Conservation Genetics
https://doi.org/10.1007/s10592-018-1136-x

RESEARCH ARTICLE

Application of microsatellite genotyping by sequencing (SSR-GBS) to measure genetic diversity of the East African *Oreochromis niloticus*

Papius Dias Tibihika[1,3] · Manuel Curto[1] · Eva Dornstauder-Schrammel[1] · Silvia Winter[1,5] · Esayas Alemayehu[2,4] · Herwig Waidbacher[2] · Harald Meimberg[1]

Received: 19 July 2018 / Accepted: 10 December 2018

Abstract

Microsatellites play an important role when investigating population and ecological genetics, although high effort in development and genotyping constitute a technical constraint and remains a major bottleneck. Here we use a microsatellite genotyping approach utilizing sequences of amplicons for allele calling (SSR-GBS) based on Illumina that requires less effort and time. The approach consist of development of highly polymorphic loci, sequencing of multiplexed PCR amplified microsatellites on an Illumina Miseq PE 300 platform and bioinformatic treatment of the sequenced data using custom scripts. The procedure allows automation in allele calling, which can be more reliably replicated and thereby removes biases that might prevent concatenation of datasets from different analyses. Additionally, the methodology enhances information content in the sequenced data beyond the traditional amplicon length (AL) approaches. Using 26 newly developed microsatellite markers and SSR-GBS we investigate the population genetic assessment of anthropogenically altered populations of East African Nile tilapia to show the potential of this genotyping approach. More precisely, we compare genotypic data generated considering AL and whole amplicon information (WAI). We found that genotypes based on WAI are not only able to recover a higher number of alleles but also a more detailed genetic structure pattern. We discuss the capability and importance of WAI allele calling and show perspectives for implementation in the future conservation genetic studies. More specifically, we demonstrate how the current markers and techniques might contribute useful information for studies concerning resources sustainable exploitation and conservation using the East African Nile tilapia.

Keywords Nile tilapia · Microsatellite markers · Amplicon sequencing · Amplicon length · Genotyping pipeline · Next generation sequencing

Introduction

Microsatellites/simple sequence repeats (SSR) or short tandem repeats (STRs) have proved reliable as ideal genetic markers in population genetic studies (Guichoux et al. 2011; Selkoe and Toonen 2006; Schlotterer et al. 1991). This is due to their important attributes including, but not limited

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10592-018-1136-x) contains supplementary material, which is available to authorized users.

✉ Papius Dias Tibihika
papiust@yahoo.com; papius.tibihika@students.boku.ac.at

[1] Institute for Integrative Nature Conservation Research, University of Natural Resources and Life Sciences, Vienna (BOKU), Gregor Mendel-Straße 33, 1180 Vienna, Austria

[2] Institute for Hydrobiology and Aquatic Ecosystems Management, University of Natural Resources and Life Sciences, Vienna (BOKU), Gregor-Mendel-Straße 33/DG, 1180 Vienna, Austria

[3] Kachwekano Zonal Agricultural Research and Development Institute, National Agricultural Research Organization, P. O. Box 421, Kabale, Uganda

[4] National Fishery and Aquatic Life Research Centre, P.O. Box 64, Sebeta, Ethiopia

[5] Division of Plant Protection, University of Natural Resources and Life Sciences, Vienna (BOKU), Gregor Mendel-Straße 33, 1180 Vienna, Austria

Published online: 17 December 2018

to, their ubiquity in the genome, co-dominant inheritance, high polymorphism and reproducibility (Sundaray et al. 2016; Selkoe and Toonen 2006). Although there is a growing set of other markers and sequencing techniques, SSRs are still one of the leading genotyping options based on their methodological simplicity. Additionally, the use of these markers is further rendered reliable following the genesis of next-generation sequencing (NGS) technologies and the identification of numerous loci in non-model species accompanied by reduced costs (Guichoux et al. 2011; Curto et al. 2013). Nevertheless, many of the previous genetic studies on Nile Tilapia, have utilised finger print methods (Mwanja et al. 1996, 2008; Agnèse et al. 1997; Seyoum and Kornfield 1992). These methods have limited information content (e.g. Liu and Cordes 2004; Miah et al. 2013; Okumuş and Çiftci 2003; Guichoux et al. 2011). Traditionally developed SSR had been employed, but frequently with low number of loci investigated and the resulting low amount of statistical power, (e.g. Bezault et al. 2012; Lee and Kocher 1996; Hassanien and Gilbey 2005; Tariq Ezaz et al. 2004).

Normally, SSRs are genotyped by using the fragment length variation, through capillary electrophoresis. In most cases, due to the high mutation rates of the repetition motifs, this is sufficient to generate adequate amounts of variation at the intraspecific level. However, considering the amplicon length (AL) variation alone might miss out other useful polymorphic information. By genotyping microsatellite markers via high throughput sequencing methods such as Illumina, it is possible to compare AL and whole amplicon information (WAI) allele calling procedures, and ascertain variations in the amount of genetic information generated (De Barba et al. 2017). Other advantages of using high throughput sequencing for microsatellite genotyping include, the automation and replicability with low cost of the genotyping process (De Barba et al. 2017; Vartia et al. 2016; Zhan et al. 2017). Using available bioinformatic tools, one can either automatize or at least significantly decrease the level of artefacts in allele calling and thereby elevating information content of the marker. This p facilitates the reproducibility of the genotyping process, in comparison to the traditional electropherogram length frequency (AL) procedures (Farrell et al. 2016; De Barba et al. 2017). In the current study, we expect this method (WAI allele calling), which we refer to as SSR-GBS (Simple Sequence Repeats-Genotyping By-Sequencing), to become more informative and subsequently widely adopted as an important tool for genotyping. We test this assumption using the East Africa Nile tilapia as a case study based on their historical dynamics (Balirwa 1992; Kaufman 1992; Ogutu-Ohwayo 1990).

The East African Nile tilapia (*Oreochromis niloticus*, L. 1758) is an important species for augmenting aquaculture and capture fisheries. The cichlid is native to lakes Albert, George, Edward, Turkana, and Baringo among others, as well as the lower Nile River (Ogutu-Ohwayo 1990). However, human activities, might have had an effect on the native populations through the introduction of potentially maladaptive traits through translocations and promoting bottlenecks by overfishing. Additionally, the activities like transplantations of Nile tilapia into numerous non-native habitats including lakes Victoria, Kyoga, Nabugabo, many satellite water bodies and other systems, like aquacultural set-ups (Welcomme 1966; Balirwa 1992) might have altered the species' genetic structure through admixture promoted by hybridization with closely related indigenous species. Broadly, the contrast between native and non-native Nile tilapia stocks in the context of anthropogenic influence remains under-documented (Mwanja 2000). Although Nile tilapia fishery is vital for the livelihoods in East Africa, management of this important resource against detrimental anthropogenic activities is unfortunately inadequate due to insufficient research (Njiru et al. 2007; Ogutu-Ohwayo 1990). Therefore, the use of more informative microsatellite datasets, might contribute to efficiently characterise the populations (Hedrick 2001; Oliveira et al. 2006).

Here, we investigate the potential of the SSR-GBS approach by describing the genetic variability and structure of East African Nile tilapia. This study is a proof of concept testing the ability of SSR-GBS in producing genotype data with a higher statistical power than the traditional SSR amplicon length genotyping due to its potential in recovering SNPs associated with SSR motif variation. This provides a higher number of alleles and reduces the length homoplasy effect that is characteristic to this marker system. We demonstrate this by comparing the importance and strengths of the two allele calling methods (WAI and the AL, similar to the traditional electropherogram methods). As a side product of this work, we develop a set of SSR markers for Nile tilapia and test their applicability on other tilapiine species.

Materials and methods

Sample collection

The East Africa (Uganda) Nile tilapia samples were collected monthly overnight between July and December 2016, as part of commercial catches of local fishermen using the East Africa member states recommended 127 mm of gill nets. The sampling sites were defined upon advice from law enforcement fisheries officers, together with local fishermen and sampling sites recorded using a Global Positioning System (GPS). On landing, muscle tissue samples were taken immediately, preserved in absolute ethanol and later stored under − 20 °C, except during transportation. The species are not protected, but rather fished commercially, and sampling was performed by members of a government

institution (National Agricultural Research Organization of Uganda), hence no other permission was required. Apart from one non-native population (Lake Victoria), the other three populations (Lakes Albert and George and River Nile) are native. In total, we collected 107 Ugandan individuals (Table 1). All animal rights were observed during the field excursions. Fish was obtained as part of commercial fishing operations by local fishermen and killed in this process.

For SSR development, we used a single Ethiopian Nile tilapia sample (Lake Ziway), which was accessed through a selective breeding project at the Institute for Integrative Nature Conservation Research-BOKU, Vienna. From this sample, fresh fin clips were taken, implying that there was no fish sacrifice. The Ethiopian sample was later used for low coverage total DNA shot gun sequencing for SSR marker discovery prior to East African sampling. We also sampled Ethiopian *Tilapia zillii* for cross-species amplification tests.

DNA extraction

A piece of ethanol-preserved tissue muscle (approximately 0.1 g) was digested overnight in 500 µl lysis buffer (2% SDS, 2%PVP 40, 250 m MNaCl, 200 M Tris–HCl, 5 mM EDTA, pH8) containing 200 ng of proteinase K enzyme. Genomic DNA extraction was carried out using magnetic beads (MagSi-DNA beadsMagnaMedics) and a magnetic separator SL-MagSep96 (Steinbrenner, Germany) following a modified MagSi-DNA Vegetal kit protocol. DNA binding was carried out by mixing 17 µl of beads with 500 µl each of binding buffer and clear lysate in a 2 ml 96 well plate. Bound DNA beads were then washed twice in 80% 600 µl cold ethanol. Later, DNA was eluted twice by adding 50 µl (first elution) and 70 µl (second elution) of elution buffer (65 °C 10 mM Tris–HCl, pH 8.3), following the above-mentioned kit protocol. The quality of the extracted DNA was inspected on 0.8% agarose gel.

