain / population as the inter-marker Pearson's squared correlation coefficient r^2 corrected for relatedness (r^2vs) using the package LDcorSV v1.3.1 (Mangin *et al*. 2012) in R v 3.5.0 (R Core Team 2014). For comparison, two MAF thresholds were applied to the data before measuring the extent of LD, MAF > 0.05 and MAF > 0.1. The average r^2 was calculated in 10-kb bins (pairwise distance between SNPs) for each Nile tilapia chromosome. The LD decay was visualized using the R package ggplot2 (Hadley 2009) by plotting the average r^2 within each bin (across all chromosomes) against inter-marker distances, which extended from zero up to 10 Mb.

Ethics statement

Data collection and sampling of the GIFT samples was performed as part of a non-profit selective breeding program run by WorldFish. The animals from this breeding population are managed in accordance with the Guiding Principles of the Animal Care, Welfare and Ethics Policy of WorldFish. Tissue sampling was carried out in accordance with the norms established by the Reporting *In Vivo* Experiments (ARRIVE) guidelines.

Data availability

Raw sequence reads from the two pools analysed for SNP discovery have been deposited in NCBI's Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) under accession number PRJNA520791. Genome

position and probes for all SNPs included in the ~65K SNP array are given in File S1. Genome position and allele frequency of array SNPs discovered in the Pool-seq data can be found in the European Variation Archive (EVA, https://www.ebi.ac.uk/eva/) under accession number PRJEB38548. The tilapia SNP array is available for commercial purchase from ThermoFisher (array number 551071, email: BioinformaticsServices@thermofisher.com).

RESULTS

SNP selection and array development

The pooled DNA sequencing resulted in 458M and 461M paired-end reads for the two DNA pools. The alignment of the quality control filtered reads against the Nile tilapia reference genome (Genbank accession GCF_001858045.2) led to the discovery of ~20 million putative polymorphisms. Of the 1,166,652 bi-allelic SNPs that remained after applying post-alignment quality control (QC) filters, 694,348 fell within genes or in the neighbouring regions of genes (i.e. within <1 kb). After additional filtering criteria related to allelic frequency thresholds (removal of SNPs with average MAF < 0.05 or > 0.45) and the type of allele polymorphism (removal of *A/T* and *G/C* variants), 351,188 SNPs were sent as 71-mer nucleotide sequences to ThermoFisher for *in silico* probe scoring. From the list of scored SNP probes provided by ThermoFisher, only those that were categorised as either 'recommended' or 'neutral' were selected.

The final ~65K SNP array contained (i) 7 sex determination markers, (ii) 6,883 SNPs discovered in our population that overlap with SNP panels identified in other strains / populations, (iii) 11,328 SNPs located in exons, and (iv) 47,239 SNPs occurring in

genes or within < 1 kb of genes. The latter set of SNPs were selected to be evenly physically spaced along the 22 chromosomes (Supplementary Figure S1) and 130 of the longest unplaced scaffolds of the Nile tilapia genome assembly.

SNP array validation

After QC of the genotyping data, seven, two and one fish were removed due low call rate from the Abbassa wild, Abbassa strain and BEST population, respectively. Therefore, 125 individual fish from across nine different strains / populations were used to validate the SNP array (Table 1). The obtained raw intensity files were imported to the Axiom Analysis Suite software v2.0.035 for quality control analysis and genotype calling. Genotypes were called following the Best Practices Workflow using the default settings for diploid organisms (Thermo Fisher Scientific Inc 2018). The SNP probe sets were classified into one of the following six category classes based on cluster properties and QC metrics: PolyHighResolution, NoMinorHom, MonoHighRes, Off Target Variant (OTV), CallRateBelowThreshold, and Other. Of the 65,450 SNPs assayed by the platform, 54,604 SNPs (83.4 %) were classified as PolyHighResolution markers, the class with the highest quality probes and presence of both the major and minor homozygous clusters. The number of SNPs that showed a good cluster resolution but no evidence of individuals with minor homozygous genotypes (NoMinorHom) was 2,122 (3.2 %). Only 374 SNPs (0.5 %) on the array were monomorphic (MonoHighResolution). Among the SNPs that failed to provide reliable genotypes, 194 SNPs (0.2 %) were OTV, 3,026 SNPs (4.6 %) had a SNP call rate below the chosen threshold of 0.97 (CallRateBelowThreshold), and 5,130 (7.8 %) were not classified into any of the above categories (Other). After applying

355 standard QC filters, 54,310 (MAF > 0.05) and 49,429 (MAF > 0.1) SNPs and 125
356 individuals were retained for the assessment of LD decay; after the pruning of
357 markers based on LD, 42,460 SNPs remained for the estimation of general
358 population statistics and population structure.

359

360 **Minor allele frequency and genetic diversity in Nile tilapia populations / strains**

361 The average observed heterozygosity of the genotyped populations was 0.29, with
362 the GIFT strain from Malaysia (i.e. the primary discovery population) having the
363 highest value (0.35), and the Kenyan population the lowest (0.21) (Table 1). Overall,
364 the observed heterozygositites (Ho) were slightly higher than expected (He), and
365 showed a similar pattern across populations. The only exception was the Kenyan
366 strain, for which the Ho was lower than the He (0.21 vs 0.24).

