2024

This project is funded by the European Union

TRUEFISH
ADVANCING AQUACULTURE

96 SNPs Panel Species Identification Tool

SPECIES IDENTIFICATION TOOL DEVELOPED FOR TILAPIA IN LAKE VICTORIA

ADAM G. CIEZAREK, ASILATU HAMISI SHECHONGE, CASSIUS ARUHO, KEVIN OBIERO, PAPIUS DIAS TIBIHIKA, BENDULA WISMEN, JOHN A.H. BENZIE, WILFRIED HAERTY

MAY 2024

Landell Mills

Food and Agriculture Organization of the United Nations

WorldFish

CGIAR

This project is funded by the European Union

96 SNPs Panel Species Identification Tool: Species Identification Tool Developed for Tilapia in Lake Victoria

Authors
Adam G. Ciezarek[1], Asilatu Hamisi Shechonge[2], Cassius Aruho[3], Kevin Obiero[4], Papius Dias Tibihika[5], Bendula Wismen[6], John A.H. Benzie[6], Wilfried Haerty[1]

[1] Earlham Institute, UK
[2] Tanzania Fisheries Research Institute, Tanzania
[3] National Fisheries Resources Research Institute, Uganda
[4] Kenya Marine and Fisheries Research Institute, Kenya
[5] WorldFish East and Southern Africa, Uganda
[6] WorldFish (HQ), Penang, Malaysia

Contents

List of Figures

List of Abbreviations

EAC	East African Community
EI	Earlham Institute
DNA	Deoxyribonucleic Acid
KMFRI	Kenya Marine and Fisheries Research Institute
LVFO	Lake Victoria Fisheries Organization
NaFIRRI	National Fisheries Resources Research Institute, Uganda
PCA	Principal Component Analysis
SNPs	Single nucleotide polymorphisms
TAFIRI	Tanzania Fisheries Research Institute
UK	United Kingdom

Challenge: The need for accurate species identification

The current tilapia species identification relies on the use of characteristic to each species (skin, eye, and fin coloration, presence, position and colour of spots or stripes, body and shape, fin characteristics). Species classification relies on the comparison of those traits to a reference specimen and have been collated in a volume printed by The Natural History Museum (Trewavas, 1983), which is rarely available outside of the UK. The major challenges associated with species identification stem from the fact that most of these species-specific traits have been described in a single or few reference specimen and existing morphological variation and plasticity among local populations can often undermine classification. Furthermore, in the countries surrounding the Lake Victoria, more than 30 species of *Oreochromis* have been described, many of which can readily hybridise further increasing the risk of misidentification. Finally, most of the traits use for classification have been described in sexually mature males, making juvenile and female identification difficult. Finally morphological species identification necessitates extensive training and expertise.

Aim

The objective of this study is to develop a tool to enable the accurate identification of tilapia species and hybrids from the wild in Lake Victoria.

Initial approach

To facilitate the development of a tool to enable tilapia species identification, we first consider the application of image recognition approaches to images of *Oreochromis* specimen to provide a predicted species name based on a large compendium of curated images used for the training of the algorithm, including the availability of genetic identification of the species of origin for the photographed specimen to be used to confirm species identification.

Application of image recognition to photographs for species / hybrids identification

We first considered to apply machine learning approaches based on image recognition to ascribe a status to a photographed individual. We selected convolutional neural network algorithms applied to set of curated tilapia photographs to assess the power of the method in identifying species of origin or hybrid status (Figure 1)