SSR discovery

High quality extracted genomic DNA from the Ethiopian sample was sent for library preparation on the Illumina MiSeq paired-end (PE) 300 sequencing platform (San Diego, USA) as described in Shendure and Ji (2008); Castoe et al. (2012). Both library preparation and sequencing were done at the Genomics Service Unit, Ludwig-Maximilian's-Universität München, Germany. Sequences generated by Illumina Miseq were quality checked using FASTQC (Andrews and FastQC 2010) and trimmed for the removal of adapter sequences and low quality regions (Phred < 20) using CUTADAPT vers. 0.11.1 (Martin 2011). Forward and reverse reads were merged using PEAR version 0.9.4 (Zhang et al. 2014) considering only minimum overlaps of 15 bp with a p-value below 0.01 for the highest observed expected alignment scores. Later, the sequences were screened for microsatellite motifs (from di to penta nucleotide repeats) using the SSR_pipeline program (Miller et al. 2013). Here, we considered sequences with at least 10 repeats for 2mers, eight for 3mers, and six for 4/5mers. A total of 6,724 SSR motif reads were revealed comprising 4,629 2mers, 818 3mers, 868 4mers and 409 5mers. For subsequent primer design, sequences of equal or greater than 350 bp long and microsatellites with flanking regions longer than 30 bp were considered. This length was chosen to facilitate detection and elimination of primer dimer and other low molecular weight artefacts from the specific amplification products using the washing method described below. Considering the inclusion of 60 bp oligonucleotides in the PCR as a multiplex assay, such artefacts are difficult to suppress. Artefacts can be unequivocally detected using gel electrophoresis and discarded using magnetic beads. In addition, longer sequence information is higher chances of recovering extra information on the flanking sites that may contribute to the increase on the number of alleles. Raw reads were submitted to Sequence Read Archive (SRA) database with the reference number SRX3398501.

Primer design for amplicon sequencing

Specific PCR primers were designed in Geneious software version 10.3 (Kearse et al. 2012) using the default Primer3 program (Untergasser et al. 2012). Manual primer3 adjustments were set at 55 °C for optimal primer melting temperature, with a GC content in the range of 20–50–80, optimal oligo length between 18–20–23 bp, and amplification product size between 350 and 450 bp. We designed primers in a way that the complete primer motif would be included in the first or last 300 bp of the amplicon being able to be covered completely by one of MiSeq's paired reads. This prevents

Table 1 Nile tilapia sample sources from East African freshwater bodies in Uganda

Population	Location	Habitat type	No.	Latitude	Longitude	Elevation
Albert	Ntoroko	Native	23	01.05206N	030.53464E	618
George	Hamukungu	Native	35	00.01739S	030.08698E	916
River Nile	Kibuye	Native	24	01.18734N	032.96865E	1062
Victoria	Kamuwunga	Non-native	25	00.12747N	031.93999E	1139

Elevation = meters above the sea level and No. = number of samples

that the repetitive unit is part of the overlap of the paired reads which would lead to difficulties with the merging step in the bioinformatics pipeline. Here, 48 primer pairs were designed. We decided to develop primers from own shot gun sequences instead of published Nile tilapia genomic information to increase the likelihood that markers fit on East African populations. However, the availability of a Nile tilapia genome from the GenBank accession number MKQE01000000, allowed us to screen the entire sequences containing microsatellites used for primer design for potentially duplicated markers, using BLASTn (Basic Local Alignment Search Tool for nucleotides). BLASTn outputs provide a list of pairwise alignment matches and sequence hits above which a statistical threshold is displayed (Xiong 2006). In the current study, BLASTn-aligned sequences were selected if the E-values were zero. The E-value is comparable to the probability value p, as the least value suggests a lower likelihood that the database matches are a result of random chance, but rather the database matches display a significant similarity (Xiong 2006). Only primers that originated from sequences showing single matches on the genome were considered because they were more likely to represent single copy regions. Of the 48 primer sequences initially mapped on the genome, 13 primers were found more than once in the genome and subsequently discarded, thus leaving 35 primer pairs. Although higher numbers of microsatellite markers can provide robust population genetics results (Capote et al. 2012; Ryman et al. 2006), generally a number of microsatellite loci in the range of 8–20 are considered adequately informative to determine genetic structure between populations (Arthofer et al. 2018; Vartia et al. 2014; Koskinen et al. 2004). In the current study, the initial 48 primer pairs resulted in a number sufficient to test genetic structure patterns in east African Nile Tilapia. Nevertheless, more markers can be easily added with the procedure and resources presented here.

For Illumina sequencing, primers were extended by part of the Illumina adapters P5 (TCTTTCCCTACACGACGC TCTTCCGATCT) and P7 (CTGGAGTTCAGACGTGTG CTCTTCCGATCT) at the 5′ end of the primer forward and reverse, respectively (Fig. 1). These correspond to the Illumina sequencing primers and served as a linker for the second PCR where the remaining parts of the adapters are added. This procedure was conducted using primers containing all the components necessary for Illumina sequencing. In this second (index) PCR, for each sample, we used a novel combination of two different indexes of 8 bp, P5-(*AAT GATACGGCGACCACCGAGATCTACAC[Index]ACACTC TTTCCCTACACGACG)* and P7-(*CAAGCAGAAGACGGC ATACGAGAT[Index]GTGACTGGAGTTCAGACGTGT*). This was vital for allowing the pooling of a large sample size in the down-stream analysis. After the second PCR,

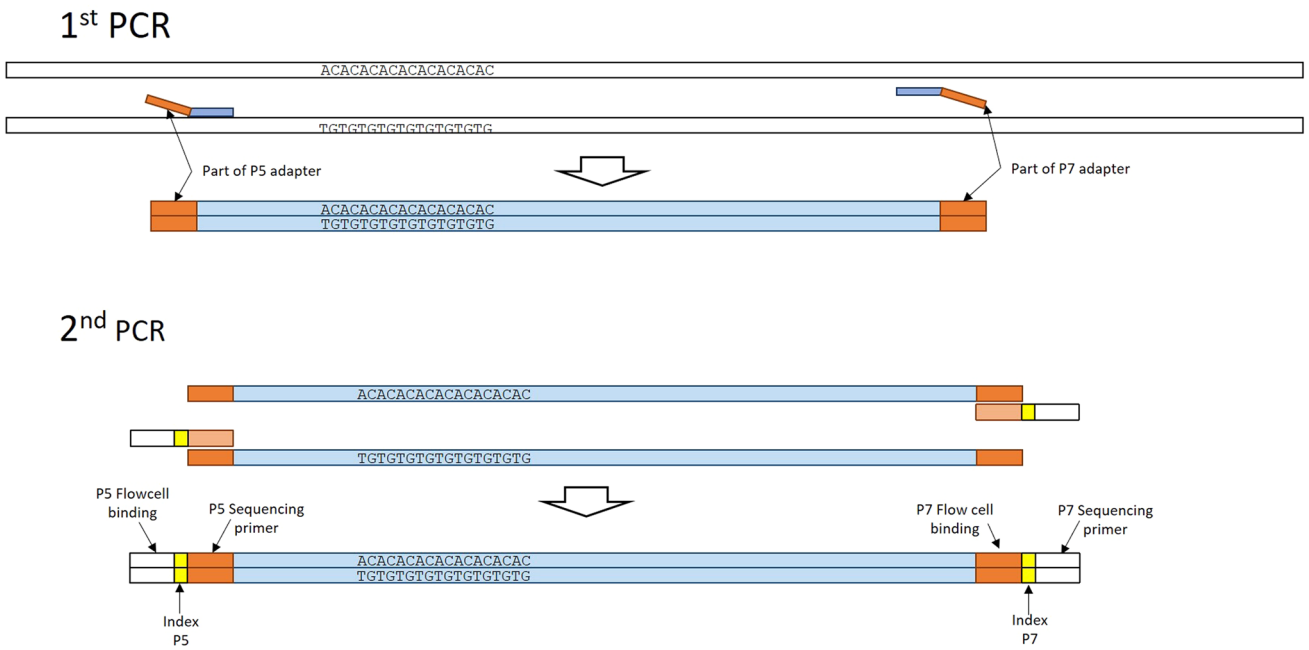

Fig. 1 Amplicon library preparation approach. We used two-step PCR procedures: In the first PCR, the microsatellite was amplified with locus specific primers (blue) and extended with part of the Illumina adapter (orange). This motif serves as linker for the second PCR where the remaining parts of the Illumina adapter were included. The primers in the second PCR from 5′ to 3′ are constituted by: a region complementary to the linker corresponding to the Illumina sequencing primer (orange), Index information used for sample identification after pooling (yellow), and Illumina flow cell binding motif (white). (Color figure online)

the resulting amplicons had following parts from 5′ to 3′ (Fig. 1): (1) P5 motif for flow cell hybridization, (2) index 1 of 8 bp, (3) P5one sequencing primer, (4) specific forward primer, (5) target DNA for sequencing; specific reverse primer; (6) P7 sequencing primer, (7) index 2 of 8 bp, and (8) P7 motif for flow cell hybridization.