367 The average MAF of all 42,460 successfully genotyped SNPs ranged from 0.23 to
368 0.26 across the six strains and the single wild population evaluated. The number of
369 informative markers (MAF > 0) in the array was highest for samples from GIFT and
370 GIFT-derived populations compared to populations with non-GIFT genetic
371 backgrounds (Figure 1). The primary discovery population had the greatest number
372 of informative markers, 40,930 SNPs (96 %). As expected, the populations
373 genetically closer to the GIFT discovery population from Malaysia had the second
374 and third highest numbers of informative markers – 40,743 (95 %) and 39,562 (93
375 %) informative SNPs in the GIFT stocks from Bangladesh and the Philippines,
376 respectively. Likewise, GIFT-derived strains exhibit similarly high levels of
377 informative SNPs, with 38,232 (90 %) markers segregating in the GET-EXCEL and
378 37,867 (89 %) in the BEST strain. The number of informative markers for the three

non-GIFT strains evaluated in this study were 30,631 (72 %) for the FaST strain, 31,061 (73 %) for the Kenyan domesticated line and 30,786 (72 %) for the Abbassa strain. A large fraction of these informative SNPs co-segregate with the GIFT strain (Figure 2). The average MAF for the markers that are common to all the different representative strains evaluated (total = 19,815 SNPs) was similar and ranged from 0.26 to 0.28. The single wild population analysed, Abbassa-wild, exhibited the lowest number of informative markers (28,421 SNPs; 66 %).

Population structure

The population stratification of the nine Nile tilapia strains / populations was visualized using a PCA to reduce the dimensions of the genotype data (Figure 3). The two first eigenvectors accounted for 22 % of the total variance. The first dimension, which explains 13 % of the variance, mainly separates GIFT and GIFT-derived populations from the Nile tilapia strains / populations of African origin (Abbassa-strain, Abbassa-wild and Kenya). The second principal component explains 9 % of the total variance and separates the strains / populations from Africa into two clusters, one comprised of Nile tilapia individuals from Egypt (Abbassa-strain and Abbassa wild) and the other comprising the Kenyan domestic line. Additionally, this dimension also separates Asian GIFT, GIFT-derived and non-GIFT strains into three distinct clusters represented by the (i) FaST strain, (ii) GIFT strains from Malaysia, Philippines and Bangladesh, and (iii) non-GIFT strains, namely GET-EXCEL and BEST. Three individuals of putative Kenyan origin did not group with the Kenyan cluster (those with negative PC1 values in Figure 3).

Linkage disequilibrium decay

The overall average LD between marker pairs was relatively low and decayed as physical distance increased. Similar patterns of LD decay were observed across Nile tilapia populations for the two MAF thresholds applied to the data, although the MAF filter of 0.1 resulted in higher magnitudes of r^2 (Figure 4). Two distinct patterns of LD decay were observed across strains / populations. A first group – composed exclusively by domestic lines (GIFT-Ma, GIFT-Ba, GIFT-Ph, GET-EXCEL, BEST, FaST, Kenya and Abbassa-strain) – showed a moderate to low LD decay over short and long-range distances. The average observed values of r^2 at the smallest inter-marker distance evaluated (10 kb bin) was ~0.2 (MAF > 0.1 dataset). Within short-range distances (< 100 kb), pairwise correlations experienced a 10 to 23% decrease when LD between markers separated by ~10 kb was compared to pairs of loci separated by ~ 100 kb. Considering long-range distances, the average r^2 dropped by 65 % from that estimated at 10 kb compared to 10,000 kb in GIFT-Ma (0.21 vs 0.08), 60 % in GIFT-Ba (0.18 vs 0.07), 61 % in GIFT-Ph (0.19 vs 0.07), 59 % in GET-EXCEL (0.19 vs 0.08), 54 % in the BEST strain (0.19 vs 0.09), 72 % in FaST (0.27 vs 0.08), 54 % in the Kenyan strain (0.17 vs 0.08) and 52 % in the Abbassa-strain (0.18 vs 0.09). By contrast, LD decayed much more slowly in the Abbassa-wild population, as initial levels of LD persist over long inter-marker distances. The reduction in r^2 between SNPs at 10 kb vs 10,000 kb distance apart from each other was of only 22 % (0.19 vs 0.15) (Figure 4).

DISCUSSION

The ~65K SNP array developed in this study is an open-access high-throughput genotyping platform for Nile tilapia. A large majority of the SNPs on the platform were of high quality and polymorphic – 87 % of the SNPs fell in either the PolyHighResolution or NoMinorHom categories. This performance value lies in the upper range of current aquaculture SNP arrays (e.g. 89 % for rainbow trout (Palti *et al*. 2015) and 77 % for the latest catfish array (Zeng *et al*. 2017)), demonstrating the efficacy of our Pool-Seq strategy for robust SNP discovery at a fraction of the sequencing effort of typical SNP chip designs.