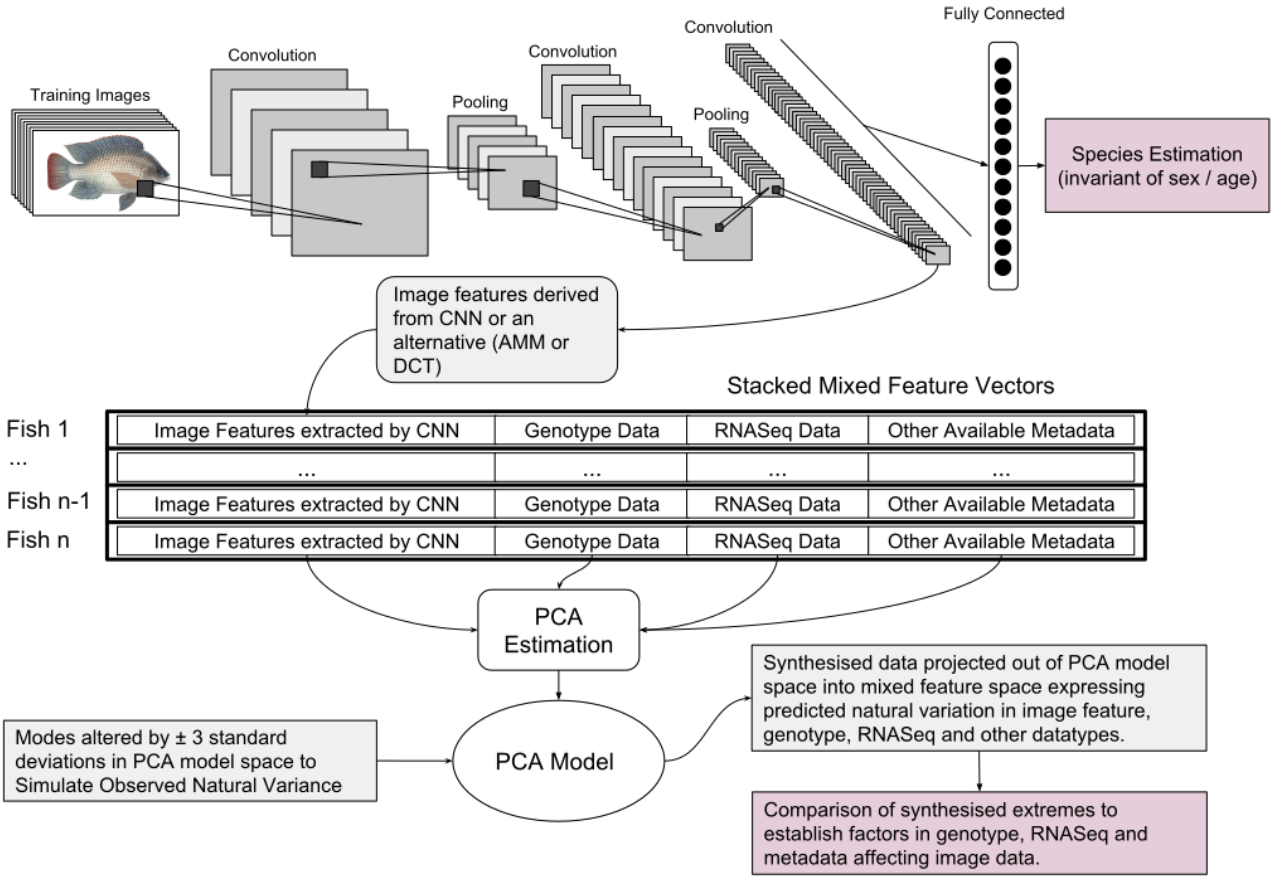

Figure 1 Application of convolutional neural networks for species identification.

Training set: We used data collected in Tanzania before the sampling performed as part of the genetic screening research for TRUEFISH, as sampling was only performed in 2022 (sequenced in 2023). We used a set of 931 expertly curated tilapia images with species and or hybrid labels. All these images are from sexually mature males. Additionally, 433 of these images were accompanied with genotyping and species / hybrid identification through low coverage sequencing (Ciezarek, et al., 2023). This subset is used as ground truth in the species / status.

Results and limitations: Overall, we obtained an f-score of 0.725 over 11 species. Although this is promising, the approach is not ready to be deployed because of its current accuracy. The lowest accuracy was achieved for hybrids and rare species.

The major limitation we encountered in further improving this score and the accuracy of species identification is the relatively low number of accurately labelled images, some species / hybrids being very poorly represented (with less than 10 images) in the training set leading to low ability in identifying individuals from these species. The second major issue we encountered is that the work developed so far was based on images of sexually mature males, severely limiting the applicability to fingerlings or females.

The acquisition of a large compendium of images across developmental stages for accurately labelled individuals will be necessary. For instance, Elhamod et al. 2021 applied a similar approach to classify fish species using a training set encompassing 63,758 photographs for 575 species. The authors demonstrated good accuracy (> 90%) (Elhamod, et al., 2021).

Finally, as the training set was performed only on sexually mature males, the classifier would not be able to identify female or juvenile specimens. To enable such developments, it is necessary to collect large training data sets including both photographs and genotypes of each individual of both sexes and across the different developmental stages. The availability of genotypes is necessary to validate the identifications and enable to facilitate the machine learning process.