SSR primer testing

To ascertain the applicability and usability of the developed primers, we first amplified them in singleplex reactions to tested two scenarios: (1) transferability of the developed primers on East African Nile tilapia, and (2) cross-species amplification of *T. zillii*. The amplification success rate of the candidate loci during PCR reactions was tested by assaying a Nile tilapia sample from Uganda. For cross-species amplification, only two genomic DNA samples for *T. zillii* were tested on a panel of 35 SSR loci. PCR reactions were conducted in a 10 µl total volume. All primer pairs were tested using the QIAGEN Multiplex PCR Master Mix (Qiagen, CA, U.S.A) kit. PCR reaction volume during Nile tilapia amplification was composed of 5 µl Master mix, 4 µl primer mix and 1 µl genomic DNA. Primer mix was a combination of 1 µl Reverse primer + 1 µl Forward primer (100 µM each), plus 98 µl of water. Finally, the cycler reaction mixtures were performed based on the following PCR profiles: initialisation at 95 °C for 15 min, followed by 35 cycles of denaturation at 95 °C for 30 s, annealing at 55 °C for 60 s, elongation at 72 °C for 60 s and final extension at 72 °C for 10 min. The success of the PCR products was tested by electrophoresis on 1.8% agarose gel. Here, 33 primer pairs were successfully identified in specific PCR products, which subsequently were used for the multiplex PCR approach on Ugandan Nile tilapia populations. Successful markers for cross-species amplification were based on the PCR singleplex gel products based on two replications.

PCR multiplex and Illumina sequencing

All 33 gel-screened primers were combined in a single multiplex reaction. PCR reactions were carried out in a 10 µl total volume containing 5 µl Master mix, 2.5 µl water, 0.5 µl primer mix (1 µM each) and 2 µl genomic DNA. Thermal cycler profiles were analogous to the single-plex PCR. The resulting PCR products were purified using magnetic bead procedures, following slight modifications from AgencourtAMPure XP PCR Purification protocol. Here, we mixed 4 µl PCR products with 2.86 µl of AMPure XP beads (Beckman Coulter, Inc, Bree, CA) and incubated for five minutes at room temperature. Bound DNA beads were captured by an inverted magnetic bead extraction device, VP 407-AM-N (V&P SCIENTIFIC, INC) and washed twice in 200 µl of 80% ethanol for 45 s. Later, the beads were dried at room temperature for five minutes and eluted in 17 µl of elution buffer (65°C10 mM Tris–Hcl, pH 8.3).

The second (index) PCR was performed in a total reaction volume of 10 µl, containing 5 µl master mix, 2 µl each of index primer (1 µM) P5 and P7, and 1 µl of template purified PCR products. PCR was run with the following thermocycler conditions: 95 °C for 15 min, followed by 10 cycles of denaturation at 95 °C for 30 s, annealing at 58 °C for 60 s, elongation at 72 °C for 60 s and final extension at 72 °C for five minutes. Finally, all the PCR products were pooled and sent for PE 300 bp sequencing in an Illumina MiSeq at the Genomics Service Unit at Ludwig Maximillian Universität, München, Germany. The samples used in this work occupied 11% of the MiSeq run.

Sequence analysis and SSR-GBS genotyping

Reads from Illumina were quality controlled and merged as described in "SSR discovery". Overlap was only possible because SSR motifs were covered completely by one of the paired reads. The resulting sequences should start with the forward primer and end with the reverse and we used this criterion to de-multiplex the sequences according to primer content, creating one fastq files per sample and locus using script 1 (Supplementary material Table S4). This script looks for mismatches between the amplification primers from the beginning (forward) and the end (reverse) of the sequences. Only reads with a mismatch to both primers below two base pairs were considered. From this step the allele calling was performed in two steps: first using the AL, which resembles the traditional SSR genotyping, and then by considering possible SNP variation within alleles of the same length recovering the whole amplicon sequence information (WAI). After each step, a codominant matrix was produced allowing for a comparison between data that would be produced by tradition SSR genotyping to the one from SSR-GBS. Allele calling based on AL incorporates length variation at the repetition motif plus possible indels in the flanking regions. All types of length variation and SNPs are used as information with WAI because two amplicon sequences were only considered as the same alleles if they were equal.

To call alleles using AL, we calculated the length distribution per sample and per locus using script 2 producing files containing the number of times each length occurs per marker and sample (Supplementary material Table S4). Then we used the script 3 to automatically call the alleles and plot histograms based on sequence length resembling the chromatograms obtained in traditional SSR genotyping (Supplementary material Table S4). Only genotypes with a minimum depth of 10 reads were considered. Automatic allele calling considered homozygous genotypes if there was a length with a frequency equal or above to 90% of the total

number of reads. Heterozygous genotypes were assumed if the frequency of two lengths was above 90% of the reads and if the frequency of both lengths did not differ by more than 20%. In case these criteria were not matched, the genotypes were marked for manual control. Nevertheless, all possible genotypes were manually controlled with the aid of the produced histograms.

For WAI allele calling, the sequences with the same length of the alleles defined in the AL step were extracted and used to produce a 70% consensus sequence per length class. The extraction of sequences per length allele was done using the script 4 and the consensus sequence the script 5. In the 70% consensus, the positions with the most common nucleotide of frequency below 70% were coded with the ambiguous base "N" and considered as potential heterozygous SNPs. These sequences were divided into two files based on the two most frequent nucleotides for that position using the script 6. In case more than one potential SNP was found, these positions were considered as linked and the two most frequent nucleotide combinations were recovered. We observed, that chimeric sequences could occur between alleles that differed by more than one SNP. This causes the occurrence of sequence states intermediate between the alleles. In each case these intermediate states were less frequent and could be unambiguously resolved visually, either by comparing them to other alleles in the sampling, or by including the sequence length as additional information. However, these occurrences were rare. Only in a few cases which were not considered due to low overall read counts more than two similarly frequent nucleotide combinations were found. Similarly, the two most frequent combinations between a SNP and length signal was called as allele in samples that were heterozygous with AL. WAI allele calling was finally done using scrip 7 where a number was attributed to each unique sequence (allele) and saving this information in a codominant matrix. All scripts are available in GitHub (https://github.com/mcurto/SSR-GBS-pipeline) and detailed description of script 1–script 7 is given in supplementary table S4. The sequence analyses resulted in 26 loci with genotypes for most of the samples that were used for further statistical assessment. Raw reads can be found in the SRA database with the references SRX3398667 to SRX3398776.

Statistical analyses

Descriptive population genetics analyses for SSR loci were determined using various programs. The software Micro-Checker version 2.2.0.3 (Van Oosterhout et al. 2004), was used to estimate the presence of null alleles, evidence of allele drop-out, or stuttering during PCR amplification. Test for deviations from Hardy–Weinberg Equilibrium (HWE) and calculations of the fixation index (Fis) were performed in GenePop version 4.6.9 (Rousset 2008). Markov chain parameters for all tests in GenePop were run at 10,000 dememorizations, 100 batches and 5000 iterations per batch and Fis values were recorded (Weir and Cockerham 1984). Fis was specifically determined to assess the type of HWE on the populations, in aspects of excess or deficiency heterozygosity (Dorak 2014). Here, positive or negative Fis values indicate excess homozygosity or excess heterozygosity (outbreeding) respectively (Dorak 2014). Observed heterozygosity (Ho), expected heterozygosity (He) and loci polymorphic information content (PIC) were determined using Cervus software version 3.0.7 (Kalinowski et al. 2007). Allelic richness and number of alleles per locus were calculated with Fstat program version 2.9.3.3 (Goudet 2001). To further assess the extent of informativeness and hence usability of the developed markers, we tested the genetic structure and principle coordinate analysis (PCoA) on the four Nile tilapia populations using STRUCTURE version 2.3.4 (Hubisz et al. 2009) and GenAlex version 6.5 (Peakall and Smouse 2006), respectively. STRUCTURE classifies populations by genetically allocating them into groups whose individuals share similar patterns of variation (Porras-Hurtado et al. 2013). The program further is rendered useful as it can identify subpopulations of the whole population by maximizing HWE linkage within potential subpopulations (Porras-Hurtado et al. 2013). STRUCTURE was set at 100,000 burn-in period and the application of Markov Chain Monte Carlo (MCMC) was run at 500,000 replications, with each cluster (K) assigned to 10 iterations. STRUCTURE default settings for the admixture model and allele frequencies correlated were implemented. For inference to the K that best suits the data, we ran STRUCTURE HARVESTER. Here, the program collates STRUCTURE results and validates multiple K values for optimal detection and thereby depicts the best K value from tens or hundreds of iterations (Earl and vonHoldt 2012), as indicated in the supplementary material Table S1, Fig. S1 and Fig. S2. Similarly, for presenting informative genetic STRUCTURE outputs, we ran the CLUMPAK clustering Markov package pipeline across the K values for summation and graphical representation of the results obtained from STRUCTURE (Kopelman et al. 2015). From these analyses, we present and compare the results regarding; PIC, number of alleles (Na), allelic richness (Ar), HWE per population, Fixation index (Fis), PCoA, and STRUCTURE based on the two allele calling methods, AL and WAI.

Results

From the 2,404,293 paired reads generated for primer design, 6,724 contained SSR motifs and complied with our quality criteria. Out of the 35 developed SSR primer pairs, only 26 successfully amplified most of the Ugandan

Nile tilapia samples used (Table 2). Both 2mers and 3mers dominated (each eight in number), followed by 4mers and 5mers; each seven in numbers. For SSR-GBS, we generated a total number of 4,783,118 paired reads, of which 1,738,315 passed our quality control procedure, yielding an average number of 181 reads per marker and individual. In total, 12% of the called alleles were flagged for manual control, between 0 and 73% per locus. However, only in four loci most genotypes could not be determined (39–73%), mainly because of a frequently occurring unspecific product or length polymorphisms out of frame of the microsatellite motif and thus easy to correct. Without these markers only about 3.5% were flagged.

Not all of the developed markers were very informative based on the genotyping results (Table 2; Fig. 2). For instance, two loci, Ti27 and Ti28, exhibited two alleles, which were the lowest number of alleles indicated by WAI, but with AL showing only one allele for locus Ti27 (Table 2). The same loci, Ti27 and Ti28, indicated similarly the least allelic richness of less than 2 considering both allele calling methods (Table 2). In general, the rest of the loci indicated allelic numbers and richness ranging from 3 to 56 and from 1.8 to 13.9 respectively (Table 2; Fig. 2a). In total for the whole dataset, AL resulted in 270 alleles compared to 407 alleles using WAI (Table 2). Similarly, WAI yielded a total of 157 allelic richness contrary to AL with a total of 140 (Table 2). The analysis consistently recovered higher values for number of alleles when we used whole sequence information rather than sequence length for a specific population per locus (Fig. 3). For instance WAI generated, for Lake Albert, 212 alleles compared to 157 for AL; for Lake George, 222 compared to 187 for AL; for River Nile, 129 compared to 110 for AL; and for Victoria, 207 compared to 170 for AL (Fig. 3; Supplementary material Table S3). Number of alleles based on the two allele calling methods was congruent to allelic richness where WAI presented higher values than AL throughout the loci (Supplementary material Table S2).