Two published SNP arrays have been developed for Nile tilapia, each of ~58K SNPs (Joshi *et al*. 2018; Yáñez *et al*. 2020). These platforms capture the genetic diversity of specific improved lines, but their efficacy has only been demonstrated in the GST® (i.e. GIFT line further improved through genomic tools) (Joshi *et al*. 2018) or GIFT and GIFT-related strains from South America (Yáñez *et al*. 2020). In our array, the bulk of SNPs were derived from a SNP discovery process performed on two DNA pools of 100 fish of the core breeding nucleus of the WorldFish GIFT strain, which underpins a large proportion of global tilapia production. However, to mitigate ascertainment bias and widen the applicability of the platform, panels from previous SNP discovery projects were cross-referenced and common SNPs were prioritised. Yet, as expected, the number of informative SNPs decreases with increasing genetic distance from the primary discovery population (e.g. ~63 % in non-GIFT strains; Figure 2). Even though a small number of individuals (~15 per strain / population) were genotyped with the array, there were at least ~30,000 SNPs segregating in each of the population samples evaluated, and near 20,000 common SNPs segregating in all non-GIFT strains tested, namely Abbassa, Kenya and FaST.

Therefore, it is proposed that this SNP array can serve as a common platform for use by the tilapia genetics and breeding community to encourage cross-study comparisons and meta-analyses of genomic datasets.

A principal component analysis demonstrated that our 65K SNP array distinguishes the four major strains evaluated in this study (GIFT, Abbassa, Kenya and FaST), indicating clear independent clusters based on the first two principal components. While the purpose of this analysis was to test the utility of the SNP array to distinguish populations, a few interesting observations were noted. First, individuals from the Abbassa genetically improved strain clustered with wild fish from the same region (i.e. Abbassa, Egypt). This pattern is consistent with a short period of artificial selection that has not yet led to significant shifts in allele frequencies. Additionally, the projection of the Kenyan cluster along a line in the PC plot may indicate the recent admixture of two populations, as suggested for this dispersion pattern by Patterson *et al*. (2006). On the other hand, GIFT and GIFT-derived strains form a loose cluster that separates in dimension 2 of the PC plot but that is not clearly maintained in dimensions 3 to 6 (Supplementary Figures S2-S3). This lack of consistency likely indicates that the population structure of this cluster may not be well represented by the first two PCs. As expected, there is a large degree of overlap among the GIFT strains, most likely due to their common origin. The GIFT-derived strains (GET-EXCEL and BEST) tend to co-cluster in the PC plot; as both strains were developed in the Philippines (Tayamen 2004; Tayamen *et al*. 2004), this concordance could reflect shared breeding goals and similar production systems and breeding practices. Interestingly, although these GIFT-derived strains are the product of selection programmes applied to base populations originating from different strains, the PCA suggests they are genetically closer to the GIFT strain. For

instance, GET-EXCEL is a synthetic strain developed based on four parental lines: the GIFT strain (8^{TH} generation), the FaST strain (13^{TH} generation), an Egyptian strain (composed by animals sourced from eight locations in Egypt) and a Kenyan strain (coming from stock collected in Lake Turkana) (Tayamen 2004). However, in the PC plot GET-EXCEL individuals group with the BEST strain, closer to the GIFT cluster, and more distant to any of the other genetic clusters they supposedly derive from (i.e. Abbassa, Kenya and FaST). This observation may suggest that the GET-EXCEL tilapia has a reduced Abbassa, Kenyan and FaST genetic component, which could be explained by an unequal contribution of parental lines during the establishment of the strain.

Linkage disequilibrium (LD) is the non-random association between the observed frequencies of a particular combination of alleles (Wall and Pritchard 2003). Adequate LD is critical for the implementation of GWAS studies and genomic selection in breeding programmes. Both methods exploit the LD that exists between markers and quantitative trait loci (QTL) or causative mutations (Flint-Garcia *et al*. 2003; Goddard and Hayes 2009). Hence, the magnitude and extent of LD decay between genetic markers can be used to predict the marker density required for QTL mapping. For all the evaluated Nile tilapia populations (GIFT, GIFT-derived and non-GIFT), overall relatively low levels of LD (r^2 ~0.2) were accompanied by a moderate to slow decay with increasing distance. Despite the small number of animals used to assess LD decay (~15 individuals per strain / population), a similar pattern was found to that reported by Yoshida *et al*. (2019b) for GIFT and GIFT-derived commercial populations in South America. The weak correlation found between SNPs is consistent with previous findings in GIFT strains (Xia *et al*. 2015; Yoshida *et al*. 2019b) and is comparatively lower than estimates obtained for other farmed fish

502 species such as Atlantic salmon (Barria *et al*. 2018; Kijas *et al*. 2017). Nevertheless,
503 it is worth noting that despite the relatively low levels of LD, the SNP density of the
504 array is in excess of requirements to obtain maximal genomic prediction accuracy in
505 the context of a typical sibling testing breeding programme in tilapia (Yoshida *et al*.
506 2019a), and indeed for the majority of aquaculture species tested to date (Houston *et*
507 *al*. 2020). Historical factors that affect effective population size (e.g. population
508 bottlenecks, admixture, selective breeding) may influence patterns of LD (Gaut and
509 Long 2003). Contrary to the expectation of domesticated lines showing longer LD
510 than wild populations (Gray *et al.* 2009; McRae *et al*. 2002), the single wild
511 population examined in this study (i.e. Abbassa wild) showed the slowest rate of
512 decrease and the highest LD at longer distances compared to all Nile tilapia strains
513 evaluated. Because it is possible that the sampled Abbassa population is not a good
514 representation of wild individuals (e.g. due to interbreeding with escapees) or LD
515 estimates are being biased by population structure (hypothesis that was not tested in
516 this study), additional wild populations should be evaluated. The general trend
517 observed across strains of overall low levels of r^2 suggests that patterns of LD in Nile
518 tilapia are complex and likely associated with particular features of the process of
519 domestication of this species (Xia *et al*. 2015).