A tool for species identification based on genotype.

To overcome the issue of the accuracy of the machine learning approach due to the size of the training set and the limitation to sexually mature males. We decided to develop another approach to provide the power to accurately identify the species or status (hybrids) of individuals regardless their sex and developmental stages.

Using genotype information for specimen, and comparison to reference genotypes for *Oreochromis* species enables species identification or hybrid status as shown as part of the TRUEFISH genetic screening research and from our previous work (Ciezarek, et al., 2022).

More specifically using the genome information, we were able to identify a small set of genetic variants (96) that enable the accurate species identification as well as hybrid status. As observed on Figure 2, panels based on 118 variants and 96 variants are much more powerful in classifying individuals to a species, or a hybrid status. Morphological identification had previously classified 43 individuals as hybrids, while only 14 of those were confirmed genetically to be hybrids, while the remaining individuals being genetically classified as either *O. niloticus*, *O. leucostictus*, or *O. urolepis*. Similarly, eight individuals previously identified as pure species based on morphological traits were found to be hybrids when treated with the variant panels. These observations demonstrate the limitations in identifying species and hybrids based on morphological parameters alone, and the strengths of applying genetic analyses methods instead. We demonstrated that those variants were sufficient to recapitulate classification based on full genome sequencing (Figure 3a), but most importantly the approach outperformed the use of microsatellites as well as morphological identification (Figure 3b).

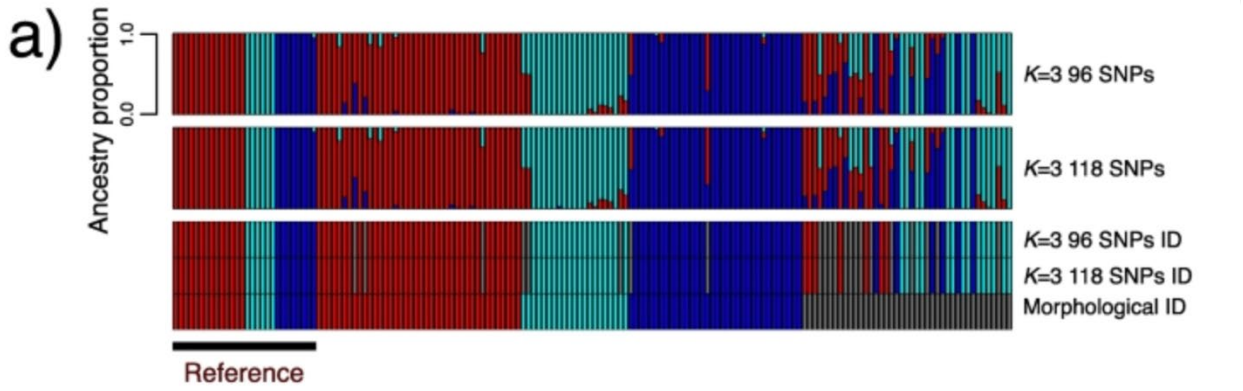

Figure 2 Targeted SNPs genotyping for species and hybrids identification outperforms morphological identification. Red: O.urolepis; blue: O. leucostictus; cyan: O. niloticus; grey: hybrids. Haplotypes identification based on 96 or 118 variants. Source: Ciezarek et al, 2022