The PIC per locus generally varied from 0.03 to 0.94 with WAI and 0.00 to 0.93 based on AL (Table 2; Fig. 2b). The WAI procedure displayed 18 loci with a polymorphic information content (PIC) ≥ 0.50, contrary to the AL with14 loci (Table 2; Fig. 2b). WAI exhibited four loci with a PIC > 0.25 and < 0.50 and four loci with a PIC < 0.25. Contrary, AL revealed five loci with a PIC > 0.25 and < 0.50 and seven loci with a PIC < 0.25 (Table 2: Fig. 2b). Based on WAI, the expected and observed heterozygosity ranged from 0.00 to 0.94 and 0.00 to 0.89, respectively (Table 2, Table S8). On average, WAI recovered, respectively, an Ho and an He of 0.51 and 0.61, while AL an Ho and an He of 0.46 and 0.51, respectively (Table 2). For cross-species amplification, a total number of 22 loci exhibited positive amplification on *T. zillii*, albeit four indicated weak gel bands. Interestingly, four loci (Ti10, Ti11, Ti21 and Ti25) that initially showed negative results following the Nile tilapia amplification reportedly tested positive with *T. zillii* (Table 2).

Test for Hardy–Weinberg Equilibrium (HWE) for WAI and AL within the four Nile tilapia populations showed in four markers consistent deviations in all populations (Ti, 22, Ti29, Ti31, Ti35). Congruence between results of tests for HWE and of Micro-Checker indicate that this stems from the presence of null alleles (Table 3 and Table S5–7). This is also supported by Fis values, indicating deviations from HWE due to heterozygosity deficiency (excess homozygosity) for almost all loci. When all samples considered, the populations GN and RNK showed deviations at additional 4 and 3 loci respectively, of which 4 occurred with both allele call methods and three only with WAL. Populations of Lake Albert and Lake Victoria indicated the majority of loci deviating from HWE (additional 7 or 12 loci respectively), 11 of the deviations occurred with both allele call methods, 5 only with WAI and 2 only with AL. The high amount of deviations from HWE in these populations is consistent with the subpopulation structure indicated in the STRUCURE analysis.

In the STRUCTURE analysis using STRUCTURE HARVESTER based on WAI and AL, $K = 5$ and $K = 2$ were, respectively, the best K values for the populations (Fig. 4; Fig. S1; Fig. S2; Table S1;). Results from $K = 2$ up to $K = 5$ are shown, to include also the population number used in this study (Fig. 4). K values from 6 to 10 were uninformative and therefore were not considered based on STRUCTURE HARVESTER outputs (Table S1, Fig. S1; Fig. S2). In comparison, STRUCTURE results in more ambiguous clustering for AL than WAI. (Fig. 4b). STRUCTURE results indicated a separation among all populations only for WAI (Fig. 4). At $K = 5$, the separation between the River Nile and Lake Albert populations for WAI was not evident in the AL results. Moreover, substructure within the Lake Albert and Victoria populations is indicated at K 4 and 5 by assigning single individuals within one of the lakes to other clusters (AN18 for Lake Albert, and VKM 10 and 13 for Lake Victoria). Generally, STRUCTURE analyses were consistent with PCoA results where a slight division between the populations is indicated (Fig. 5). Here, Lake Albert is connected to the Nile which is reflected also in the analysis. Albert and the Nile populations are assigned to the same cluster. Lake George individuals are assigned to their own cluster at $K = 2$ to $K = 5$. Lake Victoria populations seem less strongly divided from Lake Albert than George (Fig. 5).

Excluding the individuals in Lake Albert and Victoria populations that were assigned to different clusters, a test for HWE and for null alleles using Micro-Checker, indicated no pronounced deviations from HWE with no differences between the allele call methods (Table 3; Table S6). The only significant signals were for the above-mentioned four

Table 2 Information of the final set of 26 SSR markers, concerning primer sequence, repetition motif, amplicon length, variation, species cross-amplification

Loci	Primer sequence (5′–3′)	R. motif	ASR	PIC	PIC*	Na	Na*	Ho	Ho*	He	He*	Ar	Ar*	CS
Ti12	F:GCCACCAAAATATTCGTGTT R:CCATGTTCTGTCTCCTTGAA	(TAC)12	377–416	0.39	0.06	10	5	0.31	0.05	0.44	0.06	3.4	1.8	--
Ti16	F:CAGACGTAGGCGATAAATCT R:GAACACATCCATTTCCACAC	(AAC)10	377–428	0.16	0.16	8	8	0.13	0.14	0.16	0.17	2.5	2.9	**
Ti8	F:CTGAAGTCCTGCTGAGATTT R:CATTGTTCTTGGCACCTCTA	(AC)15	400–402	0.16	0.07	7	6	0.13	0.04	0.16	0.07	2.3	1.8	**
Ti34	F:GCTTACAGTACATTGTGTGC R:CTGATGAGAAAAACAGACGC	(TCTCT)10	375–415	0.90	0.86	26	12	0.84	0.80	0.91	0.88	10.5	8.8	**
Ti15	F:GCTGTGATCATCTGGAGAAA R:AGGATCTAGAACCTCCAACC	(TGC)10	394–403	0.10	0.08	6	4	0.07	0.05	0.10	0.08	2	2.0	**
Ti2	F:TTCTGGGCTAACACACAAG R:AAGGTGTCACACAGTTTAGG	(CA)18	402–434	0.77	0.79	22	17	0.75	0.86	0.80	0.81	7.4	8.5	**
Ti29	F:AGGTCAAGATCAAGCAGTTT R:CATCAACATAATTCAGTGTGGA	(TAAAA)10	345–355	0.52	0.24	12	12	0.20	0.11	0.60	0.25	4.1	4.0	**
Ti1	F:TTATCACTGCTGAACGTCTT R:GTTTTGGCTGCTACACATTC	(TG)10	399–407	0.50	0.50	8	6	0.48	0.51	0.52	0.52	3.5	3.5	**
Ti18	F:AGCAAGTGAGATAAGCACTG R:TACATAGCAGTGCAGTTTGC	(ATCT)8	397–461	0.92	0.88	33	17	0.89	0.95	0.93	0.90	11.7	10.1	**
Ti4	F:TGTGCAGAATAGAATAGCCC R:GAAAGGAAAAATGTTGGTGGT	(GT)18	403–415	0.78	0.77	11	9	0.82	0.84	0.81	0.81	6.1	6.0	--
Ti13	F:AATCCGTTAGCTGCAGATAG R:GCTGATTAAACACAAAGTTGG	(ATG)10	394–403	0.38	0.38	5	6	0.27	0.29	0.40	0.40	3.7	4.1	--
Ti26	F:ATTGCTTCATCCCTTGAGTT R:ACACGGAAAACCTAATGACA	(ACAA)10	411–427	0.64	0.63	10	10	0.56	0.58	0.68	0.68	5	5.5	--
Ti35	F:TCAACCACAAACTCCTCTTT R:AAACTAAGTGCAGCTCATGA	(AAAAG)14	364–424	0.90	0.83	25	17	0.53	0.59	0.91	0.85	11	9.1	**
Ti28	F:TGTCTTGGGATTTGAGATCA R:CGGAGGTTTCTTCCTGTTAA	(ATTCA)8	384–384	0.26	0.01	2	2	0.26	0.01	0.30	0.01	2	1.1	*
Ti17	F:AACTGAAGAAGAAGCCTTGG R:ATCATCTTCCTCTACTGCCT	(GAA)21	364–391	0.71	0.75	17	13	0.62	0.62	0.74	0.73	6.3	6.4	*
Ti5	F:AAGGAGGATGATCAGGACAC R:AGACCTCCACTGTGATCTTA	(CA)10	401–419	0.53	0.25	9	7	0.50	0.21	0.56	0.26	4.5	3.5	--
Ti33	F:GCTTATGGCTGTATGGAGTT R:CGACTTCTGTTGTGATTTGG	(TTCAA)6	382–422	0.82	0.82	9	8	0.75	0.76	0.85	0.85	6.5	6.7	--
Ti31	F:GAAACTATCCACAGAAGCCA R:AGGCTTCTTACAGTTGGATG	(CTAAT)7	302–380	0.75	0.73	13	9	0.20	0.34	0.78	0.77	6.3	5.8	--
Ti22	F:ACTGACCAAGTGCTTTGTAT R:AACTGCTGTGTTGAACTTTG	(CTAT)20	303–475	0.94	0.93	56	44	0.59	0.65	0.95	0.94	13.9	15.4	*
Ti7	F:TCTTTGTGTCAGAACTGTGT R:ACTCTGCTTTTAGCCAATCA	(AC)17	404–424	0.73	0.75	14	13	0.92	0.96	0.76	0.78	6.8	7.7	**

Table 2 (continued)