520 At the mean inter-marker spacing on the SNP array (~16 kb), the average r^2 across
521 autosomes was 0.2. According to the simulations performed by Hu and Xu (2008),
522 an r^2 of at least 0.2 is required to achieve a power above 0.8 to detect a QTL for a
523 complex trait of low heritability ($h^2 \sim 0.05$). Although overall LD levels appear to be
524 low in Nile tilapia, our preliminary results suggest that this array provides sufficient
525 genomic resolution to capture association signals in different strains, and will

therefore contribute to expand genetic research in this species and effectively support ongoing and future breeding programmes.

CONCLUSION

A high quality Nile tilapia SNP array was created and validated in several strains. The SNP array was built by prioritising markers that are evenly spaced across gene entities and their local neighbourhood (within < 1 kb), thereby potentially increasing the chance of detecting variants that alter gene expression and / or protein function. The open-access nature of the SNP array together with demonstration of its utility across multiple strains will facilitate its use in genetic research in this species. This may include studies to assess the origin of farmed populations, to track introgression of farmed genomes into the wild, and to understand the genetic architecture of traits of interest. Further, this SNP array will contribute to the management of farmed tilapia populations, and enable accelerated genetic gain and better control inbreeding in breeding programmes via genomic selection.

ACKNOWLEDGEMENTS

This publication was made possible through the support provided by the CGIAR Research Program on Fish Agri-Food Systems (FISH) led by WorldFish. The program is supported by contributors to the CGIAR Trust Fund. CP, DR, AB, PW and RDH are supported by funding from the BBSRC Institute Strategic Programme Grants BB/P013759/1 and BB/P013740/1.

LITERATURE CITED

Aljanabi, S.M., and I. Martinez, 1997 Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic. Acid. Res. 25:4692-4693.

Andrews, S., 2010 FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ansah, Y. B., E. A. Frimpong, and E.M. Hallerman, 2014 Genetically-Improved Tilapia Strains in Africa: Potential Benefits and Negative Impacts. Sustainability 6: 3697-3721.

Ardlie, K. G., L. Kruglyak, and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 3: 299-309.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A .L. Shiver *et al.*, 2008 Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS One 3: e3376.

Bangera, R., K. Correa, J. P. Lhorente, R. Figueroa, and J. M. Yáñez, 2017 Genomic predictions can accelerate selection for resistance against *Piscirickettsia salmonis* in Atlantic salmon (*Salmo salar*). BMC Genomics 18: 121.

Barria, A., M. E. López, G. Yoshida, R. Carvalheiro, J. P. Lhorente *et al.*, 2018 Population Genomic Structure and Genome-Wide Linkage Disequilibrium in Farmed Atlantic Salmon (*Salmo salar L.*) Using Dense SNP Genotypes. Front. Genet. 9: 649.

Bentsen, H. B., B. Gjerde, A. E. Eknath, M. S. P. de Vera, R. R. Velasco *et al.*, 2017 Genetic improvement of farmed tilapias: Response to five generations of

selection for increased body weight at harvest in *Oreochromis niloticus* and the further impact of the project. Aquaculture 468:206-217.

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 30: 2114-2120.

Bolivar, R. B., 1998 *Estimation of response to within-family selection for growth in Nile tilapia (O. niloticus)*. Dalhousie University, Halifax, N.S., Canada.

Chang, C. C., C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell *et al.*, 2015 Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:7-7.

Cingolani, P., A. Platts, L .L. Wang, M. Coon, T. Nguyen *et al.*, 2012 A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly 6: 80-92.

Conte, M. A., W. J. Gammerdinger, K. L. Bartie, D. J. Penman, and T. D. Kocher, 2017 A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. BMC Genomics 18: 341.

Eknath, A., M. Dey, M. Rye, B. Gjerde, T. A. Abella *et al.*, 1998 Selective Breeding of Nile Tilapia for Asia in *6th World Congress on Genetics Applied to Livestock Production*, Armidale, Australia.

Eknath, A. E., and B. O. Acosta, 1998 Genetic Improvement of Farmed Tilapias (GIFT) Project: Final report, March 1988 to December 1997.

Eknath, A. E., and G. Hulata, 2009 Use and exchange of genetic resources of Nile tilapia (*Oreochromis niloticus*). Rev. Aquacult. 1: 197-213.

Eknath, A. E., M. M. Tayamen, M. S. Palada-de Vera, J. C. Danting, R. A. Reyes *et al.*, 1993 Genetic improvement of farmed tilapias: the growth performance of eight strains of *Oreochromis niloticus* tested in different farm environments. Aquaculture 111: 171-188.

FAO, 2018 The State of World Fisheries and Aquaculture 2018 - Meeting the sustainable development goals. Food and Agriculture Organization of the United Nations (FAO), Rome, Italy.

Flint-Garcia, S. A., J. M. Thornsberry, and E. S. T. Buckler, 2003 Structure of linkage disequilibrium in plants. Annu. Rev. Plant. Biol. 54:357-374.

Frichot, E., and O. François, 2015 LEA: An R package for landscape and ecological association studies. Methods. Ecol. Evol. 6: 925-929.