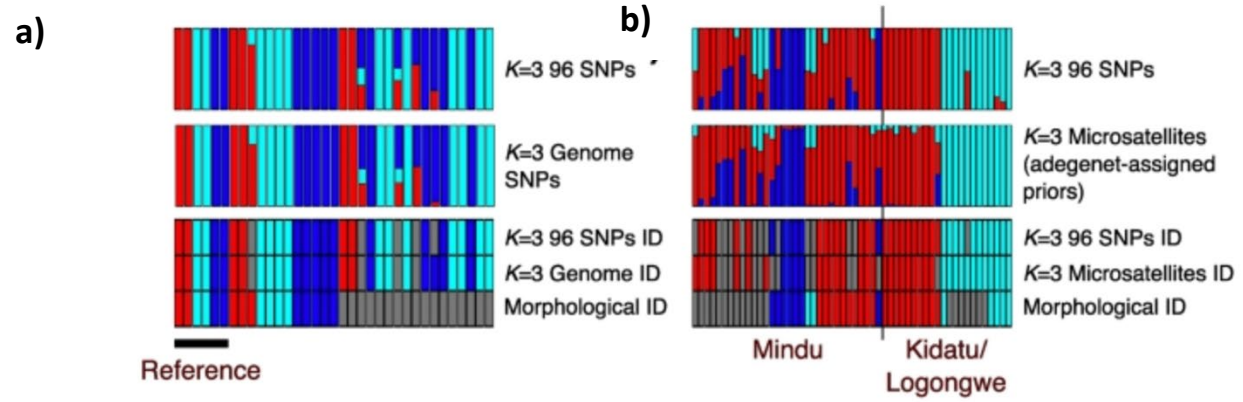

Figure 3: a) Comparison of selected variants to discriminate species and hybrids compared to genome wide variants or morphological identification, b) Comparison with the 96 SNPs with microsatellites genotyping. Source: Ciezarek et al. 2022

Identification of 96 SNPs enabling species identification

The sequencing performed as part of TRUEFISH allowed the identification of a 96 SNPs panel successfully identified all 259 pure species samples consistently with the genomic data. We demonstrated that this panel can accurately separate the species present in our data as well as to identify F1 hybrids both using Principal Component Analyses (PCA) (Figure 4) and ADMIXTURE analysis (Figure 5).

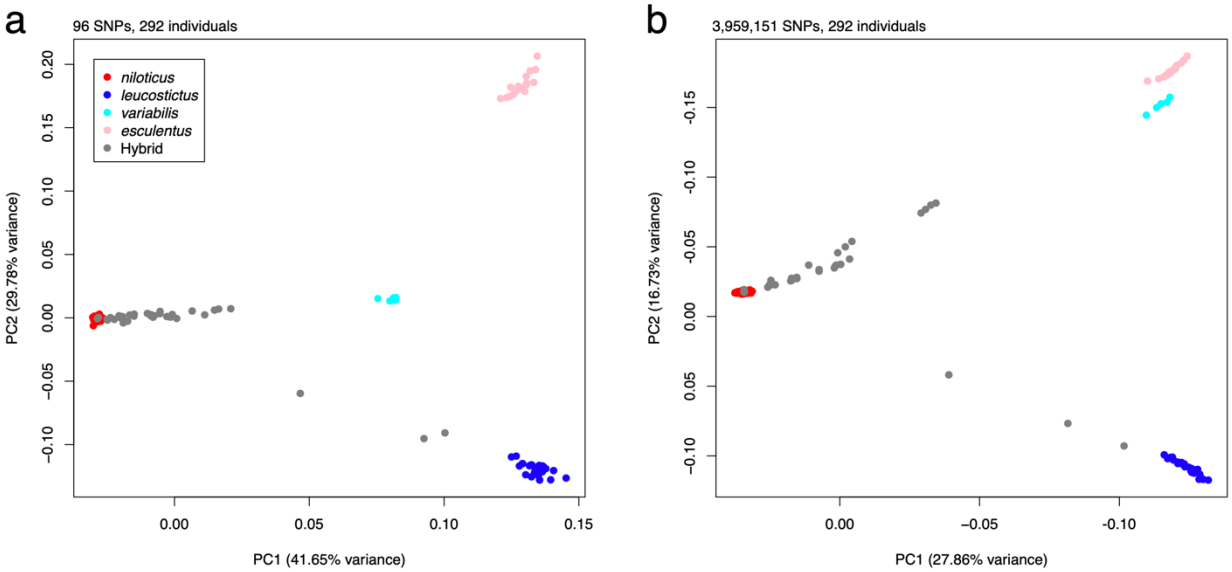

Figure 4 a) Principal Component Analysis (PCA) of all pure individuals and individuals identified genetically as hybrids for the 96 SNPs panel, b) genome-wide mapping of O. niloticus.