Loci	Primer sequence (5′–3′)	R. motif	ASR	PIC	PIC*	Na	Na*	Ho	Ho*	He	He*	Ar	Ar*	CS
Ti9	F:CTCAGTGACGAAGCCAAA R:CCTGGCAATCAAAAGAACAA	(AC)10	409–415	0.30	0.25	8	5	0.27	0.24	0.32	0.26	3.6	3.4	**
Ti14	F:TCCCTAAAATATGCCACCAA R:TAGTGCTTTAATGGCTCTGG	(TAA)19	368–374	0.54	0.34	6	3	0.43	0.28	0.58	0.38	4.3	6.0	**
Ti27	F:CTGTCTTTCTTGATGTGGGA R:ATGCACAAATTTTAAGGGCC	(TTTG)6	391–391	0.03	0.00	2	1	0.03	0.00	0.03	0.00	1.3	1.0	--
Ti6	F:CAGCTCTCATGAACACTTGA R:ACCCATAAATCACACCAGTC	(GA)23	384–418	0.75	0.69	25	16	0.97	0.98	0.78	0.73	7.1	6.5	**
Ti32	F:CAGGAAATGGCTCCAAAATG R:TTGTAGCTAGGAATCAGTGC	(AAAAT)7	330–350	0.89	0.35	31	6	0.88	0.36	0.90	0.39	10.7	3.4	**
Ti24	F:ACTGACAACATAAAGACATATGC R:CACAGTTTGAATCCACCATC	(TTAC)9	403–455	0.84	0.79	32	14	0.78	0.73	0.85	0.81	10.7	8.5	*
Ti10[n]	F:CCATTGATTAGTGCAACACC R:ACATGAATGCCAAATGACAG	(TTA)9	354–403											**
Ti11[n]	F:CATTGTTCGATGACCTTAAGT R:CCCTGCTTTTATGCCTACTT	(TAA)12	384–452											**
Ti21[n]	F:CAAAACTAAAAGCTTCCCCC R:GGATGTTTTCAGTGTGATGT	(TAGA)11	384–491											**
Ti25[n]	F:TTCAACTCAAACAGCTTCCT R:TCCATCAAAGATGAAATAGTGT	(ATCT)6	417–420											**
Total				15.1	12.8	407	267	13.1	11.9	15.8	13.4	157.2	143.5	
Ave				0.58	0.49	15.7	10.3	0.51	0.46	0.61	0.51	6.1	5.5	

Variability values were estimated using all four Ugandan Nile tilapia populations

F forward, *R* reverse, *R. motif* repeat motif, *ASR* allele size range (bp), *PIC* WAI polymorphic information content, *PIC** AL polymorphic information content, *Na* and *Na** number of alleles for WAI and AL respectively, *Ar* and *Ar**=allelic richness for WAI and AL respectively, *Ho** and *He** observed and expected heterozygosity for AL, *Ho* and *He* observed and expected heterozygosity for WAI, *Superscript n* loci which tested negative with Nile tilapia amplification, *CS* cross−species amplification success, **strong PCR products, *weak, – very weak, and *Ave* average

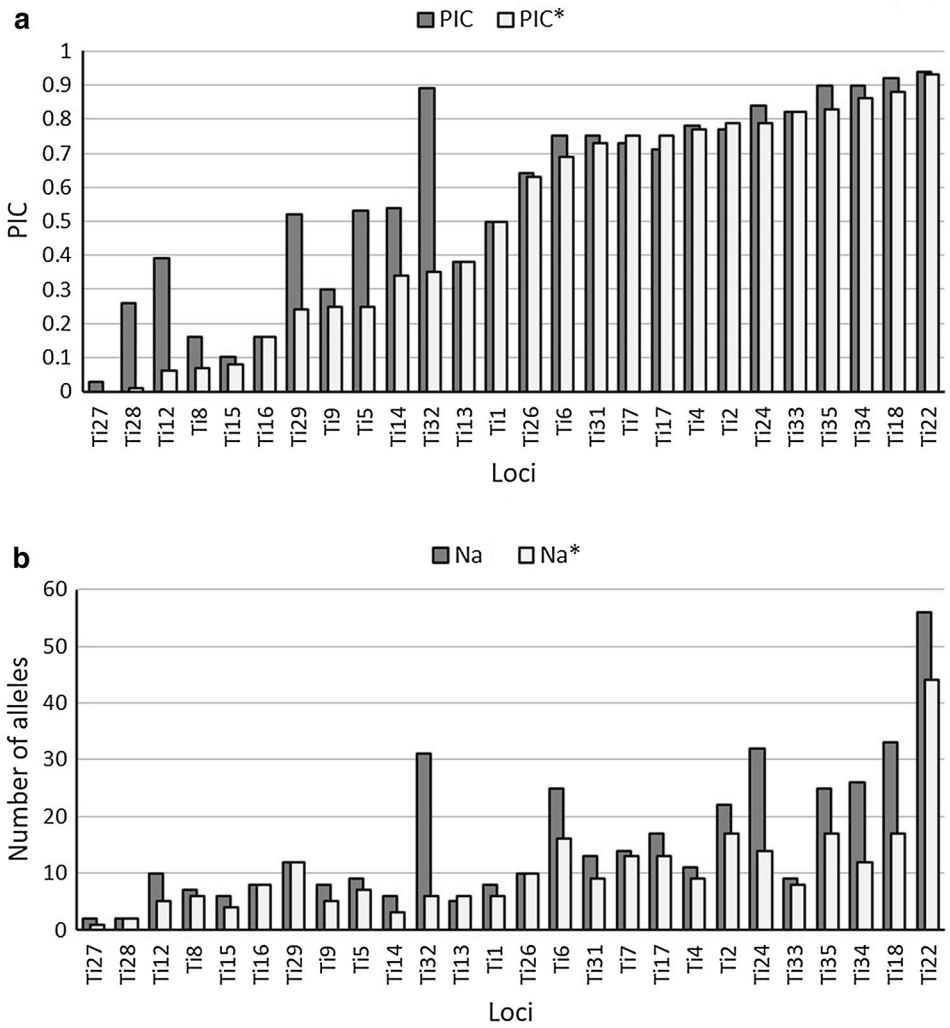

Fig. 2 Comparison between allele calling methods; whole sequence information (dark grey) and amplicon length (white). This was performed using number of alleles (Na) and polymorphic information content (PIC) for each of the 26 SSR loci. **a, b** represent PIC and Na per locus respectively

markers for all populations, besides that only three to four markers showed deviations in single populations. Of a total of 12 occurrences, 8 applied to both methods, and 3 only to WAI and 1 only to AL.

Discussion

With the genesis of high throughput sequencing methods, the classical genotyping approaches can be modified to recover a high level of information and to increase automation. This decreases the necessity for technical expertise in SSR genotyping that remains one difficulty despite the widespread use of the markers worldwide (Sunnucks 2000; De Barba et al. 2017). Constraints in allele call for SSR genotyping include the occurrence of null alleles, stutter bands as result of the slippage artefact in PCR reaction or allele dropout caused by amplification or fragment measurement (Vartia et al. 2016). In this study, we compared the information usually recovered from traditional electropherogram (AL) frequency distribution for scoring microsatellite genotypes with WAI for allele scoring/calling procedures. To a greater extent, the two allele calling procedures varied in generating the number of alleles, allelic richness, PIC, and genetic structure, of the studied Nile tilapia populations. Here, WAI approach consistently indicated to be more informative in aspects of cataloguing these herein mentioned genetic characteristics in contrast to the AL frequency procedure. However, this was unsurprising because unlike AL procedure, WAI approach systematically surveys the entire sequence hence summarising variability from the repetition motif and SNPs in the flanking regions. In our findings, the high level of variation also resulted in a genetic structure where every population could be characterized by its own cluster, which was not possible when AL was considered. One obstacle in the traditional microsatellite genotyping approach is attributed to genotyping artefacts and biases which originate from equipment specifications and different laboratory methods. With SSR-GBS, because the entire sequence information is utilised, this may no longer be a problem. This approach

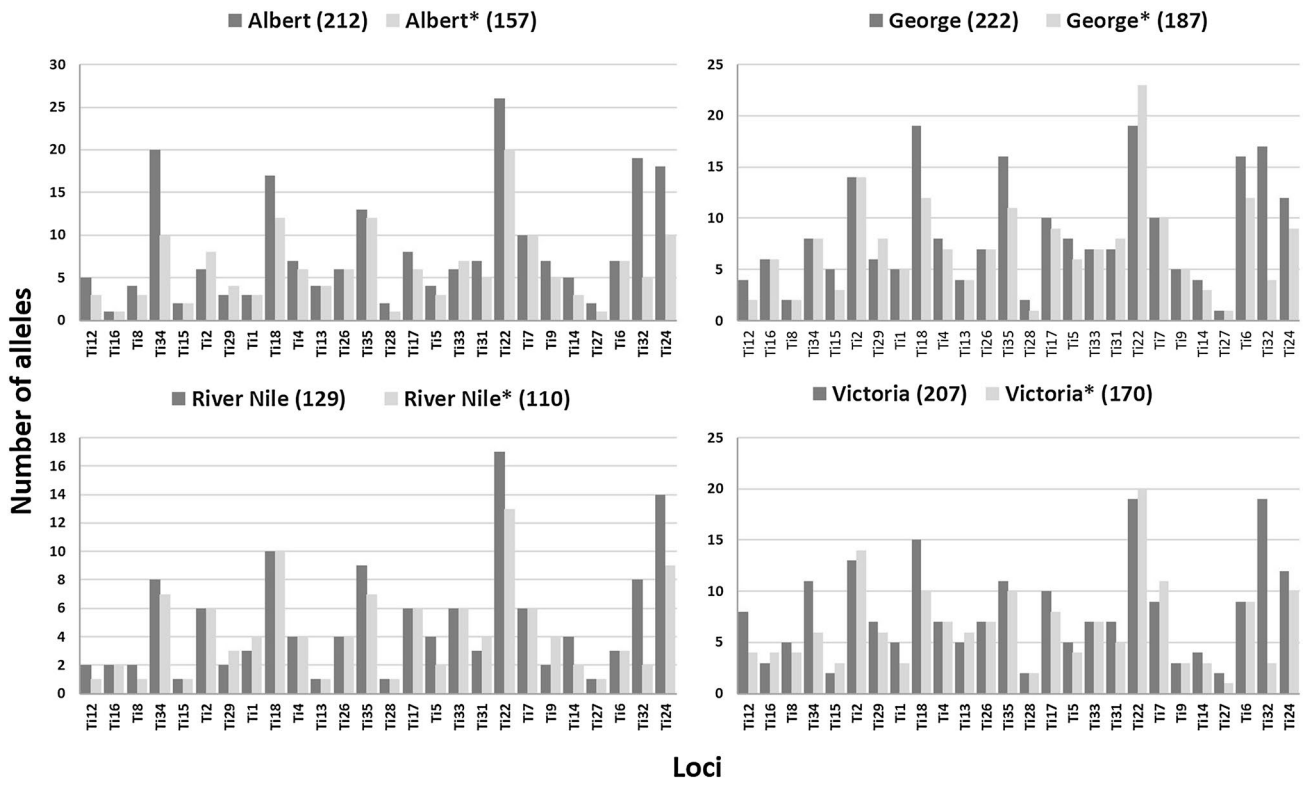

Fig. 3 Comparison between allele calling methods based on number of alleles exhibited per population for each of the 26 SSR loci. Whole amplicon information (dark grey) and amplicon length (light grey). Values in brackets indicate total number of alleles generated from the two allele calling procedures

(WAI) will allow for the creation of reliable and informative SSR genotype databases that can be used for meta-analytical studies and long-term projects with the contribution of several research groups. For the reasons stated above, it is expected that SSR-GBS will become a standard practice in conservation genetics.