Garrison, E., and G. Marth, 2012 Haplotype-based variant detection from short-read sequencing. arXiv. Available online at https://arxiv.org/abs/1207.3907 (Preprint posted July 17, 2012) .

Gaut, B. S., and A. D. Long, 2003 The lowdown on linkage disequilibrium. Plant cell 15: 1502-1506.

Goddard, M. E., and B. J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10: 381-391.

Gray, M. M., J. M. Granka, C. D. Bustamante, N. B. Sutter, A. R. Boyko *et al.*, 2009 Linkage disequilibrium and demographic history of wild and domestic canids. Genetics 181: 1493-1505.

Gutierrez, A. P., O. Matika, T. P. Bean, and R. D. Houston, 2018 Genomic Selection for Growth Traits in Pacific Oyster (*Crassostrea gigas*): Potential of Low-Density Marker Panels for Breeding Value Prediction. Front. Genet. 9: 391.

621 Gutierrez, A. P., F. Turner, K. Gharbi, R. Talbot, N. R. Lowe *et al.*, 2017
622 Development of a Medium Density Combined-Species SNP Array for Pacific
623 and European Oysters (*Crassostrea gigas and Ostrea edulis*). G3 (Bethesda)
624 7: 2209-2218.
625 Hadley, W., 2009 ggplot2. Springer-Verlag New York.
626 Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations.
627 Theor. Appl. Genet. 38: 5.
628 Houston, R. D., J. B. Taggart, T. Cézard, M. Bekaert, N. R. Lowe *et al.*, 2014
629 Development and validation of a high density SNP genotyping array for
630 Atlantic salmon (*Salmo salar*). BMC Genomics 15: 90.
631 Houston, R. D., T. P. Bean, D. J. Macqueen, M. K. Gundappa, Y. H. Jin *et al*., 2020
632 Harnessing genomics to fast-track genetic improvement in aquaculture. Nat.
633 Rev. Genet. doi: 10.1038/s41576-020-0227-y (Prepint posted April 16, 2020).
634 Hu, Z., and S. Xu, 2008 A simple method for calculating the statistical power for
635 detecting a QTL located in a marker interval. Heredity 101: 48-52.
636 Jansen, M. D., H. T. Dong, and C. V. Mohan, 2019 Tilapia lake virus: a threat to the
637 global tilapia industry? Rev. Aquacult. 11: 725-739.
638 Jorgenson, E., J.S. Witte, 2006 A gene-centric approach to genome-wide
639 association studies. Nat. Rev. Genet. 7: 885-891.
640 Joshi, R., M. Árnyasi, S. Lien, H. M. Gjøen, A. T. Alvarez *et al.*, 2018 Development
641 and Validation of 58K SNP-Array and High-Density Linkage Map in Nile
642 Tilapia (*O. niloticus*). Front. Genet. 9: 472.
643 Joshi, R., A. Skaarud, M. de Vera, A.T. Alvarez, and J. Ødegård, 2019 Genomic
644 prediction for commercial traits using univariate and multivariate approaches
645 in Nile tilapia (*Oreochromis niloticus*). Aquaculture 516: 734641.

Kijas, J., N. Elliot, P. Kube, B. Evans, N. Botwright *et al.*, 2017 Diversity and linkage disequilibrium in farmed Tasmanian Atlantic salmon. Anim. Genet. 48: 237-241.

Komen, J., and T. Trọng, 2014 Nile tilapia genetic improvement: achievements and future directions in *The 10th International Symposium on Tilapia in Aquaculture (ISTA10)*, Jerusalem, Israel.

Kumar, G., J. Langa, I. Montes, D. Conklin, M. Kocour *et al.*, 2019 A novel transcriptome-derived SNPs array for tench (*Tinca tinca* L.). PLoS One 14: e0213992.

Lapegue, S., E. Harrang, S. Heurtebise, E. Flahauw, C. Donnadieu *et al.*, 2014 Development of SNP-genotyping arrays in two shellfish species. Mol. Ecol. Resour. 14: 820-830.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics (Oxford, England) 25: 1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25 :2078-2079.

Lind, C. E., A. Kilian, and J. A. H. Benzie, 2017 Development of Diversity Arrays Technology markers as a tool for rapid genomic assessment in Nile tilapia, *Oreochromis niloticus*. Anim. Genet. 48: 362-364.

Liu, S., L. Sun, Y. Li, F. Sun, Y. Jiang *et al.*, 2014 Development of the catfish 250K SNP array for genome-wide association studies. BMC Res. Notes. 7: 135.

Mangin, B., A. Siberchicot, S. Nicolas, A. Doligez, P. This *et al.*, 2012 Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 108: 285-291.

McRae, A .F., J. C. McEwan, K. G. Dodds, T. Wilson, A. M. Crawford *et al.*, 2002 Linkage disequilibrium in domestic sheep. Genetics 160: 1113-1122.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Montanari, S., L. Bianco, B. J. Allen, P. J. Martínez-García, N. V. Bassil *et al.*, 2019 Development of a highly efficient Axiom™ 70 K SNP array for Pyrus and evaluation for high-density mapping and germplasm characterization. BMC Genomics 20: 331.

Neira, R., 2010 Breeding in Aquaculture Species: Genetic Improvement Programs in Developing Countries in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Germany.