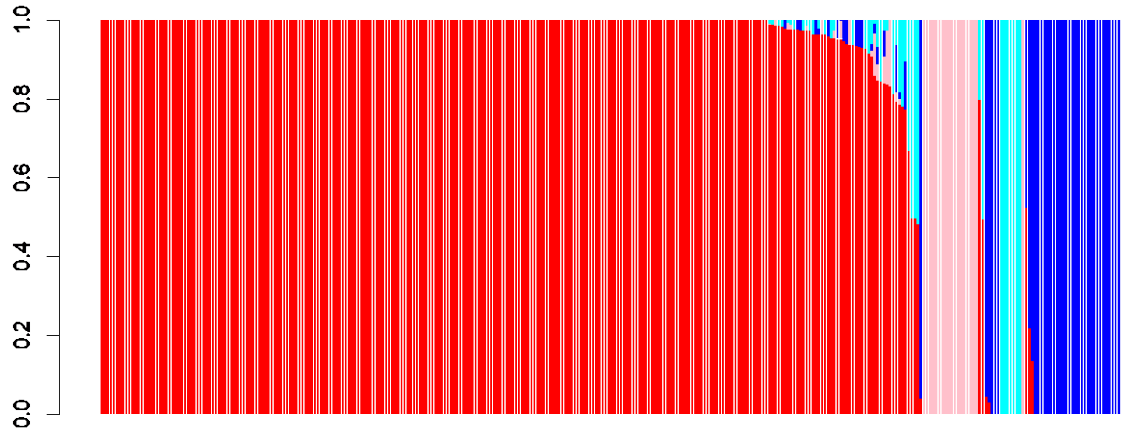

Figure 5 ADMIXTURE results based on the 96 SNPs panel for all individuals identified as Oreochromis. The coloured bars correspond to different ancestry groups (red: O. niloticus, pink: O. esculentus, cyan: O. variabilis, blue: O. leucostictus)

Because of the accuracy of the genotyping-based approach we recommend its use for species identification and hybrid status recognition. It also has the major advantage of being applicable across all development stages (eggs, fingerlings, adults) and sex, and from relatively low input material (fin clips).

Using the 96 SNPs Identification Tool

The analyses described above demonstrated the utility of analysing 96 SNPs to accurately identify tilapia species and their F1 hybrids in the Lake Victoria Basin. The tool itself is the sequence incorporating each of the SNPs and associated information which are provided in Annex 1. This provides all the information necessary for a sequencing laboratory or commercial sequence analysis company to create the 96 SNPs panel to analyse the samples sent to them. Provision of this digital tool gives the freedom for LVFO and national departments to engage analysis of their material as they desire and through provider they choose.

The information generated by the laboratory/commercial sequence service provider will be genotypes at each relevant position. The personnel trained in bioinformatics analysis will have the capability to undertake any further interpretative work that may be required.

The tool provides the means to extend the range of samples analysed and enable the monitoring biodiversity at the species level in wild and cultured populations. Although the tools will also provide some information with respect to genetic variability within species, it has primarily been designed to differentiate species and identify hybrids. It provides the critical means of gaining sequence information to be associated with images to build up the large number of validated images to needed to attempt a digital analysis tool.

General procedure for implementation of the tool

Tissue samples such as fin clips which do not require sacrifice of the fish required, are collected by applying the protocols used in the TRUEFISH study. These tissue samples would be sent to the laboratory or service provider of choice for DNA extraction and SNP marker analysis. The service provider should be aware of the DNA extraction procedures needed but, briefly, they should extract DNA from the fin clips using a DNA extraction kit (for instance: PureLink® Genomic DNA extraction kit (Life Technologies) or Quiagen QIAmp DNA Mini kit).

In order for the provider the set up their processing they would require at a minimum to have the information concerning the identity and position of the 96 SNPs – that is the information provided in Annex 1 but accessible digitally as Doi: 10.5281/zenodo.11193182 File: all.vcf.gz.

An experienced provider should be aware of the software and analysis methods available to them. However, in contacting them, they can be informed of the resources to be used in addition to the core information on the 96 SNPs variants. Design of the SNPs panel can be accessed through the following resource: https://www.agenabio.com/services/assays-by-agena/

The providers will need to access the reference genome as a fastp file as well as the positions of the variants stored in the file: variants toolkit.xlsx and available here: Doi: 10.5281/zenodo.11193182

All this information can be sent to them for them to set their process. They potentially could provide a full analysis including species identification, or they could supply the identified genotypes for further interpretation or bioinformatics analysis by the requesting team in EAC.

Further information that can assist understanding the processes to be used in the successful implementation of the tool can be found in the methods described in Ciezarek et al. (2022). A full list of software is given in Annex 2.

Conclusion

Tilapia species and hybrid identification represents a significant challenge. Morphological identification based on a set of traits is particularly challenging because of existing morphological variation and plasticity across populations within species. It also requires significant expertise and experience and is limited to sexually mature males.