Despite a few loci indicating PIC values of less than 0.25, most of the developed markers were polymorphic and informative. PIC values were congruent with the expected heterozygosity and allelic numbers/richness among the populations, translating into their capability and reliability for molecular genetic analyses. In genetic marker development, heterozygosity and PIC are seen as important attributes to estimate information content of loci and their applicability in down-stream analysis (Nagy et al. 2012; Chesnokov and Artemyeva 2015). In the current study, our analysis indicated 18 loci with PIC values above 0.5 and 4 loci with PIC values below 0.5 but above 0.25. Here, the informative loci summed up to 22. According to Botstein et al. (1980), highly informative markers should have a PIC value of greater than 0.5, reasonably informative markers a PIC of greater than 0.25 and slightly informative markers of PIC of less than 0.25. Therefore, even though 18 loci exhibited a PIC greater than 0.5, it is likely that the remaining loci (4) with a PIC greater than 0.25 are also useful if combined with the former during genotyping (Sunnucks 2000). Zane et al. (2002) showed that the use of several loci can increase the sensitivity of population genetic analyses and by using SSR-GBS the number of loci can easily be increased.

From the 26 tested loci, four loci deviated from HWE for most studied populations. Positive Fis and Micro-Checker results indicated that this might be caused by allele null alleles, a known limitation of amplification based genotyping approaches where one of the copies in a heterozygote individual is failed to be detected and genotyped as homozygote (Selkoe and Toonen 2006). WAI did recover more information and might have allowed to observe undetected subpopulation structure within the lakes investigated. Based on STRUCTURE results this was the case of the lakes Albert and Victoria, which harbour populations with individuals assigned to multiple clusters. In accordance to this, is the finding that a higher number of loci deviate from HWE in these populations. Given the translocation and artificial stoking background especially of lake Victoria, substructure might be caused within these populations due to human mediated admixture between distinct gene pools (Tibihika et al. 2018). Further studies will focus on how Nile tilapia stocks were affected by this anthropogenic activity.

Table 3 Tests for deviations of Hardy–Weinberg Equilibrium (HWE) per population and locus for both allele calling methods without considering the samples contributing to subpoputalion structure in Lakes Albert and Victoria

Populations	AN				GH				RNK				VKM			
Locus	p¶	p*	Fis¶	Fis*	p¶	p*	Fis¶	Fis*	p¶	p*	Fis¶	Fis*	p¶	p*	Fis¶	Fis*
Ti12	0.02	–	0.48	–	0.10	–	−0.08		0.01	–	0.79	–	0.03	1.00	0.32	−0.02
Ti16	–	–	–	–	0.69	0.69	−0.02	−0.02	–	–	–	–	–	–	–	–
Ti8	0.25	–	0.34	–	–	–	–	–	1.00	–	−0.10	–	–	–	–	–
Ti34	0.79	0.02	−0.02	0.12	0.62	0.61	−0.02	−0.02	0.75	0.79	−0.03	−0.04	0.36	0.57	0.06	−0.03
Ti15	–	–	–	–	1.00	1.00	−0.04	−0.02	–	–	–	–	–	–	–	–
Ti2	0.33	0.92	0.11	−0.12	0.60	0.95	−0.05	−0.09	0.34	0.94	0.08	−0.20	0.62	0.74	0.01	−0.01
Ti29	0.00	1.00	0.83	−0.03	0.00	0.00	0.59	0.58	0.08	0.02	0.38	0.50	0.00	0.00	0.73	0.82
Ti1	0.83	0.87	−0.11	−0.14	0.04	0.04	0.06	0.07	0.21	0.44	0.21	0.07	0.43	0.83	0.07	−0.11
Ti18	0.14	0.87	0.06	−0.09	0.89	0.88	−0.04	−0.05	0.01	0.86	0.12	−0.10	0.93	1.00	−0.08	−0.14
Ti4	0.19	0.63	0.09	−0.05	0.96	0.96	−0.19	−0.19	0.86	0.79	−0.22	−0.17	0.91	0.93	−0.10	−0.11
Ti13	0.07	0.07	0.32	0.32	0.14	0.14	0.23	0.23	–	–	–	–	0.36	1.00	0.03	−0.17
Ti26	0.85	1.00	−0.16	−0.28	0.26	0.27	0.06	0.06	0.00	0.00	0.34	0.34	0.00	0.00	0.43	0.36
Ti35	0.00	0.01	0.61	0.23	0.00	0.06	0.26	0.15	0.00	0.06	0.45	0.30	0.00	0.05	0.38	0.28
Ti28	1.00	–	−0.02	–	0.97		−0.25		–	–	–	–	0.21	–	0.35	–
Ti17	0.08	0.09	0.27	0.32	0.12	0.13	0.03	0.03	0.66	0.66	0.01	0.01	0.02	0.04	0.26	0.25
Ti5	0.91	1.00	−0.16	−0.06	0.20	0.23	0.19	0.13	0.07	0.21	0.18	0.35	0.80	1.00	−0.13	−0.13
Ti33	0.21	0.21	0.13	0.12	0.49	0.50	0.01	0.01	0.07	0.07	0.05	0.05	0.59	0.62	−0.08	−0.09
Ti31	0.00	0.00	0.73	0.51	0.00	0.00	0.78	0.54	0.01	0.00	0.78	0.60	0.00	0.00	0.67	0.54
Ti22	0.00	0.00	0.25	0.23	0.00	0.00	0.68	0.44	0.00	0.03	0.27	0.20	0.00	0.00	0.37	0.30
Ti7	0.98	1.00	−0.13	−0.18	1.00	1.00	−0.22	−0.22	0.99	1.00	−0.32	−0.41	1.00	1.00	−0.38	−0.40
Ti9	0.33	1.00	0.12	−0.10	0.02	0.04	0.22	0.19	–	1.00		−0.07	1.00	1.00	−0.05	−0.02
Ti14	0.00	0.05	0.40	0.45	0.01	0.01	0.20	0.43	0.13	0.43	0.21	0.17	0.03	1.00	0.09	−0.24
Ti27	–	–	–	–	–	–	–	–	–	–	–	–	1.00	–	−0.02	–
Ti6	1.00	1.00	−0.74	−0.75	0.88	0.91	−0.10	−0.13	1.00	1.00	−0.78	−0.85	1.00	1.00	−0.53	−0.54
Ti32	0.25	0.10	−0.02	−0.01	0.01	0.14	0.02	0.07	0.73	0.60	−0.01	0.06	1.00	0.41	−0.07	0.03
Ti24	0.92	0.53	−0.05	0.00	0.48	0.53	0.05	0.03	0.59	0.26	0.00	0.07	0.34	0.22	0.06	0.09
NDL	6.00	4.00			8.00	6.00			6.00	4.00			8.00	5.00		
NLwi	3.00	6.00			2.00	3.00			6.00	7.00			3.00	5.00		

AN Albert; *GH* George, *RNK* Nile, *VKM* Victoria, *NDL* number of deviated loci from HWE, *NLwi* number of loci without information

p¶ probability value for WAI allele calling, p* probability value for AL allele calling, Fis¶ and Fis* fixation indices for WAI and AL respectively

The genotyping based on WAI showed slightly more loci deviating from HWE than with AL (3 vs 1 after exclusion of individuals that cause substructure within populations). Since the WAI genotyping is able to detect a higher number of alleles, such effects might become more likely using this method.

Cross-species amplification results were unexpected given the genetic distance between Nile tilapia and *Tilapia zilli* (Elghobashy et al. 2005). A total of 18 loci strongly amplified *T. zillii* genomic DNA. Four loci (Ti10, Ti11, Ti21 and Ti25), which showed negative results for Ugandan Nile tilapia amplification, were positive with *T. zillii*. This suggests the existence of allele drop out in Ugandan Nile tilapia, while the individual from Ethiopia shared alleles with *T. zillii*. The capacity of SSR markers to cross amplify species might also be locus-dependent (Gen-Hua et al. 2010), although species' phylogenetic relationships or homoplasy could play a role (Bezault et al. 2012). The pattern of cross amplification between the species might here be explained by factors like outbreeding/hybridisation or admixed populations (Barbara et al. 2007).