Ng, W. K., and N. Romano, 2013 A review of the nutrition and feeding management of farmed tilapia throughout the culture cycle. Rev. Aquacult. 5: 220-254.

Nugent, C. M., J. S. Leong, K. A. Christensen, E. B. Rondeau, M. K. Brachmann *et al.*, 2019 Design and characterization of an 87k SNP genotyping array for Arctic charr (*Salvelinus alpinus*). PLoS One 14: e0215008.

Palaiokostas, C., M. Bekaert, M. G. Q. Khan, J. B. Taggart, K. Gharbi *et al.*, 2013 Mapping and validation of the major sex-determining region in Nile tilapia (*Oreochromis niloticus* L.) Using RAD sequencing. PLoS One 8: e68389-e68389.

Palaiokostas, C., M. Bekaert, M. G. Q. Khan, J. B. Taggart, K. Gharbi *et al.*, 2015 A novel sex-determining QTL in Nile tilapia (*Oreochromis niloticus*). BMC Genomics 16: 171-171.

694 Palti, Y., G. Gao, S. Liu, M. P. Kent, S. Lien *et al.*, 2015 The development and
695 characterization of a 57K single nucleotide polymorphism array for rainbow
696 trout. Mol. Ecol. Resour. 15: 662-672.
697 Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and
698 eigenanalysis. PLoS Genet. 2: e190.
699 Qi, H., K. Song, C. Li, W. Wang, B. Li *et al.*, 2017 Construction and evaluation of a
700 high-density SNP array for the Pacific oyster (*Crassostrea gigas*). PLoS One
701 12: e0174007.
702 R Core Team, 2014 R: A Language and Environment for Statistical Computing. R
703 Foundation for Statistical Computing, Vienna, Austria. Available online at
704 http://www.R-project.org/.
705 Robledo, D., C. Palaiokostas, L. Bargelloni, P. Martínez, and R. Houston, 2018a
706 Applications of genotyping by sequencing in aquaculture breeding and
707 genetics. Rev. Aquacult. 10: 670-682.
708 Robledo, D., O. Matika, A. Hamilton, and R. D. Houston, 2018b Genome-Wide
709 Association and Genomic Selection for Resistance to Amoebic Gill Disease in
710 Atlantic Salmon. G3 (Bethesda) 8: 1195-1203.
711 Sonesson, A. K., and T. H. Meuwissen, 2009 Testing strategies for genomic
712 selection in aquaculture breeding programs. Genet. Sel. Evol. 41:37.
713 Tayamen, M., 2004 Nationwide dissemination of GET-EXCEL tilapia in the
714 Philippines in *Proceedings of the Sixth International Symposium On Tilapia In*
715 *Aquaculture*, Manila, Philippines.
716 Tayamen, M., T. Abella, R. Reyes, J. Danting, A. Mendoza *et al.*, 2004 Development
717 of tilapia for saline waters in the Philippines in *Proceedings of the Sixth*
718 *International Symposium On Tilapia In Aquaculture*, Manila, Philippines.

719 Thermo Fisher Scientific Inc, 2018 AxiomTMAnalysis Suite (AxAS) v4.0 USER
720 GUIDE.
721 Tracy, C. A., and H. Widom, 1994 Level-spacing distributions and the Airy kernel.
722 Comm. Math. Phys. 159: 151-174.
723 Trọng, T. Q., H. A. Mulder, J. A. M. van Arendonk, and H. Komen, 2013 Heritability
724 and genotype by environment interaction estimates for harvest weight, growth
725 rate, and shape of Nile tilapia (*Oreochromis niloticus*) grown in river cage and
726 VAC in Vietnam. Aquaculture 384-387: 119-127.
727 Tsai, H. Y., A. Hamilton, A. E. Tinch, D. R. Guy, K. Gharbi *et al.*, 2015 Genome wide
728 association and genomic prediction for growth traits in juvenile farmed Atlantic
729 salmon using a high density SNP array. BMC Genomics 16: 969.
730 Tsai, H. Y., A. Hamilton, A. E. Tinch, D. R. Guy, J. E. Bron *et al.*, 2016 Genomic
731 prediction of host resistance to sea lice in farmed Atlantic salmon populations.
732 Genet. Sel. Evol. 48: 47.
733 Van Bers, N. E. M., R. P. M. A. Crooijmans, M. A. M. Groenen, B. W. Dibbits, and J.
734 Komen, 2012 SNP marker detection and genotyping in tilapia. Mol. Ecol.
735 Resour. 12: 932-941.
736 Wall, J. D., and J. K. Pritchard, 2003 Haplotype blocks and linkage disequilibrium in
737 the human genome. Nat. Rev. Genet. 4: 587-597.
738 Xia, J. H., Z. Bai, Z. Meng, Y. Zhang, L. Wang *et al.*, 2015 Signatures of selection in
739 tilapia revealed by whole genome resequencing. Sci. Rep. 5: 14168-14168.
740 Xu, J., Z. Zhao, X. Zhang, X. Zheng, J. Li *et al.*, 2014 Development and evaluation of
741 the first high-throughput SNP array for common carp (*Cyprinus carpio*). BMC
742 Genomics 15: 307.

Yáñez, J.M., S. Naswa, M.E. Lopez, L. Bassini, K. Correa *et al.*, 2016 Genomewide single nucleotide polymorphism discovery in Atlantic salmon (*Salmo salar*): validation in wild and farmed American and European populations. Mol. Ecol. Resour. 16: 1002-1011.