Our first approach was to focus on facilitating morphological identification based on machine learning methods applied to photographs of specimen. This methodology would have the advantage of rapidly identifying a species. Although our preliminary results based on a total of 931 images (including genotypes for 433 individuals) showed promise, the current accuracy of the species identification using images is not high enough to enable the deployment of this approach. To reach good accuracy using images it will be important to significantly increase the size of the training set to include several hundreds of curated photographs per species. It will be important to also expand the training set to female and juvenile specimen to broaden the application of such approach. All of which will require confirmation of the image identity by using genetic information.

Because of the limitations using images, the best approach to accurately identify species or hybrid status is to access the genotype of a specimen and compare it to reference sets.

To enable the cost-effective identification of the different species in the Lake Victoria Basin, we identified a set of 96 SNPs that together enable accurate classification of specimens. We recommend the use of those variants for population and stocks survey.

Acknowledgements

This work was undertaken with funding from the European Union (FED/2018/039-248). Additional funding was supported by the CGIAR Research Initiative on Resilient Aquatic Food Systems for Healthy People and Planet, led by WorldFish, and supported by the contributors to the CGIAR Trust Fund.
Collection of samples were made possible through the participation of EAC research institutes of KMFRI, TAFIRI and NaFIRRI, represented by Dr. Kevin Obiero, Dr. Asilatu Hamisi Shechonge and Dr. Cassius Aruho respectively. Genotyping and sequencing of the samples were made possible by the Earlham Institute, UK, implemented by Dr. Wilfried Haerty and Dr. Adam G. Ciezarek.
The first draft of the report was initiated by AGC, KO, AHS, CA, PDT and WH, and reviews were completed by JAHB and BW.

References

Ciezarek, A. G., Mehta, T. K., Man, A., Ford, A. G., Kavembe, G. D., Kasozi, N., . . . Haerty, W. (2023). Ancient and ongoing hybridization in the Oreochromis cichlid fishes. *BioRxiv*.

Ciezarek, A., Ford, A., Etherington, G., Kasozi, N., Malinsky, M., Mehta, T., . . . George, T. (2022). Whole genome resequencing data enables a targeted SNP panel for conservation and aquaculture of Oreochromis cichlid fishes. *Aquaculture, 548*.

Elhamod, M., Diamond, K. M., Yaga, A., Bakis, Y., Bart Jr., H. L., Mabee, P., . . . Karpatne, A. (2021). Hierarchy-guided neural network for species classification. *Methods in Ecology and Evolution, 13*(3), 642-652.

Trewavas, E. (1983). *Tilapine Fishes of the Genera Sarotherodon, Oreochromis and Danakilia.* London: London British Museum (Natural History Musuem).

Annex 1

The 96 SNPs species identification tool for Oreochromis fish in the Lake Victoria.

CHROMOSOME NUMBER	POSITION	REF	ALT
NC_031972.2	4498643	G	A
NC_031972.2	9596645	T	A
NC_031972.2	18213475	A	G
NC_031972.2	29733738	T	A
NC_031972.2	36477764	C	A
NC_031972.2	46538095	C	T
NC_031972.2	51584848	G	A
NC_031972.2	63639839	T	C
NC_031978.2	4692873	T	C
NC_031978.2	11446928	G	C
NC_031978.2	25739348	T	G
NC_031978.2	31491328	C	A
NC_031977.2	3123405	C	G
NC_031977.2	11549004	G	A
NC_031977.2	28360567	G	A
NC_031977.2	34455360	A	C
NC_031984.2	4339277	A	C
NC_031984.2	15330311	C	G
NC_031984.2	23199108	G	C
NC_031984.2	28990454	C	T
NC_031982.2	7868817	A	G
NC_031982.2	16050061	T	G
NC_031982.2	20237961	G	T
NC_031982.2	29010313	T	G
NC_031979.2	144710	C	G
NC_031979.2	15261946	T	C
NC_031979.2	24820394	T	G
NC_031979.2	33951328	C	A
NC_031983.2	4872897	A	G
NC_031983.2	7948801	T	C
NC_031983.2	17635728	G	A
NC_031983.2	28481112	C	T
NC_031986.2	4056073	A	G
NC_031986.2	7360814	T	G
NC_031986.2	16975381	T	C
NC_031986.2	22505668	A	T
NC_031986.2	24567761	C	A
NC_031986.2	30024018	A	G
NC_031986.2	38543990	C	T