Conclusions

The development of SSR-GBS has enhanced microsatellite analysis. The Illumina approach genotyped SSR loci successfully, demonstrated polymorphism and their usability in down-stream applications such genetic diversity and structure. We expect that also other analyses like pedigree analysis, gene flow and hybridisation, genetic linkage maps, etc. can be facilitated by the method. SSR genotyping that

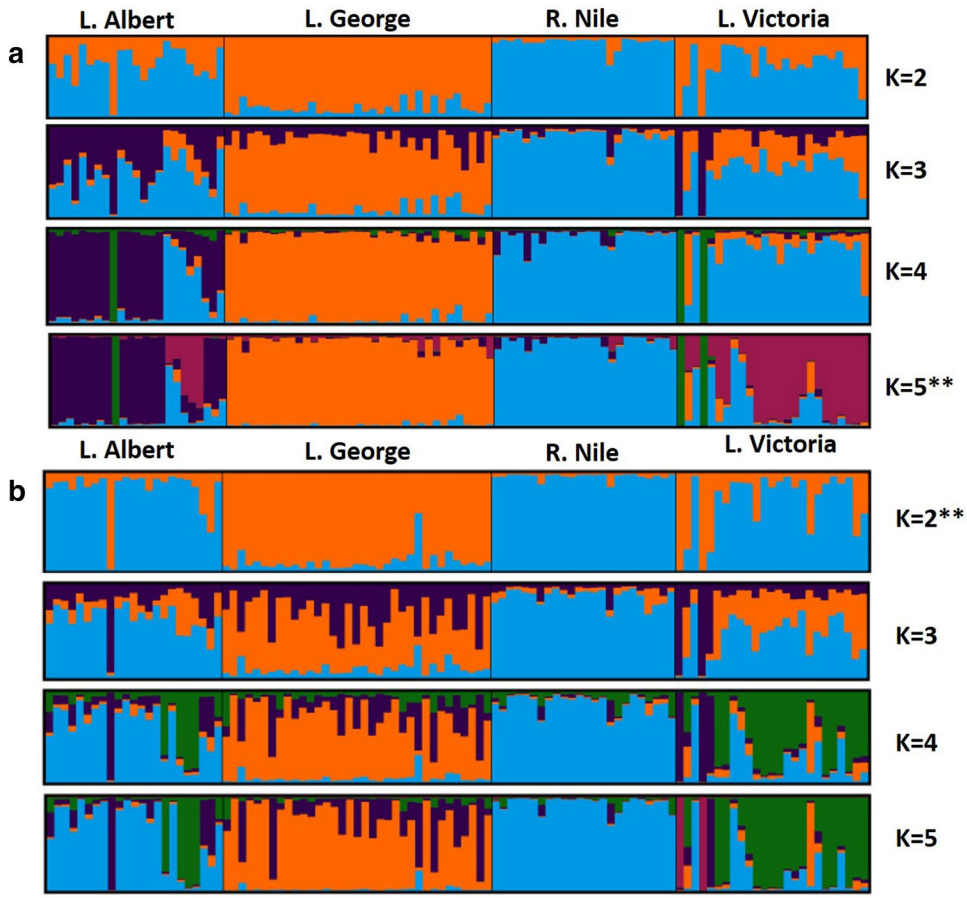

Fig. 4 Genetic structure bar plots for the four East African (Uganda) Nile tilapia populations based in cluster assignment probability calculated in SRUCTURE for the best values of K = 5 and K = 2. **a** represents Nile tilapia genetic structures inferred from whole amplicon information and **b** from amplicon length allele calling methods

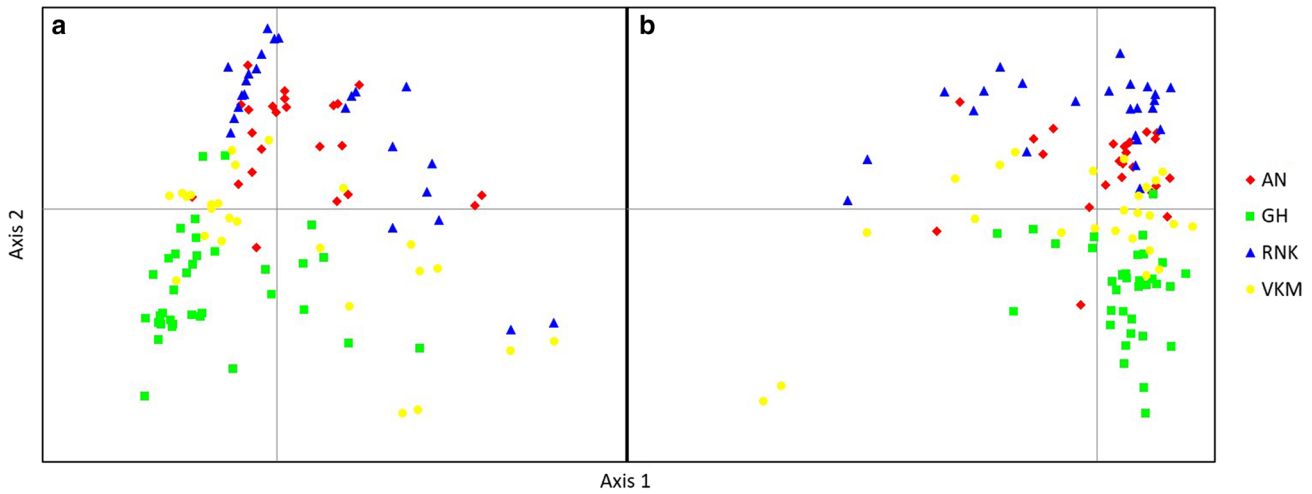

Fig. 5 Principle coordinate analysis (PCoA), depicting genetic similarities/variabilities between the four East African (Uganda) Nile tilapia populations. **a** represents PCoA results from whole amplicon information and **b** from amplicon length allele calling methods

uses automated WAI for allele calling greatly reduced bias and artefacts, and thereby yielded more information than the commonly employed traditional AL. The reproducibility of the method could be vital for Nile tilapia breeding programs, because genotyping from multiple generations across a long period of time are directly comparable. STRUCTURE, PCoA, PIC and He results suggest that the majority of the presented 26 loci, particularly the 22 polymorphic loci, are highly informative for comprehensively studying the Nile tilapia population patterns. In addition, this number

can be easily increased using this approach. The findings from the current study should aid in a more integrative and comprehensive approach towards restoration, management and conservation of the East African Nile tilapia and related species. Future studies should consider the application of SSR-GBS for Nile tilapia congenerics; *Oreochromis variabilis* and *Oreochromis esculentus* (Ngege), as the species have been recorded as endangered. Knowledge about the genetic-make of the herein studied tilapiines might be useful for restoration and subsequent conservation.

Acknowledgements Open access funding provided by University of Natural Resources and Life Sciences Vienna (BOKU). This work was funded by the Austrian Partnership Programme in Higher Education and Research for Development (APPEAR), a programme of the Austrian Development Cooperation (ADC) and implemented by the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH). The bioinformatics pipeline was supported by research funds from the Department of Integrative Biology and Biodiversity Research-University of Natural Resources and Life Sciences Vienna (BOKU). We are also indebted to Kachwekano Zonal Agricultural Research and Development Institute-National Agricultural Research Organization-Uganda for their great assistance during field activities.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All international, national or institutional guidelines for the care and use of animals were observed and adhered to. We also declare that the current work and manuscript have not been submitted to more than one journal for simultaneous consideration and have not been published elsewhere.

References

Agnèse JF, Adépo-Gourène B, Abban EK, Fermon Y (1997) Genetic differentiation among natural populations of the Nile tilapia *Oreochromis niloticus* (Teleostei, Cichlidae). Heredity 79(1):88–96. https://doi.org/10.1038/hdy.1997.126

Andrews S, FastQC A (2010) A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arthofer W, Heussler C, Krapf P, Schlick-Steiner BC, Steiner FM (2018) Identifying the minimum number of microsatellite loci needed to assess population genetic structure: a case study in fly culturing. Fly 12(1):13–22. https://doi.org/10.1080/19336934.2017.1396400

Balirwa JS (1992) The evolution of the fishery of *Oreochromis niloticus* (Pisces: *Cichlidae*) in Lake Victoria. Hydrobiologia 232(1):85–89. https://doi.org/10.1007/BF00014616

Barbara T, Palma-Silva C, Paggi GM, Bered F, Fay MF, Lexer C (2007) Cross-species transfer of nuclear microsatellite markers: potential and limitations. Mol Ecol 16(18):3759–3767. https://doi.org/10.1111/j.1365-294X.2007.03439.x

Bezault E, Rognon X, Gharbi K, Baroiller J-F, Chevassus B (2012) Microsatellites cross-species amplification across some African cichlids. Int J Evol Biol 2012:1–7. https://doi.org/10.1155/2012/870935

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32(3):314–331

Capote N, Pastrana AM, Aguado A, Sánchez-Torres P (2012) Molecular tools for detection of plant pathogenic fungi and fungicide resistance. In: Cumagun CJR (ed.) Plant pathology. InTech, Rijeka. https://doi.org/10.5772/38011

Castoe TA, Poole AW, de Koning AJ, Jones KL, Tomback DF, Oyler-McCance SJ, Fike JA, Lance SL, Streicher JW, Smith EN (2012) Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. PLoS ONE 7(2):e30953. https://doi.org/10.1371/journal.pone.0136465

Chesnokov YV, Artemyeva A (2015) Evaluation of the measure of polymorphism information of genetic diversity. Agric Biol 50(5):571–578. https://doi.org/10.15389/agrobiology.2015.5.571eng

Curto MA, Tembrock LR, Puppo P, Nogueira M, Simmons MP, Meimberg H (2013) Evaluation of microsatellites of *Catha edulis* (qat; Celastraceae) identified using pyrosequencing. Biochem Syst Ecol 49:1–9. https://doi.org/10.1016/j.bse.2013.02.002

De Barba M, Miquel C, Lobréaux S, Quenette P, Swenson J, Taberlet P (2017) High-throughput microsatellite genotyping in ecology: improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. Mol Ecol Resour 17(3):492–507. https://doi.org/10.1111/1755-0998.12594

Dorak MT (2014) Basic population genetics. http://www.dorak.info/genetics/popgen.html