Yáñez, J.M., G. Yoshida, A. Barria, R. Palma-Véjares, D. Travisany *et al.*, 2020 High-Throughput Single Nucleotide Polymorphism (SNP) Discovery and Validation Through Whole-Genome Resequencing in Nile Tilapia (*Oreochromis niloticus*). Mar. Biotechnol. 22: 109-117.

Yoshida, G. M., J. P. Lhorente, K. Correa, J. Soto, D. Salas *et al.*, 2019a Genome-Wide Association Study and Cost-Efficient Genomic Predictions for Growth and Fillet Yield in Nile Tilapia (*Oreochromis niloticus*). G3 (Bethesda) 9: 2597-2607.

Yoshida, G. M., A. Barria, K. Correa, G. Cáceres, A. Jedlicki *et al.*, 2019b Genome-Wide Patterns of Population Structure and Linkage Disequilibrium in Farmed Nile Tilapia (*Oreochromis niloticus*). Front. Genet. 10: 745.

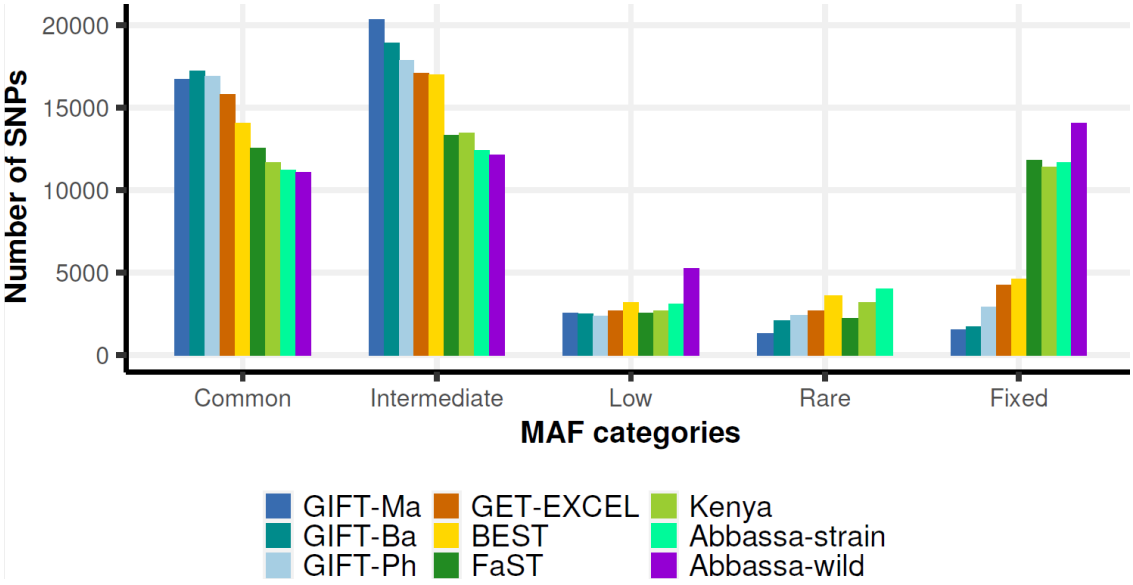
Zeng, Q., Q. Fu, Y. Li, G. Waldbieser, B. Bosworth *et al.*, 2017 Development of a 690 K SNP array in catfish and its application for genetic mapping and validation of the reference genome sequence. Sci. Rep. 7: 40347.

Zenger, K. R., M. S. Khatkar, D. B. Jones, N. Khalilisamani, D.R. Jerry *et al.*, 2019 Genomic Selection in Aquaculture: Application, Limitations and Opportunities With Special Reference to Marine Shrimp and Pearl Oysters. Front. Genet. 9: 693.

Table 1. Origin and observed (Ho) and expected (He) heterozygositites for the Nile tilapia populations used for the validation of the SNP array

Population ID	Genetic background	Type	Origin	Number of samples passing QC filters	He	Ho	95% CI (Ho)
GIFT-Ma[a]	GIFT	Domesticated	Malaysia	15	0.337	0.350	0.348-0.352
GIFT-Ba	GIFT	Domesticated	Bangladesh	15	0.334	0.347	0.346-0.349
GIFT-Ph	GIFT	Domesticated	Philippines	15	0.322	0.328	0.327-0.330
GET-EXCEL	GIFT-derived	Domesticated	Philippines	15	0.304	0.325	0.323-0.327
BEST	GIFT-derived	Domesticated	Philippines	14	0.294	0.317	0.316-0.320
FaST	Non-GIFT	Domesticated	Philippines	15	0.243	0.252	0.250-0.254
Kenyan	Non-GIFT	Domesticated	Kenya	15	0.236	0.209	0.207-0.211
Abbassa strain	Non-GIFT	Domesticated	Egypt	13	0.229	0.239	0.237-0.241
Abbassa Wild	Non-GIFT	Wild	Egypt	8	0.220	0.258	0.259-0.264

[a]discovery population



773

774 **Figure 1. MAF categories of SNPs from the ~65K SNP-chip across nine different Nile tilapia**