NC_031986.2	44490428	C	T
NC_031970.2	7325434	C	A
NC_031970.2	15715450	A	G
NC_031970.2	23559092	A	G
NC_031970.2	31865810	C	T
NC_031981.2	2950147	G	A
NC_031981.2	18187498	G	A
NC_031981.2	20085160	T	C
NC_031981.2	38532882	A	G
NC_031973.2	5969592	C	T
NC_031973.2	13322224	G	A
NC_031973.2	19428635	T	C
NC_031973.2	29338298	G	A
NC_031975.2	4148116	C	T
NC_031975.2	12856271	A	C
NC_031975.2	24687388	C	T
NC_031975.2	26741532	A	G
NC_031969.2	7208925	A	C
NC_031969.2	14406868	T	C
NC_031969.2	17840623	A	G
NC_031969.2	28873156	G	A
NC_031976.2	5452750	A	G
NC_031976.2	13325716	A	G
NC_031976.2	19852771	G	A
NC_031976.2	32495843	T	A
NC_031980.2	802836	T	C
NC_031980.2	19825693	G	A
NC_031980.2	27201092	T	C
NC_031980.2	34114015	G	T
NC_031987.2	8887217	G	A
NC_031987.2	12174975	C	T
NC_031987.2	23541857	C	T
NC_031987.2	30432365	G	C
NC_031971.2	2395326	C	T
NC_031971.2	8867471	C	T
NC_031971.2	13244136	A	G
NC_031971.2	16757040	A	T
NC_031971.2	24215657	G	A
NC_031971.2	26694833	C	A
NC_031971.2	36511168	C	T
NC_031971.2	41982418	A	G
NC_031974.2	5842361	A	C

NC_031974.2	15334284	T	C
NC_031974.2	23470295	G	C
NC_031974.2	29941906	G	T
NC_031965.2	3127883	T	C
NC_031965.2	14913852	A	G
NC_031965.2	27302878	G	A
NC_031965.2	30742605	G	C
NC_031985.2	9425778	C	T
NC_031985.2	12482583	T	C
NC_031985.2	26890659	A	C
NC_031985.2	30409341	A	T
NC_031966.2	1694146	T	C
NC_031966.2	17543836	C	T
NC_031966.2	21459700	T	G
NC_031966.2	33153489	C	G

Annex 2

The following is a list of the software resources that could be supplied to the providers to assist their choice of approach to provide a sequencing analysis using the 96 SNPs Identification Tool, or for further analysis by personnel trained in TRUEFISH if only basic genotypic information was requested from the service provider.

.

Resources necessary for analysis of the data:
Oreochromis niloticus reference genome:
https://ftp.ensembl.org/pub/current_fasta/oreochromis_niloticus/dna/Oreochromis_niloticus.O_niloticus_UMD_NMBU.dna.toplevel.fa.gz

Existing genotypes for the 96 SNPs from the TRUEFISH project to use as reference:
Doi: 10.5281/zenodo.11193182
File: all.vcf.gz

Software necessary:
Fastp: https://github.com/OpenGene/fastp
BWA: https://github.com/lh3/bwa
Samtools: https://github.com/samtools/samtools
Picard: https://github.com/broadinstitute/picard
Bcftools: https://github.com/samtools/bcftools
Plink 2: https://www.cog-genomics.org/plink/2.0/
Admixture: https://dalexander.github.io/admixture/download.html

Process flow:

1. Read trimming using fastp
2. Mapping of the reads to the genome using BWA
3. Removal of duplicates using picard MarkDuplicates
4. Variant calling for each individual using bcftools mpileup and merging of the calls using bcftools merge
5. Merging of the new calls with the existing variants (all.vcf.gz) using bctools concat
6. PCA generation using:
   bcftools view -e 'F_MISSING > 0.25' -m 2 -M 2 ../../run_March/3_SNPs/all_samples_TrueFish.vcf.gz| bcftools +prune -m 0.6 -w 20kb -Oz -o allind_prune_TrueFish.vcf.gz
   plink2 --vcf allind_prune_TrueFish.vcf.gz --allow-extra-chr --double-id --pca 20 –out allind_TrueFish
7. Generate admixture analyses:
   plink2 --vcf ../../4_Pca/Onil/Onil.vcf.gz --allow-extra-chr --double-id --make-bed --out Onil
   python /hpc-home/ciezarek/rename_chroms_admixture.py Onil.bim
   admixture --cv Onil.bed -j16 “cluster_number”