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4(2):359–361. https://doi.org/10.1007/s12686-011-9548-7

Elghobashy HA, Ramadan AA, Ibrahem IH, Rashed MA (2005) Phylogenetic relationships for some *Tilapia* species using electrophoresis. Egypt J Exp Biol (Zool) 1:57–63

Farrell ED, Carlsson JE, Carlsson J (2016) Next Gen Pop Gen: implementing a high-throughput approach to population genetics in boarfish (*Capros aper*). Open Sci 3(12):160651. doi:1098/rsos.160651

Gen-Hua Y, Kovacs B, Orban L (2010) A new problem with cross-species amplification of microsatellites: generation of non-homologous products. Zool Res 31(2):131–140. https://doi.org/10.3724/SP.J.1141.2010.02131

Goudet J (2001) FSTAT; a program to estimate and test gene diversities and fixation indices, version 2.9.3. http://www.unil.ch/izea/softwares/fstat.html

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F (2011) Current trends in microsatellite genotyping. Mol Ecol Resour 11(4):591–611. https://doi.org/10.1111/j.1755-0998.2011.03014.x

Hassanien HA, Gilbey J (2005) Genetic diversity and differentiation of Nile tilapia (*Oreochromis niloticus*) revealed by DNA microsatellites. Aquac Res 36(14):1450–1457

Hedrick PW (2001) Conservation genetics: where are we now? Trends Ecol Evol 16(11):629–636. https://doi.org/10.1016/S0169-5347(01)02282-0

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9(5):1322–1332. https://doi.org/10.1111/j.1755-0998.2009.02591.x

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol 16(5):1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Kaufman L (1992) Catastrophic change in species-rich freshwater ecosystems. Bioscience 42(11):846–858. https://doi.org/10.2307/1312084

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12):1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol Ecol Resour 15(5):1179–1191. https://doi.org/10.1111/1755-0998.12387

Koskinen MT, Hirvonen H, LANDRY PA, Primmer CR (2004) The benefits of increasing the number of microsatellites utilized in genetic population studies: an empirical perspective. Hereditas 141(1):61–67. https://doi.org/10.1111/j.1601-5223.2004.01804.x

Lee WJ, Kocher T (1996) Microsatellite DNA markers for genetic mapping in *Oreochromis niloticus*. J Fish Biol 49(1):169–171. https://doi.org/10.1111/j.1095-8649.1996.tb00014.x

Liu Z, Cordes J (2004) DNA marker technologies and their applications in aquaculture genetics. Aquaculture 238(1):1–37. https://doi.org/10.1016/j.aquaculture.2004.05.027

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1):10–12. https://doi.org/10.14806/ej.17.1.200

Miah G, Rafii MY, Ismail MR, Puteh AB, Rahim HA, Islam KN, Latif MA (2013) A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. Int J Mol Sci 14(11):22499–22528. https://doi.org/10.3390/ijms141122499

Miller MP, Knaus BJ, Mullins TD, Haig SM (2013) SSR_pipeline: a bioinformatic infrastructure for identifying microsatellites from paired-end Illumina high-throughput DNA sequencing data. J Hered 104(6):881–885. https://doi.org/10.1093/jhered/est056

Mwanja W (2000) Genetic biodiversity and evolution of two principal fisheries species groups, the labeine and tilapiine, of Lake Victoria, East Africa. PhD Thesis, Ohio State University, Cleveland

Mwanja W, Booton G, Kaufman L, Chandler M, Fuerst P, Donaldson E, MacKinlay D (1996) Population and stock characterization of Lake Victoria Tilapine fishes based on RAPD markers. In: Donaldson EM, Mackinlay DD (eds) Aquaculture biotechnology symposium proceedings of the international congress on the biology of fishes. American Fisheries Society, Washington, pp 115–124

Mwanja W, Booton GC, Kaufman L, Fuerst PA (2008) A profile of the introduced *Oreochromis niloticus* (Pisces: Teleostei) populations in Lake Victoria Region in relation to its putative origin of Lakes Edward and Albert (Uganda-E. Africa) based on random amplified polymorphic DNA analysis. Afr J Biotechnol 7(11):1769–1773. https://doi.org/10.5897/AJB08.813

Nagy S, Poczai P, Cernák I, Gorji AM, Hegedűs G, Taller J (2012) PICcalc: an online program to calculate polymorphic information content for molecular genetic studies. Biochem Genet 50(9–10):670–672. https://doi.org/10.1007/s10528-012-9509-1

Njiru M, Nzungi P, Getabu A, Wakwabi E, Othina A, Jembe T, Wekesa S (2007) Are fisheries management, measures in Lake Victoria successful? The case of Nile perch and Nile tilapia fishery. Afr J Ecol 45(3):315–323. https://doi.org/10.1111/j.1365-2028.2006.00712.x

Ogutu-Ohwayo R (1990) The decline of the native fishes of lakes Victoria and Kyoga (East Africa) and the impact of introduced species, especially the Nile perch, *Lates niloticus*, and the Nile tilapia, *Oreochromis niloticus*. Environ Biol Fishes 27(2):81–96. https://doi.org/10.1007/BF00001938

Okumuş İ, Çiftci Y (2003) Fish population genetics and molecular markers: II-molecular markers and their applications in fisheries and aquaculture. Turk J Fish Aquat Sci 3(1):51–79

Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. Genet Mol Biol 29(2):294–307. https://doi.org/10.1590/S1415-47572006000200018

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Resour 6(1):288–295. https://doi.org/10.1111/j.1471-8286.2005.01155.x

Porras-Hurtado L, Ruiz Y, Santos C, Phillips C, Carracedo Á, Lareu MV (2013) An overview of STRUCTURE: applications, parameter settings, and supporting software. Front Genet. https://doi.org/10.3389/fgene.2013.00098

Rousset F (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Resour 8(1):103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Ryman N, Palm S, André C, Carvalho GR, Dahlgren TG, Jorde PE, Laikre L, Larsson LC, PalmÉ A, Ruzzante DE (2006) Power for detecting genetic divergence: differences between statistical methods and marker loci. Mol Ecol 15(8):2031–2045. https://doi.org/10.1111/j.1365-294X.2006.02839.x

Schlotterer C, Amos B, Tautz D (1991) Conservation of polymorphic simple sequence loci in cetacean species. Nature 354(6348):63

Schoebel C, Brodbeck S, Buehler D, Cornejo C, Gajurel J, Hartikainen H, Keller D, Leys M, Říčanová Š, Segelbacher G (2013) Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequencing. J Evol Biol 26(3):600–611. https://doi.org/10.1111/jeb.12077

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol Lett 9(5):615–629. https://doi.org/10.1111/j.1461-0248.2006.00889.x

Seyoum S, Kornfield I (1992) Identification of the subspecies of *Oreochromis niloticus* (Pisces: Cichlidae) using restriction endonuclease analysis of mitochondrial DNA. Aquaculture 102(1–2):29–42

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135

Sundaray JK, Rasal KD, Chakrapani V, Swain P, Kumar D, Ninawe AS, Nandi S, Jayasankar P (2016) Simple sequence repeats (SSRs) markers in fish genomic research and their acceleration via next-generation sequencing and computational approaches. Aquacult Int 24(4):1089–1102. https://doi.org/10.1007/s10499-016-9973-4

Sunnucks P (2000) Efficient genetic markers for population biology. Trends Ecol Evol 15(5):199–203. https://doi.org/10.1016/S0169-5347(00)01825-5

Tariq Ezaz M, Sayeed S, McAndrew BJ, Penman DJ (2004) Use of microsatellite loci and AFLP markers to verify gynogenesis and clonal lines in Nile tilapia *Oreochromis niloticus* L. Aquac Res 35(15):1472–1481. https://doi.org/10.1111/j.1365-2109.2004.01175.x

Tibihika PD, Waidbacher H, Masembe C, Curto M, Sabatino S, Alemayehu E, Meulenbroek P, Akoll P, Meimberg H (2018) Anthropogenic impacts on the contextual morphological diversification and adaptation of Nile tilapia (*Oreochromis niloticus*, L. 1758) in East Africa. Environ Biol Fishes 101(3):363–381. https://doi.org/10.1007/s10641-017-0704-0

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces.

Nucleic Acids Res 40(15):e115–e115. https://doi.org/10.1093/nar/gks596

Van Oosterhout C, Hutchinson WF, Wills DP, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes 4(3):535–538. https://doi.org/10.1111/j.1471-8286.2004.00684.x

Vartia S, Collins PC, Cross TF, Fitzgerald RD, Gauthier DT, McGinnity P, Mirimin L, Carlsson J (2014) Multiplexing with three-primer PCR for rapid and economical microsatellite validation. Hereditas 151(2–3):43–54. https://doi.org/10.1111/hrd2.00044

Vartia S, Villanueva-Cañas JL, Finarelli J, Farrell ED, Collins PC, Hughes GM, Carlsson JE, Gauthier DT, McGinnity P, Cross TF (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. R Soc Open Sci 3(1):150565. https://doi.org/10.1098/rsos.150565

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38(6):1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Welcomme R (1966) Recent changes in the stocks of Tilapia in Lake Victoria. Nature 212:52–54. https://doi.org/10.1038/212052a0

Xiong J (2006) Essential bioinformatics. Cambridge University Press, New York

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Ecol 11(1):1–16. https://doi.org/10.1046/j.0962-1083.2001.01418.x

Zhan L, Paterson IG, Fraser BA, Watson B, Bradbury IR, Nadukkalam Ravindran P, Reznick D, Beiko RG, Bentzen P (2017) MEGASAT: automated inference of microsatellite genotypes from sequence data. Mol Ecol Resour 17(2):247–256. https://doi.org/10.1111/1755-0998.12561

Zhang L, Wang S, Yin S, Hong S, Kim KP, Kleckner N (2014) Topoisomerase II mediates meiotic crossover interference. Nature 511(7511):551–556. https://doi.org/10.1038/nature13442