775 **strains / populations.**

776

777

778

779

780

781

782

783

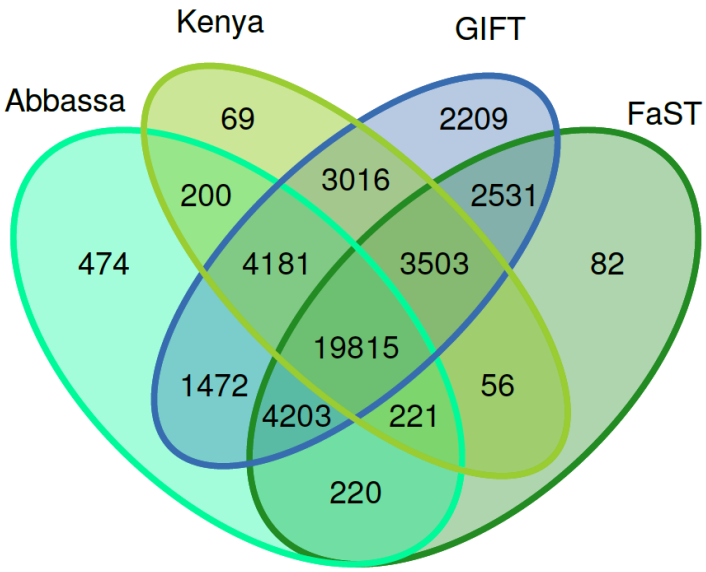
784

785

786

787

788



789

790 **Figure 2. Number of informative SNPs (MAF>0) shared among the four distinct strains**

791 **evaluated in this study: Abbassa, Kenya, GIFT and FaST.**

792

793

794

795

796

797

798

799

800

801

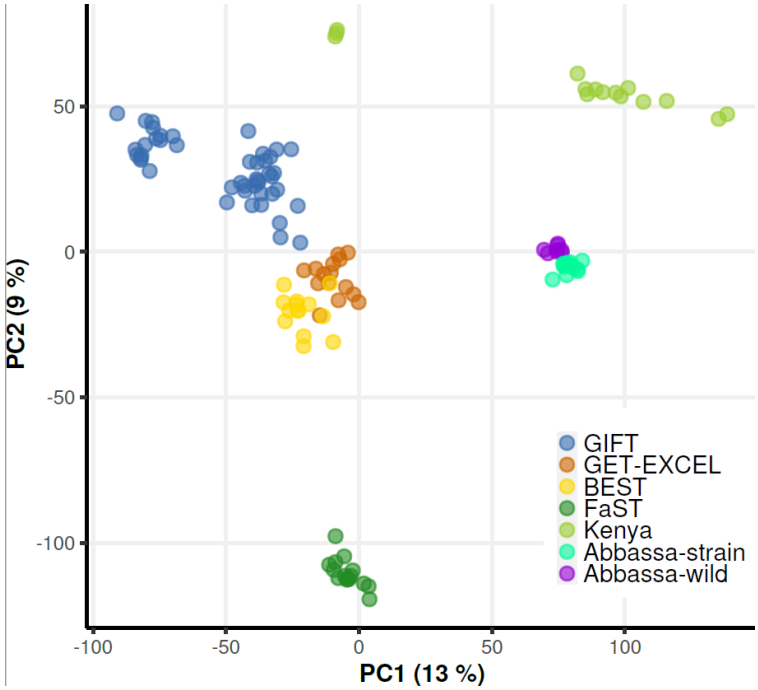
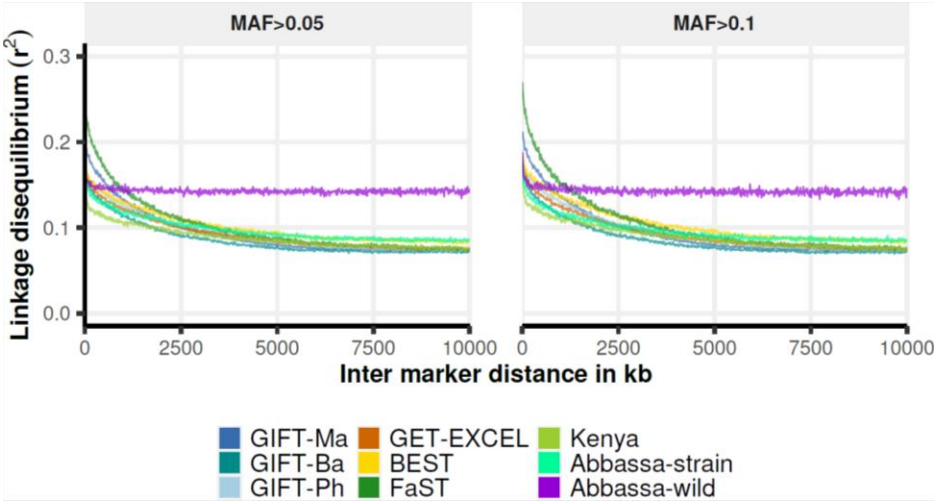


Figure 3. PCA representing the structure of nine different strains / populations used for the validation of the ~65k SNP array. The total number of individuals (dots) is 125. Each dot is colour coded according its origin, as shown in the legend at the bottom right corner of the plot.



817

818 **Figure 4. Linkage disequilibrium decay (r^2) over distance (in kb) among different Nile tilapia**

819 **strains / populations genotyped with the ~65K SNP array**. LD decay after applying a MAF

820 threshold of 0.05 (left panel) and 0.1 (right panel).

821